Mutual Information as a Function of Matrix SNR for Linear Gaussian Channels

Galen Reeves*†, Henry D. Pfister* and Alex Dytso‡
*Department of Electrical Engineering, Duke University
†Department of Statistical Science, Duke University
‡Department of Electrical Engineering, Princeton University

***Abstract*—This paper focuses on the mutual information and minimum mean-squared error (MMSE) as a function a matrix-valued signal-to-noise ratio (SNR) for a linear Gaussian channel with arbitrary input distribution. As shown by Lamarca, the mutual-information is a concave function of a positive semi-definite matrix, which we call the matrix SNR. This implies that the mapping from the matrix SNR to the MMSE matrix is decreasing monotone. Building upon these functional properties, we start to construct a unifying framework that provides a bridge between classical information-theoretic inequalities, such as the entropy power inequality, and interpolation techniques used in statistical physics and random matrix theory. This framework provides new insight into the structure of phase transitions in coding theory and compressed sensing. In particular, it is shown that the parallel combination of linear channels with freely-independent matrices can be characterized succinctly via free convolution.**

***Index Terms*—I-MMSE, entropy power inequality, conditional central limit theorem, random matrix theory, compressed sensing. Gaussian logarithmic Sobolev inequality.**

I. Introduction

The functional properties of mutual information play an important role in applications across the mathematical sciences. In some cases, these functional properties lead to simple and elegant proofs of deep mathematical results. In other cases, they provide a crucial step in the proof of new results.

The focus of this paper is on the mutual information and minimum mean-squared error (MMSE) associated with the linear Gaussian channel, which is described by

$$\boldsymbol{Y} = A\boldsymbol{X} + \boldsymbol{N}, \tag{1}$$

where A is a $k \times n$ channel matrix, $\boldsymbol{X}$ is an n-dimensional random vector (or signal), and $\boldsymbol{N} \sim \mathcal{N}(0, I_n)$ is standard Gaussian noise. Our motivation comes largely from the fact that $I(\boldsymbol{X}; A\boldsymbol{X}+\boldsymbol{N})$ is closely related to interesting open questions in compressed sensing, coding theory, and statistical physics. In particular, the inference problems associated with these applications can all have phase transitions as the dimension increases.

For example, if $\boldsymbol{X}$ is uniformly distributed on points in a codebook and $A = \sqrt{s}I$, then we have coded communication over a Gaussian channel and the error probability can have a phase transition (as the problem size grows) when s increases [1]. If instead, one chooses A to be a random diagonal matrix where each entry is 0 with probability ϵ and $\sqrt{s}$ with probability $1-\epsilon$, then this approximates (for large s) an erasure channel and the error probability can have a phase transition as ϵ increases [2]. When $\boldsymbol{X}$ is a sparse random vector (e.g., i.i.d. with a probability mass at 0), there can be a phase transition in the MMSE as the sparsity level increases [3]. The locations of these phase transitions can sometimes be predicted by the non-rigorous replica method from statistical physics [4]. An interesting open question for all of these problems is determining if and when these replica formulas are correct.

Recently, the first two authors used information-theoretic tools to prove that the replica-symmetric prediction for the MMSE function in compressed sensing is exact when the channel matrix is drawn i.i.d. from the Gaussian ensemble [5]. Subsequent work established a similar result using a different proof technique [6], which was later extended to the generalized linear model [7]. Based on this, the first author described how these ideas might be extended to right-orthogonally invariant channel matrices and multilayer linear models [8]. This paper provides some rigorous progress on this problem.

A. Matrix SNR

The key insight underlying the results in this paper is that the linear Gaussian channel can be parameterized by a positive semi-definite matrix that generalizes the scalar signal-to-ratio (SNR). This insight follows from the fact that a linear Gaussian channel with $k \times n$ channel matrix A is statistically equivalent to a linear Gaussian channel with $n \times n$ channel matrix $(A^TA)^{\frac{1}{2}}$, regardless of the signal distribution. Although this fact follows naturally from the independence and orthogonal invariance of the Gaussian noise, our literature review did not reveal any references before the 2009 paper of Lamarca [9], and we believe this result has not been fully exploited. This approach also brings the multivariate derivative formula for the mutual information closer to the well-developed theory associated with the scalar I-MMSE relationship [10].

An important consequence of the parametrization described above is that one obtains an explicit additivity rule for linear Gaussian channels. Specifically, two parallel channels with matrices A and B, respectively, are equivalent to a channel

The work of G. Reeves was supported in part by funding from the Laboratory for Analytic Sciences (LAS). The work of G. Reeves and H. Pfister was supported part by the NSF under Grant No. 1718494. Any opinions, findings, conclusions, and recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

978-1-5386-4780-6/18/$31.00©2018 IEEE

with $n \times n$ matrix $(A^TA + B^TB)^{\frac{1}{2}}$. Combining this fact with the chain rule and the low-noise expansion of mutual information [11], one finds that: 1) the mutual information $I(\boldsymbol{X}; A\boldsymbol{X} + \boldsymbol{N})$ is concave in A^TA; and 2) the minimum mean-square error matrix $\mathbb{E}[\mathsf{Cov}(\boldsymbol{X} \mid A\boldsymbol{X} + \boldsymbol{N})]$ is decreasing monotone map of A^TA. Mathematically, this means that for any n-dimensional random vector $\boldsymbol{X}$ and matrices A, B with n columns, one has

$$\begin{aligned}\lambda I(\boldsymbol{X}; A\boldsymbol{X} + \boldsymbol{N}) + (1-\lambda) I(\boldsymbol{X}; B\boldsymbol{X} + \boldsymbol{N}') \\ \leq I(\boldsymbol{X}; (\lambda A^TA + (1-\lambda)B^TB)^{\frac{1}{2}}\boldsymbol{X} + \boldsymbol{N}''),\end{aligned} \tag{2}$$

for all $0 \leq \lambda \leq 1$ where $\boldsymbol{N}, \boldsymbol{N}'$ and $\boldsymbol{N}''$ are independent standard Gaussian vectors. Furthermore, the MMSE matrix satisfies $\left\langle A^TA - B^TB, \Delta \right\rangle \leq 0$ where $\Delta = \mathbb{E}[\mathsf{Cov}(\boldsymbol{X}|A\boldsymbol{X} + \boldsymbol{N})] - \mathbb{E}[\mathsf{Cov}(\boldsymbol{X}|B\boldsymbol{X} + \boldsymbol{N}')]$ and $\langle \cdot, \cdot \rangle$ denotes the trace inner-product on matrices.

Remark 1. Although it may be tempting to reinterpret (2) in terms of differential entropy, such a decomposition becomes cumbersome due to a mismatch between problem dimensions.

Remark 2. It is important to note that (2) holds for arbitrary matrices. Increasing the scale of the matrices $A \mapsto \sqrt{s}A$ for some positive number s is equivalent to decreasing the noise power. If $\boldsymbol{X}$ has finite entropy then the low-noise limit is well defined and one obtains

$$\begin{aligned}\lambda H(\boldsymbol{X} \mid A\boldsymbol{X}) + (1-\lambda) H(\boldsymbol{X} \mid B\boldsymbol{X}) \\ \geq H(\boldsymbol{X} \mid (\lambda A^TA + (1-\lambda)B^TB)^{\frac{1}{2}}\boldsymbol{X}).\end{aligned} \tag{3}$$

B. Our Contributions

Building upon the aforementioned functional properties, we obtain the following results:

- An effective Fisher information matrix is introduced and shown to have a monotonicity property that provides a multivariate version of the single-crossing property (Theorem 2). Using this result, we obtain matrix inequalities that mimic well-known scalar bounds on the mutual information.
- A new relationship is identified between three different measures of the distance between the distribution of an n-dimensional random vector and an i.i.d. Gaussian distribution. In particular, Theorem 3 shows that the relative entropy measure can be expressed as the sum of relative entropy associated with a (random) low-dimensional projection and a second term that can be related to the deficit in the entropy power inequality (EPI).
- Using ideas from free probability theory, concentration results and functional properties are established that characterize the asymptotic behavior of right-orthogonally invariant random matrices (Theorem 4). Although these results are closely linked to the conjectured limits implied by the replica method, they are proven rigorously here under mild assumptions.

Due to space constraints many of the proofs are either sketched or omitted.

C. Notation

We use $\mathbb{S}^n$, $\mathbb{S}^n_+$ and $\mathbb{S}^n_{++}$ to denote the space $n \times n$ symmetric matrices, positive semi-definite matrices, positive definite matrices, respectively. Given a positive-definite matrix S, we use $S^{\frac{1}{2}}$ to denote the positive-definite square root.

II. Matrix-SNR Functions

For any n-dimensional random vector $\boldsymbol{X}$, the mutual information function $I_{\boldsymbol{X}} : \mathbb{S}^n_+ \to [0, \infty)$ and MMSE function $M_{\boldsymbol{X}} : \mathbb{S}^n_+ \to \mathbb{S}^n_+$ are defined by

$$I_{\boldsymbol{X}}(S) \triangleq I(\boldsymbol{X}; S^{\frac{1}{2}}\boldsymbol{X} + \boldsymbol{N}) \tag{4}$$

$$M_{\boldsymbol{X}}(S) \triangleq \mathbb{E}\Big[\mathsf{Cov}(\boldsymbol{X} \mid S^{\frac{1}{2}}\boldsymbol{X} + \boldsymbol{N})\Big], \tag{5}$$

where $\boldsymbol{N}$ is a standard Gaussian vector. The following result shows that these functions can be related to the mutual information and MMSE associated with an arbitrary $k \times n$ matrix A. For completeness, we also sketch a partial proof below.

Lemma 1 ([9])**.** *For every $k \times n$ matrix A, we have*

$$I_{\boldsymbol{X}}(A^TA) = I(\boldsymbol{X}; A\boldsymbol{X} + \boldsymbol{W}) \tag{6}$$

$$M_{\boldsymbol{X}}(A^TA) = \mathbb{E}[\mathsf{Cov}(\boldsymbol{X} \mid A\boldsymbol{X} + \boldsymbol{W})], \tag{7}$$

where $\boldsymbol{W} \sim \mathcal{N}(0, I_k)$.

Sketch of Proof. First, consider the case $k < n$. Let B be the $n \times n$ matrix whose first k rows are equal to A and whose remaining $(n-k)$ rows are equal to zero. Clearly, $I(\boldsymbol{X}; A\boldsymbol{X} + \boldsymbol{W}) = I(\boldsymbol{X}; B\boldsymbol{X} + \boldsymbol{N})$, because the extra rows in B do not convey any information about $\boldsymbol{X}$. Next, consider the singular value decomposition $B = U\Sigma V^T$ where all matrices are $n \times n$. By the orthogonal invariance of the standard Gaussian distribution, we can write

$$\begin{aligned}I(\boldsymbol{X}; B\boldsymbol{X} + \boldsymbol{N}) &= I(\boldsymbol{X}; U\Sigma V^T\boldsymbol{X} + \boldsymbol{N}) \\ &= I(\boldsymbol{X}; \Sigma V^T\boldsymbol{X} + \boldsymbol{N}) \\ &= I(\boldsymbol{X}; V\Sigma V^T\boldsymbol{X} + \boldsymbol{N}).\end{aligned}$$

Noting that $V\Sigma V^T = \left(B^TB\right)^{\frac{1}{2}} = \left(A^TA\right)^{\frac{1}{2}}$ gives the stated identity. The case of $k \geq n$ follows from similar arguments and proofs of the MMSE result (7) can be found in [9], [12]. □

Lemma 2. *For all $S, T \in \mathbb{S}^n_+$, we have*

$$I_{\boldsymbol{X}}(S + T) = I(\boldsymbol{X}; \boldsymbol{Y}, \boldsymbol{Z}) \tag{8}$$

$$M_{\boldsymbol{X}}(S + T) = \mathbb{E}[\mathsf{Cov}(\boldsymbol{X} \mid \boldsymbol{Y}, \boldsymbol{Z})], \tag{9}$$

where $\boldsymbol{Y} = S^{\frac{1}{2}}\boldsymbol{X} + \boldsymbol{N}$, $\boldsymbol{Z} = T^{\frac{1}{2}}\boldsymbol{X} + \boldsymbol{N}'$, and $\boldsymbol{N}$ and $\boldsymbol{N}'$ are independent standard Gaussian vectors.

Proof. The right-hand side of (8) can be expressed as $I(\boldsymbol{X}; A\boldsymbol{X} + \boldsymbol{N}'')$ where A is the $2n \times n$ matrix obtained by stacking $S^{\frac{1}{2}}$ and $T^{\frac{1}{2}}$ and $\boldsymbol{N}''$ is the $2n \times 1$ vector obtained by stacking $\boldsymbol{N}$ and $\boldsymbol{N}'$. Noting that $A^TA = S + T$ and invoking Lemma 1 gives the stated result. □

Lemma 3 (Scaling property)**.** *For any $k \times n$ matrix A and $S \in \mathbb{S}_+^n$, we have*

$$I_{A\boldsymbol{X}}(S) = I_{\boldsymbol{X}}(ASA^T) \tag{10}$$
$$M_{A\boldsymbol{X}}(S) = AM_{\boldsymbol{X}}(ASA^T)A^T. \tag{11}$$

Lemma 4. *The mutual information function $I_{\boldsymbol{X}}(S)$ is twice differentiable on $\mathbb{S}_{++}^n$ with gradient and Hessian given by*

$$\nabla_S I_{\boldsymbol{X}}(S) = \frac{1}{2} M_{\boldsymbol{X}}(S) \tag{12}$$
$$\nabla_S^2 I_{\boldsymbol{X}}(S) = -\frac{1}{2}\mathbb{E}_{\boldsymbol{Y}}[\Phi_{\boldsymbol{X}}(\boldsymbol{Y}) \otimes \Phi_{\boldsymbol{X}}(\boldsymbol{Y})] \tag{13}$$

where $\boldsymbol{Y} = S^{\frac{1}{2}}\boldsymbol{X} + \boldsymbol{N}$ and $\Phi_{\boldsymbol{X}}(\boldsymbol{y}) = \mathsf{Cov}(\boldsymbol{X} \mid \boldsymbol{Y} = \boldsymbol{y})$.

Remark 3. The Hessian is negative semi-definite since the covariance $\Phi_{\boldsymbol{X}}(\boldsymbol{Y})$ is positive semi-definite. This implies that (6) is a concave function on $\mathbb{S}_{++}^n$. This concavity and the results in Lemma 4 can be found in Lamarca [9] and Payaró et al. [12].

The concavity is summarized by the following theorem.

Theorem 1. *For any n-dimensional random vector $\boldsymbol{X}$, The mutual information is concave on $\mathbb{S}_+^n$. In other words,*

$$(1-\lambda)I_{\boldsymbol{X}}(S) + \lambda I_{\boldsymbol{X}}(T) \geq I_{\boldsymbol{X}}((1-\lambda)S + \lambda T) \tag{14}$$

for all $S, T \in \mathbb{S}_+^n$ and $\lambda \in [0,1]$. Furthermore, the MMSE function (7) *is a decreasing monotone mapping on $\mathbb{S}_+^n$. Thus,*

$$\operatorname{tr}((S-T)(M_{\boldsymbol{X}}(S) - M_{\boldsymbol{X}}(T))) \leq 0 \tag{15}$$

for all $S, T \in \mathbb{S}_+^n$.

III. Applications

This section describes some new results that can be obtained using properties of the mutual information and MMSE matrix as a function of the matrix SNR.

A. Bounds on the mutual information and MMSE

The Fisher information matrix of a random vector $\boldsymbol{Y}$ with density $p(\boldsymbol{y})$ is defined to be $J(\boldsymbol{Y}) \triangleq \mathsf{Cov}(\rho(\boldsymbol{Y}))$ where $\rho(\boldsymbol{y}) \triangleq \nabla \log p(\boldsymbol{y})$ is the score function. Given any random vector $\boldsymbol{X}$ with $\mathsf{Cov}(\boldsymbol{X}) \in \mathbb{S}_{++}^n$, we define the effective Fisher information matrix $K_{\boldsymbol{X}} \colon \mathbb{S}_+^n \to \mathbb{S}_+^n$ by

$$K_{\boldsymbol{X}}(S) \triangleq M_{\boldsymbol{X}}^{-1}(S) - S. \tag{16}$$

This is the Fisher information matrix of the multivariate Gaussian distribution whose MMSE matrix equals $M_{\boldsymbol{X}}(S)$. Using the matrix version of Brown's identity [13, Proposition 6]

$$J(\boldsymbol{X} + S^{-\frac{1}{2}}\boldsymbol{N})S^{-1} + SM_{\boldsymbol{X}}(S) = I, \tag{17}$$

it can be verified that

$$K_{\boldsymbol{X}}^{-1}(S) = J^{-1}(\boldsymbol{X} + S^{-\frac{1}{2}}\boldsymbol{N}) - S^{-1} \tag{18}$$

for all $S \in \mathbb{S}_{++}^n$. This characterization shows that $K_{\boldsymbol{X}}^{-1}(S)$ is well-defined even if $\mathsf{Cov}(\boldsymbol{X})$ is degenerate.

Lemma 5. *The effective Fisher information matrix $K_{\boldsymbol{X}}(S)$ is decreasing monotone on $\mathbb{S}_+^n$ with*

$$K_{\boldsymbol{X}}(0) = \mathsf{Cov}^{-1}(\boldsymbol{X}), \tag{19}$$
$$\lim_{\lambda_{\min}(S)\to\infty} K_{\boldsymbol{X}}(S) = J(\boldsymbol{X}). \tag{20}$$

Sketch of Proof. As the gradient is negative semi-definite, monotonicity follows. The limits come from (16) and (18). □

Theorem 2. *Let $\boldsymbol{X}$ be an n-dimensional random vector with positive-definite covariance matrix. For any matrices $R, S, T \in \mathbb{S}_+$ with $R \preceq S \preceq T$, the MMSE matrix satisfies*

$$(K_{\boldsymbol{X}}(T) + S)^{-1} \preceq M_{\boldsymbol{X}}(S) \preceq (K_{\boldsymbol{X}}(R) + S)^{-1}. \tag{21}$$

Sketch of Proof. This result follows from the monotonicity of $K_{\boldsymbol{X}}(S)$, established in Lemma 5, and the fact that inversion reverses the partial-order on positive semi-definite matrices. □

Theorem 2 can be thought of as a matrix generalization of the single-crossing property (see [14]). Taking the limits $\lambda_{\max}(T) \to 0$ and $\lambda_{\min}(R) \to \infty$, one obtains the bounds

$$(J(\boldsymbol{X}) + S)^{-1} \preceq M_{\boldsymbol{X}}(S) \preceq \left(\mathsf{Cov}^{-1}(\boldsymbol{X}) + S\right)^{-1}. \tag{22}$$

Note that the left-hand side of (22) is a matrix version of the Bayesian Cramer-Rao lower bound [15]. Here the lower bound is meaningful only if $\boldsymbol{X}$ has finite Fisher information. The right hand side of (22) is often called the linear MMSE.

By the multivariate I-MMSE relationship (10), we see that the mutual information can be expressed as

$$I_{\boldsymbol{X}}(S) = \frac{1}{2}\int_0^1 \operatorname{tr}\left(M_{\boldsymbol{X}}(S_t)\,\frac{\mathrm{d}}{\mathrm{d}t}S_t\right)\mathrm{d}t \tag{23}$$

where the integral is over any differentiable path $t \mapsto S_t$ with $S_0 = 0$ and $S_1 = S$. In particular, letting $S_t = tS$, swapping the integral and the trace, and using Lemma 3, leads to

$$I_{\boldsymbol{X}}(S) = \frac{1}{2}\operatorname{tr}\left(S\int_0^1 M_{\boldsymbol{X}}(tS)\,\mathrm{d}t\right) \tag{24}$$
$$= \frac{1}{2}\operatorname{tr}\left(\int_0^1 M_{S^{\frac{1}{2}}\boldsymbol{X}}(tI)\,\mathrm{d}t\right). \tag{25}$$

Interestingly, this decomposition shows that the mutual information can be viewed as the trace of an integrated MMSE matrix. A similar observation was made by Dembo [16, pg. 14], who showed that an entropy can be expressed as the trace of an integral involving the Fisher information matrix.

Combining Theorem 2 with (25) provides a lower bound on the *matrix* inside the trace in (25):

$$\begin{aligned}
\int_0^1 M_{S^{\frac{1}{2}}\boldsymbol{X}}(tI)\,\mathrm{d}t &= \int_0^1 \left(K_{S^{\frac{1}{2}}\boldsymbol{X}}(tI) + tI\right)^{-1}\mathrm{d}t \\
&\succeq \int_0^1 \left(K_{S^{\frac{1}{2}}\boldsymbol{X}}(I) + tI\right)^{-1}\mathrm{d}t \\
&= \log\left(I + K^{-1}_{S^{\frac{1}{2}}\boldsymbol{X}}(I)\right) \\
&= \log\left(J^{-1}(S^{\frac{1}{2}}\boldsymbol{X} + \boldsymbol{N})\right) \\
&\succeq \log\left(I + S^{\frac{1}{2}}J^{-1}(\boldsymbol{X})S^{\frac{1}{2}})\right).
\end{aligned}$$

Here, we recall that the matrix logarithm is well-defined on $\mathbb{S}^n_{++}$. Taking the trace of both sides recovers some well-known lower bounds on the mutual information

$$\begin{aligned} I_{\boldsymbol{X}}(S) &\geq \frac{1}{2}\log\det(J^{-1}(S^{\frac{1}{2}}\boldsymbol{X}+\boldsymbol{N})) \\ &\geq \frac{1}{2}\log\det(I+SJ^{-1}(\boldsymbol{X})). \end{aligned}$$

B. Gaussian approximation via low-dimensional projections

Many of the standard approaches to measuring relative entropy with respect to the Gaussian measure are based on the fact that adding an independent Gaussian component to a random vector decreases the relative entropy to a Gaussian (see [17, Section 3.2]).

A related but different phenomenon is that most low-dimensional projections of a high-dimensional distribution are closer to an i.i.d. Gaussian, in a relative sense, than the original distribution. Some recent information-theoretic bounds, known as conditional central limit theorems (CCLTs), are provided by the first author in [18]. One motivation for the present paper is to understand how the decrease in approximation error associated with low-dimensional projections relates to properties of the original distribution.

For concreteness, we will focus on projections on the Stiefel manifold $\mathcal{V}_k(\mathbb{R}^n)$, which is the set of $k\times n$ matrices A satisfying $AA^T = I_k$. Furthermore, we will consider distributions of the form $\sqrt{s}\boldsymbol{X}+\boldsymbol{N}$ where $s\in(0,\infty)$ and $\boldsymbol{N}$ is a Gaussian perturbation. Under regularity conditions on $\boldsymbol{X}$ (e.g., finite differential entropy) the perturbation can made negligible by taking the large s limit. Finally, let $\boldsymbol{X}^*\sim\mathcal{N}(0,\frac{1}{n}I_n)$ be an i.i.d. Gaussian vector with same power as $\boldsymbol{X}$.

An important property of the i.i.d. Gaussian distribution is that, for any $A\in\mathcal{V}_k(\mathbb{R}^n)$, the k-dimensional projection $A\boldsymbol{X}^*$ is equal in distribution to the first k entries in $\boldsymbol{X}^*$. Consequently, the mutual information function satisfies

$$I_{\boldsymbol{X}^*}(s\,A^TA) = \frac{k}{n}I_{\boldsymbol{X}^*}(sI). \tag{26}$$

In the special case where the entries of $\boldsymbol{X}$ are independent, it follows from the linear entropy power inequality of Zamir and Feder [19] that

$$I_{\boldsymbol{X}}(s\,A^TA) \geq \sum_{i=1}^n [A^TA]_{i,i}I_{X_i}(s). \tag{27}$$

Furthermore, if $\boldsymbol{A}$ is drawn uniformly at random from $\mathcal{V}_k(\mathbb{R}^n)$, then expectation of the right-hand side is given by

$$\mathbb{E}\left[\sum_{i=1}^n [\boldsymbol{A}^T\boldsymbol{A}]_{i,i}I_{X_i}(s)\right] = \sum_{i=1}^n \frac{k}{n}I_{X_i}(s) = \frac{k}{n}I_{\boldsymbol{X}}(sI). \tag{28}$$

Motivated by (28), we define the average EPI deficit to be

$$\delta_{\mathrm{EPI}} \triangleq \frac{1}{k}\mathbb{E}\big[I_{\boldsymbol{X}}(s\,\boldsymbol{A}^T\boldsymbol{A})\big] - \frac{1}{n}I_{\boldsymbol{X}}(sI), \tag{29}$$

where $\boldsymbol{A}$ is drawn uniformly at random from $\mathcal{V}_k(\mathbb{R}^n)$. If the entries of $\boldsymbol{X}$ are not independent, then (27) may not hold. However, as we will see below, δ_{EPI} is non-negative for every distribution on $\boldsymbol{X}$.

Next, we consider the relative entropy between the k-dimensional projections. For each $A\in\mathcal{V}_k(\mathbb{R}^n)$, we have

$$D\big(P_{\sqrt{s}A\boldsymbol{X}+\boldsymbol{N}}\,\|\,P_{\sqrt{s}A\boldsymbol{X}^*+\boldsymbol{N}}\big) = \frac{k}{n}I_{\boldsymbol{X}^*}(sI) - I_{\boldsymbol{X}}(sA^TA).$$

We define the average CCLT deficit to be

$$\delta_{\mathrm{CCLT}} \triangleq \frac{1}{n}I_{\boldsymbol{X}^*}(sI) - \frac{1}{k}\mathbb{E}\big[I_{\boldsymbol{X}}(s\,\boldsymbol{A}^T\boldsymbol{A})\big], \tag{30}$$

where $\boldsymbol{A}$ is drawn uniformly at random from $\mathcal{V}_k(\mathbb{R}^n)$.

The following result show that the average EPI deficit and the average CCLT deficit can both be viewed a measures of the distance between the distribution of $\boldsymbol{X}$ and the i.i.d. approximation $\boldsymbol{X}^*$, and that their sum is proportional to the relative entropy after convolution with a Gaussian.

Theorem 3. *Let $\boldsymbol{X}$ be an n-dimensional random vector with finite covariance. For all $s\in(0,\infty)$ and $k\in\{1,\ldots,n-1\}$,*

$$\frac{1}{n}D\big(P_{\sqrt{s}\boldsymbol{X}+\boldsymbol{N}}\,\|\,P_{\sqrt{s}\boldsymbol{X}^*+\boldsymbol{N}}\big) = \delta_{\mathrm{CCLT}} + \delta_{\mathrm{EPI}}. \tag{31}$$

Furthermore, each of the terms on the right-hand side is non-negative and equal to zero if and only if $\boldsymbol{X}$ is i.i.d. Gaussian.

Remark 4. Part of the significance of Theorem 3 is that it provides a direct link between the deficit in the EPI and the relative entropy with respect to the i.i.d. Gaussian distribution. Combining results from [18] with the concavity of the mutual information in Theorem 1, it can be shown that, under mild conditions on $\boldsymbol{X}$, the average CCLT deficit δ_{CCLT} is small whenever $k\ll n$.

C. Additivity of information via free probability

Suppose that one obtains outputs from two independent linear Gaussian channels:

$$\boldsymbol{Y} = A\boldsymbol{X}+\boldsymbol{N}, \qquad \boldsymbol{Z} = B\boldsymbol{X}+\boldsymbol{N}', \tag{32}$$

where A and B both have n columns and $\boldsymbol{N}$ and $\boldsymbol{N}'$ are independent standard Gaussian vectors. A natural question of interest is whether the parallel combination of these channels can be characterized in terms of the individual channels. As a direct consequence of the the additivity formula in Lemma 2, we see that this question is directly related to additivity of the corresponding matrix SNRs:

$$I(\boldsymbol{X};\boldsymbol{Y},\boldsymbol{Z}) = I_{\boldsymbol{X}}(A^TA+B^TB). \tag{33}$$

In the special case where the signal $\boldsymbol{X}$ is i.i.d. Gaussian, the mutual information depends only on the singular values of the channel matrix, or equivalently the eigenvalues of matrix SNR. Consequently, the combination of the channels is completely characterized by the eigenvalues of the sum A^TA+B^TB. In general, the eigenvalues of the sum of two matrices cannot be determined based only on the eigenvalues of the individual matrices. However, one of the central results from free probability theory in that eigenvalues of the sum of freely independent matrices can be characterized in the large system limit via free additive convolution; see [20], [21].

Assumption 1. For each n, $\boldsymbol{A}_n$ and $\boldsymbol{B}_n$ are independent right-orthogonally invariant random matrices with n columns. Furthermore, as n increases to infinity, the (random) empirical spectral distributions of $\boldsymbol{A}_n^T\boldsymbol{A}_n$ and $\boldsymbol{B}_n^T\boldsymbol{B}_n$ converge almost surely to compactly supported probability measures μ and ν.

Lemma 6 ([20])**.** *Under Assumption 1, the empirical spectral distribution of the sum $\boldsymbol{A}_n^T\boldsymbol{A}_n + \boldsymbol{B}_n^T\boldsymbol{B}_n$ converges almost surely to the probability measured given by the free additive convolution of μ and ν, which is denoted by $\mu \boxplus \nu$.*

For the general case of non-Gaussian signal priors, the mutual information depends on the right-singular vectors of the channel matrix and thus cannot be determined based only on the singular values of the channel matrix. Interestingly though, the additivity principle seen in the case of Gaussian priors is still applicable when the channel matrices are freely independent.

Given a probability measure μ on $[0,\infty)$ we define

$$\mathcal{I}_n(\mu) = \frac{1}{n}\mathbb{E}\big[I_{\boldsymbol{X}}(\boldsymbol{U}^T\Lambda\boldsymbol{U})\big], \tag{34}$$

where $\boldsymbol{U}$ distributed uniformly on the group of $n\times n$ orthogonal matrices and Λ is a diagonal matrix whose entries are i.i.d. according to μ.

Assumption 2. For each n, $\boldsymbol{X}_n$ is an n-dimensional random vector with bounded second moment: $\frac{1}{n}\mathbb{E}\big[\|\boldsymbol{X}\|^2\big] < B$

The next result shows that normalized mutual information associated with a random orthogonally invariant channel matrix converges to its expectation, and furthermore, that the combination of channels with freely independent matrices is characterized by the free additive convolution.

Theorem 4. *Under Assumptions 1 and 2, the following convergence holds almost surely in the limit as $n \to \infty$:*

$$\left|\frac{1}{n}I_{\boldsymbol{X}_n}(\boldsymbol{A}_n^T\boldsymbol{A}_n) - \mathcal{I}_n(\mu)\right| \to 0 \tag{35}$$

$$\left|\frac{1}{n}I_{\boldsymbol{X}_n}(\boldsymbol{B}_n^T\boldsymbol{B}_n) - \mathcal{I}_n(\nu)\right| \to 0 \tag{36}$$

$$\left|\frac{1}{n}I_{\boldsymbol{X}_n}(\boldsymbol{A}_n^T\boldsymbol{A}_n + \boldsymbol{B}_n^T\boldsymbol{B}_n) - \mathcal{I}_n(\mu\boxplus\nu)\right| \to 0. \tag{37}$$

Sketch of Proof. Using the multivariate I-MMSE relationship (10), we show that $\mathcal{I}_n(\mu)$ is Lipschitz with a constant that depends only on the second-moment of the signal. Concentration with respect to the eigenvectors is established using further Lipschitz properties and standard concentration of measure arguments for the Haar measure on the Stiefel manifold. □

The question of whether $\mathcal{I}_n(\mu)$ converges to a well-defined limit is still open in general. Recent work has proved the existence of the limit in the special case of i.i.d. Gaussian matrices [5], [6]. More generally, analysis based on the replica method from statistical physics has provided postulated single-letter formulas for the limit (34) associated with an arbitrary spectral distribution [22], [23]. Recent work by the first author [8] provides an alternative to the replica method that can be applied to the composition of multiple channels.

An important open problem is proving that the conjectured limit of (34) is correct. The significance of Theorem 4 is that it imposes a functional constraint on (34) for large-n that is closely related to the conjectured limit.

References

[1] T. Richardson, M. A. Shokrollahi, and R. Urbanke, "Design of capacity-approaching irregular low-density parity-check codes," *IEEE Trans. Inform. Theory*, vol. 47, pp. 619–637, Feb. 2001.

[2] S. Kudekar, S. Kumar, M. Mondelli, H. D. Pfister, E. Şaşoğlu, and R. Urbanke, "Reed-Muller codes achieve capacity on erasure channels," *IEEE Trans. Inform. Theory*, vol. 63, no. 7, pp. 4298–4316, 2017.

[3] Y. Kabashima, T. Wadayama, and T. Tanaka, "A typical reconstruction limit for compressed sensing based on ℓ_p-norm minimization," *J. of Stat. Mech.: Theory and Exper.*, vol. 2009, p. L09003, 2009.

[4] L. Zdeborová and F. Krzakala, "Statistical physics of inference: Thresholds and algorithms," *Adv. in Phys.*, vol. 65, no. 5, pp. 453–552, 2016.

[5] G. Reeves and H. D. Pfister, "The replica-symmetric prediction for compressed sensing with Gaussian matrices is exact," in *Proc. IEEE Int. Symp. Inform. Theory*, Barcelona, Spain, Jul. 2016, pp. 665 – 669.

[6] J. Barbier, M. Dia, N. Macris, and F. Krzakala, "The mutual information in random linear estimation," in *Proc. Annual Allerton Conf. on Commun., Control, and Comp.*, Monticello, IL, 2016.

[7] J. Barbier, F. Krzakala, N. Macris, L. Miolane, and L. Zdeborová, "Phase transitions, optimal errors and optimality of message-passing in generalized linear models," Aug. 2017, [Online]. Available https://arxiv.org/abs/1708.03395.

[8] G. Reeves, "Additivity of information in multilayer networks via additive Gaussian noise transforms," in *Proc. Annual Allerton Conf. on Commun., Control, and Comp.*, Monticello, IL, 2017.

[9] M. Lamarca, "Linear precoding for mutual information maximization in MIMO systems," in *Proc. of the Int. Conf. on Wireless Comm. Syst.*, Tuscany, Italy, Sep. 2009.

[10] D. Guo, S. Shamai, and S. Verdú, "Mutual information and minimum mean-square error in Gaussian channels," *IEEE Trans. Inform. Theory*, vol. 51, no. 4, pp. 1261–1282, 2005.

[11] V. V. Prelov and S. Verdú, "Second-order asymptotics of mutual information," *IEEE Trans. Inform. Theory*, vol. 50, no. 8, pp. 1567–1580, Aug. 2004.

[12] M. Payaró, M. Gregori, and D. Palomar, "Yet another entropy power inequality with an application," in *Proc. of the Intern. Conf. on Wireless Commun. and Sign. Process.*, Nanjing, China, Nov. 2011.

[13] O. Rioul, "Information theoretic proofs of entropy power inequalities," *IEEE Trans. Inform. Theory*, vol. 57, no. 1, pp. 33–55, Jan. 2011.

[14] R. Bustin, M. Payaró, D. P. Palomar, and S. Shamai, "On mmse crossing properties and implications in parallel vector Gaussian channels," *IEEE Trans. Inform. Theory*, vol. 59, no. 2, pp. 818–844, 2013.

[15] H. L. Van Trees, *Detection, estimation, and modulation theory*. John Wiley & Sons, 2004.

[16] A. Dembo, "Information inequalities and uncertainty principles," Department of Statistics, Stanford University, Stanford, CA, Tech. Rep. 75, 1990.

[17] M. Raginsky and I. Sason, *Concentration of Measure Inequalities in Information Theory, Communications, and Coding*, 2nd ed. Foundations and Trends in Communications and Information Theory, 2014.

[18] G. Reeves, "Conditional central limit theorems for Gaussian projections," in *Proc. IEEE Int. Symp. Inform. Theory*, Aachen, Germany, Jun. 2017, pp. 3055–3059.

[19] R. Zamir and M. Feder, "A generalization of the entropy power inequality with applications," *IEEE Trans. Inform. Theory*, vol. 39, no. 5, pp. 1723–1728, Sep. 1993.

[20] D. Voiculescu, "Limit laws for random matrices and free products," *Inventiones mathematicae*, vol. 104, no. 1, pp. 201–220, 1991.

[21] D. N. C. Tse, "Multiuser receivers, random matrices and free probability," in *Proc. Annual Allerton Conf. on Commun., Control, and Comp.*, 1999.

[22] K. Takeda, S. Uda, and Y. Kabashima, "Analysis of CDMA systems that are characterized by eigenvalue spectrum," *Europhysics Letters*, vol. 76, no. 6, pp. 1193– 1199, Dec. 2006.

[23] A. Tulino, G. Caire, S. Verdú, and S. Shamai, "Support recovery with sparsely sampled free random matrices," *IEEE Trans. Inform. Theory*, vol. 59, no. 7, pp. 4243–4271, Jul. 2013.