The Diversity-Multiplexing-Delay Tradeoff in MIMO ARQ Channels

Hesham El Gamal, Giuseppe Caire, and Mohamed Oussama Damen

Abstract— **In this paper, we explore the fundamental performance tradeoff of the delay-limited Multi-Input-Multi-Output (MIMO) Automatic Retransmission reQuest (ARQ) channel. In particular, we extend the diversity-multiplexing tradeoff investigated by Zheng and Tse in standard delay-limited MIMO channels with coherent detection to the ARQ scenario. We establish the three-dimensional tradeoff between reliability (i.e. diversity), throughput (i.e., multiplexing gain), and delay (i.e., maximum number of retransmissions). This tradeoff quantifies the ARQ diversity gain obtained by leveraging the retransmission delay to enhance the reliability for a given multiplexing gain. Interestingly, ARQ diversity appears even in long-term static channels where all the retransmissions take place in the same channel state. Furthermore, by relaxing the input power constraint allowing variable power levels in different retransmissions, we show that power control can be used to dramatically increase the diversity advantage. Our analysis reveals some important insights on the benefits of ARQ in slow fading MIMO channels. In particular, we show that: 1) allowing for a sufficiently large retransmission delay results in an *almost* flat diversity-multiplexing tradeoff, and hence, renders operating at high multiplexing gain more advantageous; 2) MIMO ARQ channels quickly approach the ergodic limit when power control is employed.**

I. Introduction

In this work, we extend the diversity-multiplexing, characterized by Zheng and Tse in standard coherent channels [1], to Automatic Retransmission reQuest (ARQ) MIMO channels. In this case, the receiver feeds back to the transmitter a one bit success/failure indicator. In the success case, the transmitter moves on to the next information message in the transmission queue whereas in the failure case the transmitter retransmits a (possibly different) encoded version of the same message. We refer to the successive transmissions of coded versions of the *same* information message as "ARQ protocol rounds". The ARQ protocol is allowed to use a given maximum number of rounds, denoted by L. If after L rounds no successful decoding has occurred, an error is declared. In this case, we assume that the message will be dropped from the transmission queue (i.e., delay sensitive application). Therefore, we define the probability of error as the probability of no successful decoding within L protocol rounds. We investigate and completely characterize the three dimensional diversity-multiplexing-delay tradeoff in MIMO ARQ channels[1]. This tradeoff establishes, rigorously, the fact that the ARQ re-transmission delay can be exploited as a potential source for diversity. We investigate two extreme cases of channel dynamics: long-term and short-term static channels. In the long-term static case, the MIMO channel matrix is assumed to be constant over all the ARQ rounds. This scenario applies to very fast ARQ protocols and/or very slow fading environments, such as wireless LANs [2]. In the short-term static case, the MIMO channel matrix is constant over each transmission round of the ARQ protocol but changes independently from round to round. This scenario applies to slow ARQ protocols where the time between the consecutive rounds is larger than the channel coherence time, or to frequency-selective fading, where each ARQ transmission takes place at a different frequency according to some frequency hopping scheme.

Hesham El Gamal is with the ECE Department at the Ohio State University. Giuseppe Caire is with The Mobile Communication group at Eurecom Institute. Mohamed Oussama Damen is with the ECE Department at the University of Waterloo. The work of Hesham El Gamal was funded in part by the National Science Foundation under Grants CCR 0118859, ITR 0219892, and CAREER 0346887.

II. System Model and Background

We consider a frequency-flat fading M-transmit N-receive multiple-input multiple-output (MIMO) channel with no CSI at the transmitter and perfect CSI at the receiver. The following ARQ protocol is considered. The transmitter has an infinite buffer of information messages to send. The information message to be transmitted is encoded by a space-time encoder, and mapped into a sequence of L matrices, or blocks, $\{\mathbf{X}_\ell^c \in \mathbb{C}^{M\times T} \ : \ \ell = 1,\ldots,L\}$. The transmission of each block takes T channel uses, by transmitting the matrix columns in parallel over the M transmit antennas, as in standard space-time coding. At the ℓ-th round of the current information message, $\mathbf{X}_\ell^c$ is transmitted. The decoder is allowed to process the received signal over all the ℓ received blocks, in order to decode the message. If successful decoding is detected, a positive acknowledgement signal (ACK) is sent back to the transmitter whereas a negative acknowledgement

[1]Here, delay refers to the maximum number of transmission rounds L of the ARQ protocol.

(NACK) signal is sent in case of detection of a decoding failure. The ACK/NACK one-bit message is the only feedback allowed in our model and the ARQ feedback channel is assumed to be error-free and zero-delay. Upon reception of the ACK, the transmitter sends the first block of the next message in the buffer whereas the reception of the NACK triggers the transmission of the next block of the current message, $\mathbf{X}_{\ell+1}^c$. The only exception to the above rule is when the maximum number of protocol rounds, L, is reached. In this case, a NACK bit will be interpreted as an error, the current message is removed from the transmission buffer and the transmission of the next message is started anyway. Error in the system occur either when the decoder makes a decoding error at round $\ell < L$ and it fails to detect it (undetected error event) or when the decoder makes a decoding error at round L. The complex baseband model of our channel is defined by

$$\mathbf{y}_{\ell,t}^c = \sqrt{\frac{\rho}{M}}\mathbf{H}_\ell^c \mathbf{x}_{\ell,t}^c + \mathbf{w}_{\ell,t}^c, \tag{1}$$

where the index $\ell = 1, 2, \ldots,$ counts the protocol rounds and $t = 1, \ldots, T$ counts the channel uses in each block, $\{\mathbf{x}_{\ell,t}^c \in \mathbb{C}^M : t = 1, \ldots, T\}$ are the columns of the ℓ-th block $\mathbf{X}_\ell^c$, $\{\mathbf{w}_{\ell,t}^c \in \mathbb{C}^N : t = 1, \ldots, T\}$ and $\{\mathbf{y}_{\ell,t}^c \in \mathbb{C}^N : t = 1, \ldots, T\}$ denote the channel noise and the corresponding received signal block, respectively. The channel noise is assumed to be temporally and spatially white with i.i.d. entries $\sim \mathcal{N}_{\mathbb{C}}(0, 1)$. The channel in the ℓ-th round is characterized by the matrix $\mathbf{H}_\ell^c \in \mathbb{C}^{N\times M}$ with the (i, j)-th element $h_{ij,\ell}^c$ representing the fading coefficient between the j-th transmit and the i-th receive antenna. The fading coefficients are assumed to be i.i.d. $\sim \mathcal{N}_{\mathbb{C}}(0, 1)$ and remain fixed over each block, for $t = 1, \ldots, T$.

As anticipated in the Introduction, we consider two distinct scenarios of channel dynamics: 1) long-term static channels, where the channel coefficients remain constant during all L rounds; 2) short-term static channels, where the channel remains constant during each round and changes independently at each round. In the long-term static case, $\mathbf{H}_\ell^c = \mathbf{H}^c$ (independent of ℓ) for all $\ell = 1, \ldots, L$. Also, we consider two different input power constraints: 1) short-term (or per-block) average power constraint; 2) long-term average power constraint. In the first case, we enforce

$$\mathbb{E}\left[\frac{1}{T}\|\mathbf{X}_\ell^c\|_F^2\right] \leq M, \tag{2}$$

for all $\ell = 1, \ldots, L$, where expectation is with respect to the uniform probability measure over the codebook. This means that the average transmitted power in each round of the ARQ protocol is the same, irrespective of the round index ℓ. In the second case, we enforce

$$\limsup_{\tau\to\infty} \mathbb{E}\left[\frac{1}{T\tau}\sum_{s=1}^{\tau}\|\mathbf{X}^c[s]\|_F^2\right] \leq M \tag{3}$$

where we have introduced the *absolute* index, s, of the transmitted block [2], and now $\mathbf{X}^c[s]$ denotes the s-th transmitted block since the beginning of transmission. Again, expectation is with respect to the uniform probability measure over the codebook. Clearly, in both cases the parameter ρ in (1) takes on the meaning of *average* signal-to-noise ratio (SNR) per receiver antenna.

Let b denote the size of the information messages in bits and let $B[s]$ denote the number of bits removed from the transmission buffer at slot s (absolute index). We have that $B[s] = b$ if the renewal event occurs at time s, and $B[s] = 0$ otherwise. The long-term average throughput of the ARQ protocol, expressed in *transmitted* bits per channel use (PCU), is given by [3]

$$\eta \quad = \quad \liminf_{\tau\to\infty} \frac{1}{T\tau}\sum_{s=1}^{\tau} B[s]. \tag{4}$$

The long term power constraint in (3) applies to any feasible power control rule including non-stationary and randomized algorithms. In the sequel, however, we shall restrict ourselves to the class of stationary power control policies, for which the power spent at round ℓ is just a deterministic function of ℓ. In this context, we let Γ_ℓ denote the average energy allocated to the ℓ-th round of transmission. The ARQ system incurs an error if decoding fails but it is not detected, so that an ACK is fed back, or if decoding fails at round L. Let $\mathcal{E}_\ell$ denote the event that the decoding outcome is not correct with ℓ received blocks. In order to extend Zheng-Tse formulation of the diversity-multiplexing tradeoff to the ARQ case, we consider a family of ARQ protocols where the size of the information messages $b(\rho)$ depends on the operating SNR ρ. These protocols are based on a family of space-time codes $\{\mathcal{C}_\rho\}$ with first-block rate $R_1(\rho) = b(\rho)/T$ and overall block length TL. Then, we define the *effective* ARQ multiplexing gain as

$$r_e \triangleq \lim_{\rho\to\infty} \frac{\eta(\rho)}{\log\rho} \tag{5}$$

[2] Notice that ℓ is a relative index, denoting the ℓ-th block in the transmission of the current message.

where $\eta(\rho)$ is given by (4). The *effective* ARQ diversity gain is defined as

$$d = -\lim_{\rho\to\infty} \frac{\log P_e(\rho)}{\log \rho} \tag{6}$$

where $P_e(\rho)$ is the average probability of error with SNR equal to ρ.

The optimal diversity-multiplexing tradeoff of MIMO ARQ channels yields the maximum possible SNR exponent, denoted by $d^*(r_e, L)$, for every value of r_e. As a consistency check, it is immediate to verify that these definitions reduce to the standard Zheng-Tse formulation when $L = 1$ (i.e., no ARQ). Due to space limitation, we omit all the proofs and most of the technical details in this paper and refer the interested reader to [4].

III. The Fundamental Tradeoff

Theorem 1: The optimal diversity gain of the coherent block-fading MIMO ARQ channel with M transmit, N receive antennas, maximum number of ARQ rounds L, effective multiplexing gain $0 \le r_e < \min(N, M)$, and $T \to \infty$, under the short-term power constraint, is given by:

$$d^*_{ls}(r_e, L) = f\left(\frac{r_e}{L}\right) \tag{7}$$

in the case of long-term static channels, and

$$d^*_{ss}(r_e, L) = Lf\left(\frac{r_e}{L}\right) \tag{8}$$

in the case of long-term static channels. Here $f(\cdot)$ is the piecewise linear function joining the points $(k, (M-k)(N-k))$ for $k = 0, \ldots, \min\{M, N\}$.

Theorem 1 establishes the interesting fact that retransmission delay can be exploited to significantly improve diversity, especially at high multiplexing gain. The basic idea is that the multiplexing gain is determined by the rate assuming only one round whereas the diversity gain is determined by the rate of the *composite* code received at the end of the maximum number of rounds. This can be explained by the fact that *most* packets are decoded successfully in the first round and ARQ retransmissions are used to correct the *rare* error events, and hence, pushing the probability of error down with an asymptotically vanishing price in the transmission rate. Interestingly, the ARQ diversity gain appears even in long term static channels. In fact, as shown in Fig. 1, one can approach the full diversity $d = MN$ for any multiplexing gain $r_e < \min(N, M)$ by allowing for sufficiently large maximum delay L. It is important to notice here that, in this scenario, larger values of L do not imply any increase of the *temporal diversity* (i.e., each codeword is still transmitted over a single realization of the channel matrix) and have an asymptotically vanishing effect on the *average* delay. It is also evident that, in long-term static channels, the diversity improvement due to ARQ disappears as the multiplexing gain tends to zero. In fact, we have $d^*_{ls}(0, L) = d^*(0, 1) = NM$, irrespectively of L. On the other hand, in short-term static channels ARQ provides also temporal diversity, as seen in the fact that $d^*_{ss}(r_e, L) = Ld^*_{ls}(r_e, L)$. This temporal diversity gain appears at both low and high multiplexing gains.

In Theorem 1, the achievability of the exponents $d^*_{ls}(r_e, L)$ and $d^*_{ss}(r_e, L)$ is shown in the limit of large block length. The proof hinges on the use of an *incomplete* decoder, namely, the typical set decoder, which has a built-in error detection capability. It is of practical interest to assess how large T must be in order to achieve the optimal tradeoff. In most practical ARQ schemes, optimal ML decoding is used. Since ML decoding always yields a codeword, decoding errors are detected by using an outer coding layer devoted to error detection (typically, a cyclic redundancy check (CRC)). Unfortunately, as argued in [4], the "CRC" approach requires T growing to infinity in order to operate at the optimal tradeoff. This is because the undetected error probability must go to zero as a power of SNR with the correct exponent, in order to not dominate the overall error probability, and this requires a number of CRC bits that increases linearly with log SNR, so that for any fixed block length T the CRC yields a fixed loss in the multiplexing gain (the pre-log factor of the rate). Driven by the observation, we investigate the achievability of the optimal tradeoff for finite T by using an incomplete bounded-distance decoder that mimics the behavior of the typical set decoder. In particular, we consider a decoder that accepts the message $\widehat{w}$ at round ℓ if: 1) the channel is not in outage; 2) the corresponding codeword $\widehat{\mathbf{x}}$ is the unique codeword such that $|\mathbf{y}_\ell - \mathbf{H}_\ell\widehat{\mathbf{x}}|^2 \le NT\ell(1+\delta)$ for some $\delta > 0$ (which will be determined in the sequel). On the contrary, if either there is no such codeword or there are more than one then a NACK is fed back. Since the noise $\mathbf{w}_\ell$ has dimension $2NT\ell$ and it is Gaussian i.i.d. with components $\sim \mathcal{N}(0, 1/2)$, the above condition is equivalent to say that the noise is typical *and* the channel is not in outage. The term δ will be required to grow with the SNR in order to ensure that, despite the finite block length, the probability that the noise is outside the sphere of squared radius $NT\ell(1+\delta)$ vanishes with an SNR exponent at least equal to $d^*(L, r_e)$. This result

is summarized by the following:

Theorem 2: The optimal diversity $(d^*_{ls}(r_e, L), d^*_{ss}(r_e, L))$ of the coherent block-fading MIMO ARQ channel with M transmit, N receive antennas, maximum number of ARQ rounds L and effective multiplexing gain $0 \leq r_e < \min(N, M)$ can be achieved by codes with finite block length T subject to the conditions

1. $T \geq \left\lceil \frac{M+N-1}{L} \right\rceil$ for long-term static channels.
2. $T \geq M + N - 1$ for short-term static channels.

Now, we consider the long-term power constraint and construct an asymptotically optimal power control algorithm that yields very significant diversity advantage in long-term static channels especially at low multiplexing gains. A distinguishing feature of the proposed algorithm is that it avoids the non-causal feedback assumptions adopted in many earlier works. The proposed power control algorithm is enabled by the observation that the probability of transmitting the ℓ round decays polynomially with SNR. Therefore, the energy allocated to the ℓ-th block, Γ_ℓ, can be made inversely proportional to the probability of transmitting the ℓ round, allowing for a significant increase in transmitted power without violating the long-term power constraint. The larger power level in round ℓ will result in a smaller probability of transmitting the $\ell + 1$ round, and hence, even larger $\Gamma_{\ell+1}$. Through this recursive procedure, the probability of error is minimized (since this section treats only the more interesting long-term static channels, we drop the subscript "ls" in the following for brevity).

Theorem 3: The optimal diversity gain of the coherent block-fading MIMO ARQ channel with M transmit, N receive antennas, maximum number of ARQ rounds L and effective multiplexing gain $0 \leq r_e < \min(N, M)$, under the long-term power constraint, is given by $d^*(r_e, L) = \xi_L$, where ξ_L is obtained recursively as follows. Let $\xi_0 = 0$. For $\ell = 1, \ldots, L$, let

$$\xi_\ell = \inf_{\mathbf{v} \,\in\, \mathcal{O}_\ell \cap \mathbb{R}_+^{\min\{M,N\}}} \left\{ \sum_{j=1}^{\min\{M,N\}} (2j - 1 + |M - N|)\, v_j \right\} \tag{9}$$

where $\mathcal{O}_\ell$ is the set defined by

$$\mathcal{O}_\ell = \left\{ \mathbf{v} \in \mathbb{R}^{\min\{M,N\}},\ v_1 \geq .. \geq v_{\min\{M,N\}} : \sum_{j=1}^{\min\{M,N\}} \left[\max_{k=1,\ldots,\ell} \sum_{i=1}^{k} \xi_{\ell-i} + k(1 - v_j) \right]_+ \leq r_e \right\} \tag{10}$$

Moreover, the exponent $d^*(r_e, L)$ is achievable by finite block length codes if $T \geq M + N - 1$.

Since the objective function in (9) is linear and hence convex, each of the minimizations in (9) has a well-defined unique solution that can be easily found by standard numerical optimization methods.

Unfortunately, at the moment we don't have a closed form characterization of the optimal tradeoff curve in Theorem 3. To shed more light on the power control diversity gain, we derived easily computable lower and upper bounds on the optimal diversity gain $d^*(r_e, L)$ in the following Lemma.

Lemma 4: Let $d^*(r_e, L)$ denote the optimal diversity gain under long-term power constraint given by Theorem 3. Then,

$$\max\left(d_L^{(lb1)}, d_L^{(lb2)}\right) \leq d^*(r_e, L) \leq d_L^{(ub)} \tag{11}$$

where $d_L^{(lb1)}$, $d_L^{(lb2)}$ and $d_L^{(ub)}$ are obtained recursively as follows. Let

$$d_1^{(lb1)} = d_1^{(lb2)} = d_1^{(ub)} = f(r_e) \tag{12}$$

Then, for $\ell = 2, \ldots, L$ let

$$d_\ell^{(lb1)} = \left(1 + d_{\ell-1}^{(lb1)}\right) f\left(\frac{r_e}{1 + d_{\ell-1}^{(lb1)}}\right) \tag{13}$$

$$d_\ell^{(lb2)} = \frac{\ell + \sum_{k=1}^{\ell-1} d_k^{(lb2)}}{\ell} f\left(\frac{r_e}{\ell + \sum_{k=1}^{\ell-1} d_k^{(lb2)}}\right) \tag{14}$$

and

$$d_\ell^{(ub)} = \left(1 + d_{\ell-1}^{(ub)}\right) f\left(\frac{r_e}{\ell + \sum_{i=1}^{\ell-1} d_i^{(ub)}}\right). \tag{15}$$

The lower bounds established in Lemma 4 have nice intuitive interpretations. The first lower bound, i.e., $d_\ell^{(lb_1)}$, corresponds to the outage probability achieved by only the round with the maximum power level. As a side result, this lower bound also corresponds to the diversity-multiplexing tradeoff of the power control algorithm proposed in [5] where the authors assume one round of transmission and the availability of the feedback information, needed for the power control algorithm, *a-priori* (in this setting, L takes the meaning of the number of levels in the power control algorithm). The second lower bound, i.e., $d_\ell^{(lb_2)}$, corresponds to averaging the power levels[3] used in the ℓ ARQ rounds and then deriving the tradeoff under the assumption that this level is used in all the ℓ rounds. Figure 2 depicts the upper and lower bounds on the optimal diversity advantage with power control. One can see in the figure the significant gain offered through power control,

[3]Here, we average the power computed on a logarithmic scale.

compared to ARQ with constant power, especially at low multiplexing gains. In fact, the remarkably large diversity gains observed for **all multiplexing gains** even with relatively small values of L indicates that very slow fading channels quickly approach the ergodic limit when ARQ and power control are used **jointly**. This phenomenon does not appear when only power control is used without ARQ retransmissions, as in [5] for example, since in this case the diversity advantage still approaches zero as the multiplexing gain approaches its maximum value of $\min(N, M)$. Moreover, at least in this scenario, it appears that the lower and upper bounds are very tight for a wide range of multiplexing gains.

IV. Conclusions

In this paper, we investigated the fundamental tradeoff of MIMO ARQ channels. We have shown that the ARQ retransmission delay can be leveraged for significant gains in the diversity advantage. By characterizing the three dimensional diversity-multiplexing-delay tradeoff, we have quantified this ARQ diversity gain. Our results show that, with the short-term power constraint, the ARQ diversity gain is significant only at high multiplexing gains. This limitation is overcome by combining the retransmission strategy with a carefully constructed power control policy, that allocates the power in the ℓ-th round to be inversely proportional to the probability of having to transmit ℓ rounds. In this way, very high power levels can be used to "correct" the very rare error events which determine the high-SNR behavior of error probability. We showed that the diversity gain achieved by ARQ with power control is dramatically large at all multiplexing gains, so that the performance approaches rapidly the ergodic (no-outage) behavior, according to which the multiplexing gain $\min\{M, N\}$ can be achieved with arbitrarily large reliability.

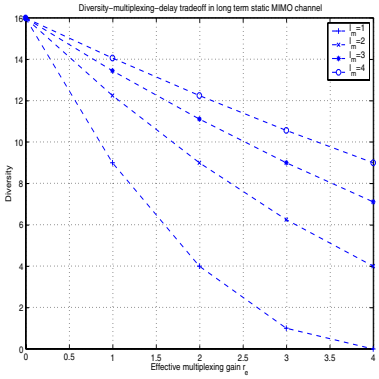

Fig. 1. The diversity-multiplexing tradeoff with different values of the maximum number of ARQ rounds "L"

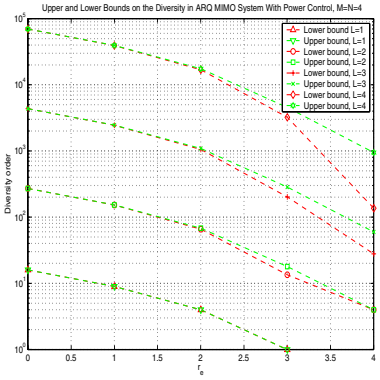

Fig. 2. The diversity-multiplexing tradeoff with power control on a log-scale (the upper and lower bound in Lemma 4).

References

[1] L. Zheng and D. N. C. Tse. Diversity and multiplexing: A fundamental tradeoff in multiple antenna channels. *IEEE Trans. Info. Theory*, 49:1073 –1096, May 2003.

[2] S. Diggavi, N. Al-Dhahir, A. Stamoulis, and A.R. Calderbank. Great expectations : The value of spatial diversity in wireless networks. *Proceedings of the IEEE (Special Issue on Gigabit Wireless)*, Feb. 2004.

[3] G. Caire and D. Tuninetti. Arq protocols for the gaussian collision channel. *IEEE Trans. on Inform. Theory*, 47:1971 –1988, July 2001.

[4] H. El Gamal, G. Caire, and M. O. Damen. The MIMO ARQ channel: Diversity-Multiplexing-Delay tradeoff. *submitted to the IEEE Trans. Inform. Theory*, Nov. 2004 (also available at www.ece.osu.edu/h̃elgamal).

[5] D. Rajan, A. Sabharwal, and B. Aazhang. Delay bounded packet scheduling of bursty sources over wireless channels. *IEEE Trans. Info. Theory*, 50:125–144, Jan. 2004.