On the Average Rate Performance of Hybrid-ARQ in Quasi-Static Fading Channels

Cong Shen, *Student Member, IEEE,* Tie Liu, *Member, IEEE,* and Michael P. Fitz, *Senior Member, IEEE*

***Abstract*—The problem of efficient communication over a scalar quasi-static fading channel is considered. The single-layer transmission (SLT) and multi-layer transmission (MLT) schemes do not require any knowledge of the channel state information (CSI) at the transmitter, but their performance is also limited. It is shown that using Hybrid-ARQ (HARQ) can significantly improve the average rate performance, provided that the rate assignment between different ARQ rounds is carefully chosen. The average rate performance of several HARQ schemes is optimized and compared. In addition, optimal power allocation among retransmissions is derived and shown to further increase the average rate. This power allocation gain is remarkable at low signal-to-noise ratio (SNR), but becomes negligible at high SNR. Comparison of two different types of limited feedback, sequential feedback (ARQ) and one-shot feedback (quantized CSI), is made from several perspectives. Although the optimization problem is formed with respect to the average rate, simulation results give a comprehensive comparison under different metrics, including average rate, outage probability, and the combination of both. Substantial performance improvement is observed with even one ARQ retransmission in all simulations. More importantly, this gain appears to be robust with respect to the fading distributions.**

***Index Terms*—Hybrid-ARQ (HARQ), incremental redundancy, fading channels, throughput, Channel State Information (CSI).**

I. INTRODUCTION

The problem studied in this paper is how to efficiently transmit information over a quasi-static wireless fading channel, where the channel gain is constant during one coherence block and changes independently from one block to another. It is assumed that the receiver can perfectly track the fading process. Depending on whether the transmitter knows about the instantaneous channel realization, different performance measures have been studied. If the transmitter has no knowledge of the channel realization other than the statistical characterization, the Shannon capacity is zero as there is always a nonzero probability that the channel is in deep fade. A useful and well-accepted performance metric is the *outage capacity* [1]. In this formulation, a fixed-rate channel code is used, and the information is reliably transmitted if the instantaneous channel gain supports the predetermined transmission rate. Otherwise, an outage is declared, and no information can be recovered at the receiver. Quasi-static fading with channel state information (CSI) available only at the receiver (CSIR) is a prime example of the detrimental effect of fading.

Further improving the performance requires an *opportunistic* view of fading: a well-designed system should be able to adapt to the channel variations, i.e., it sends some information across the channel when the channel is not-so-good and a lot of information when the channel is very good. By exploiting the "good" channel realizations, the long-term throughput can be substantially improved. However, adapting to channel fading *without* transmitter's knowledge of CSI faces some conceptual difficulties. The breakthrough was made in [2], where the author observed the similarity between communication over quasi-static fading channels and broadcasting to multiple users. This venue was later pursued in [3], [4], and [5], leading to the development of a multi-layer transmission (MLT) strategy which utilizes broadcast superposition coding. By organizing information into layers, the MLT strategy allows rate adaptation to channel fading at the expense of creating self interference during the decoding of the earlier-decoded layers in the stack. Furthermore, depending on the actual channel realization, it may occur that only part of the information can be decoded at the receiver, which may lead to some extra complications in the upper layers in the network hierarchy.

This work follows the same line of examining the long-term throughput performance of communication over quasi-static wireless fading channels [3]–[5]. Both single-layer transmission (SLT) and MLT are first briefly discussed. It is then shown that with the use of *Hybrid-ARQ (HARQ)*, the average rate performance can be dramatically improved, provided that the rate assignment of the HARQ protocol is optimally designed. This scheme is also referred as *Rate-Optimized HARQ (RO-HARQ)*. The basic idea is to *exploit* the existence of ARQ in the data link layer to increase the average rate. Roughly speaking, the initial transmission is set to be very aggressive (high rate). If the channel does not support this high rate, an ARQ will help by indicating the transmitter to reduce the rate. The average rate maximization problem of RO-HARQ is formulated, and numerical results demonstrate the remarkable gain over both SLT and MLT strategies. Moreover, this gain appears to be robust to the fading distributions. Further average

Paper approved by L. Rasmussen, the Editor for Iterative Detection, Decoding, and ARQ of the IEEE Communications Society. Manuscript received February 19, 2008; revised October 22, 2008 and April 27, 2009; accepted April 28, 2009. The work of Cong Shen and Michael P. Fitz is supported by NSF grant CCF-0431196, and by ST Microelectronics with a matching grant from the University of California Discovery Program. Part of this work was done while Cong Shen was visiting Texas A&M University.

Cong Shen is with the Department of Electrical Engineering, University of California, Los Angeles (UCLA), Los Angeles, CA 90095, USA (E-mail: congshen@ee.ucla.edu).

Tie Liu is with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843, USA (E-mail: tieliu@tamu.edu).

Michael P. Fitz was with the Department of Electrical Engineering, University of California, Los Angeles (UCLA), Los Angeles, CA 90095, USA. He is now with Northrop Grumman Space Technology, Redondo Beach, CA 90278, USA (E-mail: fitz@ee.ucla.edu).

rate increase is possible, especially in the low signal-to-noise ratio (SNR) regime, if power allocation among different retransmissions is performed. To comprehensively compare these schemes, numerical optimization and simulations are performed with respect to different performance measures, including average rate, outage probability, and the combination of both.

The idea of using ARQ to improve communication performance is not new. In fact, HARQ techniques are widely used in most of the contemporary digital communication systems. A good summary of the progress of HARQ schemes is presented in [6]. Early work regarding the HARQ system focuses on the usage of algebraic error-correction and error-detection codes [7]. Recent interests of HARQ mainly originate from the rapid progress of wireless communications, where high-rate reliable transmission faces the challenge of severe channel fluctuations. Throughput and scheduling optimization of downlink packet data systems are investigated in [8]. An information-theoretic throughput and delay analysis of several HARQ schemes in the Gaussian collision channel is reported in [9]. Throughput analysis of incremental redundancy HARQ in the block-fading additive white Gaussian noise (AWGN) channel is carried out in [10]. A general framework of diversity-multiplexing-delay tradeoff is proposed in [11] to study multiple-input multiple-output (MIMO) ARQ block fading channels. Later [12] extends this framework to incorporate discrete input distributions. From the practical implementation point of view, research interests have shifted from traditional algebraic linear block codes to the more powerful capacity-approaching modern codes. For example, the problem of designing low-density parity-check (LDPC) codes for the HARQ protocol has been addressed in [10], [13]–[17].

The rest of this paper is organized as follows. Section II defines the system model. Section III briefly discusses the average rate performance of SLT and MLT. Section IV presents the theoretical analysis of RO-HARQ. Average rate maximization is discussed in Section IV-C and IV-D, followed by the optimal power allocation in Section IV-E. Numerical comparison with several different performance metrics is reported in Section V. Section VI discusses the difference between sequential feedback and one-shot CSI feedback. Finally, Section VII concludes the paper and points out possible directions for future work.

II. SYSTEM MODEL

We consider a scalar quasi-static fading channel where the random channel gain h remains constant for a duration of T_c symbol times and then changes independently to another value according to the fading distribution. The value T_c is generally determined by the channel coherence time. The signal model can be written as

$$y[m] = hx[m] + z[m], \qquad m = 1, \cdots, L \tag{1}$$

where $\{x[m], m = 1, \cdots, L\}$ is a length-L codeword containing K information nats[1], and $z[m]$ is independent and identically distributed (i.i.d.) complex Gaussian noise with zero mean and variance N_0 (denoted as $\mathcal{CN}(0, N_0)$). We assume $T_c \gg L$ such that the transmission of K information nats only experiences one fading state. This is the worst case since no time diversity can be exploited. There is a short-term average power constraint of P on $\{x[m]\}$, which prohibits power allocation across different fading states[2]. Since each transmission experiences an AWGN channel, our analysis in this paper is restricted to the Gaussian input distribution. The channel gain and noise power are normalized to be $\mathbb{E}\left[|h|^2\right] = 1$ and $N_0 = 1$ respectively, so the average received signal-to-noise ratio is $\mathsf{SNR} \doteq P\mathbb{E}\left[|h|^2\right]/N_0 = P$. The random channel power $g \doteq |h|^2 \geq 0$ is assumed to be a continuous variable with the cumulative distribution function (CDF) $F(g)$ and the probability density function (PDF) $f(g)$. One example is the frequently-encountered Rayleigh fading with $h \sim \mathcal{CN}(0, 1)$. We also assume that the transmitter has a very large pool of information nats such that once the transmission of the current K nats ends, the transmitter starts to send the next K nats immediately.

The quasi-static fading channel is a good model for users that are stationary, or moving slowly relative to the rate of communication. Due to the slowly varying nature of the channel, channel estimation at the receiver can be performed with high accuracy. Thus, it is reasonable to assume perfect CSI at the receiver. This assumption will be made throughout this paper.

The focus of this work is on high-rate, delay-insensitive applications such as data traffic in wireless LAN. For such applications the use of capacity-approaching channel codes with long block length can be justified, and information-theoretic results are good approximations of real-world performance. This motivates us to take an information-theoretic view and consider the capacity related measures on the system performance in this paper.

III. SINGLE-LAYER AND MULTI-LAYER TRANSMISSIONS

A. Figure of Merit

Our main figure of merit is the *long-term throughput*. Let us use n to count the number of time slots, T_n for the channel uses of the n-th slot, $T^{(n)} = \sum_{i=1}^{n} T_i$ for the total number of channel uses at the end of the n-th slot, and $k^{(n)}$ for the total number of information nats successfully decoded up to slot n. Then the long-term throughput (in nats per channel use, npcu) is defined as [9]

$$\eta \doteq \lim_{n\to\infty} \frac{k^{(n)}}{T^{(n)}} = \frac{\mathbb{E}\{k\}}{\mathbb{E}\{\mathcal{T}\}} \tag{2}$$

where k is the number of successfully decoded information nats at the end of each transmission, and $\mathcal{T}$ is the number of channel uses.

Both SLT and MLT belong to the general category of *fixed-length* coding schemes, in which the code length is constant regardless of the fading state. For such fixed-length codes, the

---

[1] The information unit is nat throughout the paper, except for the numerical results.

[2] When discussing power allocation for RO-HARQ in Section IV-E, unequal power allocation between the primary transmission and ARQ retransmissions is allowed (still within one channel state).

denominator of (2) is constant, and the long-term throughput degenerates to the *average rate*:

$$\bar{R} \doteq \mathbb{E}_g\left[R(g)\right] \tag{3}$$

where g is the channel state, and $R(g)$ is the successful transmission rate when channel state is g. See [18] for a formal definition of average rate. It should be noted that average rate is different from the *ergodic capacity*; here, coding across different fading states is prohibited, c.f. channel model (1).

B. SLT and MLT schemes

Assuming that the transmitter does not know the channel state g, a simple and well-adopted scheme for communicating over a slow fading channel is to send the data at a fixed rate R. If the instantaneous channel realization supports the rate R, the receiver gets a successful transmission; otherwise it declares an outage.

The average rate of SLT can be calculated as [19, Equation (7)]

$$\begin{aligned} \bar{R}_{SLT} &= R\,\mathbb{P}\left\{\log\left(1+gP\right) \geq R\right\} \\ &= R\left(1 - F\left(\frac{e^R - 1}{P}\right)\right). \end{aligned} \tag{4}$$

In order to maximize the average rate, one needs to solve the following optimization problem:

$$\begin{array}{ll} \text{maximize} & R\left(1 - F\left(\frac{e^R-1}{P}\right)\right) \\ \text{subject to} & R \geq 0. \end{array} \tag{5}$$

This optimization problem has been studied in [19, Section III], where the Karush-Kuhn-Tucker (KKT) condition is derived. Here, we mention that with Rayleigh fading the KKT condition simplifies to

$$Re^R = P \Longleftrightarrow R^*_{SLT,Rayleigh} = \mathcal{L}(P) \tag{6}$$

where $\mathcal{L}(\cdot)$ is the Lambert W function [20] which is defined as the solution to $ye^y = x$, and

$$\bar{R}^*_{SLT,Rayleigh} = \mathcal{L}(P)\exp\left(-\frac{e^{\mathcal{L}(P)} - 1}{P}\right). \tag{7}$$

The problem with SLT is that it evaluates the transmission in an "on-off" fashion: the transmission is either entirely successful or totally failed. This scheme suffers from both over-utilizing and under-utilizing the channel since it uses a non-adaptive transmission strategy. The MLT scheme is developed to overcome the disadvantages of the SLT scheme without transmitter's knowledge of CSI. The MLT scheme adopts the multi-user broadcast superposition code for a single-user slow fading channel. It is proposed in [4] for infinitely-many layers, and in [5] for finite layers. As an example, let us consider the case with $N = 2$ layers. In this case, the transmitter splits the message w into two sub-messages w_1 and w_2. They are separately encoded into $\{x_1[m]\}$ and $\{x_2[m]\}$ respectively, assuming Gaussian signaling, and finally superposed into the transmit signal $x[m] = x_1[m] + x_2[m]$, where $\{x_1[m]\}$ has a power αP and rate R_1 npcu, $\{x_2[m]\}$ has a power $(1-\alpha)P$ and rate R_2 npcu, and the scalar $\alpha \in [0,1]$ determines the power allocation between these two layers. At the receiver, the decoding is carried out based on the channel realization h. The first step is trying to decode $\{x_1[m]\}$, treating $\{x_2[m]\}$ as adding to the noise floor. If $\{x_1[m]\}$ is not successfully decoded, the receiver will give up and declare an error. On the other hand, if $\{x_1[m]\}$ is successfully decoded, the decoding procedure continues to the second step: subtracting the successfully decoded $\{x_1[m]\}$ and then decoding $\{x_2[m]\}$. After some manipulation [5], the two-layer MLT average rate maximization problem reduces to:

$$\begin{array}{ll} \text{maximize} & R_1\left(1 - F\left(s_1\right)\right) + R_2\left(1 - F\left(s_2\right)\right) \\ \text{subject to} & R_1 \geq 0 \\ & R_2 \geq 0 \\ & \frac{e^{R_1+R_2} - e^{R_2}}{e^{R_1+R_2} - 1} \leq \alpha \leq 1 \end{array} \tag{8}$$

where $s_1 \doteq \frac{e^{R_1}-1}{P(1-(1-\alpha)e^{R_1})}$ and $s_2 \doteq \frac{e^{R_2}-1}{P(1-\alpha)}$.

The average rate of MLT can be further boosted by increasing the number of layers in the superposition code, although such gain has been shown to be insignificant in Rayleigh fading [5]. In [4], Shamai and Steiner derived the maximum average rate of a superposition code with infinitely-many layers for the Rayleigh fading:

$$\bar{R}^{\infty}_{MLT,Rayleigh} = 2E_i(S_0) - 2E_i(1) - \left(e^{-S_0} - e^{-1}\right) \tag{9}$$

where $S_0 = 2/\left(1 + \sqrt{1+4P}\right)$ and $E_i(x) = \int_x^\infty \frac{e^{-t}}{t}\mathrm{d}t$, $x \geq 0$ is the exponential integral function.

One property of the layering strategy is that it creates *self interference* to the decoding. Specifically, when the receiver decodes the n-th layer, it treats all the not-yet-decoded layers $n+1, \cdots, N$ as interference and thus decreases the effective receive SNR for the n-th layer to

$$\mathsf{SNR}_n = \frac{gP_n}{1 + g\sum_{i=n+1}^{N} P_i}.$$

As the number of layers increases, this self interference affects more layers in the decoding; in fact, only the last decoded layer is interference-free. This self interference has been the main concern of using broadcast superposition code in a single-user slow fading channel [21]. Thus, although the MLT scheme improves the performance over SLT, further performance gain can be expected if we are able to eliminate such self interference. This motivated the development of a Rate-Optimized Hybrid-ARQ (RO-HARQ) which will be described next.

IV. RO-HARQ: THEORETICAL ANALYSIS

A. HARQ Schemes

In this section, we study several HARQ schemes with optimally designed rate and power for each transmission. It will be shown that optimized HARQ can eliminate all of the aforementioned problems. With the help of ARQ, the decoding status at the receiver will be reported back to the transmitter, which indicates the successful decoding of received signal by acknowledgement (ACK) and failed decoding by negative acknowledgement (NACK). It is assumed that the ARQ feedback channel is delay and error free. The maximum number of retransmissions of HARQ is denoted by N, i.e., the total transmissions (including the initial one) cannot exceed $N+1$.

In the ARQ literature, $N+1$ is also called the maximum allowable ARQ rounds. The resulting protocol is sometimes referred to as *ARQ with a deadline*. The choice of N reflects the *worse-case delay* caused by the ARQ retransmissions, and is generally determined by the system delay requirement. For example, choosing small N models certain delay critical situation. In order to be consistent with channel model (1), the maximum ARQ rounds should limit the overall code length to be within L. Note that the HARQ protocol is typically available in almost all existing wireless systems, and thus exploiting HARQ will not need additional designs.

We consider the following three HARQ protocols[3] in this paper.

1) ALO. The transmitter encodes the K information nats at rate R npcu, and then keeps sending the same encoded packet in every retransmission. The receiver only decodes the most recently received packet. This loop continues until the ACK is declared by the receiver, or the maximum retransmissions are used.
2) RTD. The transmitter is the same as ALO. The receiver performs a maximum ratio combining (MRC) of all the received packets. In the HARQ literature, this scheme is also referred as *Chase Combining* [23].
3) INR. The transmitter encodes the K information nats into a codeword of length $T^{(N+1)}$. Then it *serially* punctures the length-L codeword into $N+1$ sub-codewords with strictly decreasing rates $K/T^{(1)} > K/T^{(2)} > \cdots > K/T^{(N+1)}$. The lengths of sub-codewords are the design parameters, which are determined by the code rate optimization and will be addressed in Section IV-C. At the n-th transmission ($n = 1, \cdots, N$), the transmitter reduces the total rate to[4] $R_1 + R_2 + \cdots + R_{N+2-n} = K/T^{(n)}$ by sending additional redundancy symbols. The receiver tries to decode based on all the packets it receives up until this moment. At the last retransmission (round $N+1$), rate R_1 is tried without any ARQ feedback. If decoding is still unsuccessful, a decoding failure is declared.

This works focuses on INR due to the following two reasons.

1) ALO scheme does not work for slow fading channels. The reason is that there is no time diversity to exploit, as the channel gain remains constant over ARQ retransmissions. Thus to keep sending the same packet and only decoding the most recently received one will not increase the chance of successful decoding.
2) RTD has an average rate performance that is inferior to INR. This will be shown both analytically and numerically. Note that although RTD has inferior performance, it has some practical advantages which make it attractive in some applications. For example, the retransmitted packet in RTD is always the same, which is easier to implement than INR, in which additional parity symbols have to be generated and transmitted. As another example, having the same size for the retransmitted packets in RTD makes the packetization simple, while in INR the length of the retransmitted parity symbols may vary from one retransmission to another.

The novelty of the proposed RO-HARQ scheme is the optimal rate assignment[5] of each transmission. Traditionally, HARQ is used in a *passive* manner in a wireless system. The purpose of HARQ is to indicate the decoding status to the transmitter, such that the transmitter can protect the "bad" packet by retransmitting. The existence of HARQ is typically not *exploited* by the transmitter; it is treated only as a binary indicator of the decoding status, and other modules in the system (e.g., channel coding, modulation, etc.) are operating as if HARQ were not available. Due to this reason, the rate of each HARQ transmission and the (equal) power allocation is usually predetermined. This work proposes the different view that HARQ can be exploited by the transmitter to provide better average rate and outage performance, i.e., it can be utilized in an *active* manner. This new view suggests that the rate and power associated with each HARQ transmission can be optimized according to the channel fading statistics to achieve better performance. The main idea is that since the transmitter is aware of the HARQ link in the system, it can transmit very aggressively (at very high rate) even if it does not know the random channel state g before transmitting. Then if the channel is not good, this high-rate transmission will fail, and the HARQ can save it by indicating this failure to the transmitter so that it can adjust to a lower rate. On the other hand, if it is "lucky" that the channel is strong enough, such "gambling" will bring high return: the strong channel realization is (almost) fully utilized, i.e., there is (almost) no waste of the good channel. This is the basic idea behind RO-HARQ.

It is now clear why RO-HARQ can eliminate the problems of SLT and MLT. By transmitting aggressively, the "good" channel realizations are fully used; with the ARQ feedback, the transmission rate can be reduced for the "bad" channel realizations. At the same time, as there is no superposition or layering in the channel coding, self interference does not exist.

B. Average Rate for HARQ

The general expression (2) characterizes the long-term throughput for both fixed and variable length coding schemes. How to evaluate and optimize the throughput performance of variable-length coding schemes depends on the specific applications and the assumptions on how to use the channel. If we assume that the channel remains constant during the transmission of K information nats, and changes independently when the transmission of current packet[6] is done and a new packet is ready to transmit, we need to evaluate

---

[3] The acronyms are borrowed from [9], [22], although the schemes and applications may be different. For example, [9], [22] studied a *slotted* system with equal length for each retransmission. They considered a multi-user TDMA system, where it is reasonable to assume each retransmission experiences a different fading state. Both are different from our assumptions.

[4] The reason of using this seemingly redundant expression for the data rate is that it simplifies equations (11), (12), (13), and the proof of Lemma 2.

[5] Another novelty is the optimal power allocation among transmissions to further improve the average rate performance. This is discussed in Section IV-E.

[6] The transmission of one packet includes several possible retransmissions, according to the HARQ protocol. Different packets contain different information.

both the average number of successfully decoded information nats (numerator of (2)) and the average delay (denominator of (2)) with respect to the fading distribution. Throughput analysis based on such assumptions is given in Appendix A and numerical performance is reported in Section V. This assumption can be valid for *burst* communication, where there is a long idle period in between transmissions of different packets such that the channel gain changes independently from one transmission to another.

A different scenario is that the transmitter has a very large pool of information nats such that the communication is "continuous". For this scenario, multiple packets (with each packet having multiple ARQ rounds) will be transmitted within one coherent period. When the channel is "good", more packets can be transmitted within one coherent period; when the channel is "bad", only a small number of packets can be transmitted within the same coherent period. As a result, each fading state has the same length of channel uses, and the empirical channel distribution will match the true one. This is the key difference from the burst communication scenario. The long-term throughput (2) for this scenario can be computed as follows. We use $R(g)$ to denote the instantaneous rate of the HARQ scheme for a given channel realization g. The total number of information nats that can be successfully transmitted is $k(g) = T_c R(g)$ over the channel state g. Hence we have

$$\eta = \frac{\mathbb{E}\{k\}}{\mathbb{E}\{\mathcal{T}\}} = \frac{\mathbb{E}\{T_c R(g))\}}{T_c} = \mathbb{E}\{R(g)\} = \bar{R} \tag{10}$$

which again degenerates to the definition of average rate. Thus in the following we shall focus on optimizing the average rate performance of different HARQ schemes.

C. Average Rate Maximization of INR and RTD

The INR scheme is described in Section IV-A. With INR, the random variable $R(g)$ is

$$R(g) = \begin{cases} \sum_{i=1}^{N+1} R_i, & \text{if } \log(1+gP) \geq \sum_{i=1}^{N+1} R_i \\ \sum_{i=1}^{n} R_i, & \text{if } \begin{array}{l} \sum_{i=1}^{n+1} R_i > \log(1+gP) \geq \\ \sum_{i=1}^{n} R_i, \forall n = 1, \cdots, N \end{array} \\ 0, & \text{if } R_1 > \log(1+gP). \end{cases} \tag{11}$$

The average rate of INR can thus be computed as

$$\bar{R}_{INR}^{N} = \sum_{n=1}^{N+1} R_n \mathbb{P}\left\{\log(1+gP) \geq \sum_{k=1}^{n} R_k\right\} \tag{12}$$

$$= \sum_{n=1}^{N+1} R_n \left(1 - F(g_n)\right) \tag{13}$$

where $g_n \doteq \frac{e^{\sum_{k=1}^{n} R_k} - 1}{P}$.

The problem is to choose the nonnegative design parameters $\{R_1, \cdots, R_{N+1}\}$ to optimize the average rate (13). Clearly, the non-negativity constraints on $\{R_n\}$ can be dropped, and the INR average rate maximization problem becomes

$$\text{maximize} \quad \sum_{n=1}^{N+1} R_n \left(1 - F(g_n)\right). \tag{14}$$

A direct evaluation of the KKT condition gives

$$1 - F(g_n) - \frac{R_n e^{\sum_{k=1}^{n} R_k}}{P} f(g_n) - \sum_{i=n+1}^{N+1} \frac{R_i e^{\sum_{k=1}^{i} R_k}}{P} f(g_i) = 0 \tag{15}$$

for $n = 1, \cdots, N$, and

$$1 - F(g_{N+1}) - \frac{R_{N+1} e^{\sum_{k=1}^{N+1} R_k}}{P} f(g_{N+1}) = 0. \tag{16}$$

In some simple cases (e.g., small N), equations (15) and (16) can be numerically solved to give the optimal rate assignment.

The average rate of RTD can be similarly computed. In the n-th transmission of RTD, $n = 1, \cdots, N+1$, the receiver performs maximum ratio combining of the n received packets. This processing effectively increases the receive SNR to $\mathsf{SNR}_n = nP$, while reducing the data rate to R/n. Thus the random variable $R(g)$ is

$$R(g) = \begin{cases} \frac{R}{n}, & \text{if } \begin{array}{l} \log(1+ngP) \geq R > \\ \log(1+(n-1)gP), \forall n = 1, \cdots, N+1 \end{array} \\ 0, & \text{if } R > \log(1+(N+1)gP). \end{cases} \tag{17}$$

The average rate can be computed as

$$\bar{R}_{RTD}^{N} = \sum_{n=1}^{N+1} \frac{R}{n} \mathbb{P}\{q_{n-1} > g \geq q_n\} \tag{18}$$

$$= \sum_{n=1}^{N+1} \frac{R}{n} \left(F(q_{n-1}) - F(q_n)\right) \tag{19}$$

where $q_n \doteq \frac{e^R - 1}{nP}$ for $n = 1, \cdots, N+1$, and $q_0 \doteq \infty$. Similarly, the RTD average rate maximization problem can be formulated as

$$\text{maximize} \quad \sum_{n=1}^{N+1} \frac{R}{n} \left(F(q_{n-1}) - F(q_n)\right). \tag{20}$$

Numerical evaluation and comparison of the optimal average rate of (14) and (20) are given in Section V, where we will see that RTD has a worse average rate performance than INR. This can also be analytically proved as follows.

Lemma 1: Denote the optimal solution to problem (14) as $(\{R_1^*, \cdots, R_{N+1}^*\}, \bar{R}_{INR}^{N^*})$, and the optimal solution to problem (20) as $(R^*, \bar{R}_{RTD}^{N^*})$. Then

$$\bar{R}_{RTD}^{N^*} \leq \bar{R}_{INR}^{N^*} \tag{21}$$

Proof: See Appendix B. ■

In order to fully enjoy the benefits of RO-HARQ, the corresponding channel code design for INR has to satisfy the following two requirements.

1) A good mother code that can be *serially* punctured into several different optimized code rates.
2) A single decoder that can handle all the rates such that the coding performance of each sub-code is capacity-approaching.

There are several existing code designs, developed in both academia (e.g., [10], [14], [15], [17], [24]) and industry (e.g., [25]) that satisfy these requirements.

D. Asymptotic Average Rate of INR

It is of interest to ask what is the asymptotic performance limit of the INR scheme (by allowing N to go to infinity, i.e., infinitely-many retransmissions). Such asymptotic performance gives the ultimate limit of the proposed scheme. It is also a good performance indication for delay insensitive applications. The following lemma states that the average rate of INR asymptotically converges to the *ergodic capacity*, even though coding across fading blocks is prohibited in INR.

Lemma 2: As $N \to \infty$, the average rate of INR with equal power allocation among transmissions converges to the ergodic capacity of the channel.

Proof: See Appendix C. ■

There are two interesting observations of Lemma 2. First, Lemma 2 applies to *any* fading distribution. This can also be seen from the proof. Secondly, in order to achieve this asymptotic limit, there is no requirement of optimal rate allocation. This observation suggests that as N becomes large, the gain of optimal rate design diminishes.

In a slow fading environment, if one is not allowed to perform coding over different fading states, it is well accepted that the outage performance is a good measure. In INR, the coding is still within one fading state. However, with the use of ARQ feedback, the performance measure switches from outage capacity to ergodic capacity, as the amount of ARQ retransmissions increases to infinity. This result, however, is not surprising in the following two aspects. First, with infinitely-many retransmissions allowed, in each fading block the transmission rate is gradually reduced (by NACK from the receiver) until it arrives at the *exact* rate that the current channel gain g can support: $\log(1+gP)$. That is, eventually the transmission happens with a rate that perfectly matches to the instantaneous channel realization. Thus all the information nats can theoretically be decoded without any error, and the average rate is the ergodic capacity. Secondly, allowing ARQ feedback effectively informs the transmitter not only the decoding status, but also the partial CSI. As the number of ARQ retransmissions goes to infinity, the informed CSI becomes perfect eventually.

From the information-theoretic viewpoint, a receiver decoding failure is equivalent to a channel outage with capacity-achieving codes. Thus, an ACK/NACK feedback essentially informs the transmitter a channel quantization, i.e., whether the instantaneous channel gain is greater than a given threshold. With multiple ARQ rounds, this process becomes a *sequential feedback* scheme, where the entire channel state space is sequentially quantized with more and more ACK/NACK feedback, and thus the transmitter gets a finer and finer knowledge of the channel. Another widely used limited feedback scheme is to perform a *global quantization* of the channel state space, and then indicate to the transmitter the interval index in which the channel realization falls. Such a scheme has been well studied in [19], [26]–[28]. We will call this scheme *one-shot feedback* and a detailed comparison with the ARQ sequential feedback is made in Section VI.

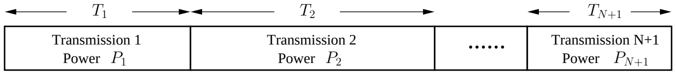


Fig. 1. Illustration of unequal power allocation for INR with N retransmissions.

E. Power Allocation for INR

In the previous sections, the general idea of RO-HARQ is discussed and the average rate performance is optimized, under the assumption that the transmit power is constant throughout the entire ARQ process. Intuitively, allowing power allocation among the $N+1$ transmissions could further improve the average rate. For example, boosting the power of primary transmission will increase the probability that the primary transmission succeeds. However this comes at the price of decreasing the power of ARQ retransmissions and hence the probability for success, if the average transmit power is kept constant. There is obviously a tradeoff in allocating power among these transmissions, and this idea is pursued in the following. Figure 1 gives a graphical illustration of this process. We shall first formulate the problem for general N, and then focus on the simplest case of $N=1$ both analytically and numerically.

Consider the INR scheme with K information nats and a maximum $N+1$ transmissions. We assume that the n-th transmission takes place with power P_n, and use $\mathcal{A}_n$ to denote the event of a successful decoding at the end of transmission n. We will approach this problem in two steps. First we derive the average rate for a fixed power allocation, and then show how to maintain a constant average power.

1) Average rate for a given power allocation $(P_1, \cdots, P_{N+1})$: Similar to (11), the rate of INR with given power allocation policy $(P_1, \cdots, P_{N+1})$ is a random variable:

$$R(g) = \begin{cases} \sum_{i=1}^{N+2-n} R_i, & \text{if } \begin{array}{c} \overline{\mathcal{A}_1}, \cdots, \overline{\mathcal{A}_{n-1}}, \mathcal{A}_n; \\ \forall n = 1, \cdots, N+1 \end{array} \\ 0, & \text{if } \overline{\mathcal{A}_1}, \cdots, \overline{\mathcal{A}_{N+1}} \end{cases} \tag{22}$$

where we define $\mathcal{A}_0$ as the empty set.

In order to evaluate $\mathbb{P}\{\mathcal{A}_n\}$, we need to derive the achievable rate of sending a long Gaussian code where different portions of the code have different power. This is not obvious but with a random coding and typical set decoding argument [29], this achievable rate can be shown to be a TDMA-type one, and thus

$$\begin{aligned} & \mathbb{P}\{\mathcal{A}_n\} \\ = \; & \mathbb{P}\left\{\sum_{i=1}^{n} \frac{T_i}{\sum_{j=1}^{n} T_j} \log(1+gP_i) \geq \sum_{i=1}^{N+2-n} R_i\right\} \\ = \; & \mathbb{P}\left\{\sum_{i=1}^{n} \frac{\frac{1}{\sum_{l=1}^{N+2-i} R_l}}{\sum_{j=1}^{n} \frac{1}{\sum_{l=1}^{N+2-j} R_l}} \log(1+gP_i) \geq \sum_{i=1}^{N+2-n} R_i\right\}. \end{aligned} \tag{23}$$

Finally the average rate is given by

$$\bar{R}_{INR}^{N} = \sum_{n=1}^{N+1} \left(\sum_{i=1}^{N+2-n} R_i \right) \mathbb{P}\left\{\overline{\mathcal{A}_1}, \cdots, \overline{\mathcal{A}_{n-1}}, \mathcal{A}_n\right\}. \tag{24}$$

2) Average power constraint: The power allocation $(P_1, \cdots, P_{N+1})$ is a design parameter. In order to make a fair comparison to the scheme discussed in Section IV-C with a constant power P, it is reasonable to put a constraint on $(P_1, \cdots, P_{N+1})$ such that the average power does not exceed P. The interesting observation of the HARQ scheme is that with unequal power allocation the *actual* power consumed in the entire transmission is a random variable. The reason is that the n-th transmission takes place only if transmissions $1, \cdots, n-1$ all fail, which is a random event determined by the random channel gain. Denote the actual consumed power as $\beta(g)$, then this discrete random variable is given by

$$\beta(g) = \begin{cases} \frac{\sum_{i=1}^{n} P_i T_i}{\sum_{i=1}^{n} T_i}, & \text{if } \overline{\mathcal{A}_1}, \cdots, \overline{\mathcal{A}_{n-1}}, \mathcal{A}_n; \forall n = 1, \cdots, N \\ \frac{\sum_{i=1}^{N+1} P_i T_i}{\sum_{i=1}^{N+1} T_i}, & \text{if } \overline{\mathcal{A}_1}, \cdots, \overline{\mathcal{A}_N}. \end{cases} \tag{25}$$

Thus the average power constraint that $(P_1, \cdots, P_{N+1})$ should satisfy is

$$\begin{aligned} &\textstyle\sum_{n=1}^{N} \frac{\sum_{i=1}^{n} P_i T_i}{\sum_{i=1}^{n} T_i} \mathbb{P}\left\{\overline{\mathcal{A}_1}, \cdots, \overline{\mathcal{A}_{n-1}}, \mathcal{A}_n\right\} \\ &+ \frac{\sum_{i=1}^{N+1} P_i T_i}{\sum_{i=1}^{N+1} T_i} \mathbb{P}\left\{\overline{\mathcal{A}_1}, \cdots, \overline{\mathcal{A}_N}\right\} \leq P. \end{aligned} \tag{26}$$

Notice that in the case of constant power allocation $P_n = P, \forall n$, the randomness of $\beta(g)$ disappears: $\beta(g) = P$ with probability 1.

3) Optimal power allocation: Finally, the average rate maximization problem under optimal power allocation can be formulated as

$$\begin{aligned} \text{maximize} \quad & \textstyle\sum_{n=1}^{N+1} \left(\sum_{i=1}^{N+2-n} R_i\right) \mathbb{P}\left\{\overline{\mathcal{A}_1}, \cdots, \overline{\mathcal{A}_{n-1}}, \mathcal{A}_n\right\} \\ \text{subject to} \quad & \textstyle\sum_{n=1}^{N} \frac{\sum_{i=1}^{n} P_i T_i}{\sum_{i=1}^{n} T_i} \mathbb{P}\left\{\overline{\mathcal{A}_1}, \cdots, \overline{\mathcal{A}_{n-1}}, \mathcal{A}_n\right\} \\ & + \frac{\sum_{i=1}^{N+1} P_i T_i}{\sum_{i=1}^{N+1} T_i} \mathbb{P}\left\{\overline{\mathcal{A}_1}, \cdots, \overline{\mathcal{A}_N}\right\} \leq P. \end{aligned} \tag{27}$$

This optimization problem is difficult to solve for general N. For simplicity let us consider the simplest case of $N = 1$. In this case Equation (24) becomes

$$\begin{aligned} \bar{R}_{INR}^{1} \quad &= (R_1 + R_2) \mathbb{P}\left\{\log(1 + gP_1) \geq R_1 + R_2\right\} \\ &+ R_1 \mathbb{P}\Big\{ \log(1 + gP_1) < R_1 + R_2, \\ &\frac{R_1}{R_1+R_2} \log(1 + gP_1) + \frac{R_2}{R_1+R_2} \log(1 + gP_2) \geq R_1 \Big\} \end{aligned} \tag{28}$$

and the average power constraint (26) is

$$\begin{aligned} &P \geq P_1 \mathbb{P}\left\{\log(1 + gP_1) \geq R_1 + R_2\right\} \\ &+ \left(\frac{R_1}{R_1+R_2} P_1 + \frac{R_2}{R_1+R_2} P_2\right) \mathbb{P}\left\{\log(1 + gP_1) < R_1 + R_2\right\}. \end{aligned} \tag{29}$$

Numerical results for $N = 1$ is reported in Section V.

Optimal average rate of problem (27) is better than that of problem (14): choosing $P_n = P, \forall n$ makes (24) equal to (13). It is then interesting to ask how big the power allocation gain is. Numerical examples in Section V shows that for the Rayleigh fading channel, this gain is remarkable in the low SNR regime, while it is negligible for medium to high SNR. This is a reasonable result for most of the known power allocation schemes, e.g., water-filling.

V. RO-HARQ: NUMERICAL RESULTS

The ARQ feedback link allows the transmitter to partially adapt to the channel conditions. Due to the lack of full channel state information at the transmitter, such adaptation is not guaranteed to support reliable transmission all the time. Thus the instantaneous successful transmission rate is a random variable, whose distribution is determined by the random channel fading and the rate assignment. As a performance metric, the average rate characterizes the mean value of this random variable. On the other hand, it is arguable that the *outage probability* (P_{out}) serves as a *worst-case* performance measure for the RO-HARQ scheme, as it describes the probability that the transmission fails after the maximum number of ARQ retransmissions are used. Thus, although the analytical discussion of this paper is focused on the average rate maximization, numerical results for different performance measures are reported in this section. To be specific, three performance metrics, average rate [3]–[5], [18], outage probability [1], [30], and average rate versus outage probability [4], are used to numerically compare the proposed RO-HARQ with SLT and MLT. Discussion of the comparison to another quantized feedback scheme is deferred to Section VI.

A. Average Rate

Numerical optimizations are performed to maximize the average rate of the schemes analyzed in Section III and IV. In the case of slow Rayleigh fading channel with $h \sim \mathcal{CN}(0, 1)$, Figure 2 reports the average rate comparison among SLT, MLT, INR and RTD, together with the optimal INR throughput for bursty communications derived in Appendix A. Ergodic capacity is also shown as the upper limit. It is clear that allowing INR ARQ feedback increases the average rate substantially. For example, even INR with $N = 1$ outperforms the infinite-layer MLT by half a bit per channel use over a wide range of SNRs (15 to 35 dB), and is 1.5 bits better than the SLT scheme with optimized rate. Notice that this average rate advantage of INR over MLT does not come with much higher complexity. MLT requires complicated encoding and decoding processes to handle the multiple layers, and this complexity increases with the number of layers. Although ARQ requires feedback and some overhead in the protocol design, typically its complexity is not as high as MLT. The average rate is boosted by another 0.5 bits if N increases to 2. As N further increases, the average rate continues to increase up until the ergodic capacity. At the same time, RTD type HARQ is shown to be inefficient in terms of the average rate performance. It is better than SLT[7], but in some configurations is even worse than MLT. This also numerically confirms Lemma 1. For this suboptimality, RTD is not considered in the remaining of the numerical simulations. Another observation from Figure 2 is a comparison of the INR throughput based on the two different assumptions (burst communication vs continuous communication) in Section IV-B. It can be seen that the average rate is always better than the optimal throughput in Appendix A. This is due to the fact (see Section IV-B) that

---

[7] Analytically, this can also be easily proved by looking at Equation (18). When $n = 1$ the component has the same form as SLT.

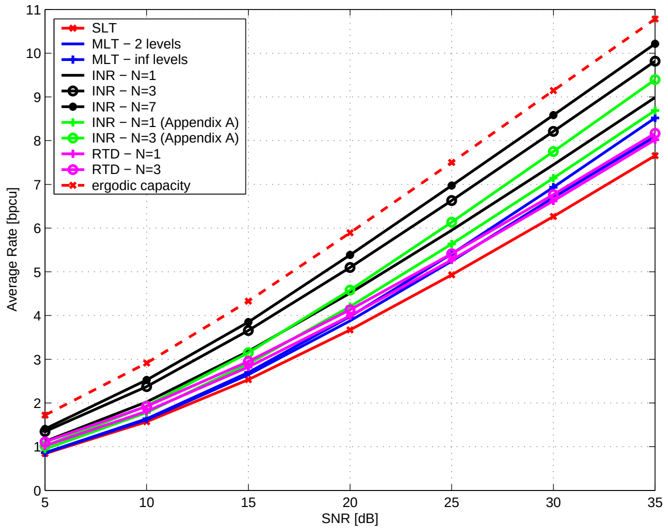


Fig. 2. Average rate (bpcu) versus receive SNR (dB) of SLT, MLT, INR, and RTD in a quasi-static Rayleigh fading channel. Equal power allocation among transmissions is performed. Optimal throughput derived in Appendix A is also plotted.

more packets are transmitted in "good" channels in continuous communication than in burst communication.

Figure 3 compares the optimized average rate performance of SLT, MLT and INR in a slow Ricean fading channel. Two different K factors are considered: Figure 3(a) for $K = 5$ and Figure 3(b) for $K = 10$. Ergodic capacity is plotted as the performance upper bound. As opposed to the Rayleigh fading case, Ricean fading channel has a high-power line-of-sight (LOS) path and thus is "less random". Numerical results show that in the Ricean fading environment, MLT with 2 levels has almost negligible gain over SLT, while INR still performs much better than both SLT and MLT. Combined with its high complexity, this result indicates the inefficiency of MLT in less random channel environment such as Ricean distributions. However, INR continues to perform very well even with $N = 1$, so it is more robust to the channel fading distribution than MLT.

The transmission strategies studied in Section III and IV are for single-antenna systems. These strategies can be readily extended to single-input multiple-output (SIMO) fading channels. Consider a SIMO Rayleigh fading channel with L_r receive antennas. It is shown in [31], [32] that the PDF of the total channel gain is

$$f_{L_r}(g) = \frac{1}{\Gamma(L_r)} g^{L_r - 1} e^{-g}, \qquad g \geq 0 \tag{30}$$

where $\Gamma(L_r) = \int_0^{\infty} t^{L_r - 1} e^{-t} \mathrm{d}t$ is the Gamma function. Figure 4 shows the average rate performance of SLT, MLT, INR, and ergodic capacity of a SIMO Rayleigh fading channel with $L_r = 2$ and $L_r = 4$. Several interesting observations can be made from these plots. First, similar to the previous case, MLT with two levels has negligible gain over SLT: the performance difference is almost indistinguishable. This again seems to suggest that the gain of MLT with two levels is not important in the "less random" fading distributions, especially considering that MLT is much more complicated than SLT. Secondly, the INR scheme performs extraordinary well: the gap between $N = 1$ INR with the ergodic capacity is only 0.6 bits and 0.7 bits at the medium to high SNR regime with $L_r = 2$ and $L_r = 4$, respectively. This complies with our previous observation that INR is robust to the channel fading distribution.

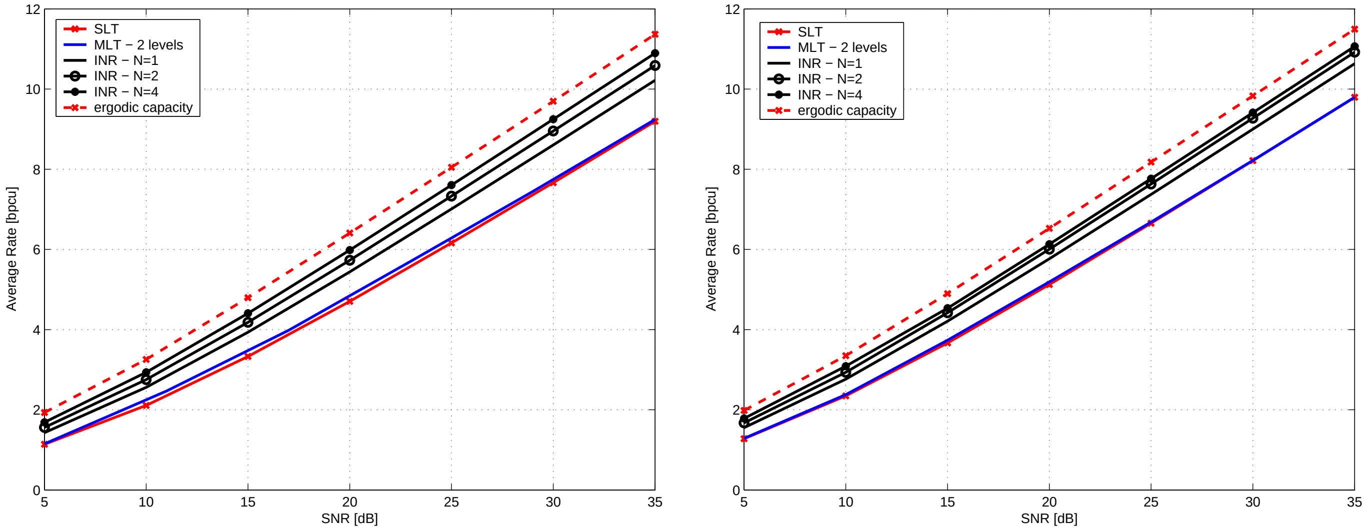
Numerical results for optimal power allocation of INR with $N = 1$ and the comparison to equal power allocation are reported in Figure 5(a) (medium-to-high SNR regime) and Figure 5(b) (low SNR regime) for the Rayleigh fading distribution. The advantage of optimal power allocation is mainly reflected in the low SNR regime. This gain diminishes as SNR increases, and becomes negligible in the medium-to-high SNR regime.

B. *Outage Probability*

The outage probability comparison is reported in Figure 6 for both SLT and INR in a Rayleigh fading channel. It should be noted that the outage probability does not apply to MLT, where the requirement that *all* transmitted data must be decoded is dropped, and thus the concept of "outage" does not hold anymore. Both SLT and INR are still optimized in terms of the average rate. Thus, the outage event of SLT is $\mathbb{P}\left\{\log(1 + gP) < R^*_{SLT,Rayleigh}\right\}$ where $R^*_{SLT,Rayleigh}$ is given in (6), and the outage event of INR with N is $\mathbb{P}\left\{\log(1 + gP) < R^*_1\right\}$ where R^*_1 is the solution of R_1 in problem (14). The advantage of INR is now more obvious: not only does it increase the average rate, it also decreases the outage probability simultaneously. In fact, with the argument made in Section IV-D, the asymptotic outage probability will be zero, as every transmission will eventually match perfectly with the instantaneous channel condition.

C. *Average Rate versus Outage Probability*

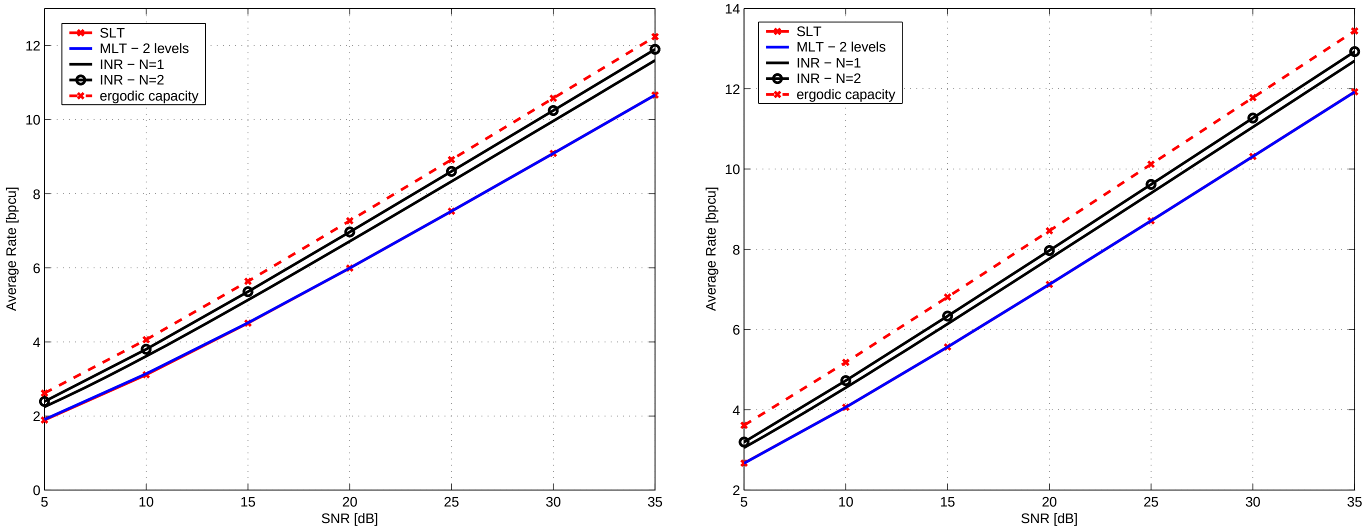
When transmitting over a slow fading channel, the successfully transmitted data rate is a random variable. Its instantaneous value depends on both the instantaneous channel gain and the communication scheme (e.g., SLT, MLT, or HARQ). Roughly speaking, the average rate describes the "mean value" of the random performance, while the outage probability characterizes the "variance" in the sense that it gives the *worst-case* performance. Thus, to have a comprehensive view and comparison of several schemes, both average rate and outage probability should be jointly considered. *Average rate versus outage probability* was proposed in [4] as a meaningful association between average rate and outage probability. This metric requires examining the average rate when the channel gain g is known to exceed some threshold g_{th}. This can be viewed as a conditional average rate where the distribution of channel gain $F(g)$ is replaced with the conditional CDF $F_{g_{th}}(u) \doteq \mathbb{P}\{g \leq u | g \geq g_{th}\}$. Such a conditional average rate examines the average rate with a given outage probability, and thus effectively relates these two metrics. Figure 7 reports the average rate performance where the threshold g_{th} is chosen such that the outage probability is 1% and 30%, respectively. The use of INR still provides remarkable gain with respect to this metric.



(a) K=5

(b) K=10

Fig. 3. Average rate (bpcu) versus receive SNR (dB) of SLT, MLT, and INR in a quasi-static Ricean fading channel with different K factors and equal power allocation among transmissions.



(a) 2 Receive Antennas

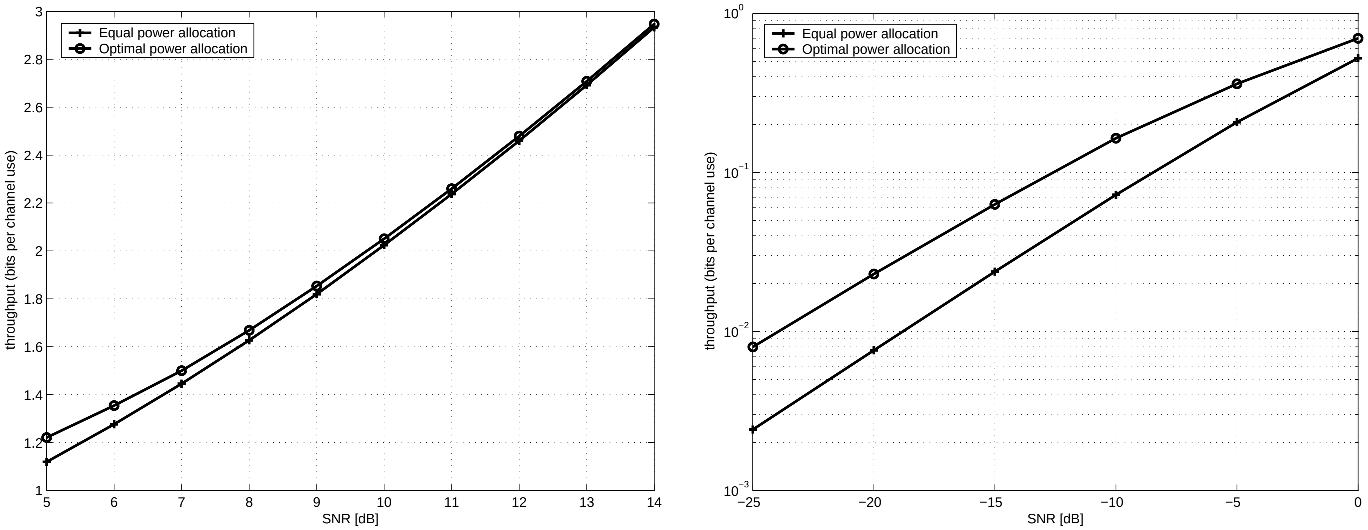
(b) 4 Receive Antennas

Fig. 4. Average rate (bpcu) versus receive SNR (dB) of SLT, MLT, and INR in a quasi-static SIMO Rayleigh fading channel, $L_r = 2$ and $L_r = 4$. Equal power allocation among transmissions is performed.

VI. HOW TO USE THE FEEDBACK CHANNEL: SEQUENTIAL VERSUS ONE-SHOT

From the information-theoretic point of view, the proposed RO-HARQ is one form of utilizing the feedback link in a wireless communication system. It falls into the general category of *quasi-static fading channel with quantized feedback*. The capacity (under different definitions, e.g., outage capacity, expected capacity, etc.) of this channel is still unknown in general. Thus it is difficult to quantify how well RO-HARQ performs in the absolute sense. A reasonable approach would be to compare RO-HARQ with other schemes that utilize the quantized feedback to improve performance, which is the purpose of this section.

One well-known approach is the quantized CSI feedback scheme [19], [26], [27]: the receiver sends an M-bit quantization of the channel state information to the transmitter before the transmission takes place, and the transmitter adjusts its rate and power according to this imperfect CSIT. To be specific, the set of all possible channel gain $G = [0, \infty)$ is divided into 2^M nonoverlapping subsets $G = \bigcup_{i=1}^{2^M} G_i$, where $G_i = [d_{i-1}, d_i)$, $d_0 = 0, d_{2^M} = \infty$. If the instantaneous channel gain $g \in G_i$, the receiver sends the index i to the transmitter using the M-bit feedback channel, and the transmitter selects a codeword with rate R_i and power P_i. Due to the imperfect CSI feedback, the transmission could fail if the data rate R_i is not supported by the instantaneous channel realization. The average rate performance can be optimized



(a) Medium-to-high SNR regime

(b) Low SNR regime

Fig. 5. Average rate (bpcu) versus receive SNR (dB) of $N = 1$ INR with optimal power allocation in a quasi-static Rayleigh fading channel.

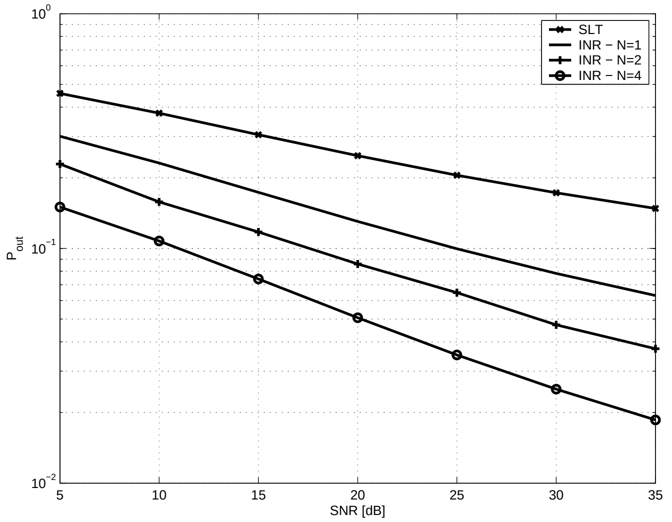


Fig. 6. Outage probability versus receive SNR (dB) of INR (N =1, 2, and 4) compared to SLT in a quasi-static Rayleigh fading channel. The two schemes are optimized in terms of the average rate.

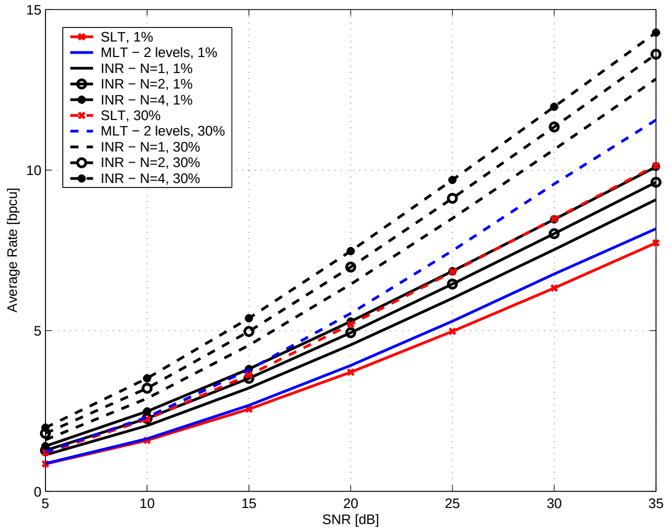


Fig. 7. Average rate versus outage probability for SLT, MLT and INR in a quasi-static Rayleigh fading channel. Two thresholds corresponding to 1% and 30% outage probability, respectively, are simulated.

[19], [26] by adjusting $\{G_i, R_i, P_i\}_{i=1}^{2^M}$.

Compared with this *one-shot feedback* scheme, RO-HARQ is in fact a *sequential feedback* scheme. It is then natural to ask the following question: since sequential and one-shot are different forms of utilizing feedback to inform quantized CSI to the transmitter, what are the advantages/disadvantages of them, and which one is preferred? A few perspectives are discussed in this section.

A. System implementation

ARQ is a technique in the *data link layer*, which is implemented in many wireless protocols. Thus exploiting ARQ feedback does not require additional system implementations, other than that the physical layer channel coding should be able to provide incremental redundancy, which has been well studied [10], [13]–[17], [25]. One-shot CSI feedback, on the other hand, belongs to the *physical layer* techniques. To exploit CSIT, typically the physical layer needs some additional designs, especially at the transmitter, to adjust the transmission based on the one-shot CSI feedback. For example, adaptive coding and modulation requires the data rate and transmit power to be frequently adjusted according to the channel variation. As another example, matrix precoder design is needed to rotate the transmit signal in case of multiple transmit antennas [33]. This increases the system complexity and cost. From this perspective, utilizing ARQ seems to be more favorable, since ARQ is already provided by the wireless protocol in the system and no additional closed-loop design is needed.

B. Position of feedback

In the RO-HARQ scheme, the ACK/NACK feedbacks are scattered into the entire transmission. There are several handshakings between the transmitter and receiver. Each such coordination adds additional overhead. The typical form of the one-shot CSI feedback, however, is to send the entire CSI bits *before* the data transmission takes place, such that the transmitter adjusts the parameters to match the partially-known channel. From the practical point of view, the latter is more favorable, since all the feedback bits are sent in one shot, which simplifies the protocol and reduces the processing delay, system overhead, and the implementation complexity.

C. Amount of feedback and performance

To make a fair comparison of the amount of feedback required for INR and one-shot CSI feedback, it is assumed that both schemes are optimized for maximum average rates, and the maximum average rates are set to be equal. This is reasonable since the resulting comparison indicates that *in order to achieve the same optimal performance, how much feedback is needed for each scheme*. The following lemma gives the relationship between N and M under this condition.

Lemma 3: For any given fading distribution, if the optimal average rates of

(1) INR scheme with maximum N retransmissions,
(2) one-shot CSI feedback scheme with an M-bit feedback channel

are the same, i.e., $\bar{R}_{INR}^{N^*} = \bar{R}_{CSI}^{M^*}$, then

$$N + 1 = 2^M. \tag{31}$$

Proof: See Appendix D. ■

One direct consequence of Lemma 3 is that we can easily make a numerical comparison of the average rate performance between the RO-HARQ scheme and the one-shot CSI feedback scheme studied in [19], [26], [27]. For example, in Figure 2 we have reported the optimal INR performance with $N = 1$, 3 and 7 in a Rayleigh fading channel. These directly correspond to the optimal performance of the one-shot CSI feedback scheme with $M = 1$, 2 and 3 bits of feedback, respectively.

Another important benefit from Lemma 3 is that since the average rate optimization problems for both one-shot CSI feedback and RO-HARQ can be directly related, existing numerical solutions for the previous problem, e.g., [26], can be used to help solve Problem (14) as long as the parameters are chosen according to (31). One should note that solutions in [26] are efficient but sub-optimal. Also they cannot help with the dynamic power allocation problem (27).

From Lemma 3, it seems that INR requires more feedback than the one-shot CSI feedback scheme: with condition (31) satisfied, $M = N$ if and only if $M = N = 1$, and $N > M$ strictly holds in all other cases. However, $N > M$ does not necessarily translate to the conclusion that INR requires more feedback. First of all, the *actual* number of retransmissions associated with INR is a random variable, which is determined by the instantaneous channel realization and the rate assignment. The number N only denotes the *maximum* number of retransmissions, which is a worst-case constraint. On the other hand, M is a *fixed* number in the one-shot CSI feedback scheme. It is not affected by the channel realization or the rate assignment. Secondly, N retransmissions is not necessarily equivalent to N bits feedback, as the number of physical bits required for N retransmissions is determined by many considerations. Wicker [7, Chapter 15.2] discussed several practical issues that might require different number of bits for ACK/NACK. For example, NACK may be set as a default and only ACK is transmitted in the feedback channel. This certainly reduces the feedback of INR. Another example is to use the recently developed *rateless codes* in HARQ [34], such as LT-codes [35] and Raptor codes [36]. With rateless codes, the receiver does not need to send any feedback to the transmitter until it accumulates enough data for successful decoding [37]. In this case the actual amount of feedback is very small.[8]

D. Asymptotic performance

It is seen in Lemma 2 that the ergodic capacity is achieved with $N \to \infty$. On the other hand, it is well known that if the one-shot CSI feedback is perfect (which requires $M \to \infty$), ergodic capacity can also be achieved[9]. Thus, moving the infinitely many bits of feedback from the beginning of transmission as one-shot CSI feedback to within the transmission as ARQ feedback gives the same asymptotic performance.

VII. CONCLUSIONS

HARQ with optimized rate assignment is shown to significantly improve the performance over the single-layer and multi-layer transmission schemes under several different performance metrics. The key idea behind this work is to *exploit* the existence of HARQ protocol to adapt the transmission to the instantaneous channel realization. Average rate is chosen as the performance metric, and the optimal rate assignment is studied. Several aspects of the RO-HARQ scheme are studied, including power allocation, asymptotic performance limit, and comparison to one-shot quantized CSI feedback. Simulation results show that even one HARQ retransmission gives a remarkable gain over conventional schemes, and this gain is robust in different fading distributions.

Possible future work includes the combination of HARQ feedback with the multi-layer scheme. With such combination the ARQ feedback will indicate which layers cannot be decoded, and hence the transmitter can resend those layers together with possibly new information in the next round. Another direction is to study the combination of sequential and one-shot feedback, in which the receiver will send back a *coarse* CSI before the data transmission with one-shot feedback, and use ARQ feedback to *refine* the quantized CSI in the sequel.

---

[8] The sequential feedback scheme fits nicely with the use of rateless code, while the one-shot feedback does not have this advantage.

[9] Note that there is a constraint of *fixed* transmit power; thus waterfilling power allocation over different fading states is not permitted.

APPENDIX A
THROUGHPUT ANALYSIS FOR BURSTY COMMUNICATIONS

A. INR

For each possible fading state g, we assume that K information nats are to be sent with a maximum of N retransmissions, with the length of the n-th transmission T_n, $\forall n = 1, \cdots, N+1$. The rate after transmission n is then $R^{(n)} = K/T^{(n)}$. We need to separately evaluate the numerator and denominator of the throughput expression (2) with respect to the fading distribution. Notice that the overall code length $T^{(N+1)}$ is a function of the fading state.

We begin with the average number of successfully transmitted information nats:

$$\mathbb{E}\{k\} = K\mathbb{P}\left\{\log(1+gP) \geq R^{(N+1)}\right\}. \tag{32}$$

To evaluate the average length of transmission for K information nats, we first define $T^{(0)} = 0$ and thus $R^{(0)} = \infty$. Then

$$\begin{aligned}\mathbb{E}\{\mathcal{T}\} &= T^{(1)}\mathbb{P}\left\{\log(1+gP) \geq R^{(1)}\right\} \\ &\quad + T^{(2)}\mathbb{P}\left\{R^{(1)} > \log(1+gP) \geq R^{(2)}\right\} \\ &\quad + \cdots \\ &\quad + T^{(N+1)}\mathbb{P}\left\{R^{(N)} > \log(1+gP)\right\} \\ &= \sum_{n=1}^{N} T^{(n)}\mathbb{P}\left\{R^{(n-1)} > \log(1+gP) \geq R^{(n)}\right\} \\ &\quad + T^{(N+1)}\mathbb{P}\left\{R^{(N)} > \log(1+gP)\right\}. \end{aligned} \tag{33}$$

The throughput of INR can now be written as Equation (34) at the top of the next page, and the INR throughput optimization problem is formulated as

$$\begin{aligned} \text{maximize} \quad & \frac{1-F_G(g_{N+1})}{\sum_{n=1}^{N}\left(\frac{1}{R^{(n+1)}}-\frac{1}{R^{(n)}}\right)F_G(g_n)} \\ \text{subject to} \quad & R^{(n)} \geq 0; \forall n = 1, \cdots, N+1 \\ & R^{(n)} \geq R^{(n+1)}; \forall n = 0, \cdots, N. \end{aligned} \tag{35}$$

Numerical results of this optimization problem for $N = 1$ and 3 are reported in Figure 2.

B. RTD

With the fact that Chase Combining effectively increases the receive SNR to ngP and reduces the overall rate to R/n after the n-th transmission, we can similarly compute the throughput of RTD.

Numerator:

$$\begin{aligned}\mathbb{E}\{k\} &= K\mathbb{P}\left\{\log(1+(N+1)gP) \geq R\right\} \\ &= K\left(1-F_G(q_{N+1})\right)\end{aligned} \tag{36}$$

Denominator:

$$\begin{aligned}&\mathbb{E}\{\mathcal{T}\} = \textstyle\sum_{n=1}^{N} nT\mathbb{P}\{\log(1+ngP) \geq R > \\ &\log(1+(n-1)gP)\} + (N+1)T\mathbb{P}\left\{R > \log(1+NgP)\right\} \\ &= \textstyle\sum_{n=1}^{N} nT\mathbb{P}\left\{q_{n-1} \geq g > q_n\right\} + (N+1)T\mathbb{P}\left\{q_N \geq g\right\} \\ &= \textstyle\sum_{n=1}^{N} nT\left[F_G(q_{n-1}) - F_G(q_n)\right] + (N+1)TF_G(q_N) \\ &= T\textstyle\sum_{n=0}^{N} F_G(q_n)\end{aligned} \tag{37}$$

RTD throughput:

$$\eta_{RTD}^{N+1} = \frac{R\left(1-F_G(q_{N+1})\right)}{\sum_{n=0}^{N} F_G(q_n)} = \frac{R\left(1-F_G(q_{N+1})\right)}{1+\sum_{n=1}^{N} F_G(q_n)} \tag{38}$$

RTD throughput optimization problem:

$$\text{maximize} \quad \frac{R\left(1-F_G(q_{N+1})\right)}{1+\sum_{n=1}^{N} F_G(q_n)}. \tag{39}$$

APPENDIX B
PROOF OF LEMMA 1

The proof is done by showing that the optimal average rate of RTD can be achieved by INR. For the simplicity of discussion in this appendix, we re-denote the boundary points of each fading region for INR and RTD to be $a_n = \frac{e^{\sum_{k=1}^{N+2-n} R_k}-1}{P}$, $b_n = \frac{e^R-1}{nP}$, $\forall n = 1, \cdots, N+1$, and $a_0 = b_0 \doteq \infty$, respectively. Let us rewrite the average rate expression for INR and RTD as:

$$\bar{R}_{INR}^{N}(R_1, \cdots, R_{N+1}) = \sum_{n=1}^{N+1}\left(\sum_{i=1}^{N+2-n} R_i\right)\mathbb{P}\left\{g \in [a_n, a_{n-1})\right\}, \tag{40}$$

$$\bar{R}_{RTD}^{N}(R) = \sum_{n=1}^{N+1}\frac{R}{n}\mathbb{P}\left\{b_{n-1} > g \geq b_n\right\}. \tag{41}$$

Consider the optimal solution for RTD: $R = R^*$, $\bar{R}_{RTD}^{N}(R) = \bar{R}_{RTD}^{N^*}$. Denote the corresponding optimal boundary points of the fading regions as

$$b_n^* = \frac{e^{R^*}-1}{nP} \tag{42}$$

for $n = 1, \cdots, N+1$ and $b_0^* = b_0$. Then choose $(R_1, \cdots, R_{N+1})$ such that

$$a_n = b_n^*, \forall n = 1, \cdots, N+1. \tag{43}$$

Condition (43) forces the event $\{a_{n-1} > g \geq a_n\}$ to be equivalent to $\left\{b_{n-1}^* > g \geq b_n^*\right\}$ for $n = 1, \cdots, N+1$. Thus, the proof is set if we can show that

$$\sum_{i=1}^{N+2-n} R_i \geq \frac{R^*}{n}, \qquad \forall n = 1, \cdots, N+1. \tag{44}$$

Condition (43) can be equivalently written as

$$\sum_{i=1}^{N+2-n} R_i = \log\frac{e^{R^*}+n-1}{n}. \tag{45}$$

We need the following simple lemma.

Lemma 4: The function

$$f(n) = n\left(e^{\frac{c}{n}}-1\right), \qquad n \geq 1. \tag{46}$$

is monotonically decreasing, where $c > 0$ is a constant.

This can be easily proved by checking that $\frac{\mathrm{d}f(n)}{\mathrm{d}n} > 0$ for positive and finite n. With Lemma 4 and Equation (45), the following inequalities can be obtained:

$$\begin{aligned}\text{Lemma 4} \quad &\Longrightarrow \quad e^{R^*}-1 \geq n\left(e^{\frac{R^*}{n}}-1\right) \\ &\Longrightarrow \quad \log\frac{e^{R^*}+n-1}{n} \geq \frac{R^*}{n} \\ &\overset{(45)}{\Longrightarrow} \quad (44)\end{aligned}$$

which concludes the proof.

$$\begin{aligned}
\eta_{INR}^{N+1} &= \frac{K\mathbb{P}\left\{\log\left(1+gP\right) \geq R^{(N+1)}\right\}}{\sum_{n=1}^{N} T^{(n)}\mathbb{P}\left\{R^{(n-1)} > \log\left(1+gP\right) \geq R^{(n)}\right\} + T^{(N+1)}\mathbb{P}\left\{R^{(N)} > \log\left(1+gP\right)\right\}} \\
&= \frac{\mathbb{P}\left\{\log\left(1+gP\right) \geq R^{(N+1)}\right\}}{\sum_{n=1}^{N} \frac{1}{R^{(n)}}\mathbb{P}\left\{R^{(n-1)} > \log\left(1+gP\right) \geq R^{(n)}\right\} + \frac{1}{R^{(N+1)}}\mathbb{P}\left\{R^{(N)} > \log\left(1+gP\right)\right\}} \\
&= \frac{1-F_G(g_{N+1})}{\sum_{n=1}^{N} \frac{1}{R^{(n)}}\left[F_G(g_{n-1}) - F_G(g_n)\right] + \frac{1}{R^{(N+1)}}F_G(g_N)} \\
&= \frac{1-F_G(g_{N+1})}{\sum_{n=1}^{N} \left(\frac{1}{R^{(n+1)}} - \frac{1}{R^{(n)}}\right)F_G(g_n)} \qquad (34)
\end{aligned}$$

APPENDIX C
PROOF OF LEMMA 2

The average rate of the INR scheme with a fixed N is given in Equation (12) as

$$\bar{R}_{INR}^{N} = \sum_{n=1}^{N+1} R_n \mathbb{P}\left\{\log\left(1+gP\right) \geq \sum_{k=1}^{n} R_k\right\}.$$

As $N \to \infty$, this average rate converges to

$$\begin{aligned}
\bar{R}_{INR}^{\infty} &= \int_0^{\infty} \mathbb{P}\left\{\log\left(1+gP\right) \geq R\right\} \mathrm{d}R \\
&= \mathbb{E}\left[\log\left(1+gP\right)\right] \qquad (47)
\end{aligned}$$

where Equation (47) is due to the fact that

$$\mathbb{E}\left[X\right] = \int_0^{\infty} \left(1 - F_X(x)\right) \mathrm{d}x$$

for a nonnegative random variable X.

APPENDIX D
PROOF OF LEMMA 3

The average rate of the INR scheme with maximum N retransmissions is given in Equation (12). By defining $r_n^* = \sum_{i=1}^{n} R_i^*$ for $n = 1, \cdots, N+1$, the maximum average rate of INR can be rewritten as

$$\bar{R}_{INR}^{N^*} = \sum_{n=1}^{N+1} r_n^* \mathbb{P}\left\{r_{n+1}^* > \log\left(1+gP\right) \geq r_n^*\right\}. \qquad (48)$$

For the CSI feedback scheme defined in Section VI, the average rate can be computed as

$$\bar{R}_{CSI}^{M} = \sum_{n=1}^{2^M} R_n \mathbb{P}\left\{\log\left(1+d_n P\right) > \log\left(1+gP\right) \geq R_n\right\}. \qquad (49)$$

It is observed in [19] that in order to maximize the average rate of the CSI feedback scheme, R_n must satisfy

$$R_n^* = \log\left(1 + d_{n-1}^* P\right). \qquad (50)$$

Substituting (50) into (49), the optimal average rate of the CSI feedback scheme is

$$\bar{R}_{CSI}^{M^*} = \sum_{n=1}^{2^M} R_n^* \mathbb{P}\left\{R_{n+1}^* > \log\left(1+gP\right) \geq R_n^*\right\}. \qquad (51)$$

Compare (51) to (48), the result $N+1 = 2^M$ follows immediately.

REFERENCES

[1] L. H. Ozarow, S. Shamai, and A. D. Wyner, "Information theoretic considerations for cellular mobile radio," *IEEE Trans. Veh. Tech.*, vol. 43, no. 2, pp. 359–378, May 1994.

[2] T. M. Cover, "Comments on broadcast channels," *IEEE Trans. Info. Theory*, vol. 44, no. 6, pp. 2524–2530, Oct. 1998.

[3] S. Shamai, "A broadcast strategy for the Gaussian slowly fading channel," in *Proceedings of International Symposium on Information Theory*, Jun.-Jul. 1997, p. 150.

[4] S. Shamai and A. Steiner, "A broadcast approach for a single-user slowly fading MIMO channel," *IEEE Trans. Info. Theory*, vol. 49, no. 10, pp. 2617–2635, Oct. 2003.

[5] Y. Liu, V. K. N. Lau, O. Y. Takeshita, and M. P. Fitz, "Optimal rate allocation for superposition coding in quasi-static fading channels," in *Proceedings of International Symposium on Information Theory*, Jun.-Jul. 2002, p. 111.

[6] D. J. Costello, J. Hagenauer, H. Imai, and S. B. Wicker, "Applications of error-control coding," *IEEE Trans. Info. Theory*, vol. 44, no. 6, pp. 2531–2560, Oct. 1998.

[7] S. Wicker, *Error Control Systems for Digital Communication and Storage*. Englewood Cliffs, NJ: Prentice-Hall, 1995.

[8] H. Zheng and H. Visvanathan, "Optimizing ARQ performance in high speed downlink systems with scheduling," *IEEE Trans. Wireless Commun.*, vol. 4, no. 2, pp. 495–506, Mar. 2005.

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

[10] S. Sesia, G. Caire, and G. Vivier, "Incremental redundancy hybrid ARQ schemes based on low-density parity-check codes," *IEEE Trans. Commun.*, vol. 52, no. 8, pp. 1311–1321, Aug. 2004.

[11] H. El Gamal, G. Caire, and M. O. Damen, "The MIMO ARQ Channel: diversity-multiplexing-delay tradeoff," *IEEE Trans. Info. Theory*, vol. 52, no. 8, pp. 3601–3621, Aug. 2006.

[12] A. Chuang, A. G. i Fabregas, L. Rasmussen, and I. Collings, "Optimal throughput-diversity-delay tradeoff in MIMO ARQ block-fading channels," *IEEE Trans. Info. Theory*, vol. 54, no. 9, pp. 3968–3986, Sept. 2008.

[13] J. Li and K. Narayanan, "Rate-compatible low density parity check (RC-LDPC) codes for capacity-approaching ARQ schemes in packet data communications," in *Proceedings of International Conference on on Communications, Internet and Information Technology (CIIT)*, US Virgin Islands, Nov. 2002, pp. 201–206.

[14] J. Ha, J. Kim, and S. W. McLaughlin, "Rate-compatible puncturing of low-density parity-check codes," *IEEE Trans. Info. Theory*, vol. 50, no. 11, pp. 2824–2836, Nov. 2004.

[15] J. Kim, W. Hur, A. Ramamoorthy, and S. W. McLaughlin, "Design of rate-compatible irregular LDPC codes for incremental redundancy hybrid ARQ systems," in *Proceedings of International Symposium on Information Theory*, Jul. 2006, pp. 1139–1143.

[16] J. Kim, A. Ramamoorthy, and S. W. McLaughlin, "Design of efficiently-encodable rate-compatible irregular LDPC codes," in *Proceedings of IEEE International Conference on Communications*, Jun. 2006, pp. 1131–1136.

[17] J. Ha, J. Kim, D. Klinc, and S. W. McLaughlin, "Rate-compatible punctured low-density parity-check codes with short block lengths," *IEEE Trans. Info. Theory*, vol. 52, no. 2, pp. 728–738, Feb. 2006.

[18] M. Effros, A. Goldsmith, and Y. Liang, "Capacity definitions of general channels with receiver side information," in *Proceedings of International Symposium on Information Theory*, Jun. 2007, pp. 921–925.

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

[20] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth, "On the Lambert W function," *Advances in Computational Mathematics*, vol. 5, pp. 329–359, 1996.

[21] X. Wang and T. Orchard, "Design of superposition coded modulation for unequal error protection," in *Proceedings of IEEE International Conference on Communications*, June 2001, pp. 412–416.

[22] D. Tuninetti, "Transmitter channel state information and repetition protocols in block fading channels," in *Proceedings of IEEE Information Theory Workshop*, Sept. 2007, pp. 505–510.

[23] D. Chase, "Code combining–a maximum-likelihood decoding approach for combining an arbitrary number of noisy packets," *IEEE Trans. Commun.*, vol. 33, no. 5, pp. 385–393, May 1985.

[24] N. Varnica, E. Soljanin, and P. Whiting, "LDPC code ensembles for incremental redundancy hybrid ARQ," in *Proceedings of International Symposium on Information Theory*, Sept. 2005, pp. 995–999.

[25] TrellisWare Technologies, *White papers on FlexiCodes and F-LDPC*. see: www.trellisware.com.

[26] L. Lin, R. D. Yates, and P. Spasojevic, "Adaptive transmission with discrete code rates and power levels," *IEEE Trans. Commun.*, vol. 51, no. 12, pp. 2115–2125, Dec. 2003.

[27] ——, "Adaptive transmission with finite code rates," *IEEE Trans. Info. Theory*, vol. 52, no. 5, pp. 1847–1860, May 2006.

[28] T. Kim and M. Skoglund, "Diversity-Multiplexing tradeoff in MIMO channels with partial CSIT," *IEEE Trans. Info. Theory*, vol. 53, no. 8, pp. 2743–2759, Aug. 2007.

[29] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley Interscience, 1992.

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

[31] A. Edelman, "Eigenvalues and condition numbers of random matrices," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, 1989.

[32] E. Telatar, "Capacity of multi-antenna Gaussian channels," AT&T–Bell Labs, Tech. Rep., June 1995.

[33] M. Vu, "Exploiting transmit channel side information in MIMO wireless systems," Ph.D. dissertation, Stanford University, Stanford, CA, 2006.

[34] E. Soljanin, N. Varnica, and P. Whiting, "Punctured vs rateless codes for hybrid ARQ," in *Proceedings of IEEE Information Theory Workshop*, Punta del Este, Uruguay, Mar. 2006.

[35] M. Luby, "LT codes," in *Proceedings of The 43rd Annual IEEE Symposium on Foundations of Computer Science*, 2002, pp. 271–282.

[36] A. Shokrollahi, "Raptor codes," *IEEE Trans. Info. Theory*, vol. 52, no. 6, pp. 2551–2567, Jun. 2006.

[37] Y. Liu, "A lossless layered coding strategy for MIMO systems," in *Proceedings of International Symposium on Information Theory*, Sept. 2005, pp. 1947–1951.

Cong Shen (S'01) received the B.S. and M.S. degrees, in 2002 and 2004 respectively, from the Department of Electronic Engineering, Tsinghua University, Beijing, China. He is currently working towards the Ph.D. degree in the Electrical Engineering Department, University of California, Los Angeles (UCLA). His research interest is on general communication theory with emphasis on wireless communications.

Tie Liu received his B.S. (1998) and M.S. (2000) degrees, both in Electrical Engineering, from the Tsinghua University, Beijing, China and M.S. degree in Mathematics (2004) and Ph.D. degree in Electrical and Computer Engineering (2006) from the University of Illinois at Urbana-Champaign. Since August 2006, he has been with the Texas A&M University where he is currently an Assistant Professor in Electrical and Computer Engineering. His research interests are in the field of information theory, wireless communication, and statistical signal processing. He is a recipient of the M. E. Van Valkenburg Graduate Research Award (2006) from the University of Illinois at Urbana-Champaign and the Best Paper Award (2008) from the Third International Conference on Cognitive Radio Oriented Wireless Networks and Communications.

Michael P. Fitz (S'82-M'83-SM'02) received the B.E.E. degree (summa cum laude) from the University of Dayton, Dayton, OH, in 1983 and the M.S. and Ph.D. degrees in electrical engineering from the University of Southern California, Los Angeles, in 1984 and 1989, respectively.

From 1983 to 1989, he was a Communication Systems Engineer with Hughes Aircraft and TRW Inc. Since 1989, he has been with the Faculty of Purdue University, West Lafayette, IN, The Ohio State University (OSU), Columbus, and the University of California, Los Angeles. He is currently with Northrop Grumman Corp. as a Senior Systems Engineer working on satellite communications. His research is in the broad area of statistical communication theory and experimentation. He is the author of *Fundamentals of Communications Systems* (New York: McGraw Hill, 2007). His research group at UCLA currently is interested in the theory of space-time modems and operates an experimental wireless wide-area network and a space-time coding testbed.

Prof. Fitz received the 2001 IEEE Communications Society Leonard G. Abraham Prize Paper Award in the field of communications systems.