Analysis of Rate Optimized Throughput for Large-Scale MIMO-(H)ARQ Schemes

Peter Larsson Student Member, IEEE, Lars K. Rasmussen Senior Member, IEEE, Mikael Skoglund, Senior Member, IEEE

Abstract—In this paper, we consider throughput and rate optimized throughput for large-scale MIMO-(H)ARQ systems in i.i.d. complex Gaussian block fading channels. Exact analysis of large-scale MIMO is generally intractable, yet the field of random matrix theory has provided asymptotic expressions for the instantaneous channel capacity, which we propose for studying MIMO-(H)ARQ systems. Even with those expressions, closed-form results for the maximum throughput and optimal rate are not known. We therefore provide i) tight asymptotic lower and upper bounds, with a diminishing gap, for the optimal rate that is useful for the design of practical link adaptation algorithms, and ii) a tight asymptotic lower bound for the maximized throughput, suggesting the practical sufficiency of MIMO-ARQ, without any packet combining, over more complex MIMO-HARQ methods.

I. INTRODUCTION

Modern wireless systems require high-reliability and high data-rate communication. High reliability can be provided by automatic-repeat-request (ARQ) [1], yet repetition redundancy (RR) [2] and incremental redundancy (IR) [3] hybrid-ARQ (HARQ) are often used for improved data rates. In addition, (code and modulation) rate adaptation allows for performance optimization [4]. High data-rate, is enabled by communication over multiple-input-multiple-output (MIMO) channels [5]. By a joint MIMO and (H)ARQ design [6][7], both high reliability and high data-rate are possible. While (H)ARQ operating in single-input-single-output (SISO) channels is well-studied, less is known about (H)ARQ operating in MIMO channels. For the latter, analytical (and optimized) performance expressions are hard to derive, and to our knowledge not known.

In this paper, we consider MIMO-ARQ, -RR, and -IR, and adopt information theoretical [4] and large-scale analysis approaches to study the throughput and its optimization. The contributions are: i) to suggest a highly accurate random matrix theory (RMT) [5], [8] approach for studying MIMO-(H)ARQ, ii) to give RMT based throughput expressions, iii) to construct tight RMT based bounds for the optimal throughput values and the optimum rate points, e.g. (24)(49)(60), iv) to consider, analyze, and compare the three main MIMO-retransmission schemes in one common framework and paper. We find that, with equal number of transmit and receive antennas, throughput optimal ARQ is practically as good as IR and RR, but with less complexity, may render MIMO-ARQ to be an interesting choice for future wireless systems.

II. SYSTEM MODEL

We consider the baseband receiver model

\[ y = \sqrt{\Gamma/N_t} H x + w, \]  

where \( \Gamma \) is the mean signal-to-noise-ratio (SNR), \( H \in \mathbb{C}^{N_t \times N_r} \) is the channel matrix, \( w, y \in \mathbb{C}^{N_r \times 1} \) are the output and noise vectors, \( x \in \mathbb{C}^{N_t \times 1} \) is the input vector, and \( N_t, N_r \) are the number of transmit and receive antennas, respectively. We will assume \( N_t = N_r \), for the analysis, but the overall approach is not limited here. Further, \( H, x \) and \( w \) are entrywise and transmission-to-transmission independent identically distributed (i.i.d.) complex circularly symmetric Gaussian random variables with zero mean and unit variance. Codeword \( x \) satisfies \( \mathbb{E}\{x^H x\} \leq 1 \), have covariance matrix \( R_{xx} = I_{N_t} \), drawn from a discrete set of messages, have rate \( R \) [nats/Hz/s] and are capacity achieving. For IR, \( x \) is block-to-block independent, whereas for RR, \( x \) is identical. The receiver knows the full CSI whereas the sender knows the mean SNR.

For HARQ, redundancy blocks are transmitted up to the point that a packet is correctly decoded, and then a subsequent packet is transmitted. Similar for ARQ, but without combining of redundancy blocks. Hence, we study lossless-(H)ARQ (with all-packet delivery) that provides an upper throughput bound for (lossy) truncated-(H)ARQ (with retransmission limit). We further assume, ideal (H)ARQ operation with error- and delay-free feedback, negligible protocol overhead, and that at least one packet is always awaiting transmission.

III. MIMO-ARQ

The instantaneous MIMO channel capacity is

\[ C = \ln \det (I_{N_t} + \Gamma N_t^{-1} H H^H). \]  

It has been shown in [9], through the use of RMT, that the instantaneous MIMO channel capacity, with \( H \) entry-wise i.i.d and \( N_t \to \infty, N_t/N_t = c \), is asymptotically Gaussian

\[ (C + N_t I_c)/S_c \to N(0,1), \]  

where the asymptotic channel capacity per antenna is

\[ I_c = c \ln (1 + \Gamma(1 - m_c)) + \ln (1 + \Gamma(c - m_c)) - m_c, \]  

the asymptotic channel capacity variance is

\[ S_c^2 = -\ln \left(1 - m_c^2/c\right), \]  

and \( m_c \) is

\[ m_c = \frac{1}{2} \left(1 + c + 1/\Gamma - \sqrt{(1 + c + 1/\Gamma)^2 - 4c}\right). \]  

We note that the ratio \( I_c/S_c \) is a monotonically increasing function in \( \Gamma \), and that the limits are

\[ \lim_{\Gamma \to 0} I_c/S_c = 1/\sqrt{c}, \quad \lim_{\Gamma \to \infty} I_c/S_c = \infty. \]
Henceforth, we assume \( N = N_t = N_r, \ c = 1, \ I/S > 1, \) and omit index \( c, \) until we treat RR in section V. The probability of successfully decoding a message, encoded with a capacity achieving channel code, is the probability that the instantaneous capacity exceeds the code rate \( R \)

\[
P = P\{C > R\}. \tag{7}
\]

The throughput for ARQ is defined as

\[
T \triangleq R/M = RP, \tag{8}
\]

where \( M \) is the mean number of transmissions until correctly decoding a message. For ARQ, it is well known that \( M = 1/P. \) Due to (3), the asymptotic decoding probability is

\[
P \simeq Q((R - NI)/S), \tag{9}
\]

where \( Q(x) = \int_x^\infty e^{-t^2/2}/\sqrt{2\pi} \, dt \) is the Q-function. Therefore, the asymptotic ARQ throughput is

\[
\tilde{T} \simeq RQ((R - NI)/S). \tag{10}
\]

In Fig. 1-4, we show the RMT-based analytical (Ana.) throughput vs. SNR for (10) and Monte Carlo simulated (Sim.) results of (7)(8). In Fig. 5, various rates are also considered. We note a great match between analytical and simulated results in Fig. 1-5, when \( N \geq 4. \) We also show related results for IR and RR in Fig. 1-5, which will be derived and discussed in section IV and V. Note that all plots use \( R \) and \( T \) in [b/Hz/s]. We note that like (9), [4] also assumed a Gaussian distributed capacity, but then only for IR in a SISO channel.

We now consider throughput optimization. This is of course well-studied for SISO-ARQ and -RR [4], [10], [11], but less so for SISO-IR [12] and our study items MIMO-(H)ARQ. We first see that (10) is maximized for \( R^* < NI \) if \( NI/S > \sqrt{2} \) (or \( N \geq 2, \) since \( \tilde{T}_{IR}(0) > 0 \) and \( \tilde{T}_{RR}(R = NI) < 0. \) We use \( x \triangleq NI/S \), giving \( \tilde{T} = (NI - Sx)(1 - Q(x)). \) Taking the derivative of \( \tilde{T}, \) wrt \( x, \) leads to the optimality condition

\[
\frac{d\ln(\tilde{T})}{dx} = \frac{-S}{NI - Sx^*} + \frac{Q'(x^*)}{1 - Q(x^*)} = 0 \tag{11}
\]

\[
\Rightarrow f(x^*) = NI/S, \ x^* > 0, \tag{12}
\]

where \( f(x) \triangleq \sqrt{2\pi} e^{x^2/2}(1 - Q(x)) + x. \) As the paper was originally completed, Sep. 2013, we found an independent work [14] that studied MIMO-ARQ, considered throughput optimization, and approximated the capacity by a Gaussian r.v. with identical mean and variance. As we propose using RMT for MIMO-(H)ARQ analysis, we get a Gaussian distributed capacity approximation, and our optimality criteria (12) be-
comes similar to [14, (6)], albeit [14, (6)] is expressed in the somewhat less convenient inverse Q-functions, and also to the earlier [13, (11)] that [14] extended to MIMO. However, with RMT, and in contrast to [14], the capacity mean and variance are non-identical. Moreover, we study (12) and optimize analytically, whereas [14, (6)] was evaluated numerically. Note also that we study all three core schemes, ARQ, RR, and IR, for completeness, whereas [14] only considered ARQ. Now, we would like to solve (12), and the optimal \( x^* \) is then given by

\[
x^* = f^{-1}(N/\Sigma),
\]

with \( f^{-1}(x) \) denoting the inverse of \( f(x) \). Hence, the optimal rate is \( \hat{R}^* = N \Sigma - Sx^* \). However, since \( f^{-1}(x) \) is not easily computed, we lower and upper bound \( f(x) \), and then solve for their inverses. First, we bound according to

\[
f_{\text{ll}}(x) \leq f(x) \leq f_{\text{uu}}(x), \quad \text{with}
\]

\[
f_{\text{ll}}(x) = \sqrt{2\pi}e^{x^2/2} - \frac{x}{1 + x},
\]

\[
f_{\text{uu}}(x) = \sqrt{2\pi}e^{x^2/2} - \frac{1}{x} + x,
\]

using to the Q-function bounds

\[
\sqrt{\pi}e^{-x^2/2} \leq \sqrt{2\pi}Q(x) \leq \frac{1}{\sqrt{\pi}}e^{-x^2/2},
\]

\( x \geq 0 \). As (15) and (16) do not have simple inverses, we carefully craft two simpler, but invertible bound, as

\[
f_{\text{ll}}(x) \leq f(x) \leq f_{\text{uu}}(x), \quad \text{with}
\]

\[
f_{\text{ll}}(x) = \sqrt{2\pi}e^{x^2/2} + \frac{x^2}{2},
\]

\[
f_{\text{uu}}(x) = \sqrt{2\pi}e^{x^2/2}.
\]

The upper bound (18), with the additive term \( x^2/2 \), is ad
dvertently designed to enable an inverse and solving (13) with Lambert’s-W function [15]. After some algebra, we get

\[
x_{\text{uu}} = \sqrt{2}\left(\frac{N}{\Sigma} - W_0\left(\sqrt{2\pi}\exp\left(\frac{N}{\Sigma}\right)\right)\right)
\]

\[
= \frac{2\ln\left(W_0\left(\sqrt{2\pi}\exp\left(\frac{N}{\Sigma}\right)\right)\right)}{2\sqrt{\pi}},
\]

where \( W_0(x) \) is the principal branch solution to the equation \( x = W(x)e^{W(x)} \). and the last step is proved in appendix. Solving (13) for the lower bound (19), valid for \( x_1 \geq 1 \), gives

\[
x_{\text{ll}} = \sqrt{2\ln\left(\frac{N/\Sigma}{\sqrt{2\pi}}\right)}.
\]

Note that some care is required for the bounds. For the lower bound \( f_{\text{ll}}(x) \), based on \( x_1 \geq 1 \), we require \( \frac{N}{\Sigma} \geq \sqrt{2\pi}e \approx 4.13 \). Hence, \( f_{\text{ll}}(x) \leq f(x), \forall \Gamma \) if \( N \geq 5 \). At a closer look, we numerically find that \( f_{\text{ll}}(x) \leq f(x) \), if \( x_1 \approx 0.7518 \), giving \( \frac{N}{\Sigma} > \sqrt{2\pi}\exp\left(x_1^2/2\right) \approx 3.325 \). Hence, \( f_{\text{ll}}(x) \leq f(x), \forall \Gamma \) if \( N \geq 4 \). For (20) to be real-valued, we also require \( \frac{N}{\Sigma} > \sqrt{2\pi} \approx 2.51 \), which is ensured \( \forall \Gamma \) if \( N \geq 3 \). Fig. 6 illustrates where those conditions on the bounds applies.

We now have the lower and upper bounds for the optimal rate \( \hat{R}^* \) that maximizes the throughput. The rate bounds are

\[
R_{\text{ll}} \leq \hat{R}^* \leq R_{\text{uu}},
\]

where

\[
R_{\text{uu}} = \frac{N}{\Sigma} - Sx_{\text{uu}},
\]

\[
R_{\text{ll}} = \frac{N}{\Sigma} - Sx_{\text{ll}}.
\]

In Fig. 7, we show throughput vs. rate, exhibiting throughput maxima, together with the rate bounds (23)(24). It is evident that the bounds on optimal rate are exceedingly tight.

We show in appendix that for fixed SNR, the rate gap between the optimal rate bounds vanishes asymptotically

\[
\lim_{N \to \infty} \Delta R \triangleq \lim_{N \to \infty} \left( R_{\text{uu}} - R_{\text{ll}} \right) = 0.
\]

In Fig. 8, we show that \( \Delta R \) is small and decreases with \( N \). Note that, the more vital metric, \( \Delta R/N \) vanishes even faster.

The throughput at the bounds for the optimal rate are then

\[
T_{\text{uu}} = R_{\text{uu}}(1 - Q(x_{\text{uu}})),
\]

\[
T_{\text{ll}} = R_{\text{ll}}(1 - Q(x_{\text{ll}})).
\]

Now, let \( \hat{T}^* \sim \hat{R}^*Q(\frac{\hat{R}^* - \hat{R}}{N}) \) denote the rate maximized throughput for the true optimal rate \( \hat{R}^* \). It is clear that

\[
T_{\text{ll}} \leq \hat{T}^* \quad \text{and} \quad T_{\text{uu}} \leq \hat{T}^*.
\]
and since $\Delta R \to 0$ with $N$, it follows that

\[
\begin{align*}
\lim_{N \to \infty} \Delta T_{uu} & \triangleq \lim_{N \to \infty} T_{uu} - \tilde{T}^* = 0, \\
\lim_{N \to \infty} \Delta T_{ll} & \triangleq \lim_{N \to \infty} \tilde{T}^* - T_{ll} = 0.
\end{align*}
\]

In Fig. 9, we plot the throughput (10) vs. SNR for $N = 8$ and a set of rates. We show the ergodic capacity (simulated), the rate bounds $R_{uu}$ (23) and $R_{ll}$ (24) for the optimal rate, and their corresponding lower bounding on the optimal throughput $T_{uu}$ (26) and $T_{ll}$ (27). We observe that $T_{ll}$ and $T_{uu}$ exhibit practically ideal performance over the full SNR range.

The maximum throughput can also be approximated by

\[
T^* \simeq T_{ll} \simeq (NI - x_{ll}^2 S) \cdot (1 - S/NI x_{ll}) = S \left( (NI/S) + S/NI \right) - (x_{ll}^2 + 1/\gamma) \tag{31},
\]

where we used $Q(x) \sim e^{-x^2/2}$ and (21) in the second step. Using (31), for increasing $N$ but fixed $\Gamma$, we note that the throughput approaches the ergodic capacity asymptotically

\[
(C_{\text{erg}} - T^*)/N \to I - T_{uu}^*/N \to 0. \tag{32}
\]

IV. MIMO-IR-HARQ

The throughput, and its optimization, is derived similarly for IR as for ARQ, but several important details differ. First, the channel capacity for $k$ blocks is now

\[
C_k^{IR} = \sum_{j=1}^{k} \ln \det \left( I_{N_1} + \frac{\Gamma}{N_1} H_j H_j^H \right). \tag{33}
\]

In contrast to ARQ, we now consider the outage probability

\[
Q_k^{IR} = \mathbb{P} \{ C_k^{IR} < R \}. \tag{34}
\]

The generic throughput for IR/RR-HARQ is given in [4] as

\[
\tilde{T}^{\text{HARQ}} = \frac{R}{M^{\text{HARQ}}} = \frac{R}{1 + \sum_{k=1}^{\infty} Q_k^{\text{HARQ}}} \tag{35}
\]

Eq. (33) is asymptotically Gaussian distributed with mean $kNI$ and variance $kS^2$. This leads to the outage probability

\[
\tilde{Q}_k^{IR} \simeq Q \left( -(R - kNI)/\sqrt{k}S \right), \tag{36}
\]

expressed in the $Q$-function, which inserted in (35) leads to

\[
\tilde{T}^{\text{IR}} \simeq \frac{R}{1 + \sum_{k=1}^{\infty} Q \left( -(R - kNI)/\sqrt{k}S \right)} \tag{37}
\]

As $N$ increases, $\tilde{T}^{\text{IR}}$ vs. $\Gamma$ gets successively more step-like, see Fig. 1-5. This is so, since for a step $K$ (indexed from high-to-low SNR) $Q_K \sim 1$, $k < K$, and $Q_K \sim 0$, $k > K$. Then for $N$ large enough, $\tilde{T}^{(IR)}$ is piece-wise approximated by

\[
\tilde{T}_K^{\text{IR}} \approx \frac{R}{K + Q \left( -(R - kNI)/\sqrt{k}S \right)} \tag{38}
\]

We now maximize the throughput for a targeted $K$ redundancy blocks per packet. Let $x_K \triangleq \frac{R}{\sqrt{k}S}$, where we now leave out the superscript (IR) for readability. The derivative of $\tilde{T}_K^{\text{IR}}$ wrt $x_K$, equating to zero, yields the new optimality condition

\[
x_K = f^{-1}(\sqrt{K} N/S), \tag{39}
\]

\[
f(x_K) = \sqrt{2\pi} e^{x_K^2/(2(K + Q(x_K)))} + x_K. \tag{40}
\]
Bounding $f(x_K)$, which differ in sign and $K$ wrt ARQ, gives

$$f_l(x_K) \leq f(x_K) \leq f_u(x_K), \quad \text{with}$$

$$f_u(x_K) = \sqrt{2\pi e} x_K / K + 1 / x_K,$$  \hspace{1cm} (42)

$$f_l(x_K) = \sqrt{2\pi e} x_K / K + x_K,$$  \hspace{1cm} (43)

where $x_u$ and $x_1$ solves $\frac{x_1^2 + x_1}{2} + 1 / x_1 = \frac{x^2}{2} + \sqrt{x^2 / 2},$ \hspace{1cm} (44)

and we carefully selected the offset to $\sqrt{\pi / 2},$ which gives $f_u(0) = f_u(x_K) < f(x_K), x > 0$, and $f_u(x_K) \approx f(x_K), 0 \leq x_K \leq x_u$. We also have $f_u(x_K) > f(x_K), x_K \geq \tilde{x}_a$, where $\tilde{x}_a$ is the largest solution to $\sqrt{2\pi e} x_K / K + x_K = \frac{\sqrt{\pi} K}{2},$ \hspace{1cm} (45)

where $\tilde{x}_a \approx x_u \approx 0.770$. While not critical, we note a near ideal match for the optimized throughput. Since $K = 2$, the rate bounds are about twice the throughput.

V. MIMO-RR-HARQ

We now proceed as in IV, but only indicate the differences here. The channel capacity for RR with $k$-blocks is now expressed in the alternate form

$$C_k^{RR} = \ln \det \left( I_N + \frac{1}{N_k} \tilde{H}^H \tilde{H} \right),$$  \hspace{1cm} (53)

where $\tilde{H} = [H_{1}; H_{2}; \ldots; H_{k}]$ stacks $k$ channel matrices, and RR now becomes equivalent to a $kN_s \times N_s$ MIMO system. The outage probability, for $N = N_s = N_r$, is then

$$Q_k^{RR} \simeq Q \left( -R - N I_k / S_k \right),$$  \hspace{1cm} (54)

where $I_k = I_{(c=k)}$ and $S_k = S_{(c=k)}$. From (35), we formulate

$$\tilde{T}_k^{RR} \simeq \frac{R}{1 + \sum_{k=1}^{\infty} Q \left( -R - N I_k / S_k \right)},$$  \hspace{1cm} (55)

We use the substitution $x_K \overset{d}{=} N I_k / S_k$, and leave out the superscript (RR), take the derivative of the $K$-step approximation $\tilde{T}_K^{RR}$ from (55) wrt $x_K$, and we arrive at

$$x_K = f^{-1}(N I_k / S_k),$$  \hspace{1cm} (56)

where

$$f(x_K) = \sqrt{2\pi e} x_K / K + Q(x_K) + x_K.$$  \hspace{1cm} (57)
Thus, we can use the IR-HARQ bounds (44)(45) to determine the bounds for the optimum rate. We then find that

\[ R_{K,u}^{RR} = N I_K - S_K x_{K,u}, \quad \text{with} \]

\[ x_{K,u} = \frac{2 \ln \left( W_0 \left( \sqrt{2 \pi e} \frac{NI_S}{S_K} \right) / K \sqrt{2 \pi} \right)}{2 \ln \left( W_0 \left( \sqrt{2 \pi e} \frac{NI_S}{S_K} \right) / K \sqrt{2 \pi} \right)}. \]

\[ R_{K,ll}^{RR} = N I_K - S_K x_{K,ll}, \quad \text{with} \]

\[ x_{K,ll} = \frac{2 \ln \left( \frac{NI_K/\sqrt{2 \pi} S_K}{K} \right)}{2 \ln \left( \frac{NI_K/\sqrt{2 \pi} S_K}{K} \right)} - \frac{1}{2K}. \]

In Fig. 12, the optimum rate bounds (58)(60) are shown, and in Fig. 13, the (optimal) throughput value(s) are shown for \( R/N = 4 \) and \( K = 2 \), with a great match at the \( K = 2 \)-step.

VI. SUMMARY AND CONCLUSIONS

We studied throughput, and its optimization, for MIMO-ARQ, -RR and -IR, and derived closed-form asymptotic bounds for the optimal rate points and throughput.

We found that the maximized throughput of ARQ, RR, and IR was practically the same for \( N \geq 4 \). We also saw that the maximum throughput of IR improved with more redundancy blocks, yet with increased delay, whereas RR deteriorated. HARQ, especially IR, has higher complexity than ARQ. With large-scale MIMO, the complexity difference may worsen. Thus, given the performance measure and system assumptions, we think that future large-scale MIMO systems may aim, not for IR-HARQ, nor for RR-HARQ, but for ARQ.

APPENDIX

Derivation of (20): Set \( t \triangleq \sqrt{2 \pi} e^{x/N I_S} \), and we then get \( t - W_0(\sqrt{2 \pi e} t) = \ln \left( e^t - W_0(\sqrt{2 \pi e} t) \right) = \ln \left( W_0(\sqrt{2 \pi e} t) / \sqrt{2 \pi} \right) \), where we used \( W(x) = e^x \) in the second step.

Derivation of (25): To show that \( \Delta I \to 0 \) as \( N \) increases, is equivalent to show that \( W_0(\sqrt{2 \pi} e^{x/N I_S}) \approx N I_S/N \). For this, we use the expansion for Lambert’s-W function, \( W_0(x) \approx L_1 - L_2 + L_3 \), \( L_3 = \ln(x) \), \( L_2 = \ln(\ln(\ln(x))) \) and \( W_0(\sqrt{2 \pi} e^{x/N I_S}) = (\ln(2\pi)/2 + N I_S/N) / \ln(\ln(\ln(x))) \), or \( (N I_S/N + \mathcal{O}(\ln(\ln(x)))) \).

ACKNOWLEDGEMENT

The first author would like to thank Maksym Ginyk for a seminar on RMT and MIMO that in part inspired this work.

REFERENCES