An Achievable Rate for the MIMO Individual Channel

Yuval Lomnitz, Meir Feder
Tel Aviv University, Dept. of EE-Systems
Email: {yuvall,meir}@eng.tau.ac.il

***Abstract*—We consider the problem of communicating over a multiple-input multiple-output (MIMO) real valued channel for which no mathematical model is specified, and achievable rates are given as a function of the channel input and output sequences known a-posteriori. This paper extends previous results regarding individual channels by presenting a rate function for the MIMO individual channel, and showing its achievability in a fixed transmission rate communication scenario.**

I. Introduction

We consider a channel, termed an *individual channel*, where no specific probabilistic or mathematical relation between the input and the output is assumed. This channel is an extreme case of an unknown channel. Achievable rates are characterized using only the input and output sequences, which capture the actual (a-posteriori) channel behavior. This point of view is similar to the approach used in universal source coding of individual sequences where the goal is to asymptotically attain for each sequence the same coding rate achieved by the best encoder from a model class, tuned to the sequence. This framework is an evolution of those considered in Shayevitz and Feder [1] and Eswaran et. al. [2] and is presented in more detail in our papers [3] and [4], together with the relevant background. We will give a brief introduction below.

The setting we consider includes a single encoder receiving a message to transmit and emitting symbols $x_i \in \mathcal{X}, i = 1,2,..,n$ and a decoder receiving a sequence of symbols $y_i \in \mathcal{Y}, i = 1,2,..,n$ and attempting to reconstruct the message. In the present paper the input and output symbols are real-valued vectors, i.e. $\mathcal{X} = \mathbb{R}^t$ and $\mathcal{Y} = \mathbb{R}^r$. The relation between $\mathbf{x} = [x_1,...,x_n]^T$ and $\mathbf{y} = [y_1,...,y_n]^T$ is unknown to the encoder and decoder. We consider two communication scenarios: with feedback (possibly imperfect) and without feedback. For the case in which there is no feedback the communication system transmits in a constant rate, and outage is unavoidable, i.e. one cannot guarantee a small probability of error in all circumstances. In the case feedback exists, the communication rate may be dynamically adapted and outage may be prevented. In both cases we assume common randomness exists in the encoder and the decoder. The results in the current paper extend the previous results, yet only for the first case, of transmission in a constant rate.

The performance is measured by a rate function $R_{\text{emp}} : \mathcal{X}^n \times \mathcal{Y}^n \to \mathbb{R}$ representing an empirical measure of the achievable rate between the channel input and channel output, over n channel uses. In examples here and in [3][4], R_{emp} can be viewed as the mutual information achieved in a certain family of statistical models (in the current scope, all zero mean Gaussian channels), when the model parameters match the empirical ones. In communication without feedback we say that a given rate function $R_{\text{emp}}(\mathbf{x},\mathbf{y})$ is achievable with an input distribution $Q(\mathbf{x})$ if for large block size $n \to \infty$, it is possible to communicate at any rate R and an arbitrarily small error probability is obtained whenever $R_{\text{emp}}(\mathbf{x},\mathbf{y}) > R$. The communication system is required to emit blocks with probability distribution $Q(\mathbf{x})$, which is possible due to the use of randomization. By placing this additional constraint we leave aside the question of adapting the input distribution, so that the current framework attempts at achieving the empirical "mutual information" rather than the empirical "capacity". Another reason for the fixed prior is avoiding degenerate systems which may transmit only "bad" sequences with low (or zero) R_{emp}. This constraint is further discussed in [3], section VIII.C.

The main result of this paper is that for the multiple-input multiple-output (MIMO) channel $\mathbb{R}^t \to \mathbb{R}^r$ (i.e. with t transmit and r receive antennas) the rate function defined below is asymptotically achievable, in the fixed rate case:

$$R_{\text{emp}}(\mathbf{X},\mathbf{Y}) = \frac{1}{2}\log\left(\frac{\left|\hat{\mathbf{R}}_{XX}\right| \cdot \left|\hat{\mathbf{R}}_{YY}\right|}{\left|\hat{\mathbf{R}}_{(XY)(XY)}\right|}\right) \qquad (1)$$

where the $n \times t$ matrix $\mathbf{X}$ denotes the channel input over n symbols, and the $n \times r$ matrix $\mathbf{Y}$ denotes the output. $\hat{\mathbf{R}}_{XX} = \frac{1}{n}\mathbf{X}^T\mathbf{X}$, $\hat{\mathbf{R}}_{YY} = \frac{1}{n}\mathbf{Y}^T\mathbf{Y}$ and $\hat{\mathbf{R}}_{(XY)(XY)} = \frac{1}{n}[\mathbf{XY}]^T[\mathbf{XY}]$ are the input, the output and the joint empirical correlation matrices, respectively. This is a generalization of the result of [3] where the rate function $R_{\text{emp}} = \frac{1}{2}\log\left(\frac{1}{1-\hat{\rho}(\mathbf{x},\mathbf{y})^2}\right)$ was proved to be achievable for real valued SISO channel $\mathbb{R} \to \mathbb{R}$ ($\hat{\rho}$ denotes empirical correlation). As in [3], the proof is geometrically intuitive. The results easily extend to the *complex* MIMO case, and to rate function using the empirical *covariance* (rather than the correlation), but we focus here on the simpler case.

The paper is organized as follows: in Section II we explain the motivation for this rate function and its relation to the probabilistic Gaussian channel, in Section III we present in detail the main result, which is proven in Section IV. Section V is devoted to comments and further research items.

We use lowercase boldface letters to denote vectors, and uppercase boldface letters to denote matrices. We use the same

notation for random variables and their sample values, and the distinction should be clear from the context.

II. Origin of the rate function

Consider the channel from $\mathbf{x} \in \mathbb{R}^t$ to $\mathbf{y} \in \mathbb{R}^r$ which are real valued vectors. For the additive white Gaussian noise (AWGN) MIMO channel $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{v}$ with $\mathbf{v} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$, and $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I})$ it is well known that the mutual information is

$$I(\mathbf{x};\mathbf{y}) = \frac{1}{2}\log\left|\mathbf{I} + \frac{1}{\sigma^2}\mathbf{H}^T\mathbf{H}\right| \tag{2}$$

see for example [5][6]. This reflects the maximum achievable rate with the fixed covariance matrix $E\mathbf{x}\mathbf{x}^T = \mathbf{I}$, and is sometimes termed the *open-loop MIMO capacity*, since equal power is a reasonable choice when the transmitter does not know the channel. A more general form of the mutual information is obtained by assuming $\mathbf{x}, \mathbf{y}$ are any jointly Gaussian random vectors and writing:

$$h(\mathbf{x}) = \frac{1}{2}\log|2\pi e \cdot \mathrm{cov}(\mathbf{x})| \tag{3}$$

$$h(\mathbf{y}) = \frac{1}{2}\log|2\pi e \cdot \mathrm{cov}(\mathbf{y})| \tag{4}$$

$$h(\mathbf{x},\mathbf{y}) = \frac{1}{2}\log\left|2\pi e \cdot \mathrm{cov}\left(\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}\right)\right| \tag{5}$$

Therefore:

$$I(\mathbf{x};\mathbf{y}) = h(\mathbf{x})+h(\mathbf{y})-h(\mathbf{x},\mathbf{y}) = \frac{1}{2}\log\left[\frac{|\mathrm{cov}(\mathbf{x})|\cdot|\mathrm{cov}(\mathbf{y})|}{|\mathrm{cov}\left(\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}\right)|}\right] \tag{6}$$

where the factors $2\pi e$ cancel out since the dimension of the covariance matrix in the denominator is the sum of the dimensions in the nominator. The expression (6) is more general than (2) since it does not assume the noise is white, and is suitable for our purpose since it expresses the mutual information through properties of the input and output vectors without using an explicit channel structure. For the case of the AWGN MIMO channel it yields the same value as (2). For the particular scalar case where $\mathbf{x}$, $\mathbf{y}$ are scalars with variances σ_X^2, σ_Y^2 and correlation factor ρ, Equation (6) evaluates to $I(\mathbf{x};\mathbf{y}) = \frac{1}{2}\log\left(\frac{1}{1-\rho^2}\right)$, as previously obtained for the SISO case.

The empirical rate function we defined in (1) is an empirical version of the mutual information expression in (6), except that the covariance matrices are replaced by empirical *correlation* (rather than covariance) matrices, i.e. we do not cancel the mean. When $|\hat{\mathbf{R}}_{XX}| = 0$ or $|\hat{\mathbf{R}}_{YY}| = 0$ (which leads also to $|\hat{\mathbf{R}}_{(XY)(XY)}| = 0$) , the rate function will be defined by removing the columns of $\mathbf{X}$ or $\mathbf{Y}$ (respectively), which are linearly dependant on the others, until these determinants become positive. It is not important which columns are removed to break the linear dependence, due to this function's invariance to linear transformation (Property 2 below). For the case of $\mathbf{Y} = 0$ or $\mathbf{X} = 0$ we define $R_{\mathrm{emp}} = 0$.

The rate function has the following properties which are expected from an empirical metric of the "mutual information":

1) **Non-negativity:** $R_{\mathrm{emp}}(\mathbf{X},\mathbf{Y}) \geq 0$. This is evident from the fact $R_{\mathrm{emp}}(\mathbf{X},\mathbf{Y})$ is the mutual information between two Gaussian vectors with the respective covariances. It will also be shown in passing as part of the derivation in Section IV.
2) **Invariance under linear transformations:** Any invertible linear matrix operation on the input or output (for example, multiplying any of the input or output signals by a factor, adding signals, etc) does not change $R_{\mathrm{emp}}(\mathbf{X},\mathbf{Y})$, i.e. $R_{\mathrm{emp}}(\mathbf{X}\mathbf{G}_x, \mathbf{Y}\mathbf{G}_y) = R_{\mathrm{emp}}(\mathbf{X},\mathbf{Y})$. *Proof:* Suppose we multiply $\mathbf{X}$ and $\mathbf{Y}$ by arbitrary matrices $G_{x,t\times t}$ and $G_{y,r\times r}$ respectively. Define $\mathbf{X}' = \mathbf{X}\mathbf{G}_x$ then $\left|\hat{\mathbf{R}}'_{XX}\right| = \left|\frac{1}{n}\mathbf{X}'^T\mathbf{X}'\right| = \left|\mathbf{G}_x\hat{\mathbf{R}}_{XX}\mathbf{G}_x\right| = \left|\hat{\mathbf{R}}_{XX}\right|\cdot|\mathbf{G}_x|^2$. And likewise for $\mathbf{Y}$. Since $[\mathbf{X}',\mathbf{Y}'] = [\mathbf{X},\mathbf{Y}]\cdot\begin{bmatrix}\mathbf{G}_x & 0\\ 0 & \mathbf{G}_y\end{bmatrix}$ then from the same considerations we will have $\left|\hat{\mathbf{R}}'_{(XY)(XY)}\right| = \left|\hat{\mathbf{R}}_{(XY)(XY)}\right|\cdot\left|\begin{matrix}\mathbf{G}_x & 0\\ 0 & \mathbf{G}_y\end{matrix}\right|^2 = \left|\hat{\mathbf{R}}_{(XY)(XY)}\right|\cdot|\mathbf{G}_x|^2\cdot|\mathbf{G}_y|^2$, therefore the factors cancel out and $R_{\mathrm{emp}}(\mathbf{X}',\mathbf{Y}') = R_{\mathrm{emp}}(\mathbf{X},\mathbf{Y})$
3) **Symmetry:** $R_{\mathrm{emp}}(\mathbf{X},\mathbf{Y}) = R_{\mathrm{emp}}(\mathbf{Y},\mathbf{X})$

III. The main result

Theorem 1 (Non-adaptive, continuous MIMO channel)**.** *Given the channel $\mathbb{R}^t \to \mathbb{R}^r$, define the input over n symbols as an $n \times t$ matrix $\mathbf{X}$, and the output as an $n \times r$ matrix $\mathbf{Y}$. Let $\hat{\mathbf{R}}_{XX} = \frac{1}{n}\mathbf{X}^T\mathbf{X}$, $\hat{\mathbf{R}}_{YY} = \frac{1}{n}\mathbf{Y}^T\mathbf{Y}$ and $\hat{\mathbf{R}}_{(XY)(XY)} = \frac{1}{n}[\mathbf{X}\mathbf{Y}]^T[\mathbf{X}\mathbf{Y}]$ be the input, the output and the joint empirical correlation matrices, respectively. Define the rate function*

$$R_{\mathrm{emp}}(\mathbf{X},\mathbf{Y}) = \frac{1}{2}\log\left(\frac{\left|\hat{\mathbf{R}}_{XX}\right|\cdot\left|\hat{\mathbf{R}}_{YY}\right|}{\left|\hat{\mathbf{R}}_{(XY)(XY)}\right|}\right) \tag{7}$$

Then for every $P_e > 0$, a positive definite $t \times t$ matrix Λ_x and $n \geq t + r$ there exists random encoder-decoder pair of rate R over block size n, such that the distribution of the input sequence is $\mathbf{X} \sim \mathcal{N}^n(0, \Lambda_x)$ and for any $\gamma < 1 - \frac{t+r-1}{n}$ the probability of error for any message given an input sequence $\mathbf{X}$ and output sequence $\mathbf{Y}$ is not greater than P_e if:

$$R \leq \gamma\cdot R_{\mathrm{emp}}(\mathbf{X},\mathbf{Y}) + \frac{\log(P_e)}{n} - t\lceil r/2\rceil\frac{\log(n)}{n} - \frac{\log(C_L)}{n} \tag{8}$$

where

$$C_L = \frac{1}{\Gamma^t\left(\frac{r}{2}\right)2^{t\lceil r/2\rceil}}\cdot\left(\frac{2}{(1-\gamma)n - t - r + 1} + \frac{2}{r}\right)^t \tag{9}$$

Specifically, for every $\delta > 0$ and $\gamma < 1$ there exists n large enough so that the probability of error is not greater than P_e if:

$$R \leq \gamma\cdot R_{\mathrm{emp}}(\mathbf{X},\mathbf{Y}) - \delta \tag{10}$$

The theorem almost directly follows from the next lemma which we will prove subsequently:

Lemma 1. *For any $n \times r$ matrix $\mathbf{Y}$, the probability of $R_{\text{emp}}(\mathbf{X},\mathbf{Y}) \geq T$ where $\mathbf{X}$ is randomly drawn $\mathbf{X} \sim \mathcal{N}^n(0,\Lambda_x)$ is bounded by:*

$$\Pr\{R_{\text{emp}}(\mathbf{X},\mathbf{Y}) \geq T\} \leq C_L \cdot n^{t\lceil r/2\rceil} \exp(-\gamma \cdot n \cdot T) \quad (11)$$

For any γ in the range $0 \leq \gamma < 1 - \frac{t+r-1}{n}$, and where C_L is defined in (9).

Note that the bound does not depend on Λ_x. To prove Theorem 1, the codebook $\{\mathbf{X}_m\}_{m=1}^{\exp(nR)}$ is randomly generated by i.i.d. selection of each codeword from the Gaussian matrix distribution $\mathcal{N}^n(0,\Lambda_x)$. The common randomness is the codebook itself. The encoder sends the w-th codeword, and the decoder uses maximum empirical rate decoder i.e. chooses:

$$\hat{w} = \underset{m}{\text{argmax}}\{R_{\text{emp}}(\mathbf{X}_m;\mathbf{Y})\} \quad (12)$$

where ties are broken arbitrarily. By using Lemma 1 and the union bound, the probability of error given $\mathbf{X}_w, \mathbf{Y}$ is bounded by:

$$\begin{aligned} &P_e^{(w)}(\mathbf{X}_w,\mathbf{Y}) \leq \\ &\leq \Pr\left\{ \bigcup_{m\neq w} (R_{\text{emp}}(\mathbf{X}_m;\mathbf{Y}) \geq R_{\text{emp}}(\mathbf{X}_w;\mathbf{Y})) \middle| \mathbf{X}_w \right\} \leq \\ &\leq \exp(nR)\cdot C_L \cdot n^{t\lceil r/2\rceil} \exp(-\gamma\cdot n \cdot R_{\text{emp}}(\mathbf{X}_w;\mathbf{Y})) = \\ &\quad = C_L \cdot n^{t\lceil r/2\rceil} \exp[n(R - \gamma \cdot R_{\text{emp}}(\mathbf{X}_w;\mathbf{Y}))] \quad (13) \end{aligned}$$

Therefore if (8) is satisfied, then $P_e^{(w)}(\mathbf{X}_w,\mathbf{Y}) \leq P_e$, which proves the first part of the theorem. The second part follows directly from the first part. For any $\gamma < 1$ and $\delta > 0$ there is n large enough so that the condition $\gamma < 1 - \frac{t+r-1}{n}$ is satisfied, and then n could be increased till the redundancy in (8), $\frac{\log(P_e)}{n} - t\lceil r/2\rceil\frac{\log(n)}{n} - \frac{\log(C_L)}{n}$ would be smaller than δ (note that C_L is decreasing in n), therefore $P_e^{(w)}(\mathbf{X}_w,\mathbf{Y}) \leq P_e$ will be satisfied if (10) is satisfied. □

IV. Proof of Lemma 1

To prove Lemma 1 we use the Chernoff bound:

$$\begin{aligned} &\Pr\{R_{\text{emp}}(\mathbf{X},\mathbf{Y}) \geq T\} = \\ &\quad = \Pr\{\exp(n\gamma R_{\text{emp}}(\mathbf{X},\mathbf{Y})) \geq \exp(n\gamma T)\} \leq \\ &\quad \leq \frac{E\exp(n\gamma R_{\text{emp}}(\mathbf{X},\mathbf{Y}))}{\exp(n\gamma T)} \equiv L\exp(-n\gamma T) \quad (14) \end{aligned}$$

To prove the lemma we need to calculate

$$L = E\exp(n\gamma R_{\text{emp}}(\mathbf{X},\mathbf{Y})) = E\left(\frac{\left|\hat{\mathbf{R}}_{XX}\right|\cdot\left|\hat{\mathbf{R}}_{YY}\right|}{\left|\hat{\mathbf{R}}_{(XY)(XY)}\right|}\right)^{\frac{\gamma\cdot n}{2}} \quad (15)$$

where the expected value is taken with respect to $\mathbf{X}$. The remainder of this section is devoted to upper bounding L. We will first assume that $\Lambda_x = \mathbf{I}_{t\times t}$, i.e. $\mathbf{X} \sim \mathcal{N}^n(0,\mathbf{I})$, and then extend to general Λ_x.

Define $\mathbf{Z} = [\mathbf{Y},\mathbf{X}]$. We perform a QR decomposition of $\mathbf{X}, \mathbf{Y}$ and $\mathbf{Z}$ in order to obtain more friendly expressions. As a reminder, QR decomposition of a matrix $\mathbf{A}_{n\times k} = \mathbf{Q}_{n\times k}\mathbf{R}_{k\times k}$ (with $\mathbf{Q}^T\mathbf{Q} = I$ and $\mathbf{R}$ upper triangular) is performed by Gram-Schmidt process. We start from the left column of $\mathbf{A}$ and work our way to the last one. At each time we take a column of $\mathbf{A}$ and split it to the part which can be represented by a linear combination of the columns to the left of it (equivalently, to the columns of $\mathbf{Q}$ already generated), and the "innovation", i.e. the part which is orthogonal to the subspace generated by the previous columns. The vector representing the innovation is normalized, and becomes the respective column of $\mathbf{Q}$, and its power becomes the diagonal element in $\mathbf{R}$. The coefficients representing the part of the vector which is in the subspace of previous columns become the elements of $\mathbf{R}$ above the diagonal. Another important property of QR decomposition is that the determinant of $\mathbf{A}^T\mathbf{A}$ can be written in terms of the diagonal elements in $\mathbf{R}$: $\left|\mathbf{A}^T\mathbf{A}\right| = \left|\mathbf{R}^T\mathbf{Q}^T\mathbf{Q}\mathbf{R}\right| = \left|\mathbf{R}^T\mathbf{R}\right| = \left|\mathbf{R}\right|^2 = \prod_{i=1}^k R_{ii}^2$.

Now define the diagonal of the upper triangular matrix in the QR decomposition of the matrices $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ respectively to be the vectors $\mathbf{a}$, $\mathbf{b}$ and $[\mathbf{c},\mathbf{d}]$. I.e. if $\mathbf{X} = \mathbf{Q}_x\mathbf{R}_x$, $\mathbf{Y} = \mathbf{Q}_y\mathbf{R}_y$ and $\mathbf{Z} = \mathbf{Q}_z\mathbf{R}_z$ then $\mathbf{a} = \text{diag}(\mathbf{R}_x)$, $\mathbf{b} = \text{diag}(\mathbf{R}_y)$, and $[\mathbf{c},\mathbf{d}] = \text{diag}(\mathbf{R}_z)$. The lengths of the vectors $\mathbf{c},\mathbf{d}$ are r,t respectively, so that they overlap with the columns of $\mathbf{Y}$ and $\mathbf{X}$ in the matrix $\mathbf{Z}$. We have:

$$\frac{\left|\hat{\mathbf{R}}_{XX}\right|\cdot\left|\hat{\mathbf{R}}_{YY}\right|}{\left|\hat{\mathbf{R}}_{(XY)(XY)}\right|} = \frac{\left|\frac{1}{n}\mathbf{X}^T\mathbf{X}\right|\cdot\left|\frac{1}{n}\mathbf{Y}^T\mathbf{Y}\right|}{\left|\frac{1}{n}\mathbf{Z}^T\mathbf{Z}\right|} = \frac{\prod_{i=1}^t a_i^2\prod_{i=1}^r b_i^2}{\prod_{i=1}^r c_i^2\prod_{i=1}^t d_i^2} \quad (16)$$

Note that the $\frac{1}{n}$ factors cancel out because the matrix dimensions are t and r in the nominator and $t+r$ in the denominator. Since the Gram-Schmidt process operates sequentially from the first column to the last, and the first r columns of $\mathbf{Z}$ and $\mathbf{Y}$ are equal, we will have $\mathbf{b} = \mathbf{c}$. Therefore we can write:

$$\frac{\left|\hat{\mathbf{R}}_{XX}\right|\cdot\left|\hat{\mathbf{R}}_{YY}\right|}{\left|\hat{\mathbf{R}}_{(XY)(XY)}\right|} = \prod_{i=1}^t\left(\frac{a_i}{d_i}\right)^2 \quad (17)$$

Note that a_i and d_i both relate to the same vector, the i-th column of $\mathbf{X}$. The ratio $\frac{a_i}{d_i}$ is the ratio between the innovation of the i-th column of $\mathbf{X}$ with respect to the subspace spanned by previous columns of $\mathbf{X}$ alone (nominator) or these columns together with the columns of $\mathbf{Y}$ (denominator). Obviously from this reason $|d_i| \leq |a_i|$ (and therefore $R_{\text{emp}}(\mathbf{X},\mathbf{Y}) \geq 0$ - Property 1).

The key observation in this derivation is as follows: consider a sequential drawing of the columns of $\mathbf{X}$ and calculation of the factors $\frac{a_i}{d_i}$. Since the i-th column of $\mathbf{X}$ is chosen isotropically and independently of the previous columns, the value of previous columns does not affect the distribution of the innovations d_i, a_i (only the number of dimensions in previous columns does). Using this observation which we will prove subsequently, we would be able to break L represented as the expected value of a product (17) into a product of expected values (equations (19)-(20)), and the proof is completed by a (tedious) calculation of these expected values.

To show the independence of a_i, d_i in previously drawn values, denote by $\mathbf{X}_m^i$ a matrix including the columns m to i of $\mathbf{X}$, and by $\mathbf{x}_i$ the i-th column. Define a a unitary $n \times n$ matrix $\mathbf{Q}$ whose first $i-1$ columns span the subspace spanned by the first $i-1$ columns of $\mathbf{X}$, its next r columns extend this subspace to cover also the columns subspace of $\mathbf{Y}$, and the next $n-(i-1)-r$ columns complete it to an orthonormal basis. This matrix does not depend on $\mathbf{X}_i^t$ and specifically on the column i. We assume that the columns of $\mathbf{Y}$ are linearly independent (we will relax this assumption later). Also, in probability one, assuming $n \geq t+r$, the columns of $\mathbf{X}_1^{i-1}$ are linearly independent of each other and of the columns of $\mathbf{Y}$. To see this, it is easy to show that the projection of each column in any direction orthogonal to the subspace already spanned by previous ones (including $\mathbf{Y}$), is also Gaussian therefore has probability 0 to be 0, as long as there exists such an orthogonal vector, i.e. the number of previously generated vectors is smaller than n.

Now define $\mathbf{z} = \mathbf{Q}^T\mathbf{x}_i$. Since $\mathbf{x}_i \sim \mathcal{N}(0, \mathbf{I}_{n\times n})$ also $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I}_{n\times n})$. The first $i-1$ elements of $\mathbf{z}$ represent the projections of $\mathbf{x}_i$ to the subspace spanned by previous columns of $\mathbf{X}$, and the next r elements represent the projections to the subspace spanned by columns of $\mathbf{Y}$. So a_i^2 collects the energy of all elements except the first $i-1$, and d_i^2 collects the energy of all elements except the first $i-1+r$. To see this formally, in the Gram-Schmidt process the coefficients of the projection of $\mathbf{x}_i$ on the subspace spanned by $\mathbf{X}_1^{i-1}$ are $\mathbf{Q}_1^{i-1^T}\mathbf{x}_i$, and the projection itself is $\mathbf{Q}_1^{i-1}\mathbf{Q}_1^{i-1^T}\mathbf{x}_i$, therefore the innovation is $\mathbf{v}_i = \mathbf{x}_i - \mathbf{Q}_1^{i-1}\mathbf{Q}_1^{i-1^T}\mathbf{x}_i$. Since $\mathbf{Q}_1^{i-1^T}\mathbf{v}_i = 0$ and $\mathbf{Q}_\mathbf{i}^{\mathbf{n}T}\mathbf{v}_i = \mathbf{Q}_\mathbf{i}^{\mathbf{n}T}\mathbf{x}_i$ we have $a_i^2 = \|\mathbf{v}_i\|^2 = \left\|\mathbf{Q}^T\mathbf{v}_i\right\|^2 = \left\|\begin{bmatrix} \mathbf{Q}_1^{i-1^T} \\ \mathbf{Q}_\mathbf{i}^{\mathbf{n}T} \end{bmatrix}\mathbf{v}_i\right\|^2 = \left\|\begin{bmatrix} 0 \\ \mathbf{Q}_i^{nT}\mathbf{x}_i \end{bmatrix}\right\|^2 = \|\mathbf{z}_i^n\|^2$ and similarly, $d_i^2 = \left\|\mathbf{x}_i - \mathbf{Q}_1^{i-1+r}\mathbf{Q}_1^{i-1+r^T}\mathbf{x}_i\right\|^2 = \left\|\mathbf{z}_{i+r}^n\right\|^2$. Therefore a_i, d_i are independent of $\mathbf{Y}$ and the previous columns of $\mathbf{X}$, and can be given by norms over parts of a Gaussian i.i.d. vector of length n. Defining

$$D_i \equiv E\left(\left(\frac{a_i}{d_i}\right)^{\gamma n}\middle|\mathbf{X}_1^{i-1}\right) = E\left(\left(\frac{a_i}{d_i}\right)^{\gamma n}\right) \tag{18}$$

Where the equality is due to the independence shown above, we have for any $k = 1, 2, ..., t$:

$$\begin{aligned} E\prod_{i=1}^{k}\left(\frac{a_i}{d_i}\right)^{\gamma n} &= E\left[E\left(\prod_{i=1}^{k}\left(\frac{a_i}{d_i}\right)^{\gamma n}\middle|\mathbf{X}_1^{k-1}\right)\right] = \\ &= E\left[\prod_{i=1}^{k-1}\left(\frac{a_i}{d_i}\right)^{\gamma n}\cdot E\left(\left(\frac{a_k}{d_k}\right)^{\gamma n}\middle|\mathbf{X}_1^{k-1}\right)\right] = \\ &= E\left[\prod_{i=1}^{k-1}\left(\frac{a_i}{d_i}\right)^{\gamma n}\cdot D_k\right] = E\left(\prod_{i=1}^{k-1}\left(\frac{a_i}{d_i}\right)^{\gamma n}\right)\cdot D_k \end{aligned} \tag{19}$$

Therefore by induction:

$$L = E\left[\frac{\left|\hat{\mathbf{R}}_{XX}\right|\cdot\left|\hat{\mathbf{R}}_{YY}\right|}{\left|\hat{\mathbf{R}}_{(XY)(XY)}\right|}\right]^{\frac{\gamma\cdot n}{2}} = E\prod_{i=1}^{t}\left(\frac{a_i}{d_i}\right)^{\gamma n} \overset{(19)}{=} \prod_{i=1}^{t} D_i \tag{20}$$

Now we bound D_i (using the previously defined Gaussian vector $\mathbf{z}$):

$$\begin{aligned} D_i = E\left(\frac{a_i^2}{d_i^2}\right)^{\gamma n/2} &= E\left(\frac{\|\mathbf{z}_i^n\|^2}{\left\|\mathbf{z}_{i+r}^n\right\|^2}\right)^{\gamma n/2} = \\ &= E\left(1+\frac{\sum_{j=i}^{i+r-1} z_j^2}{\sum_{j=i+r}^{n} z_j^2}\right)^{\gamma n/2} = \\ &\overset{(a)}{=} \underset{\substack{h\sim\chi^2(r)\\ s\sim\chi^2(n-i-r+1)}}{E}\left(1+\frac{h}{s}\right)^{\gamma n/2} = \\ = \iint_0^\infty \left(1+\frac{h}{s}\right)^{\frac{\gamma n}{2}} &\frac{h^{\frac{r}{2}-1}e^{-\frac{h}{2}}}{2^{r/2}\Gamma\left(\frac{r}{2}\right)}\cdot\frac{s^{\frac{n-i-r+1}{2}-1}e^{-\frac{s}{2}}}{2^{\frac{n-i-r+1}{2}}\Gamma\left(\frac{n-i-r+1}{2}\right)}\cdot ds\cdot dh = \\ &= \underbrace{\frac{1}{2^{(n-i+1)/2}\Gamma\left(\frac{r}{2}\right)\Gamma\left(\frac{n-i-r+1}{2}\right)}}_{c_1}\cdot \\ \cdot \iint_0^\infty (s+h)^{\frac{\gamma n}{2}} h^{\frac{r}{2}-1} s^{\frac{(1-\gamma)n-i-r-1}{2}} e^{-\frac{s+h}{2}}\cdot ds\cdot dh &= \\ \overset{(b)}{=} c_1 \iint_0^\infty w^{\frac{\gamma n}{2}}\left(\frac{1}{v+1}w\right)^{\frac{r}{2}-1}\left(\frac{v}{v+1}w\right)^{\frac{(1-\gamma)n-i-r-1}{2}}&\cdot \\ \cdot e^{-w/2}\frac{w}{(v+1)^2}\cdot dw\cdot dv = c_1\underbrace{\int_{w=0}^{\infty} w^{\frac{n-i-1}{2}}e^{-w/2}dw}_{c_w}&\cdot \\ \cdot \underbrace{\int_{v=0}^{\infty}\left(\frac{1}{v+1}\right)^{\frac{(1-\gamma)n-i+1}{2}} v^{\frac{(1-\gamma)n-i-r-1}{2}}\cdot dv}_{c_v} & \end{aligned} \tag{21}$$

where in (a) we used independent Chi-Squared distributed h, s, and in (b) we changed variables from s, h to $w = s+h, v = s/h$, with inverse transformation $s = \frac{v}{v+1}w, h = \frac{1}{v+1}w$ and Jacobian $J^{-1} = \frac{\partial w,v}{\partial s,h} = \begin{vmatrix} 1 & 1 \\ 1/h & -s/h^2 \end{vmatrix} = \frac{s+h}{h^2} = \frac{(v+1)^2}{w}$. The first integral in the expression above evaluates to:

$$c_w = 2^{\frac{n-i+1}{2}}\Gamma\left(\frac{n-i+1}{2}\right) \tag{22}$$

By definition of $\Gamma(\cdot)$. The second integral behaves like $v^{\frac{(1-\gamma)n-i-r-1}{2}}$ near $v = 0$ and like $v^{\frac{(1-\gamma)n-i-r-1}{2}-\frac{(1-\gamma)n-i+1}{2}} = v^{\frac{-r-2}{2}}$ at $v \to \infty$. Therefore it will exist (converge) iff the power of v near 0 is larger than -1 and at ∞ is smaller than -1. The first condition is $\frac{(1-\gamma)n-i-r-1}{2} > -1 \Rightarrow (1-\gamma)n > i+r-1 \Rightarrow \gamma < 1-\frac{i+r-1}{n}$. The other condition always holds since $r > 0$. Note that since the power of $\frac{1}{v+1}$ is larger by more than 1 than the power of v it is positive (when the first condition holds). Therefore we can bound:

$$c_v < \int_{v=0}^{\infty} \left(\frac{1}{\max(v,1)}\right)^{\frac{(1-\gamma)n-i+1}{2}} v^{\frac{(1-\gamma)n-i-r-1}{2}} \cdot dv = \\ = \frac{2}{(1-\gamma)n-i-r+1} + \frac{2}{r} \leq \frac{2}{(1-\gamma)n-t-r+1} + \frac{2}{r} \quad (23)$$

Combining (21), (22) and (23) we obtain:

$$D_i < \frac{\Gamma\left(\frac{n-i+1}{2}\right)}{\Gamma\left(\frac{r}{2}\right)\Gamma\left(\frac{n-i-r+1}{2}\right)} \cdot \left(\frac{2}{(1-\gamma)n-t-r+1} + \frac{2}{r}\right) \quad (24)$$

Since L results in a rate loss of $\frac{1}{n}\log L$, and $\Gamma\left(\frac{n-i+1}{2}\right)$ is superexponential in n, we would like to express more explicitly the dependence on n. Using $\Gamma(t+1) = t\Gamma(t)$ with $t = \frac{n-t+1-2i}{2}$, $i = 1, 2, ..\lceil r/2 \rceil$ we can obtain the bound

$$\frac{\Gamma\left(\frac{n-t+1}{2}\right)}{\Gamma\left(\frac{n-t+1-r}{2}\right)} \leq \left(\frac{n}{2}\right)^{\lceil r/2 \rceil} \quad (25)$$

therefore

$$D_i < \frac{\left(\frac{n}{2}\right)^{\lceil r/2 \rceil}}{\Gamma\left(\frac{r}{2}\right)} \cdot \left(\frac{2}{(1-\gamma)n-t-r+1} + \frac{2}{r}\right) \quad (26)$$

and

$$L = \prod_{i=1}^{t} D_i < \frac{\left(\frac{n}{2}\right)^{t\lceil r/2 \rceil}}{\Gamma^t\left(\frac{r}{2}\right)} \cdot \left(\frac{2}{(1-\gamma)n-t-r+1} + \frac{2}{r}\right)^t = \\ = C_L \cdot n^{t\lceil r/2 \rceil} \quad (27)$$

Substituting the above into (14) proves Lemma 1 for $\Lambda_x = \mathbf{I}$. The two assumptions on the parameters of the problem we have made in order for L to be bounded are (a) $n \geq t + r$ which was needed in order that each new column of $\mathbf{X}$ would not be spanned by the previous columns and the columns of $\mathbf{Y}$ in probability 1, and (b) $\forall i \leq t : \gamma < 1 - \frac{i+r-1}{n} \Rightarrow \gamma < 1 - \frac{t+r-1}{n}$, is needed for the existence of $\{D_i\}_{i=1}^{t}$.

Suppose now that $\mathbf{X} \sim \mathcal{N}^n(0, \Lambda_x)$. Using the Cholesky decomposition we can define a coloring matrix $\mathbf{W}$, $\mathbf{W}^T\mathbf{W} = \Lambda_x$ so that $\mathbf{X} = \mathbf{W} \cdot \mathbf{X}_w$ and $\mathbf{X}_w \sim \mathcal{N}^n(0, \mathbf{I})$. Since by Property 2 the rate function is invariant to a linear transformation of $\mathbf{X}$ we would have $R_{\text{emp}}(\mathbf{X}_w, \mathbf{Y}) = R_{\text{emp}}(\mathbf{X}, \mathbf{Y})$, therefore if Lemma 1 holds with respect to the white signal $\mathbf{X}_w$ it also holds with respect to $\mathbf{X}$. With regard to the assumption that the columns of $\mathbf{Y}$ are linearly independent: if they are not, then the rate function is defined with respect to a smaller matrix $\mathbf{Y}'_{n \times r'}$ containing only the independent columns. Comparing with a full rank $\mathbf{Y}$, the random variables d_i increase (i.e. $d'_i \geq d_i$) due to the smaller dimension of $\mathbf{Y}'$, therefore $L' \leq L$ and the lemma still holds. □

V. Comments and extensions

Comparison with the SISO case: Comparing Lemma 1 with Lemma 4 of [3] for the SISO case $r = t = 1$, which is proven by a direct calculation, the bound here is slightly worse due to the limitation $\gamma < 1 - \frac{t+r-1}{n} = (n-1)/n$ which stems from the use of the Chernoff bound.

Comparison with MIMO capacity: The scheme above achieves the mutual information of a Gaussian MIMO channel but not its capacity. Achieving the capacity requires adaptation of the input distribution, which for the known AWGN channel $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{v}$ is performed by SVD and water pouring [5]. The strength of the scheme is in the lack of any assumptions about the probability distribution, which make it applicable for example for non Gaussian noise or one that depends on the transmitted signal.

Exploiting temporal correlation: In the current results, as in previous ones [3], the rate function depends on the zero order empirical probability, and lacks the ability to exploit temporal correlation. However the results can be used to exploit such correlation in the SISO or MIMO channel, in a crude way, by applying the scheme on blocks of k channel uses. The rate function over blocks is always superior to the single letter case, and the penalty is an increase in the fixed redundancy.

Using empirical covariance instead of correlation: When the matrices $\hat{\mathbf{R}}$ in (1) are replaced with the empirical correlation $\hat{\mathbf{C}}$ (where $\hat{\mathbf{C}}_{\mathbf{X}} \equiv n^{-1}(\mathbf{X} - n^{-1}\mathbf{1}^T\mathbf{X})^T(\mathbf{X} - n^{-1}\mathbf{1}^T\mathbf{X})$), the derivation is similar, except projection on an additional dimension (the all-ones vector) precedes the other projections. The results are the same with a loss of one dimension: $\gamma < 1 - \frac{t+r}{n}$ and $n > t + r$ are required, and there is a small variation in C_L.

The complex MIMO channel: The results easily extend to the *complex*-valued MIMO channel, using the same technique. The main difference is a double number of degrees of freedom in the derivation of D_i, which doubles the rate compared to Equation 1.

Adaptivity: In [3][4] we presented a communication scheme using a low rate feedback, which dynamically adapts the transmission rate and achieves the rate functions without outage. Such schemes are of higher practical interest. It is possible to show that the adaptive scheme of [3][4] achieves R_{emp} of (1) up asymptotically vanishing redundancy, and up to a set of $\mathbf{x}$ sequences having vanishing probability.

References

[1] O. Shayevitz and M. Feder, "Achieving the Empirical Capacity Using Feedback: Memoryless Additive Models",IEEE Transactions on Information Theory, Vol.55 No.3, March 2009, pp.1269 -1295

[2] K. Eswaran, A.D. Sarwate, A. Sahai, and M. Gastpar, "Zero-rate feedback can achieve the empirical capacity," IEEE Transactions on Information Theory, Vol.58, No.1, January 2010

[3] Y. Lomnitz and M. Feder, "Communication over Individual Channels," arXiv:0901.1473v1 [cs.IT], 11 Jan 2009, http://arxiv.org/abs/0901.1473v1

[4] Y. Lomnitz and M. Feder, "Feedback communication over Individual Channels," IEEE International Symposium on Information Theory (ISIT), 2009, pp.1506-1510, June 28 2009-July 3 2009

[5] I. Telatar, "Capacity of Multi-antenna Gaussian Channels," AT&T Technical Memorandum, June 1995

[6] A. Goldsmith, S.A. Jafar, N. Jindal, and S. Vishwanath, "Capacity Limits of MIMO Channels, " IEEE Journal on Selected Areas in Communications, Vol. 21, No. 5, June 2003