An Extremal Inequality Motivated by Multiterminal Information Theoretic Problems

Tie Liu and Pramod Viswanath*

November 8, 2006

Abstract

We prove a new extremal inequality, motivated by the vector Gaussian broadcast channel and the distributed source coding with a single quadratic distortion constraint problems. As a corollary, this inequality yields a generalization of the classical entropy-power inequality (EPI). As another corollary, this inequality sheds insight into maximizing the differential entropy of the sum of two dependent random variables.

Keywords: Differential entropy, distributed source coding, entropy-power inequality (EPI), Fisher information, vector Gaussian broadcast channel

1   Introduction

Like many other important results in information theory, the classical entropy-power inequality (EPI) was discovered by Shannon [1] (even though the first rigorous proof was given by Stam [2] and was later simplified by Blachman [3]). In [1, p. 641], Shannon used the EPI to prove a lower bound on the capacity of additive noise channels. While this first application was on a point-to-point scenario, the real value of the EPI showed up much later in the

---

*Tie Liu is with the Department of Electrical and Computer Engineering at the Texas A&M University, College Station, TX 77843, USA (e-mail: tieliu@ece.tamu.edu). Pramod Viswanath is with the Department of Electrical and Computer Engineering at the University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA (e-mail: pramodv@uiuc.edu).

multiterminal source/channel coding problems where the tension among users of different interests cannot be resolved by Fano's inequality alone. The most celebrated examples include Bergman's solution [4] to the scalar Gaussian broadcast channel problem, Oohama's solution [5] to the scalar quadratic Gaussian CEO problem, and Ozarow's solution [6] to the scalar Gaussian two-description problem.

Denote the set of real numbers by $\mathcal{R}$. Let $\mathbf{X}$, $\mathbf{Z}$ be two independent random vectors with densities in $\mathcal{R}^n$. The classical EPI states that

$$\exp\left[\frac{2}{n}h(\mathbf{X}+\mathbf{Z})\right] \geq \exp\left[\frac{2}{n}h(\mathbf{X})\right] + \exp\left[\frac{2}{n}h(\mathbf{Z})\right]. \tag{1}$$

Here $h(\mathbf{X})$ denotes the differential entropy of $\mathbf{X}$, and the equality holds if and only if $\mathbf{X}$, $\mathbf{Z}$ are Gaussian *and with proportional covariance matrices*.

Fix $\mathbf{Z}$ to be Gaussian with covariance matrix $\mathbf{K}_Z$. Assume that $\mathbf{K}_Z$ is strictly positive definite. Consider the optimization problem

$$\max_{p(\mathbf{x})} \left\{ h(\mathbf{X}) - \mu h(\mathbf{X}+\mathbf{Z}) \right\}, \tag{2}$$

where $\mu \in \mathcal{R}$, and the maximization is over all random vector $\mathbf{X}$ independent of $\mathbf{Z}$. The classical EPI can be used to show that for any $\mu > 1$, a Gaussian $\mathbf{X}$ with a covariance matrix proportional to $\mathbf{K}_Z$ is an optimal solution of this optimization problem. This can be done as follows. By the classical EPI,

$$h(\mathbf{X}) - \mu h(\mathbf{X}+\mathbf{Z}) \leq h(\mathbf{X}) - \frac{\mu n}{2}\log\left(\exp\left[\frac{2}{n}h(\mathbf{X})\right] + \exp\left[\frac{2}{n}h(\mathbf{Z})\right]\right). \tag{3}$$

For any fixed $a \in \mathcal{R}$ and $\mu > 1$, the function

$$f(t;a) = t - \frac{\mu n}{2}\log\left(\exp\left[\frac{2}{n}t\right] + \exp\left[\frac{2}{n}a\right]\right), \tag{4}$$

is concave in t and has a global maxima at

$$t = a - \frac{n}{2}\log(\mu-1). \tag{5}$$

Hence the right-hand side of (3) can be further bounded from above as

$$h(\mathbf{X}) - \frac{\mu n}{2}\log\left(\exp\left[\frac{2}{n}h(\mathbf{X})\right] + \exp\left[\frac{2}{n}h(\mathbf{Z})\right]\right) \leq f\left(h(\mathbf{Z}) - \frac{n}{2}\log(\mu-1); h(\mathbf{Z})\right). \tag{6}$$

The equality conditions of (3) and (6) imply that a Gaussian $\mathbf{X}$ with covariance matrix $(\mu-1)^{-1}\mathbf{K}_Z$ is an optimal solution of the optimization problem (2).

2

Note that in solving the above optimization problem, the classical EPI not only forces the optimal solution to be Gaussian, but also imposes a certain covariance structure on the Gaussian optimal solution. Hence a natural question to ask is what happens if there is an extra covariance constraint such that the original Gaussian optimal solution is no longer admissible. In that case, the classical EPI can still be used; however, the equality condition may no longer be met by the new optimal Gaussian solution because it may no longer have the required proportionality. In particular, one would be interested in finding out whether under the extra covariance constraint, a *Gaussian* $\mathbf{X}$ is still an optimal solution to optimization problems such as (2).

One particular type of covariance constraint is the following matrix covariance constraint:

$$\mathrm{Cov}(\mathbf{X}) \preceq \mathbf{S}. \tag{7}$$

Here $\mathrm{Cov}(\mathbf{X})$ denotes the covariance matrix of $\mathbf{X}$, "$\preceq$" represents "less or equal to" in the positive semidefinite partial ordering of real symmetric matrices, and $\mathbf{S}$ is a positive semidefinite matrix. The reason for considering such a matrix covariance constraint is largely due to its generality: it subsumes many other covariance constraints including the important trace constraint.

The focus of this paper is the following slightly more general optimization problem:

$$\begin{aligned} &\max_{p(\mathbf{x})} && h(\mathbf{X} + \mathbf{Z}_1) - \mu h(\mathbf{X} + \mathbf{Z}_2) \\ &\text{subject to} && \mathrm{Cov}(\mathbf{X}) \preceq \mathbf{S}, \end{aligned} \tag{8}$$

where $\mathbf{Z}_1$, $\mathbf{Z}_2$ are Gaussian vectors with strictly positive definite covariance matrix $\mathbf{K}_{Z_1}$ and $\mathbf{K}_{Z_2}$, respectively, and the maximization is over all random vector $\mathbf{X}$ independent of $\mathbf{Z}_1$ and $\mathbf{Z}_2$. As we shall see, such an optimization problem appears naturally when one is to evaluate certain genie-aided outer bounds on the capacity/rate region for the vector Gaussian broadcast channel and the distributed source coding with a single quadratic distortion constraint problems. Our main result is summarized in the following theorem.

Theorem 1 *For any $\mu \geq 1$ and any positive semidefinite $\mathbf{S}$, a Gaussian $\mathbf{X}$ is an optimal solution of the optimization problem* (8).

The rest of the paper is organized as follows. In Section 2, we prove our main result. We give two proofs: a direct proof using the classical EPI, and a strengthened proof following the perturbation approach of Stam [2] and Blachman [3]. In Section 3, we discuss some ramifications of the main result. In Section 4, we apply our main result to the vector Gaussian broadcast channel and the distributed source coding with a single quadratic distortion constraint problems. For the former problem, our main result leads to an *exact* characterization of the capacity region. Finally, in Section 5, we conclude by summarizing our contribution in the context of the applications of information theoretic inequalities in resolving multiterminal transmission/compression problems.

2 Proofs of the Main Result

2.1 A Direct Proof

In this first proof, we show that the classical EPI can be *appropriately* used to give a direct proof to Theorem 1. The fact that the classical EPI is relevant here is not surprising, considering that the objective function of the optimization problem (8) involves the entropy of the sum of two independent random vectors. Nonetheless, based on our discussion in Section 1, a direct use of the classical EPI might be loose because the covariance matrix of the optimal Gaussian solution might not have the required proportionality.

Our approach to resolve this issue is inspired by the mathematical import of an interesting technique, called *enhancement*, introduced by Weingarten et al. [7]. Our proof combines the idea of enhancement with the worst additive noise lemma [8], [9, Lemma II.2] stated as follows.

Lemma 2 (Worst additive noise lemma) *Let $\mathbf{Z}$ be a Gaussian vector with covariance matrix $\mathbf{K}_Z$, and let $\mathbf{K}_X$ be a positive semidefinite matrix. Consider the following optimization problem:*

$$\begin{aligned} \min_{p(\mathbf{x})} \quad & I(\mathbf{Z}; \mathbf{Z} + \mathbf{X}) \\ \text{subject to} \quad & \mathrm{Cov}(\mathbf{X}) = \mathbf{K}_X, \end{aligned} \tag{9}$$

where $I(\mathbf{Z}; \mathbf{Z} + \mathbf{X})$ denotes the mutual information between $\mathbf{Z}$ and $\mathbf{X} + \mathbf{Z}$, and the maximization is over all random vector $\mathbf{X}$ independent of $\mathbf{Z}$. A Gaussian $\mathbf{X}$ is an optimal solution of this optimization problem (no matter $\mathbf{K}_X$ and $\mathbf{K}_Z$ are proportional or not).

The details of the direct proof are in Appendix A.

2.2 A Perturbation Proof

From the optimization theoretic point of view, the power of the classical EPI lies in its ability to find global optima in nonconvex optimization problems such as (2). Hence one can imagine that proof of the classical EPI cannot be accomplished by any local optimization techniques. Indeed, in their classical proof Stam [2] and Blachman [3] used a *perturbation* approach, which amounts to find a *monotone* path from any distributions of the participating random vectors (i.e., $\mathbf{X}$ and $\mathbf{Z}$ in (1)) to the optimal distributions (Gaussian distributions with proportional covariance matrices) for which the classical EPI holds with equality. The monotonicity guarantees that *any* distributions along the path satisfy the desired inequality,

and hence the ones to begin with. A different perturbation was later used by Dembo et al. [10, p. 1509]. The main idea, however, remains the same as that of Stam and Blachman's.

Proving monotonicity needs isoperimetric inequalities. In case of the classical EPI, it needs the classical Fisher information inequality (FII) [10, Theorem 13]. Fisher information is an important quantity in statistical estimation theory. An interesting estimation theoretic proof using the data processing inequality for Fisher information was given by Zamir [11]. (The classical FII can also be proved by using the standard data processing inequality for mutual information, invoking a connection between Fisher information and mutual information explicitly established by Guo et al. [12, Corollary 2].) This connection between the EPI and the FII is usually thought of as the estimation view of the classical EPI.

We can use the perturbation idea to give a stronger proof to Theorem 1. We construct a monotone path using the "covariance-preserving" transformation, which was previously used by Dembo et al. [10, p. 1509] in their perturbation proof of the classical EPI. To prove the monotonicity, we need the following results on Fisher information matrix.

Lemma 3 *Denote by* $\mathbf{J}(\mathbf{X})$ *the Fisher information matrix of random vector* $\mathbf{X}$.

1. *(Cramér-Rao inequality) For any random vector* $\mathbf{U}$ *(of which the Fisher information matrix is well defined) with a strictly positive definite covariance matrix,*

$$\mathbf{J}(\mathbf{U}) \succeq \mathrm{Cov}^{-1}(\mathbf{U}). \tag{10}$$

2. *(Matrix FII) For any independent random vectors* $\mathbf{U}$, $\mathbf{V}$ *and any square matrix* $\mathbf{A}$,

$$\mathbf{J}(\mathbf{U} + \mathbf{V}) \preceq \mathbf{A}\mathbf{J}(\mathbf{U})\mathbf{A}^t + (\mathbf{I} - \mathbf{A})\mathbf{J}(\mathbf{V})(\mathbf{I} - \mathbf{A})^t. \tag{11}$$

Here $\mathbf{I}$ *is the identity matrix.*

For completeness, a proof of the above lemma using the properties of *score function* is provided in Appendix B. The details of the perturbation proof are in Appendix C.

3   Ramifications of the Main Result

In this section, we discuss two special cases of the optimization problem (8) to demonstrate the breadth of our main result. We term these two scenarios as the *degraded* case and the *extremely-skewed* case. By considering the degraded case, we prove a generalization of the classical EPI. By considering the extremely-skewed case, we establish a connection between our result and the classical result of Cover and Zhang [13] on the maximum differential entropy of the sum of two *dependent* random variables.

3.1 The Degraded Case

In the degraded case, we have either $\mathbf{K}_{Z_1} \preceq \mathbf{K}_{Z_2}$ or $\mathbf{K}_{Z_1} \succeq \mathbf{K}_{Z_2}$. First consider the case $\mathbf{K}_{Z_1} \preceq \mathbf{K}_{Z_2}$. We have the following results.

Corollary 4 *Let $\mathbf{Z}_1$, $\mathbf{Z}$ be two independent Gaussian vectors with covariance matrix $\mathbf{K}_{Z_1}$ and $\mathbf{K}_Z$, respectively. Assume that $\mathbf{K}_{Z_1}$ is strictly positive definite. Consider the following optimization problem:*

$$\begin{array}{ll} \max_{p(\mathbf{x})} & h(\mathbf{X} + \mathbf{Z}_1) - \mu h(\mathbf{X} + \mathbf{Z}_1 + \mathbf{Z}) \\ \text{subject to} & \mathrm{Cov}(\mathbf{X}) \preceq \mathbf{S}, \end{array} \tag{12}$$

where the maximization is over all random vector $\mathbf{X}$ independent of $\mathbf{Z}_1$ and $\mathbf{Z}$. For any $\mu \in \mathcal{R}$ and any positive semidefinite $\mathbf{S}$, a Gaussian $\mathbf{X}$ is an optimal solution of this optimization problem.

Proof. For $\mu \geq 1$, the corollary is a special case of Theorem 1 with $\mathbf{Z}_2 = \mathbf{Z}_1 + \mathbf{Z}$. For $\mu \leq 0$, the corollary also holds because $h(\mathbf{X} + \mathbf{Z}_1)$ and $h(\mathbf{X} + \mathbf{Z}_1 + \mathbf{Z})$ are *simultaneously* maximized when $\mathbf{X}$ is Gaussian with covariance matrix $\mathbf{S}$. This left us the only case where $\mu \in (0, 1)$, which we prove next.

The objective function of optimization problem (12) can be written as

$$(1 - \mu)h(\mathbf{X} + \mathbf{Z}_1) - \mu I(\mathbf{Z}; \mathbf{X} + \mathbf{Z}_1 + \mathbf{Z}). \tag{13}$$

Here $h(\mathbf{X} + \mathbf{Z}_1)$ is maximized when $\mathbf{X}$ is Gaussian with covariance matrix $\mathbf{S}$. By the worst noise result of Lemma 2, $I(\mathbf{Z}; \mathbf{X} + \mathbf{Z}_1 + \mathbf{Z})$ is minimized when $\mathbf{X}$ is Gaussian. Further within the Gaussian class, the one with the full covariance matrix $\mathbf{S}$ minimizes $I(\mathbf{Z}; \mathbf{Z} + \mathbf{X} + \mathbf{Z}_1)$. For $\mu \in (0, 1)$, both μ and $1 - \mu$ are positive. We conclude that the objective function (13) is maximized when $\mathbf{X}$ is Gaussian with covariance matrix $\mathbf{S}$. This completes the proof. $\square$

Corollary 5 *Let $\mathbf{Z}$ be a Gaussian vector with covariance matrix $\mathbf{K}_Z$. Assume that $\mathbf{K}_Z$ is strictly positive definite. Consider the following optimization problem*

$$\begin{array}{ll} \max_{p(\mathbf{x})} & h(\mathbf{X}) - \mu h(\mathbf{X} + \mathbf{Z}) \\ \text{subject to} & \mathrm{Cov}(\mathbf{X}) \preceq \mathbf{S}, \end{array} \tag{14}$$

where the maximization is over all random vector $\mathbf{X}$ independent of $\mathbf{Z}$. For any $\mu \in \mathcal{R}$ and any positive semidefinite $\mathbf{S}$, a Gaussian $\mathbf{X}$ is an optimal solution of this optimization problem.

Observe that the optimization problem (14) is simply a constrained version of the optimization problem (2). Recall from Section 1 that the optimization problem (2) can be solved by the classical EPI. Conversely, it can be shown that the special case of the classical EPI with one of the participant random vectors (say, $\mathbf{Z}$ in (1)) fixed to be Gaussian can also be obtained from the fact that a Gaussian $\mathbf{X}$ is an optimal solution of the optimization problem (2). This can be done as follows. Choosing

$$\mu = 1 + \exp\left[\frac{2}{n}\left(h(\mathbf{Z}) - h(\mathbf{X})\right)\right],\tag{15}$$

we have from (5) that

$$h(\mathbf{X}_G^*) = h(\mathbf{Z}) - \frac{n}{2}\log(\mu - 1) = h(\mathbf{X}).\tag{16}$$

Since $\mathbf{X}_G^*$ is an optimal solution of the optimization problem (2) (recall that $\mathbf{X}_G^*$ has a special covariance structure of being proportional to $\mathbf{K}_Z$), we have

$$h(\mathbf{X}) - \mu h(\mathbf{X} + \mathbf{Z}) \le h(\mathbf{X}_G^*) - \mu h(\mathbf{X}_G^* + \mathbf{Z}).\tag{17}$$

Substituting (16) into (17), we have $h(\mathbf{X} + \mathbf{Z}) \ge h(\mathbf{X}_G^* + \mathbf{Z})$ for any random vector $\mathbf{X}$ independent of $\mathbf{Z}$ and satisfying $h(\mathbf{X}) = h(\mathbf{X}_G^*)$. This is precisely the Costa-Cover form of the classical EPI [10, Theorem 6], so we have proved the converse statement.

In light of the above statements, Corollary 5 can be thought of a generalization of the classical EPI. For technical reasons, we were not able to prove Corollary 5 directly from Corollary 4 by letting $\mathbf{K}_{Z_1}$ vanish. Instead, we can resort to arguments (direct and perturbation ones) similar to those for Theorem 1 to prove Corollary 5. Observe that in the optimization problem (14) the lower constraint $\mathbf{K}_X \succeq 0$ never bites, so no enhancement is needed in the perturbation proof. The details of the proof is omitted from the paper.

We now turn to the other degraded case where $\mathbf{K}_{Z_1} \succeq \mathbf{K}_{Z_2}$. Consider the optimization problem

$$\begin{array}{ll} \max_{p(\mathbf{x})} & h(\mathbf{X} + \mathbf{Z}_2 + \mathbf{Z}) - \mu h(\mathbf{X} + \mathbf{Z}_2) \\ \text{subject to} & \mathrm{Cov}(\mathbf{X}) \preceq \mathbf{S}, \end{array}\tag{18}$$

where the maximization is over all random vector $\mathbf{X}$ independent of $\mathbf{Z}_2$ and $\mathbf{Z}$. For any $\mu \ge 1$, by Theorem 1 a Gaussian $\mathbf{X}$ is an optimal solution of this optimization problem. For $\mu \le 0$, this is also true because $h(\mathbf{X} + \mathbf{Z}_2 + \mathbf{Z})$ and $h(\mathbf{X} + \mathbf{Z}_2)$ are simultaneously maximized when $\mathbf{X}$ is Gaussian with covariance matrix $\mathbf{S}$. However, as we shall see next, this is generally *not* the case for $\mu \in (0, 1)$.

Consider the cases where

$$0 \prec \frac{\mu}{1 - \mu}\mathbf{K}_Z - \mathbf{K}_{Z_2} \prec \mathbf{S}.\tag{19}$$

7

(Note that this can only happen when $\mu \in (0, 1)$ and also depends on the realizations of $\mathbf{K}_Z$, $\mathbf{K}_{Z_2}$ and $\mathbf{S}$.) Under this assumption, we can verify that the covariance matrix $\mathbf{K}_X^*$ of $\mathbf{X}_G^*$ must satisfy:

$$\mathbf{K}_X^* = \frac{\mu}{1-\mu}\mathbf{K}_Z - \mathbf{K}_{Z_2}. \tag{20}$$

Let $\mathbf{X}$ be a *non-Gaussian* random vector satisfying:

1. $h(\mathbf{X} + \mathbf{Z}_2) = h(\mathbf{X}_G^* + \mathbf{Z}_2)$;

2. $\mathrm{Cov}(\mathbf{X}) \preceq \mathbf{S}$.

Such an $\mathbf{X}$ exists because by the assumption, $\mathbf{K}_X^*$ is *strictly* between 0 and $\mathbf{S}$. Since $\mathbf{X}$ is non-Gaussian, by the Costa-Cover form of the classical EPI, we have

$$h(\mathbf{X} + \mathbf{Z}_2 + \mathbf{Z}) > h(\mathbf{X}_G^* + \mathbf{Z}_2 + \mathbf{Z}). \tag{21}$$

We thus conclude that at least for the cases where the condition (19) holds, the optimal Gaussian solution $\mathbf{X}_G^*$ cannot be an optimal solution of the optimization problem (18).

3.2 The Extremely-Skewed Case

Suppose that $\mathbf{Z}_1$, $\mathbf{Z}_2$ are in $\mathcal{R}^2$. Let

$$\mathbf{K}_{Z_1} = \mathbf{V}_1\mathbf{\Sigma}_1\mathbf{V}_1^t, \quad \mathbf{K}_{Z_2} = \mathbf{V}_2\mathbf{\Sigma}_2\mathbf{V}_2^t, \tag{22}$$

where $\mathbf{V}_1$, $\mathbf{V}_2$ are orthogonal matrices and

$$\mathbf{\Sigma}_1 = \mathrm{Diag}(\lambda_{11}, \lambda_{12}), \quad \mathbf{\Sigma}_2 = \mathrm{Diag}(\lambda_{21}, \lambda_{22}) \tag{23}$$

are diagonal matrices. Consider the limiting situation where λ_{12}, $\lambda_{21} \to \infty$, while λ_{11}, λ_{22} are kept fixed. Compared with the degraded case where $\mathbf{K}_{Z_1}$ dominates $\mathbf{K}_{Z_2}$ in every possible direction (or vice versa), this situation between $\mathbf{K}_{Z_1}$ and $\mathbf{K}_{Z_2}$ is extremely skewed. We have the following result.

Corollary 6 *Let Z be a Gaussian random variable, and let $\mathbf{v}_1$, $\mathbf{v}_2$ be two deterministic vectors in $\mathcal{R}^2$. Consider the optimization problem*

$$\begin{array}{ll} \max_{p(\mathbf{x})} & h(\mathbf{v}_1^t\mathbf{X} + Z) - \mu h(\mathbf{v}_2^t\mathbf{X} + Z) \\ \text{subject to} & \mathrm{Cov}(\mathbf{X}) \preceq \mathbf{S}, \end{array} \tag{24}$$

where the maximization is over all random vector $\mathbf{X}$ (in $\mathcal{R}^2$) independent of Z. For any $\mu \geq 1$ and any positive semidefinite $\mathbf{S}$, a Gaussian $\mathbf{X}$ is an optimal solution of this optimization problem.

Proof. See Appendix D. □

Next, we use Corollary 6 to solve an optimization problem that involves maximizing the differential entropy of the sum of two *dependent* random variables. To put it in perspective, let us first consider the following simple optimization problem:

$$\begin{aligned} \max_{p(x_1,x_2)} \quad & h(X_1 + X_2) \\ \text{subject to} \quad & \mathrm{Var}(X_1) \leq a_1, \quad \mathrm{Var}(X_2) \leq a_2, \end{aligned} \quad (25)$$

where $a_1, a_2 \geq 0$ are real numbers, $\mathrm{Var}(X)$ denotes the variance of X, and the maximization is over all jointly distributed random variables (X_1, X_2). The solution to this optimization problem is clear: $h(X_1 + X_2)$ is maximized when X_1, X_2 are jointly Gaussian with variance a_1 and a_2, respectively, and are aligned, i.e., $X_1 = \sqrt{a_1/a_2} X_2$ almost surely.

Replacing both variance constraints in the optimization problem (25) by the entropy constraints, we have the following optimization problem:

$$\begin{aligned} \max_{p(x_1,x_2)} \quad & h(X_1 + X_2) \\ \text{subject to} \quad & h(X_1) \leq a_1, \quad h(X_2) \leq a_2, \end{aligned} \quad (26)$$

where $a_1, a_2 \in \mathcal{R}$, and the maximization is over all jointly distributed random variables (X_1, X_2). Different from the optimization problem (25), a jointly Gaussian (X_1, X_2) is *not* always an optimal solution of (26). This can seen as follows. Consider the case $a_1 = a_2$. Let (X_{1G}^*, X_{2G}^*) be the optimal Gaussian solution of the optimization problem (26). We have $X_{1G}^* = X_{2G}^*$ almost surely, i.e., X_{1G}^* and X_{2G}^* are aligned and have the same marginal distribution. Consider all jointly distributed random variables (X_1, X_2) for which X_1, X_2 have the same marginal density function f which satisfies:

1. $h(X_1) = h(X_{1G}^*)$;

2. f is *not* log-concave.

The classical result of Cover and Zhang [13] asserts that among all (X_1, X_2) satisfying the above conditions, there is at least one that satisfies

$$h(X_1 + X_2) > h(2X_1) = h(2X_{1G}^*) = h(X_{1G}^* + X_{2G}^*). \quad (27)$$

We thus conclude that a jointly Gaussian (X_1, X_2) is not always an optimal solution of the optimization problem (26).

Between (25) and (26) is the following optimization problem:

$$\begin{aligned} \max_{p(x_1,x_2)} \quad & h(X_1 + X_2) \\ \text{subject to} \quad & \mathrm{Var}(X_1) \leq a_1, \quad h(X_2) \leq a_2, \end{aligned} \quad (28)$$

9

where a_1, a_2 are real numbers with $a_1 \geq 0$, and the maximization is over all jointly distributed random variables (X_1, X_2). The question whether a Gaussian (X_1, X_2) is an optimal solution of this optimization problem remains, to our best knowledge, an open problem. The following result, however, can be proved using Corollary 6.

Corollary 7 *Let Z be a Gaussian variable, and let a_1, a_2 be real numbers with $a_1 \geq 0$. Consider the optimization problem*

$$
\begin{array}{ll}
\max_{p(x_1,x_2)} & h(X_1 + X_2 + Z) \\
\text{subject to} & \mathrm{Var}(X_1) \leq a_1, \quad h(X_2 + Z) \leq a_2,
\end{array}
\tag{29}
$$

where the maximization is over all jointly distributed random variables (X_1, X_2) independent of Z. A Gaussian (X_1, X_2) is an optimal solution of this optimization problem for any $a_1 \geq 0$ and any $h(Z) \leq a_2 \leq a_2^*$ where

$$
a_2^* = \frac{1}{2} \log \left(2\pi e \left(\mathrm{Var}(Z) + \frac{1}{4} \left(\sqrt{a_1 + 4\mathrm{Var}(Z)} - \sqrt{a_1} \right)^2 \right) \right).
\tag{30}
$$

Proof. See Appendix E. $\qquad\square$

4 Applications in Multiterminal Information Theory

4.1 The Vector Gaussian Broadcast Channel

We now use our main result to give an exact characterization of the capacity region of the vector Gaussian broadcast channel. The capacity region of the vector Gaussian broadcast channel was first characterized by Weigarten et al. [7].

Consider the following two-user discrete-time vector Gaussian broadcast channel:

$$
\mathbf{Y}_k[m] = \mathbf{X}[m] + \mathbf{Z}_k[m], \quad k = 1, 2,
\tag{31}
$$

where $\{\mathbf{X}[m]\}$ is the channel input subject to an average matrix power constraint

$$
\frac{1}{N} \sum_{m=1}^{N} \mathbf{X}[m]\mathbf{X}^t[m] \preceq \mathbf{S},
\tag{32}
$$

and the noise $\{\mathbf{Z}_k[m]\}$ is i.i.d. Gaussian with zero mean and strictly positive definite covariance matrix $\mathbf{K}_{Z_k}$ and is independent of $\{\mathbf{X}[m]\}$. The covariance structure of the Gaussian

noise models a scalar Gaussian broadcast channel with memory. Alternatively, it can also model the downlink of a cellular system with multiple antennas; this was the motivation of [7].

A vector Gaussian broadcast is in general a *nondegraded* broadcast channel. An exact characterization of the capacity region had been a long-standing open problem in multi-terminal information theory, particularly when viewed in the context of a scalar Gaussian broadcast channel with memory. Prior to [7], only bounds were known. An outer bound, derived by Marton and Körner [14, Theorem 5], is given by $\mathcal{O} = \mathcal{O}_1 \cap \mathcal{O}_2$, where $\mathcal{O}_1$ is the set of rate pairs (R_1, R_2) satisfying

$$
\begin{aligned}
R_1 &\leq I(\mathbf{X}; \mathbf{Y}_1 | U) \\
R_2 &\leq I(U; \mathbf{Y}_2)
\end{aligned}
\tag{33}
$$
$$
\tag{34}
$$

for some $p(\mathbf{y}_1, \mathbf{y}_2, \mathbf{x}, u) = p(\mathbf{y}_1, \mathbf{y}_2 | \mathbf{x}) p(\mathbf{x}, u)$ such that $p(\mathbf{y}_1, \mathbf{y}_2 | \mathbf{x})$ is the channel transition matrix and $p(\mathbf{x})$ satisfies the constraint $\mathbb{E}[\mathbf{X}\mathbf{X}^t] \preceq \mathbf{S}$, and $\mathcal{O}_2$ is the set of rate pairs (R_1, R_2) satisfying

$$
\begin{aligned}
R_1 &\leq I(V; \mathbf{Y}_1) \\
R_2 &\leq I(\mathbf{X}; \mathbf{Y}_2 | V)
\end{aligned}
\tag{35}
$$
$$
\tag{36}
$$

for some $p(\mathbf{y}_1, \mathbf{y}_2, \mathbf{x}, v) = p(\mathbf{y}_1, \mathbf{y}_2 | \mathbf{x}) p(\mathbf{x}, v)$ such that $p(\mathbf{y}_1, \mathbf{y}_2 | \mathbf{x})$ is the channel transition matrix and $p(\mathbf{x})$ satisfies the constraint $\mathbb{E}[\mathbf{X}\mathbf{X}^t] \preceq \mathbf{S}$.

Next, we derive a tight upper bound on the achievable weighted sum rate

$$
\mu_1 R_1 + \mu_2 R_2,
\tag{37}
$$

using the Marton-Körner outer bound as the starting point. Since a capacity region is always convex (per time-sharing argument), an exact characterization of all the achievable weighted sum rates for all nonnegative μ_1, μ_2 provides an exact characterization of the entire capacity region. First consider the case $\mu_2 \geq \mu_1 \geq 0$. By the Marton-Körner outer bound, any achievable rate pair (R_1, R_2) must satisfy:

$$
\begin{aligned}
\mu_1 R_1 + \mu_2 R_2 &\leq \mu_1 \cdot \max \{ I(\mathbf{X}; \mathbf{Y}_1 | U) + \mu I(U; \mathbf{Y}_2) \} \\
&= \mu_1 \cdot \max \{ -h(\mathbf{Z}_1) + \mu h(\mathbf{X} + \mathbf{Z}_2) + [h(\mathbf{X} + \mathbf{Z}_1 | U) - \mu h(\mathbf{X} + \mathbf{Z}_2 | U)] \}.
\end{aligned}
\tag{38}
$$
$$
\tag{39}
$$

Here $\mu = \frac{\mu_2}{\mu_1} \geq 1$, and the maximization is over all $(U, \mathbf{X})$ independent of $(\mathbf{Z}_1, \mathbf{Z}_2)$ and satisfying the matrix constraint $\mathbb{E}[\mathbf{X}\mathbf{X}^t] \preceq \mathbf{S}$. Consider the terms $h(\mathbf{Z}_1)$, $h(\mathbf{X} + \mathbf{Z}_2)$ and $h(\mathbf{X} + \mathbf{Z}_1 | U) - \mu h(\mathbf{X} + \mathbf{Z}_2 | U)$ separately. We have

$$
h(\mathbf{Z}_1) = \frac{1}{2} \log \left((2\pi e)^n |\mathbf{K}_{Z_1}| \right)
\tag{40}
$$

and
$$h(\mathbf{X} + \mathbf{Z}_2) \leq \frac{1}{2} \log \left((2\pi e)^n |\mathbf{S} + \mathbf{K}_{Z_2}| \right). \tag{41}$$

Further note that maximizing $h(\mathbf{X} + \mathbf{Z}_1|U) - \mu h(\mathbf{X} + \mathbf{Z}_2|U)$ is simply a conditional version of the optimization problem (8). We have the following result, which is a conditional version of Theorem 1.

Theorem 8 *Let $\mathbf{Z}_1$, $\mathbf{Z}_2$ be two Gaussian vectors with strictly positive definite covariance matrices $\mathbf{K}_{Z_1}$ and $\mathbf{K}_{Z_2}$, respectively. Let $\mu \geq 1$ be a real number, $\mathbf{S}$ be a positive semidefinite matrix, and U be a random variable independent of $\mathbf{Z}_1$ and $\mathbf{Z}_2$. Consider the optimization problem*

$$\begin{array}{ll} \max_{p(\mathbf{x}|u)} & h(\mathbf{X} + \mathbf{Z}_1|U) - \mu h(\mathbf{X} + \mathbf{Z}_2|U) \\ \text{subject to} & \mathrm{Cov}(\mathbf{X}|U) \preceq \mathbf{S}, \end{array} \tag{42}$$

where the maximization is over all conditional distribution of $\mathbf{X}$ given U independent of $\mathbf{Z}_1$ and $\mathbf{Z}_2$. A Gaussian $p(\mathbf{x}|u)$ with the same covariance matrix for each u is an optimal solution of this optimization problem.

The result of the above theorem has two parts. The part that says a Gaussian $p(\mathbf{x}|u)$ is an optimal solution follows directly from Theorem 1; the part that says the optimal Gaussian $p(\mathbf{x}|u)$ has the same covariance matrix for each u is equivalent to that the optimal value of the optimization problem (8) is a *concave* function of $\mathbf{S}$. Despite being a matrix problem, a direct proof of the concavity turns out to be difficult. Instead, Theorem 8 can be proved following the same footsteps as those for Theorem 1, except that we need to replace the classical EPI by a conditional version proved by Bergmans [4, Lemma II]. Let $\mathbf{Z}$ be a Gaussian vector. Bergmans' conditional EPI states that

$$\exp\left[\frac{2}{n}h(\mathbf{X} + \mathbf{Z}|U)\right] \geq \exp\left[\frac{2}{n}h(\mathbf{X}|U)\right] + \exp\left[\frac{2}{n}h(\mathbf{Z})\right] \tag{43}$$

for any $(\mathbf{X}, U)$ independent of $\mathbf{Z}$. The equality holds if and only if conditional on $U = u$, $\mathbf{X}$ is Gaussian with a covariance matrix proportional to that of $\mathbf{Z}$ and has the same covariance matrix for each u. The details of the proof are omitted from the paper.

By Theorem 8, we have

$$h(\mathbf{X}+\mathbf{Z}_1|U) - \mu h(\mathbf{X}+\mathbf{Z}_2|U) \leq \max_{0 \preceq \mathbf{K}_X \preceq \mathbf{S}} \left\{ \frac{1}{2} \log \left((2\pi e)^n |\mathbf{K}_X + \mathbf{K}_{Z_1}| \right) - \frac{\mu}{2} \log \left((2\pi e)^n |\mathbf{K}_X + \mathbf{K}_{Z_2}| \right) \right\}. \tag{44}$$

Substituting (40), (41) and (44) into (39), we obtain

$$\mu_1 R_1 + \mu_2 R_2 \leq \max_{0 \preceq \mathbf{K}_X \preceq \mathbf{S}} \left\{ \frac{\mu_1}{2} \log \left| \frac{\mathbf{K}_X + \mathbf{K}_{Z_1}}{\mathbf{K}_{Z_1}} \right| + \frac{\mu_2}{2} \log \left| \frac{\mathbf{S} + \mathbf{K}_{Z_2}}{\mathbf{K}_X + \mathbf{K}_{Z_2}} \right| \right\}. \tag{45}$$

Note that the weighted sum rates give by (45) can be achieved by *dirty-paper coding* [16, 17], so (45) is an *exact* characterization of all the achievable weighted sum rates for $\mu_2 \geq \mu_1 \geq 0$.

For $\mu_1 \geq \mu_2 \geq 0$, we have from the Marton-Körner bound that

$$\mu_1 R_1 + \mu_2 R_2 \leq \mu_2 \cdot \max_{p(\mathbf{X}, V)} \left\{ \mu I(V; \mathbf{Y}_1) + I(\mathbf{X}; \mathbf{Y}_2 | V) \right\}. \tag{46}$$

Here $\mu = \frac{\mu_1}{\mu_2} \geq 1$, and the maximization is over all $(V, \mathbf{X})$ independent of $(\mathbf{Z}_1, \mathbf{Z}_2)$ and satisfying the matrix constraint $\mathbb{E}[\mathbf{X}\mathbf{X}^t] \preceq \mathbf{S}$. Relabeling V as U, the optimization problem becomes identical to that in (38). We thus conclude that

$$\mu_1 R_1 + \mu_2 R_2 \leq \max_{0 \preceq \mathbf{K}_X \preceq \mathbf{S}} \left\{ \frac{\mu_1}{2} \log \left| \frac{\mathbf{S} + \mathbf{K}_{Z_1}}{\mathbf{K}_X + \mathbf{K}_{Z_1}} \right| + \frac{\mu_2}{2} \log \left| \frac{\mathbf{K}_X + \mathbf{K}_{Z_2}}{\mathbf{K}_{Z_2}} \right| \right\}. \tag{47}$$

is an exact characterization of all the achievable weighted sum rates for $\mu_1 \geq \mu_2 \geq 0$. This settles the problem of characterizing the entire capacity region of the vector Gaussian broadcast channel.

4.2 Distributed Source Coding with a Single Quadratic Distortion Constraint

Our result is also relevant in the following distributed source coding problem. Let $\{\mathbf{Y}_1[m]\}$, $\{\mathbf{Y}_2[m]\}$ be two i.i.d. vector Gaussian sources with strictly positive definite covariance matrix $\mathbf{K}_{Y_1}$ and $\mathbf{K}_{Y_2}$, respectively. At each time m, $\mathbf{Y}_1[m]$ and $\mathbf{Y}_2[m]$ are jointly Gaussian. The encoder is only allowed to perform *separate* encoding on the sources. The decoder, on the other hand, can reconstruct the sources based on *both* encoded messages. We wish to characterize the entire rate region for which the quadratic distortion for reconstructing $\{\mathbf{Y}_1[m]\}$ at the decoder

$$\frac{1}{N} \sum_{m=1}^{N} \left(\mathbf{Y}_1[m] - \widehat{\mathbf{Y}}_1[m] \right) \left(\mathbf{Y}_1[m] - \widehat{\mathbf{Y}}_1[m] \right)^t \preceq \mathbf{D}. \tag{48}$$

(There is no distortion constraint on the source $\{\mathbf{Y}_2[m]\}$.) This is the so-called *distributed source coding with a single quadratic distortion constraint problem*.

Note that $\mathbf{Y}_1[m]$, $\mathbf{Y}_2[m]$ are jointly Gaussian, so without loss of generality we can write

$$\mathbf{Y}_1[m] = \mathbf{A}\mathbf{Y}_2[m] + \mathbf{Z}[m], \tag{49}$$

where $\mathbf{A}$ is an invertible matrix and $\mathbf{Z}[m]$ is Gaussian and independent of $\mathbf{Y}_2[m]$. Since there is no distortion constraint on $\{\mathbf{Y}_2[m]\}$, we can always assume that $\mathbf{Y}_1[m]$ is a degraded

version of $\mathbf{Y}_2[m]$ by relabeling $\mathbf{A}\mathbf{Y}_2[m]$ as $\mathbf{Y}_2[m]$. In this case, an outer bound can be obtained similarly to that for the discrete memoryless degraded broadcast channel [19]:

$$
\begin{aligned}
R_1 &\geq I(\mathbf{Y}_1; \widehat{\mathbf{Y}}_1 | U) \\
R_2 &\geq I(U; \mathbf{Y}_2)
\end{aligned}
\tag{50}
$$

for some $p(u, \widehat{\mathbf{y}}_1, \mathbf{y}_1, \mathbf{y}_2) = p(\widehat{\mathbf{y}}_1 | u, \mathbf{y}_1) p(u | \mathbf{y}_2) p(\mathbf{y}_1, \mathbf{y}_2)$, where $p(\mathbf{y}_1, \mathbf{y}_2)$ is the joint distribution of the sources and $p(\widehat{\mathbf{y}}_1 | u, \mathbf{y}_1)$ satisfies the matrix constraint $\mathbb{E}[(\mathbf{Y}_1 - \widehat{\mathbf{Y}}_1)(\mathbf{Y}_1 - \widehat{\mathbf{Y}}_1)^t] \preceq \mathbf{D}$. The proof is deferred to Appendix F. Next, we derive a lower bound on all the achievable weighted sum rates $\mu_1 R_1 + \mu_2 R_2$ for all nonnegative μ_1, μ_2, using this outer bound as the starting point.

By the outer bound (50), all the achievable rate pairs (R_1, R_2) must satisfy:

$$
\mu_1 R_1 + \mu_2 R_2 \geq \mu_1 \cdot \min_{p(u, \widehat{\mathbf{y}} | \mathbf{y}_1, \mathbf{y}_2)} \left\{ \mu I(\mathbf{Y}_1; \widehat{\mathbf{Y}}_1 | U) + I(U; \mathbf{Y}_2) \right\}
\tag{51}
$$

$$
= \mu_1 \cdot \min_{p(u, \widehat{\mathbf{y}} | \mathbf{y}_1, \mathbf{y}_2)} \left\{ h(\mathbf{Y}_2) - \mu h(\mathbf{Y}_1 | \widehat{\mathbf{Y}}_1, U) - [h(\mathbf{Y}_2 | U) - \mu h(\mathbf{Y}_2 + \mathbf{Z} | U)] \right\}.
\tag{52}
$$

Here $\mu = \frac{\mu_2}{\mu_1} \geq 0$, and the minimization is over all $p(u, \widehat{\mathbf{y}} | \mathbf{y}_1, \mathbf{y}_2)$ such that U is independent of $\mathbf{Z}$ and $\mathbb{E}[(\mathbf{Y}_1 - \widehat{\mathbf{Y}}_1)(\mathbf{Y}_1 - \widehat{\mathbf{Y}}_1)^t] \preceq \mathbf{D}$ is satisfied. Consider the terms $h(\mathbf{Y}_2)$, $h(\mathbf{Y}_1 | \widehat{\mathbf{Y}}_1, U)$ and $h(\mathbf{Y}_2 | U) - \mu h(\mathbf{Y}_2 + \mathbf{Z} | U)$ separately. We have

$$
h(\mathbf{Y}_2) = \frac{1}{2} \log \left((2\pi e)^n |\mathbf{K}_{Y_2}| \right)
\tag{53}
$$

and

$$
h(\mathbf{Y}_1 | \widehat{\mathbf{Y}}_1, U) = h(\mathbf{Y}_1 - \widehat{\mathbf{Y}}_1 | \widehat{\mathbf{Y}}_1, U) \leq h(\mathbf{Y}_1 - \widehat{\mathbf{Y}}_1) \leq \frac{1}{2} \log \left((2\pi e)^n |\mathbf{D}| \right).
\tag{54}
$$

Hence we only need to maximize $h(\mathbf{Y}_2 | U) - \mu h(\mathbf{Y}_2 + \mathbf{Z} | U)$ subject to the constraints

$$
\mathrm{Cov}(\mathbf{Y}_2 | U) \preceq \mathbf{K}_{Y_2} \quad \text{and} \quad \mathrm{Cov}(\mathbf{Y}_1 | U) \succeq \mathbf{D}.
\tag{55}
$$

In case that the constraint $\mathrm{Cov}(\mathbf{Y}_1 | U) \succeq \mathbf{D}$ does not bite, we can use (a conditional version of) Corollary 5 to show that a Gaussian test channel between $\mathbf{Y}_2$ and $\mathbf{U}$ is a maximizer:

$$
h(\mathbf{Y}_2 | U) - \mu h(\mathbf{Y}_2 + \mathbf{Z} | U) \leq \max_{0 \preceq \mathbf{K} \preceq \mathbf{K}_{Y_2}} \left\{ \frac{1}{2} \log \left((2\pi e)^n |\mathbf{K}| \right) - \frac{\mu}{2} \log \left((2\pi e)^n |\mathbf{K} + \mathbf{K}_{Y_1} - \mathbf{K}_{Y_2}| \right) \right\}.
\tag{56}
$$

Substituting (53), (54) and (56) into (52), we have

$$
\mu_1 R_1 + \mu_2 R_2 \geq \max_{0 \preceq \mathbf{K} \preceq \mathbf{K}_{Y_2}} \left\{ \frac{\mu_1}{2} \log \left| \frac{\mathbf{K}_{Y_2}}{\mathbf{K}} \right| + \frac{\mu_2}{2} \log \left| \frac{\mathbf{K} + \mathbf{K}_{Y_1} - \mathbf{K}_{Y_2}}{\mathbf{D}} \right| \right\}.
\tag{57}
$$

On the other hand, this weighted sum rate can be achieved by the following natural Gaussian separation scheme:

14

1. Quantize $\{\mathbf{Y}_1[m]\}$ and $\{\mathbf{Y}_2[m]\}$ separately using Gaussian codebooks;

2. Use *Slepian-Wolf coding* [18] on the quantized version of $\{\mathbf{Y}_1[m]\}$, treating the quantized version of $\{\mathbf{Y}_2[m]\}$ as decoder side information.

This would have settled the rate region for the distributed source coding with a single quadratic constraint problem.

Unfortunately, there are indeed instances where the constraint $\mathrm{Cov}(\mathbf{Y}_1|U) \succeq \mathbf{D}$ cannot be ignored; in such cases, the outer bound studied here will be strictly inside the inner bound achieved by the natural Gaussian separation scheme.

5   Concluding Remarks

The classical EPI is an important inequality with interesting connections to statistical estimation theory. In information theory, it has been key to the proof of the converse coding theorem in several important scalar Gaussian multiterminal problems [4, 5, 6]. In the vector situation, the equality condition of the classical EPI is stringent: the equality requires the participating random vectors not only be Gaussian but also have proportional covariance matrices. In several instances, this coupling between the Gaussianity and the proportionality is the main cause that prevents the classical EPI from being directly useful in extending the converse proof from the scalar case to the vector situation.

In this paper, we proved a new extremal inequality involving entropies of random vectors. In one special case, this inequality can be seen as a *robust* version of the classical EPI. By "robust", we refer to the fact that in the new extremal inequality, the optimality of a Gaussian distribution does not couple with a specific covariance structure, i.e. proportionality. We show that the new extremal inequality is useful in evaluating certain genie-aided outer bounds for the capacity/rate region for the vector Gaussian broadcast channel and the distributed source coding with a single quadratic constraint problems.

We offered two proofs to the new extremal inequality: one by appropriately using the classical EPI, and the other by the perturbation approach of Stam [2] and Blachman [3]. The perturbation approach gives more insights: it takes the problem (via the de Bruijn identity) to the Fisher information domain where the proportionality no longer seems a hurdle. Whereas the advantage of the perturbation approach is not crucial for the entropy inequalities discussed in this paper, it becomes crucial in some other situations [20] where the enhancement technique of Weingarten et al. does not suffice.

A   A Direct Proof of Theorem 1

We now show that the classical EPI can be appropriately used to prove Theorem 1. We first give the outline of the proof.

Proof Outline. We first show that without loss of generality, we can assume that $\mathbf{S}$ is *strictly* positive definite. Next, we denote the optimization problem (8) by P and the optimal value of P by (P). To show that a Gaussian $\mathbf{X}$ is an optimal solution of P, it is sufficient to show that $(\mathrm{P}) = (P_G)$, where P_G is the Gaussian version of P by restricting the solution space within Gaussian distributions:

$$
\begin{aligned}
\max_{\mathbf{K}_X} \quad & \tfrac{1}{2}\log\left((2\pi e)^n \left|\mathbf{K}_X + \mathbf{K}_{Z_1}\right|\right) - \tfrac{\mu}{2}\log\left((2\pi e)^n \left|\mathbf{K}_X + \mathbf{K}_{Z_2}\right|\right) \\
\text{subject to} \quad & \mathbf{0} \preceq \mathbf{K}_X \preceq \mathbf{S}.
\end{aligned}
\tag{58}
$$

Since restricting the solution space can only decrease the optimal value of a maximization problem, we readily have $(P) \geq (P_G)$. To prove the reverse inequality $(P) \leq (P_G)$, we shall consider an auxiliary optimization problem $\widetilde{P}$ and its Gaussian version $\widetilde{P}_G$. In particular, we shall construct a $\widetilde{P}$ such that:

$$
(P) \leq (\widetilde{P}), \quad (\widetilde{P}) = (\widetilde{P}_G), \quad (\widetilde{P}_G) = (P_G).
\tag{59}
$$

We will then have $(P) \leq (P_G)$ and hence $(P) = (P_G)$.

The proof is rather long, so we divide it into several steps.

Step 1: $\mathbf{S} \succeq 0$, $|\mathbf{S}| = 0$. We show that for any $\mathbf{S} \succeq 0$ but $|\mathbf{S}| = 0$, there is an *equivalent* optimization problem of type (8) in which the the upper bound on $\mathbf{X}$ is strictly positive definite.

Suppose that the rank of $\mathbf{S}$ is $r < n$, i.e., $\mathbf{S}$ is rank deficient. Let

$$
\mathbf{S} = \mathbf{Q}_S \mathbf{\Sigma}_S \mathbf{Q}_S^t,
\tag{60}
$$

where $\mathbf{Q}_S$ is an orthogonal matrix, and

$$
\mathbf{\Sigma}_S = \mathrm{Diag}(\lambda_1, \cdots, \lambda_r, 0, \cdots, 0)
\tag{61}
$$

is a diagonal matrix. For any $\mathbf{X} \preceq \mathbf{S}$, let $\overline{\mathbf{X}} = \left(\overline{\mathbf{X}}_a^t, \overline{\mathbf{X}}_b^t\right)^t = \mathbf{Q}_S^t \mathbf{X}$ where $\overline{\mathbf{X}}_a$ is of a length r. We have

$$
\mathrm{Cov}(\overline{\mathbf{X}}) = \mathbf{Q}_S^t \mathrm{Cov}(\mathbf{X}) \mathbf{Q}_S \preceq \mathbf{Q}_S^t \mathbf{S} \mathbf{Q}_S = \mathbf{\Sigma}_S,
\tag{62}
$$

which implies that $\mathrm{Cov}(\overline{\mathbf{X}}_b) = 0$, i.e., $\overline{\mathbf{X}}_b$ is deterministic. Without loss of generality, let us assume that $\overline{\mathbf{X}}_b = 0$. So an optimization over $\mathrm{Cov}(\mathbf{X}) \preceq \mathbf{S}$ is the same as an optimization over $\mathrm{Cov}(\overline{\mathbf{X}}_a) \preceq \mathrm{Diag}(\lambda_1, \cdots, \lambda_r)$.

16

Next, let

$$\mathbf{Q}_S^t \mathbf{K}_{Z_i} \mathbf{Q}_S = \begin{pmatrix} \mathbf{A}_i & \mathbf{B}_i^t \\ \mathbf{B}_i & \mathbf{C}_i \end{pmatrix} \tag{63}$$

where $\mathbf{A}_i$, $\mathbf{B}_i$ and $\mathbf{C}_i$ are submatrices of size $r \times r$, $(n-r) \times r$, and $(n-r) \times (n-r)$, respectively, and let

$$\mathbf{D}_i = \begin{pmatrix} \mathbf{I} & -\mathbf{B}_i^t \mathbf{C}_i^{-1} \\ 0 & \mathbf{I} \end{pmatrix}. \tag{64}$$

We have

$$\mathbf{D}\mathbf{Q}_S^t \mathbf{X} = \begin{pmatrix} \mathbf{I} & -\mathbf{B}_i^t \mathbf{C}_i^{-1} \\ 0 & \mathbf{I} \end{pmatrix} \begin{pmatrix} \overline{\mathbf{X}}_a \\ 0 \end{pmatrix} = \begin{pmatrix} \overline{\mathbf{X}}_a \\ 0 \end{pmatrix}, \tag{65}$$

and

$$\mathrm{Cov}(\mathbf{D}\mathbf{Q}_S^t \mathbf{Z}_i) = \begin{pmatrix} \mathbf{I} & -\mathbf{B}_i^t \mathbf{C}_i^{-1} \\ 0 & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{A}_i & \mathbf{B}_i^t \\ \mathbf{B}_i & \mathbf{C}_i \end{pmatrix} \begin{pmatrix} \mathbf{I} & 0 \\ -\mathbf{C}_i^{-1}\mathbf{B}_i & \mathbf{I} \end{pmatrix} = \begin{pmatrix} \mathbf{A}_i - \mathbf{B}_i^t \mathbf{C}_i^{-1}\mathbf{B}_i & 0 \\ 0 & \mathbf{C}_i \end{pmatrix}. \tag{66}$$

Hence if we let $\mathbf{D}\mathbf{Q}_S^t \mathbf{Z}_i = (\overline{\mathbf{Z}}_{i,a}^t, \overline{\mathbf{Z}}_{i,b}^t)^t$ where $\overline{\mathbf{Z}}_{i,a}$ is of a length r, then $\overline{\mathbf{Z}}_{i,a}$ and $\overline{\mathbf{Z}}_{i,b}$ are statistically independent. It follows that

$$h(\mathbf{X} + \mathbf{Z}_i) = h(\mathbf{D}\mathbf{Q}_S^t \mathbf{X} + \mathbf{D}\mathbf{Q}_S^t \mathbf{Z}_i) = h(\mathbf{X}_a + \overline{\mathbf{Z}}_{i,a}, \overline{\mathbf{Z}}_{i,b}) = h(\mathbf{X}_a + \overline{\mathbf{Z}}_{i,a}) + h(\overline{\mathbf{Z}}_{i,b}). \tag{67}$$

So maximizing $h(\mathbf{X} + \mathbf{Z}_1) - \mu h(\mathbf{X} + \mathbf{Z}_2)$ is equivalent to maximizing $h(\overline{\mathbf{X}}_a + \overline{\mathbf{Z}}_{1,a}) - \mu h(\overline{\mathbf{X}}_a + \overline{\mathbf{Z}}_{2,a}) + h(\overline{\mathbf{Z}}_{1,b}) - \mu h(\overline{\mathbf{Z}}_{2,b})$. Note that $h(\overline{\mathbf{Z}}_{i,b})$, $i = 1, 2$, are constants. Hence to show that (8) has a Gaussian optimal solution for a rank deficient $\mathbf{S}$, it is sufficient to show that

$$\begin{aligned} \max_{p(\mathbf{x_a})} \quad & h(\overline{\mathbf{X}}_a + \overline{\mathbf{Z}}_{1,a}) - h(\overline{\mathbf{X}}_a + \overline{\mathbf{Z}}_{1,a}) \\ \text{subject to} \quad & \mathrm{Cov}(\overline{\mathbf{X}}_a) \preceq \mathrm{Diag}(\lambda_1, \cdots, \lambda_r), \end{aligned} \tag{68}$$

has a Gaussian optimal solution. Since $\mathrm{Diag}(\lambda_1, \cdots, \lambda_r)$ now has a full rank, we conclude that without loss of generality, we may assume that $\mathbf{S}$ in (8) is strictly positive definite.

Step 2: Construction of $\widetilde{P}$. Let $\mathbf{X}_G^*$ be an optimal Gaussian solution of P, and let $\mathbf{K}_X^*$ be the covariance matrix of $\mathbf{X}_G^*$. Then $\mathbf{K}_X^*$ is an optimal solution to the optimization problem (58). Although this conic program is generally nonconvex, it was shown in [7, Lemma 5] that for $\mathbf{S} \succ 0$, $\mathbf{K}_X^*$ must satisfy the following KKT-like conditions:

$$\frac{1}{2}(\mathbf{K}_X^* + \mathbf{K}_{Z_1})^{-1} + \mathbf{M}_1 = \frac{\mu}{2}(\mathbf{K}_X^* + \mathbf{K}_{Z_2})^{-1} + \mathbf{M}_2 \tag{69}$$

$$\mathbf{M}_1 \mathbf{K}_X^* = 0 \tag{70}$$

$$\mathbf{M}_2(\mathbf{S} - \mathbf{K}_X^*) = 0, \tag{71}$$

where $\mathbf{M}_1, \mathbf{M}_2 \succeq 0$ are Lagrange multipliers corresponding to $\mathbf{K}_X \succeq 0$ and $\mathbf{K}_X \preceq \mathbf{S}$, respectively. Let $\mathbf{K}_{\widetilde{Z}_1}$, $\mathbf{K}_{\widetilde{Z}_2}$ be two real symmetric matrices satisfying

$$\frac{1}{2}(\mathbf{K}_X^* + \mathbf{K}_{Z_1})^{-1} + \mathbf{M}_1 = \frac{1}{2}(\mathbf{K}_X^* + \mathbf{K}_{\widetilde{Z}_1})^{-1}, \tag{72}$$

$$\frac{\mu}{2}(\mathbf{K}_X^* + \mathbf{K}_{Z_2})^{-1} + \mathbf{M}_2 = \frac{\mu}{2}(\mathbf{K}_X^* + \mathbf{K}_{\widetilde{Z}_2})^{-1}. \tag{73}$$

17

We have the following results on $\mathbf{K}_{\widetilde{Z}_1}$ and $\mathbf{K}_{\widetilde{Z}_2}$ proved in [7, Lemma 11,12].

Lemma 9 *For* $\mathbf{K}_X^*$, $\mathbf{K}_{Z_i}$, $\mathbf{K}_{\widetilde{Z}_i}$, $\mathbf{M}_i$, $i = 1, 2$, *related through* (69) *to* (73), *and* $\mu \geq 1$, *we have*

$$0 \;\preceq\; \mathbf{K}_{\widetilde{Z}_1} \;\preceq\; \mathbf{K}_{Z_1}, \tag{74}$$

$$\mathbf{K}_{\widetilde{Z}_1} \;\preceq\; \mathbf{K}_{\widetilde{Z}_2} \;\preceq\; \mathbf{K}_{Z_2}. \tag{75}$$

The matrices $\mathbf{K}_{\widetilde{Z}_1}$, $\mathbf{K}_{\widetilde{Z}_2}$ are positive semidefinite, so they can serve as covariance matrices. Let $\widetilde{\mathbf{Z}}_1$, $\widetilde{\mathbf{Z}}_2$ be two Gaussian vectors with covariance matrix $\mathbf{K}_{\widetilde{Z}_1}$ and $\mathbf{K}_{\widetilde{Z}_2}$, respectively. Let us define the auxiliary optimization problem $\widetilde{P}$ as:

$$\begin{array}{ll} \max_{p(\mathbf{x})} & h(\mathbf{X} + \widetilde{\mathbf{Z}}_1) - \mu h(\mathbf{X} + \widetilde{\mathbf{Z}}_2) + F \\ \text{subject to} & \mathrm{Cov}(\mathbf{X}) \preceq \mathbf{S}, \end{array} \tag{76}$$

where the constant

$$F := h(\mathbf{Z}_1) - h(\widetilde{\mathbf{Z}}_1) + \mu \left(h(\mathbf{X}_G^{(S)} + \widetilde{\mathbf{Z}}_2) - h(\mathbf{X}_G^{(S)} + \mathbf{Z}_2) \right), \tag{77}$$

$\mathbf{X}_G^{(S)}$ is a Gaussian vector with covariance matrix $\mathbf{S}$ and independent of $\mathbf{Z}_2$ and $\widetilde{\mathbf{Z}}_2$, and the maximization is over all random vector $\mathbf{X}$ independent of $\widetilde{\mathbf{Z}}_1$ and $\widetilde{\mathbf{Z}}_2$.

In [7, p. 3937], the authors call the process of replacing $\mathbf{Z}_1$ and $\mathbf{Z}_2$ with $\widetilde{\mathbf{Z}}_1$ and $\widetilde{\mathbf{Z}}_2$, respectively, *enhancement*. Next, we show that the auxiliary optimization problem $\widetilde{P}$ defined in (76) satisfies the desired chain of relationships (59).

Step 3: Proof of $(P) \leq (\widetilde{P})$. Note that P and $\widetilde{P}$ have the same solution space. So to show that $(P) \leq (\widetilde{P})$, it is sufficient to show that for each admissible solution, the value of the objective function of P is less or equal to that of $\widetilde{P}$.

The difference between the objective functions of P and $\widetilde{P}$ can be written as

$$h(\mathbf{X} + \mathbf{Z}_1) - h(\mathbf{Z}_1) - h(\mathbf{X} + \widetilde{\mathbf{Z}}_1) + h(\widetilde{\mathbf{Z}}_1)$$
$$- \mu \left(h(\mathbf{X} + \mathbf{Z}_2) - h(\mathbf{X} + \widetilde{\mathbf{Z}}_2) - h(\mathbf{X}_S + \mathbf{Z}_2) + h(\mathbf{X}_S + \widetilde{\mathbf{Z}}_2) \right). \tag{78}$$

By Lemma 9, $\mathbf{K}_{Z_i} \succeq \mathbf{K}_{\widetilde{Z}_i}$ for $i = 1, 2$. So we can write $\mathbf{Z}_i = \widetilde{\mathbf{Z}}_i + \widehat{\mathbf{Z}}_i$, where $\widehat{\mathbf{Z}}_i$ is a Gaussian vector independent of $\widetilde{\mathbf{Z}}_i$. We have

$$\begin{aligned} h(\mathbf{X} + \mathbf{Z}_1) - h(\mathbf{Z}_1) - h(\mathbf{X} + \widetilde{\mathbf{Z}}_1) + h(\widetilde{\mathbf{Z}}_1) &= I(\mathbf{X}; \mathbf{X} + \mathbf{Z}_1) - I(\mathbf{X}; \mathbf{X} + \widetilde{\mathbf{Z}}_1) & (79) \\ &= I(\mathbf{X}; \mathbf{X} + \widetilde{\mathbf{Z}}_1 + \widehat{\mathbf{Z}}_1) - I(\mathbf{X}; \mathbf{X} + \widetilde{\mathbf{Z}}_1) & (80) \\ &\leq 0, & (81) \end{aligned}$$

18

where the inequality is due to the Markov chain

$$\mathbf{X} \to \mathbf{X} + \widetilde{\mathbf{Z}}_1 \to \mathbf{X} + \widetilde{\mathbf{Z}}_1 + \widehat{\mathbf{Z}}_1. \tag{82}$$

Further, let $\mathbf{X}_G$ be a Gaussian random vector with the same covariance matrix as that of $\mathbf{X}$. Assume that $\mathbf{X}_G$ is independent of $\widetilde{\mathbf{Z}}_2$ and $\widehat{\mathbf{Z}}_2$. Note that both $\mathbf{X}_G$ and $\mathbf{X}_G^{(S)}$ are Gaussian and that

$$\mathrm{Cov}(\mathbf{X}_G) = \mathrm{Cov}(\mathbf{X}) \preceq \mathbf{S} = \mathrm{Cov}(\mathbf{X}_G^{(S)}). \tag{83}$$

So we can write $\mathbf{X}_G^{(S)} = \mathbf{X}_G + \widehat{\mathbf{X}}_G$, where $\widehat{\mathbf{X}}_G$ is a Gaussian random vector independent of $\mathbf{X}_G$. We have

$$
\begin{aligned}
h(\mathbf{X} + \mathbf{Z}_2) &- h(\mathbf{X} + \widetilde{\mathbf{Z}}_2) - h(\mathbf{X}_G^{(S)} + \mathbf{Z}_2) + h(\mathbf{X}_G^{(S)} + \widetilde{\mathbf{Z}}_2) \\
&= h(\mathbf{X} + \widetilde{\mathbf{Z}}_2 + \widehat{\mathbf{Z}}_2) - h(\mathbf{X} + \widetilde{\mathbf{Z}}_2) - (h(\mathbf{X}_G^{(S)} + \widetilde{\mathbf{Z}}_2 + \widehat{\mathbf{Z}}_2) - h(\mathbf{X}_G^{(S)} + \widetilde{\mathbf{Z}}_2)) &(84) \\
&= I(\widehat{\mathbf{Z}}_2; \mathbf{X} + \widetilde{\mathbf{Z}}_2 + \widehat{\mathbf{Z}}_2) - I(\widehat{\mathbf{Z}}_2; \mathbf{X}_G^{(S)} + \widetilde{\mathbf{Z}}_2 + \widehat{\mathbf{Z}}_2) &(85) \\
&\geq I(\widehat{\mathbf{Z}}_2; \mathbf{X}_G + \widetilde{\mathbf{Z}}_2 + \widehat{\mathbf{Z}}_2) - I(\widehat{\mathbf{Z}}_2; \mathbf{X}_G^{(S)} + \widetilde{\mathbf{Z}}_2 + \widehat{\mathbf{Z}}_2) &(86) \\
&= I(\widehat{\mathbf{Z}}_2; \mathbf{X}_G + \widetilde{\mathbf{Z}}_2 + \widehat{\mathbf{Z}}_2) - I(\widehat{\mathbf{Z}}_2; \widehat{\mathbf{X}}_G + \mathbf{X}_G + \widetilde{\mathbf{Z}}_2 + \widehat{\mathbf{Z}}_2) &(87) \\
&\geq 0, &(88)
\end{aligned}
$$

where inequality (86) follows from

$$I(\widehat{\mathbf{Z}}_2; \mathbf{X} + \widetilde{\mathbf{Z}}_2 + \widehat{\mathbf{Z}}_2) \geq I(\widehat{\mathbf{Z}}_2; \mathbf{X}_G + \widetilde{\mathbf{Z}}_2 + \widehat{\mathbf{Z}}_2) \tag{89}$$

which is due to the worst noise result of Lemma 2, and inequality (88) follows from the Markov chain

$$\widehat{\mathbf{Z}}_2 \to \mathbf{X}_G + \widetilde{\mathbf{Z}}_2 + \widehat{\mathbf{Z}}_2 \to \widehat{\mathbf{X}}_G + \mathbf{X}_G + \widetilde{\mathbf{Z}}_2 + \widehat{\mathbf{Z}}_2. \tag{90}$$

Substituting (81) and (88) into (78), we conclude that the difference between the objective functions of P and $\widetilde{\mathrm{P}}$ is nonpositive for any admissible $\mathbf{X}$ (i.e., $\mathrm{Cov}(\mathbf{X}) \preceq \mathbf{S}$) and any $\mu > 1$.

Step 4: Proof of $(\widetilde{P}) = (\widetilde{P}_G)$. To show that $(\widetilde{P}) = (\widetilde{P}_G)$, it is sufficient to show that $\mathbf{X}_G^*$, the optimal solution of P_G, is also an optimal solution of $\widetilde{P}$. We consider the cases $\mu = 1$ and $\mu > 1$ separately.

First assume that $\mu > 1$. By Lemma 9, $\mathbf{K}_{\widetilde{Z}_2} \succeq \mathbf{K}_{\widetilde{Z}_1}$. So we can write $\widetilde{\mathbf{Z}}_2 = \widetilde{\mathbf{Z}}_1 + \widetilde{\mathbf{Z}}$, where $\widetilde{\mathbf{Z}}$ is Gaussian and independent of $\widetilde{\mathbf{Z}}_1$. We have

$$
\begin{aligned}
h(\mathbf{X} + \widetilde{\mathbf{Z}}_1) - \mu h(\mathbf{X} + \widetilde{\mathbf{Z}}_2) &= h(\mathbf{X} + \widetilde{\mathbf{Z}}_1) - \mu h(\mathbf{X} + \widetilde{\mathbf{Z}}_1 + \widetilde{\mathbf{Z}}) &(91) \\
&\leq h(\mathbf{X} + \widetilde{\mathbf{Z}}_1) - \frac{\mu n}{2} \log\left(\exp\left[\frac{2}{n}h(\mathbf{X} + \widetilde{\mathbf{Z}}_1)\right] + \exp\left[\frac{2}{n}h(\widetilde{\mathbf{Z}})\right]\right) &(92) \\
&\leq f\left(h(\widetilde{\mathbf{Z}}) - \frac{n}{2}\log(\mu - 1); h(\widetilde{\mathbf{Z}})\right), &(93)
\end{aligned}
$$

19

where (92) follows from the classical EPI, and the function f in (93) was defined in (4). Next, we verify that the upper bound on the right-hand side of (93) is achieved by $\mathbf{X}_G^*$. Substituting (72) and (73) into the KKT-like condition (69), we obtain

$$(\mathbf{K}_X^* + \mathbf{K}_{\widetilde{Z}_1})^{-1} = \mu(\mathbf{K}_X^* + \mathbf{K}_{\widetilde{Z}_2})^{-1}, \tag{94}$$

which gives

$$\mathbf{K}_X^* + \mathbf{K}_{\widetilde{Z}_1} = (\mu - 1)^{-1}\mathbf{K}_{\widetilde{Z}}. \tag{95}$$

Hence, $\mathbf{X}_G^* + \widetilde{\mathbf{Z}}_1$ and $\widetilde{\mathbf{Z}}$ have proportional covariance matrices and inequality (92) holds with equality. Further by (95),

$$h(\mathbf{X}_G^* + \widetilde{\mathbf{Z}}_1) = h(\widetilde{\mathbf{Z}}) - \frac{n}{2}\log(\mu - 1). \tag{96}$$

A comparison of (96) and (5) confirms that $h(\mathbf{X}_G^* + \widetilde{\mathbf{Z}}_1)$ achieves the global maxima of function $f(t; h(\widetilde{\mathbf{Z}}))$, i.e., inequality (93) becomes equality with $\mathbf{X}_G^*$. We thus conclude that $\mathbf{X}_G^*$ is an optimal solution of $\widetilde{P}$ for all $\mu > 1$.

For $\mu = 1$, we have from (94) that $\mathbf{K}_{\widetilde{Z}_1} = \mathbf{K}_{\widetilde{Z}_2}$. So the objective function of $\widetilde{P}$ is constant, and $\mathbf{X}_G^*$ is trivially an optimal solution of $\widetilde{P}$.

Step 5: Proof of $(\widetilde{P}_G) = (P_G)$. Note that $\mathbf{X}_G^*$ is an optimal solution of both $\widetilde{P}_G$ and P_G. So to show that $(\widetilde{P}_G) = (P_G)$, we only need to compare the objective functions of $\widetilde{P}_G$ and P_G evaluated at $\mathbf{X}_G^*$. The following result, which is a minor generalization of [7, Lemma 11,12], shows that the objective functions of $\widetilde{P}_G$ and P_G take equal values at $\mathbf{X}_G^*$.

Lemma 10 *For* $\mathbf{K}_X^*$, $\mathbf{K}_{Z_i}$, $\mathbf{K}_{\widetilde{Z}_i}$, $\mathbf{M}_i$, $i = 1, 2$, *defined through* (69) *to* (73) *and* $\mu \geq 1$, *we have*

$$\begin{align}
(\mathbf{K}_X^* + \mathbf{K}_{\widetilde{Z}_1})^{-1}\mathbf{K}_{\widetilde{Z}_1} &= (\mathbf{K}_X^* + \mathbf{K}_{Z_1})^{-1}\mathbf{K}_{Z_1}, \tag{97} \\
(\mathbf{K}_X^* + \mathbf{K}_{\widetilde{Z}_2})^{-1}(\mathbf{S} + \mathbf{K}_{\widetilde{Z}_2}) &= (\mathbf{K}_X^* + \mathbf{K}_{Z_2})^{-1}(\mathbf{S} + \mathbf{K}_{Z_2}). \tag{98}
\end{align}$$

Combining Steps 1-5, we conclude that for any $\mu \geq 1$ and any positive semidefinite $\mathbf{S}$, a Gaussian $\mathbf{X}$ is an optimal solution of (8). This completes the direct proof of Theorem 1.

A few comments on why we need the auxiliary optimization problem $\widetilde{P}$ are now in place. For the classical EPI to be tight, we need $\mathbf{K}_X^* + \mathbf{K}_{Z_1}$ and $\mathbf{K}_X^* + \mathbf{K}_{Z_2}$ to be proportional to each other. However, by the KKT-like condition (69), a guarantee of proportionality requires both multipliers $\mathbf{M}_1$ and $\mathbf{M}_2$ be zero. The purpose of enhancement is to absorb the (possibly) nonzero Lagrange multipliers $\mathbf{M}_1$, $\mathbf{M}_2$ into the covariance matrices of $\mathbf{Z}_1$ and $\mathbf{Z}_2$, creating a new optimization problem which can be solved directly by the classical EPI. The constant F is needed to make sure that $(P_G) = (\widetilde{P}_G)$; the choice of F is motivated by the vector Gaussian broadcast channel problem.

B Proof of Lemma 3

We first give some preliminaries on Fisher information and score function. This material can be found, for example, in [15, Chapter 3.2].

Definition 11 *For a random vector $\mathbf{U}$ with a differentiable density function f_U in $\mathcal{R}^n$, the Fisher information matrix $\mathbf{J}(\cdot)$ is defined as*

$$\mathbf{J}(\mathbf{U}) := \mathbb{E}[\boldsymbol{\rho}_U(\mathbf{U})\boldsymbol{\rho}_U^t(\mathbf{U})], \tag{99}$$

where the vector-valued score function $\boldsymbol{\rho}_U(\cdot)$ is defined as

$$\boldsymbol{\rho}_U(\mathbf{u}) := \nabla \log f_U(\mathbf{u}) = \left(\frac{\partial}{\partial u_1} \log f_U(\mathbf{u}), \cdots, \frac{\partial}{\partial u_n} \log f_U(\mathbf{u}) \right)^t. \tag{100}$$

The following results on score function are known.

Lemma 12 *The following statements on score function are true.*

1. *(Gaussian Distribution) If $\mathbf{U}$ is a Gaussian vector with zero mean and positive definite covariance matrix $\mathbf{K}_U$, then*
$$\boldsymbol{\rho}_U(\mathbf{u}) = -\mathbf{K}_U^{-1}\mathbf{u}. \tag{101}$$

2. *(Stein Identity) For any smooth scalar-valued function g well behaved at infinity, we have*
$$\mathbb{E}[g(\mathbf{U})\boldsymbol{\rho}_U(\mathbf{U})] = -\mathbb{E}[\nabla g(\mathbf{U})]. \tag{102}$$
In particular, we have
$$\mathbb{E}[\boldsymbol{\rho}_U(\mathbf{U})] = 0 \quad and \quad \mathbb{E}[\mathbf{U}\boldsymbol{\rho}_U^t(\mathbf{U})] = -\mathbf{I}, \tag{103}$$
where $\mathbf{I}$ is the identity matrix.

3. *(Behavior on Convolution) If $\mathbf{U}$, $\mathbf{V}$ are two independent random vectors and $\mathbf{W} = \mathbf{U} + \mathbf{V}$, then*
$$\boldsymbol{\rho}_W(\mathbf{w}) = \mathbb{E}[\boldsymbol{\rho}_U(\mathbf{U})|\mathbf{W} = \mathbf{w}] = \mathbb{E}[\boldsymbol{\rho}_V(\mathbf{V})|\mathbf{W} = \mathbf{w}]. \tag{104}$$

We now use the above properties of score function to prove Lemma 3. We first prove the Cramér-Rao inequality. The Fisher information matrix $\mathbf{J}(\mathbf{U})$ has nothing to do with the mean of $\mathbf{U}$, so without loss of generality we can assume that $\mathbf{U}$ has zero mean. We have

$$
\begin{align}
0 \quad &\preceq \quad \mathbb{E}[(\boldsymbol{\rho}_U(\mathbf{U}) + \mathbf{K}_U^{-1}\mathbf{U})(\boldsymbol{\rho}_U(\mathbf{U}) + \mathbf{K}_U^{-1}\mathbf{U})^t] \tag{105} \\
&= \quad \mathbb{E}[\boldsymbol{\rho}_U(\mathbf{U})\boldsymbol{\rho}_U(\mathbf{U})^t] + \mathbf{K}_U^{-1}\mathbb{E}[\mathbf{U}\boldsymbol{\rho}_U^t(\mathbf{U})] + \mathbb{E}[\boldsymbol{\rho}_U(\mathbf{U})\mathbf{U}^t]\mathbf{K}_U^{-1} + \mathbf{K}_U^{-1}\mathbb{E}[\mathbf{U}\mathbf{U}^t]\mathbf{K}_U^{-1} \tag{106} \\
&= \quad \mathbf{J}(\mathbf{U}) - \mathbf{K}_U^{-1} - \mathbf{K}_U^{-1} + \mathbf{K}_U^{-1} \tag{107} \\
&= \quad \mathbf{J}(\mathbf{U}) - \mathbf{K}_U^{-1}. \tag{108}
\end{align}
$$

Here in (107) we use the facts that

$$
\mathbb{E}[\boldsymbol{\rho}_U(\mathbf{U})\boldsymbol{\rho}_U(\mathbf{U})^t] = \mathbf{J}(\mathbf{U}) \tag{109}
$$

by the definition of Fisher information matrix and that

$$
\mathbb{E}[\mathbf{U}\boldsymbol{\rho}_U^t(\mathbf{U})] = \mathbb{E}[\boldsymbol{\rho}_U(\mathbf{U})\mathbf{U}^t] = \mathbf{I} \tag{110}
$$

by the Stein identity. We conclude that $\mathbf{J}(\mathbf{U}) \succeq \mathbf{K}_U^{-1}$ for any random vector $\mathbf{U}$ with a strictly positive definite covariance matrix $\mathbf{K}_U$.

The matrix FII can be proved similarly:

$$
\begin{align}
0 \quad &\preceq \quad \mathbb{E}[(\boldsymbol{\rho}_W(\mathbf{W}) - \mathbf{A}\boldsymbol{\rho}_U(\mathbf{U}) - (\mathbf{I} - \mathbf{A})\boldsymbol{\rho}_V(\mathbf{V}))(\boldsymbol{\rho}_W(\mathbf{W}) - \mathbf{A}\boldsymbol{\rho}_U(\mathbf{U}) - (\mathbf{I} - \mathbf{A})\boldsymbol{\rho}_V(\mathbf{V}))^t] \tag{111} \\
&= \quad \mathbb{E}[\boldsymbol{\rho}_W(\mathbf{W})\boldsymbol{\rho}_W^t(\mathbf{W})] + \mathbf{A}\mathbb{E}[\boldsymbol{\rho}_U(\mathbf{U})\boldsymbol{\rho}_U^t(\mathbf{U})]\mathbf{A}^t + (\mathbf{I} - \mathbf{A})\mathbb{E}[\boldsymbol{\rho}_V(\mathbf{V})\boldsymbol{\rho}_V^t(\mathbf{V})](\mathbf{I} - \mathbf{A})^t \\
&\qquad -\mathbb{E}[\boldsymbol{\rho}_W(\mathbf{W})\boldsymbol{\rho}_U^t(\mathbf{U})]\mathbf{A}^t - \mathbf{A}\mathbb{E}[\boldsymbol{\rho}_U(\mathbf{U})\boldsymbol{\rho}_W^t(\mathbf{W})] \\
&\qquad -\mathbb{E}[\boldsymbol{\rho}_W(\mathbf{W})\boldsymbol{\rho}_V^t(\mathbf{V})](\mathbf{I} - \mathbf{A})^t - (\mathbf{I} - \mathbf{A})\mathbb{E}[\boldsymbol{\rho}_V(\mathbf{V})\boldsymbol{\rho}_W^t(\mathbf{W})] \\
&\qquad +\mathbf{A}\mathbb{E}[\boldsymbol{\rho}_U(\mathbf{U})\boldsymbol{\rho}_V^t(\mathbf{V})](\mathbf{I} - \mathbf{A})^t + (\mathbf{I} - \mathbf{A})\mathbb{E}[\boldsymbol{\rho}_V(\mathbf{V})\boldsymbol{\rho}_U^t(\mathbf{U})]\mathbf{A}^t. \tag{112}
\end{align}
$$

By the definition of Fisher information matrix,

$$
\mathbb{E}[\boldsymbol{\rho}_W(\mathbf{W})\boldsymbol{\rho}_W^t(\mathbf{W})] = \mathbf{J}(\mathbf{W}), \quad \mathbb{E}[\boldsymbol{\rho}_U(\mathbf{U})\boldsymbol{\rho}_U^t(\mathbf{U})] = \mathbf{J}(\mathbf{U}), \quad \mathbb{E}[\boldsymbol{\rho}_V(\mathbf{V})\boldsymbol{\rho}_V^t(\mathbf{V})] = \mathbf{J}(\mathbf{V}). \tag{113}
$$

By the convolution behavior of score function,

$$
\mathbb{E}[\boldsymbol{\rho}_W(\mathbf{W})\boldsymbol{\rho}_U^t(\mathbf{U})] = \mathbb{E}[\boldsymbol{\rho}_W(\mathbf{W})\mathbb{E}[\boldsymbol{\rho}_U^t(\mathbf{U})|\mathbf{W}]] = \mathbb{E}[\boldsymbol{\rho}_W(\mathbf{W})\boldsymbol{\rho}_W^t(\mathbf{W})] = \mathbf{J}(\mathbf{W}) \tag{114}
$$

and similarly

$$
\mathbb{E}[\boldsymbol{\rho}_U(\mathbf{U})\boldsymbol{\rho}_W^t(\mathbf{W})] = \mathbf{J}(\mathbf{W}), \quad \mathbb{E}[\boldsymbol{\rho}_W(\mathbf{W})\boldsymbol{\rho}_V^t(\mathbf{V})] = \mathbb{E}[\boldsymbol{\rho}_V(\mathbf{V})\boldsymbol{\rho}_W^t(\mathbf{W})] = \mathbf{J}(\mathbf{W}). \tag{115}
$$

Finally, since $\mathbf{U}$, $\mathbf{V}$ are independent and by the Stein identity with $f = 1$, we have

$$
\mathbb{E}[\boldsymbol{\rho}_U(\mathbf{U})\boldsymbol{\rho}_V^t(\mathbf{V})] = \mathbb{E}[\boldsymbol{\rho}_U(\mathbf{U})]\mathbb{E}[\boldsymbol{\rho}_U^t(\mathbf{V})] = 0 \tag{116}
$$

and similarly
$$\mathbb{E}[\boldsymbol{\rho}_V(\mathbf{V})\boldsymbol{\rho}_U^t(\mathbf{U})] = 0. \tag{117}$$
Substituting (113)-(117) into (112), we obtain
$$
\begin{aligned}
0 &\preceq \mathbf{J}(\mathbf{W}) + \mathbf{A}\mathbf{J}(\mathbf{U})\mathbf{A}^t + (\mathbf{I} - \mathbf{A})\mathbf{J}(\mathbf{V})(\mathbf{I} - \mathbf{A})^t - \mathbf{J}(\mathbf{W})\mathbf{A}^t - \mathbf{A}\mathbf{J}(\mathbf{W}) - \mathbf{J}(\mathbf{W})(\mathbf{I} - \mathbf{A})^t \\
&\quad - (\mathbf{I} - \mathbf{A})\mathbf{J}(\mathbf{W}) \tag{118} \\
&= -\mathbf{J}(\mathbf{W}) + \mathbf{A}\mathbf{J}(\mathbf{U})\mathbf{A}^t + (\mathbf{I} - \mathbf{A})\mathbf{J}(\mathbf{V})(\mathbf{I} - \mathbf{A})^t, \tag{119}
\end{aligned}
$$
which gives
$$\mathbf{J}(\mathbf{W}) \preceq \mathbf{A}\mathbf{J}(\mathbf{U})\mathbf{A}^t + (\mathbf{I} - \mathbf{A})\mathbf{J}(\mathbf{V})(\mathbf{I} - \mathbf{A})^t \tag{120}$$
for any square matrix $\mathbf{A}$. This completes the proof.

C   A Perturbation Proof of Theorem 1

We first give the outline of the proof.

Proof Outline. Without loss of generality, let us assume that $\mathbf{S} \succ 0$. To show that a Gaussian $\mathbf{X}$ is an optimal solution of P, it is sufficient to show that $(P) = (P_G)$. We have $(P) \geq (P_G)$ (for free); we only need to show that $(P) \leq (P_G)$. For that purpose we shall consider the auxiliary optimization problem $\overline{P}$:

$$
\begin{aligned}
&\max_{p(\mathbf{x})} \quad h(\mathbf{X} + \widetilde{\mathbf{Z}}_1) - \mu h(\mathbf{X} + \mathbf{Z}_2) + h(\mathbf{Z}_1) - h(\widetilde{\mathbf{Z}}_1) \\
&\text{subject to} \quad \text{Cov}(\mathbf{X}) \preceq \mathbf{S},
\end{aligned} \tag{121}
$$

where the maximization is over all random vector $\mathbf{X}$ independent of $\widetilde{\mathbf{Z}}_1$ and $\mathbf{Z}_2$. Compared with the auxiliary optimization problem $\widetilde{P}$ in the direct proof, this enhancement is only on $\mathbf{Z}_1$. Following the same footsteps as those in the direct proof, we can show that $(P) \leq (\overline{P})$ and $(\overline{P}_G) = (P_G)$. (In proving $(P) \leq (\overline{P})$, only the equations (79)-(81) and the Markov chain (82) in Appendix A are needed.) All we need to show now is that $(\overline{P}) = (\overline{P}_G)$.

Proof of $(\overline{P}) = (\overline{P}_G)$. To show that $(\overline{P}) = (\overline{P}_G)$, we shall show that $\mathbf{X}_G^*$ is a global optimal solution of $\overline{P}$. For that we shall prove the following strong result: for any admissible random vector $\mathbf{X}$ there is a monotone increasing path connecting $\mathbf{X}$ and $\mathbf{X}_G^*$ (see Figure 1).

We consider the "covariance-preserving" transformation of Dembo et al. [10]:

$$\mathbf{X}_\lambda = \sqrt{1 - \lambda}\mathbf{X} + \sqrt{\lambda}\mathbf{X}_G^*, \quad \lambda \in [0, 1]. \tag{122}$$

Then $\{\mathbf{X}_\lambda\}$ is a family of distributions indexed by $\lambda \in [0, 1]$ and connecting $\mathbf{X}$ (when $\lambda = 0$) with $\mathbf{X}_G^*$ (when $\lambda = 1$). Let $\overline{g}(\lambda)$ be the objective function of $\overline{P}$ evaluated along the path $\{\mathbf{X}_\lambda\}$:

$$\overline{g}(\lambda) := h(\mathbf{X}_\lambda + \widetilde{\mathbf{Z}}_1) - \mu h(\mathbf{X}_\lambda + \mathbf{Z}_2) + h(\mathbf{Z}_1) - h(\widetilde{\mathbf{Z}}_1). \tag{123}$$

$\mathbf{X}_G^*$

$\mathbf{X}_\lambda := \sqrt{1-\lambda}\mathbf{X} + \sqrt{\lambda}\mathbf{X}_G^*$

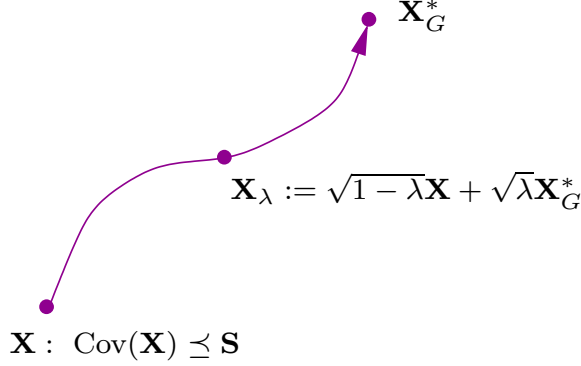$\mathbf{X}:\ \mathrm{Cov}(\mathbf{X}) \preceq \mathbf{S}$

Figure 1: A monotone path connecting $\mathbf{X}$ and $\mathbf{X}_G^*$.

Next, we calculate the derivative of $\overline{g}$ over λ.

Note that $\widetilde{\mathbf{Z}}_1$ is Gaussian and that a Gaussian distribution is stable under convolution. We can write

$$\widetilde{\mathbf{Z}}_1 = \sqrt{1-\lambda}\widetilde{\mathbf{Z}}_{1,1} + \sqrt{\lambda}\widetilde{\mathbf{Z}}_{1,2}, \tag{124}$$

where $\widetilde{\mathbf{Z}}_{1,1}$, $\widetilde{\mathbf{Z}}_{1,2}$ are independent and have the same distribution as that of $\widetilde{\mathbf{Z}}_1$. We have

$$\begin{aligned}
h(\mathbf{X}_\lambda + \widetilde{\mathbf{Z}}_1) &= h(\sqrt{1-\lambda}\mathbf{X} + \sqrt{\lambda}\mathbf{X}_G^* + \widetilde{\mathbf{Z}}_i) && (125)\\
&= h(\sqrt{1-\lambda}(\mathbf{X} + \widetilde{\mathbf{Z}}_{1,1}) + \sqrt{\lambda}(\mathbf{X}_G^* + \widetilde{\mathbf{Z}}_{1,2})) && (126)\\
&= h(\mathbf{X} + \widetilde{\mathbf{Z}}_{1,1} + \sqrt{\lambda(1-\lambda)^{-1}}(\mathbf{X}_G^* + \widetilde{\mathbf{Z}}_{1,2})) + (n/2)\log(1-\lambda). && (127)
\end{aligned}$$

By the (vector) de Bruijn identity [10, Theorem 14],

$$\begin{aligned}
2(1-\lambda)&\frac{d}{d\lambda}h(\mathbf{X}_\lambda + \widetilde{\mathbf{Z}}_1)\\
&= (1-\lambda)^{-1}\mathrm{Tr}\left((\mathbf{K}_X^* + \mathbf{K}_{\widetilde{Z}_1})\mathbf{J}\left(\mathbf{X} + \widetilde{\mathbf{Z}}_{1,1} + \sqrt{\lambda(1-\lambda)^{-1}}(\mathbf{X}_G^* + \widetilde{\mathbf{Z}}_{1,2})\right)\right) - n && (128)\\
&= \mathrm{Tr}\left((\mathbf{K}_X^* + \mathbf{K}_{\widetilde{Z}_1})\mathbf{J}\left(\sqrt{1-\lambda}(\mathbf{X} + \widetilde{\mathbf{Z}}_{1,1}) + \sqrt{\lambda}(\mathbf{X}_G^* + \widetilde{\mathbf{Z}}_{1,2})\right)\right) - n && (129)\\
&= \mathrm{Tr}\left((\mathbf{K}_X^* + \mathbf{K}_{\widetilde{Z}_1})\mathbf{J}\left(\mathbf{X}_\lambda + \widetilde{\mathbf{Z}}_1\right)\right) - n. && (130)
\end{aligned}$$

Similarly, we have

$$2(1-\lambda)\frac{d}{d\lambda}h(\mathbf{X}_\lambda + \mathbf{Z}_2) = \mathrm{Tr}\left((\mathbf{K}_X^* + \mathbf{K}_{Z_2})\mathbf{J}(\mathbf{X}_\lambda + \mathbf{Z}_2)\right) - n. \tag{131}$$

Combining (130) and (131), we have

$$2(1-\lambda)\overline{g}'(\lambda) = \mathrm{Tr}\left((\mathbf{K}_X^* + \mathbf{K}_{\widetilde{Z}_1})\mathbf{J}(\mathbf{X}_\lambda + \widetilde{\mathbf{Z}}_1) - \mu(\mathbf{K}_X^* + \mathbf{K}_{Z_2})\mathbf{J}(\mathbf{X}_\lambda + \mathbf{Z}_2)\right) + n(\mu - 1). \tag{132}$$

By the definition of $\mathbf{K}_{\widetilde{Z}_1}$ and the KKT-like condition (69), we have

$$\frac{1}{2}(\mathbf{K}_X^* + \mathbf{K}_{\widetilde{Z}_1})^{-1} = \frac{\mu}{2}(\mathbf{K}_X^* + \mathbf{K}_{Z_2})^{-1} + \mathbf{M}_2. \tag{133}$$

By the facts that $\mu \geq 1$ and $\mathbf{M}_2 \succeq 0$, we obtain from (133) that

$$\frac{1}{2}(\mathbf{K}_X^* + \mathbf{K}_{\widetilde{Z}_1})^{-1} \succeq \frac{1}{2}(\mathbf{K}_X^* + \mathbf{K}_{Z_2})^{-1} \tag{134}$$

and hence that

$$\mathbf{K}_{Z_2} \succeq \mathbf{K}_{\widetilde{Z}_1}. \tag{135}$$

We can now write $\mathbf{Z}_2 = \widetilde{\mathbf{Z}}_1 + \overline{\mathbf{Z}}$, where $\overline{\mathbf{Z}}$ is Gaussian and independent of $\widetilde{\mathbf{Z}}_1$. Applying the matrix FII of Lemma 3 with

$$\mathbf{A} = (\mathbf{K}_X^* + \mathbf{K}_{Z_2})^{-1}(\mathbf{K}_X^* + \mathbf{K}_{\widetilde{Z}_1}) \quad \text{and} \quad \mathbf{I} - \mathbf{A} = (\mathbf{K}_X^* + \mathbf{K}_{Z_2})^{-1}\mathbf{K}_{\overline{Z}}, \tag{136}$$

we have

$$
\begin{aligned}
\mathbf{J}(\mathbf{X}_\lambda + \mathbf{Z}_2) &= \mathbf{J}(\mathbf{X}_\lambda + \widetilde{\mathbf{Z}}_1 + \overline{\mathbf{Z}}) && (137)\\
&\leq (\mathbf{K}_X^* + \mathbf{K}_{Z_2})^{-1}(\mathbf{K}_X^* + \mathbf{K}_{\widetilde{Z}_1})\mathbf{J}(\mathbf{X}_\lambda + \widetilde{\mathbf{Z}}_1)(\mathbf{K}_X^* + \mathbf{K}_{\widetilde{Z}_1})(\mathbf{K}_X^* + \mathbf{K}_{Z_2})^{-1} \\
&\quad + (\mathbf{K}_X^* + \mathbf{K}_{Z_2})^{-1}\mathbf{K}_{\overline{Z}}\mathbf{J}(\overline{\mathbf{Z}})\mathbf{K}_{\overline{Z}}(\mathbf{K}_X^* + \mathbf{K}_{Z_2})^{-1} && (138)\\
&= (\mathbf{K}_X^* + \mathbf{K}_{Z_2})^{-1}(\mathbf{K}_X^* + \mathbf{K}_{\widetilde{Z}_1})\mathbf{J}(\mathbf{X}_\lambda + \widetilde{\mathbf{Z}}_1)(\mathbf{K}_X^* + \mathbf{K}_{\widetilde{Z}_1})(\mathbf{K}_X^* + \mathbf{K}_{Z_2})^{-1} \\
&\quad + (\mathbf{K}_X^* + \mathbf{K}_{Z_2})^{-1}\mathbf{K}_{\overline{Z}}(\mathbf{K}_X^* + \mathbf{K}_{Z_2})^{-1}, && (139)
\end{aligned}
$$

where the last equality follows from the fact that $\overline{\mathbf{Z}}$ is Gaussian so

$$\mathbf{K}_{\overline{Z}}\mathbf{J}(\overline{\mathbf{Z}}) = \mathbf{I}. \tag{140}$$

Substituting (139) into (132) and using the fact that $\mathbf{K}_{\overline{Z}} = \mathbf{K}_{Z_2} - \mathbf{K}_{\widetilde{Z}_1}$, we obtain

$$
\begin{aligned}
2(1 - \lambda)&\overline{g}'(\lambda) \\
&\geq \operatorname{Tr}((\mathbf{K}_X^* + \mathbf{K}_{\widetilde{Z}_1})\mathbf{J}(\mathbf{X}_\lambda + \widetilde{\mathbf{Z}}_1) - \mu(\mathbf{K}_X^* + \mathbf{K}_{\widetilde{Z}_1})\mathbf{J}(\mathbf{X}_\lambda + \widetilde{\mathbf{Z}}_1)(\mathbf{K}_X^* + \mathbf{K}_{\widetilde{Z}_1})(\mathbf{K}_X^* + \mathbf{K}_{Z_2})^{-1} \\
&\quad - \mu\mathbf{K}_{\overline{Z}}(\mathbf{K}_X^* + \mathbf{K}_{Z_2})^{-1}) + n(\mu - 1) && (141)\\
&= 2\operatorname{Tr}\left(\left((\mathbf{K}_X^* + \mathbf{K}_{\widetilde{Z}_1})\mathbf{J}(\mathbf{X}_\lambda + \widetilde{\mathbf{Z}}_1)(\mathbf{K}_X^* + \mathbf{K}_{\widetilde{Z}_1}) - (\mathbf{K}_X^* + \mathbf{K}_{\widetilde{Z}_1})\right)\mathbf{M}_2\right), && (142)
\end{aligned}
$$

where the equality follows from (133). Further by the Cramér-Rao inequality of Lemma 3,

$$
\begin{aligned}
(\mathbf{K}_X^* + \mathbf{K}_{\widetilde{Z}_1})&\mathbf{J}(\mathbf{X}_\lambda + \widetilde{\mathbf{Z}}_1)(\mathbf{K}_X^* + \mathbf{K}_{\widetilde{Z}_1}) - (\mathbf{K}_X^* + \mathbf{K}_{\widetilde{Z}_1}) \\
&\succeq (\mathbf{K}_X^* + \mathbf{K}_{\widetilde{Z}_1})\operatorname{Cov}^{-1}(\mathbf{X}_\lambda + \widetilde{\mathbf{Z}}_1)(\mathbf{K}_X^* + \mathbf{K}_{\widetilde{Z}_1}) - (\mathbf{K}_X^* + \mathbf{K}_{\widetilde{Z}_1}) && (143)\\
&= (\mathbf{K}_X^* + \mathbf{K}_{\widetilde{Z}_1})\left((1 - \lambda)\operatorname{Cov}(\mathbf{X}) + \lambda\mathbf{K}_X^* + \mathbf{K}_{\widetilde{Z}_1}\right)^{-1}(\mathbf{K}_X^* + \mathbf{K}_{\widetilde{Z}_1}) - (\mathbf{K}_X^* + \mathbf{K}_{\widetilde{Z}_1}) && (144)\\
&\succeq (\mathbf{K}_X^* + \mathbf{K}_{\widetilde{Z}_1})\left((1 - \lambda)\mathbf{S} + \lambda\mathbf{K}_X^* + \mathbf{K}_{\widetilde{Z}_1}\right)^{-1}(\mathbf{K}_X^* + \mathbf{K}_{\widetilde{Z}_1}) - (\mathbf{K}_X^* + \mathbf{K}_{\widetilde{Z}_1}) && (145)\\
&= -(1 - \lambda)(\mathbf{K}_X^* + \mathbf{K}_{\widetilde{Z}_1})\left((1 - \lambda)\mathbf{S} + \lambda\mathbf{K}_X^* + \mathbf{K}_{\widetilde{Z}_1}\right)^{-1}(\mathbf{S} - \mathbf{K}_X^*). && (146)
\end{aligned}
$$

25

Substitute (146) into (142) and recall from the KKT-like condition (71) that $(\mathbf{S}-\mathbf{K}_X^*)\mathbf{M}_2 = 0$. We have

$$\mathrm{Tr}\left(\left((\mathbf{K}_X^* + \mathbf{K}_{\widetilde{Z}_1})\mathbf{J}(\mathbf{X}_\lambda + \widetilde{\mathbf{Z}}_1)(\mathbf{K}_X^* + \mathbf{K}_{\widetilde{Z}_1}) - (\mathbf{K}_X^* + \mathbf{K}_{\widetilde{Z}_1}) \right) \mathbf{M}_2 \right)$$

$$\geq \ -(1-\lambda)\,\mathrm{Tr}\left((\mathbf{K}_X^* + \mathbf{K}_{\widetilde{Z}_1}) \left((1-\lambda)\mathbf{S} + \lambda\mathbf{K}_X^* + \mathbf{K}_{\widetilde{Z}_1} \right)^{-1} (\mathbf{S} - \mathbf{K}_X^*)\mathbf{M}_2 \right) \tag{147}$$

$$= \ 0. \tag{148}$$

We conclude that

$$\overline{g}'(\lambda) \geq 0, \quad \forall\, \lambda \in [0,1], \tag{149}$$

i.e., $\{\mathbf{X}_\lambda\}$ is a monotone increasing path connecting $\mathbf{X}$ and $\mathbf{X}_G^*$. We have found a monotone increasing path for every admissible $\mathbf{X}$, so $\mathbf{X}_G^*$ is an optimal solution of $\overline{P}$. This completes the perturbation proof of Theorem 1.

A few comments on the difference between the direct proof and the perturbation proof of Theorem 1 are now in place. In the direct proof, we enhance both $\mathbf{Z}_1$ and $\mathbf{Z}_2$ to obtain the proportionality so that the classical EPI can be applied to solve the auxiliary optimization problem $\widetilde{P}$. For the perturbation proof, however, we only need to enhance $\mathbf{Z}_1$. (If the lower constraint $\mathbf{K}_X \succeq 0$ does not bite, i.e., $\mathbf{M}_1 = 0$, no enhancement is needed at all.) A direct perturbation is then used to show that $\mathbf{X}_G^*$ is an optimal solution of the auxiliary optimization problem $\overline{P}$. Neither the classical EPI nor the worst noise result of Lemma 2 is needed in the perturbation proof.

D   Proof of Corollary 6

For any random vector $\mathbf{X}$ in $\mathcal{R}^2$ such that $\mathrm{Cov}(\mathbf{X}) \leq \mathbf{S}$ and any $\mu \geq 1$, we have from Theorem 1 that

$$h(\mathbf{X}+\mathbf{Z}_1)-\mu h(\mathbf{X}+\mathbf{Z}_2) \leq \max_{\mathbf{0} \preceq \mathbf{K}_X \preceq \mathbf{S}} \left\{ \frac{1}{2}\log((2\pi e)^2|\mathbf{K}_X + \mathbf{K}_{Z_1}|) - \frac{\mu}{2}\log((2\pi e)^2|\mathbf{K}_X + \mathbf{K}_{Z_2}|) \right\}. \tag{150}$$

Adding a constant term $\mu h(\mathbf{Z}_2) - h(\mathbf{Z}_1)$ to both sides of (150), we obtain

$$I(\mathbf{X};\mathbf{X}+\mathbf{Z}_1) - \mu I(\mathbf{X};\mathbf{X}+\mathbf{Z}_2) \leq \max_{\mathbf{0} \preceq \mathbf{K}_X \preceq \mathbf{S}} \left\{ \frac{1}{2}\log\left|\mathbf{I} + \mathbf{K}_{Z_1}^{-1}\mathbf{K}_X\right| - \frac{\mu}{2}\log\left|\mathbf{I} + \mathbf{K}_{Z_2}^{-1}\mathbf{K}_X\right| \right\}. \tag{151}$$

Let $\mathbf{K}_{Z_i} = \mathbf{V}_i\mathbf{\Sigma}_i\mathbf{V}_i^t$, where $\mathbf{V}_i = (\mathbf{v}_{i1}, \mathbf{v}_{i2})$ is an orthonormal matrix and $\mathbf{\Sigma}_i = \mathrm{Diag}(\lambda_{i1}, \lambda_{i2})$ is a diagonal matrix. Next, we consider taking the limits of both sides of (151) as $\lambda_{12}, \lambda_{21} \to \infty$.

First consider the limit of the left-hand side of (151). We need the following simple lemma.

Lemma 13 *Let $\mathbf{Z} = (Z_1, Z_2)^t$ where Z_1, Z_2 are two independent Gaussian variables with variance σ_1^2 and σ_2^2, respectively. For any random vector $\mathbf{X} = (X_1, X_2)^t$ with finite variances and independent of $\mathbf{Z}$, we have*

$$\lim_{\sigma_2^2 \to \infty} I(\mathbf{X}; \mathbf{X} + \mathbf{Z}) = I(X_1; X_1 + Z_1). \tag{152}$$

Proof. By the chain rule of mutual information,

$$I(\mathbf{X}; \mathbf{X} + \mathbf{Z}) = I(X_1; X_1 + Z_1) + I(X_1; X_2 + Z_2 | X_1 + Z_1) + I(X_2; \mathbf{X} + \mathbf{Z} | X_1). \tag{153}$$

Due to the Markov chains $X_1 + Z_1 \to X_1 \to X_2 + Z_2$ and $X_1 \to X_2 \to X_2 + Z_2$, we have

$$I(X_1; X_2 + Z_2 | X_1 + Z_1) \le I(X_1; X_2 + Z_2) \le I(X_2; X_2 + Z_2). \tag{154}$$

Furthermore, we have

$$
\begin{aligned}
I(X_2; \mathbf{X} + \mathbf{Z} | X_1) &= I(X_2; X_2 + Z_2 | X_1) + I(X_2; X_1 + Z_1 | X_1, X_2 + Z_2) & (155)\\
&= I(X_2; X_2 + Z_2 | X_1) + I(X_2; Z_1 | X_1, X_2 + Z_2) & (156)\\
&= I(X_2; X_2 + Z_2 | X_1) & (157)\\
&\le I(X_2; X_2 + Z_2), & (158)
\end{aligned}
$$

where (157) follows from the fact that Z_1 is independent of Z_2 and $\mathbf{X}$ so $I(X_2; Z_1 | X_1, X_2 + Z_2) = 0$, and (158) is due to the Markov chain $X_1 \to X_2 \to X_2 + Z_2$. Note that

$$\lim_{\sigma_2^2 \to \infty} I(X_2; X_2 + Z_2) \le \lim_{\sigma_2^2 \to \infty} \frac{1}{2} \log \left(1 + \frac{\mathrm{Var}(X_2)}{\sigma_2^2} \right) = 0 \tag{159}$$

with finite $\mathrm{Var}(X_2)$. We thus have from (154) and (158) that both $I(X_1; X_2 + Z_2 | X_1 + Z_1)$ and $I(X_2; \mathbf{X} + \mathbf{Z} | X_1)$ tend to zero in the limit as $\sigma_2^2 \to \infty$. The desired result (152) follows by taking the limit $\sigma_2^2 \to \infty$ on both sides of (153), which completes the proof. $\square$

Let $\overline{\mathbf{Z}}_i = (\overline{Z}_{i,1}, \overline{Z}_{i,2})^t = \mathbf{V}_i^t \mathbf{Z}_i$. Then, $\overline{Z}_{i,1}$ and $\overline{Z}_{i,2}$ are independent. By Lemma 13,

$$\lim_{\lambda_{12} \to \infty} I(\mathbf{X}; \mathbf{X} + \mathbf{Z}_1) = \lim_{\lambda_{12} \to \infty} I(\mathbf{V}_1^t \mathbf{X}; \mathbf{V}_1^t \mathbf{X} + \mathbf{V}_1^t \mathbf{Z}_1) = I(\mathbf{v}_{11}^t \mathbf{X}; \mathbf{v}_{11}^t \mathbf{X} + \overline{Z}_{11}) \tag{160}$$

$$\lim_{\lambda_{21} \to \infty} I(\mathbf{X}; \mathbf{X} + \mathbf{Z}_2) = \lim_{\lambda_{21} \to \infty} I(\mathbf{V}_2^t \mathbf{X}; \mathbf{V}_2^t \mathbf{X} + \mathbf{V}_2^t \mathbf{Z}_1) = I(\mathbf{v}_{22}^t \mathbf{X}; \mathbf{v}_{22}^t \mathbf{X} + \overline{Z}_{22}), \tag{161}$$

which gives

$$\lim_{\lambda_{12}, \lambda_{21} \to \infty} I(\mathbf{X}; \mathbf{X} + \mathbf{Z}_1) - \mu I(\mathbf{X} + \mathbf{Z}_2) = I(\mathbf{v}_{11}^t \mathbf{X}; \mathbf{v}_{11}^t \mathbf{X} + \overline{Z}_{11}) - \mu I(\mathbf{v}_{22}^t \mathbf{X}; \mathbf{v}_{22}^t \mathbf{X} + \overline{Z}_{22}). \tag{162}$$

Next, we consider the limit of the right-hand side of (151). For any semidefinite $\mathbf{K}_X$, we have

$$\lim_{\lambda_{12},\lambda_{21}\to\infty} \left\{ \frac{1}{2}\log\left|\mathbf{I}+\mathbf{K}_{Z_1}^{-1}\mathbf{K}_X\right| - \frac{\mu}{2}\log\left|\mathbf{I}+\mathbf{K}_{Z_2}^{-1}\mathbf{K}_X\right| \right\}$$

$$= \lim_{\lambda_{12},\lambda_{21}\to\infty} \left\{ \frac{1}{2}\log\left|\mathbf{I}+\mathbf{\Sigma}_1^{-1}\mathbf{V}_1^t\mathbf{K}_X\mathbf{V}_1\right| - \frac{\mu}{2}\log\left|\mathbf{I}+\mathbf{\Sigma}_2^{-1}\mathbf{V}_2^t\mathbf{K}_X\mathbf{V}_2\right| \right\} \quad (163)$$

$$= \frac{1}{2}\log\left(1+\lambda_{11}^{-1}\mathbf{v}_{11}^t\mathbf{K}_X\mathbf{v}_{11}\right) - \frac{\mu}{2}\log\left(1+\lambda_{22}^{-1}\mathbf{v}_{22}^t\mathbf{K}_X\mathbf{v}_{22}\right) \quad (164)$$

due to the continuity of $\log|\mathbf{I}+\mathbf{A}|$ over the semidefinite $\mathbf{A}$. Moreover, the convergence of (164) is *uniform* in $\mathbf{K}_X$, because the continuity of $\log|\mathbf{I}+\mathbf{A}|$ over $\mathbf{A}$ is uniform and $\mathbf{V}_i^t\mathbf{K}_X\mathbf{V}_i$, $i=1,2$, are bounded for $0 \preceq \mathbf{K}_X \preceq \mathbf{S}$. we thus have

$$\lim_{\lambda_{12},\lambda_{21}\to\infty} \left\{ \max_{\mathbf{0}\preceq\mathbf{K}_X\preceq\mathbf{S}} \left\{ \frac{1}{2}\log\left|\mathbf{I}+\mathbf{K}_{Z_1}^{-1}\mathbf{K}_X\right| - \frac{\mu}{2}\log\left|\mathbf{I}+\mathbf{K}_{Z_2}^{-1}\mathbf{K}_X\right| \right\} \right\}$$

$$= \max_{\mathbf{0}\preceq\mathbf{K}_X\preceq\mathbf{S}} \left\{ \frac{1}{2}\log\left(1+\lambda_{11}^{-1}\mathbf{v}_{11}^t\mathbf{K}_X\mathbf{v}_{11}\right) - \frac{\mu}{2}\log\left(1+\lambda_{22}^{-1}\mathbf{v}_{22}^t\mathbf{K}_X\mathbf{v}_{22}\right) \right\}. \quad (165)$$

Substituting (162) and (165) into (150), we obtain

$$I(\mathbf{v}_{11}^t\mathbf{X};\mathbf{v}_{11}^t\mathbf{X}+\overline{Z}_{11}) - \mu I(\mathbf{v}_{22}^t\mathbf{X};\mathbf{v}_{22}^t\mathbf{X}+\overline{Z}_{22})$$

$$\leq \max_{\mathbf{0}\preceq\mathbf{K}_X\preceq\mathbf{S}} \left\{ \frac{1}{2}\log\left(1+\lambda_{11}^{-1}\mathbf{v}_{11}^t\mathbf{K}_X\mathbf{v}_{11}\right) - \frac{\mu}{2}\log\left(1+\lambda_{22}^{-1}\mathbf{v}_{22}^t\mathbf{K}_X\mathbf{v}_{22}\right) \right\} \quad (166)$$

and hence

$$h(\mathbf{v}_{11}^t\mathbf{X}+\overline{Z}_{11}) - \mu h(\mathbf{v}_{22}^t\mathbf{X}+\overline{Z}_{22})$$

$$\leq \max_{\mathbf{0}\preceq\mathbf{K}_X\preceq\mathbf{S}} \left\{ \frac{1}{2}\log\left(2\pi e\left(\mathbf{v}_{11}^t\mathbf{K}_X\mathbf{v}_{11}+\lambda_{11}\right)\right) - \frac{\mu}{2}\log\left(2\pi e\left(\mathbf{v}_{22}^t\mathbf{K}_X\mathbf{v}_{22}+\lambda_{22}\right)\right) \right\} \quad (167)$$

for any random vector $\mathbf{X}$ such that $\mathrm{Cov}(\mathbf{X}) \preceq \mathbf{S}$ and any $\mu \geq 1$. This completes the proof.

E Proof of Corollary 7

Let $\mathbf{v}_1 = (1,1)^t$ and $\mathbf{v}_2 = (0,1)^t$. Consider $\{\mathbf{X} : \mathrm{Var}(X_1) \leq a_1\} = \bigcup_{\mathbf{S}} \{\mathbf{X} : \mathrm{Cov}(\mathbf{X}) \preceq \mathbf{S}\}$ where the union is over all $\mathbf{S}$ such that $(\mathbf{S})_{11} = a_1$. By Corollary 6, a Gaussian $(X_1,X_2$ is an optimal solution to the optimization problem

$$\begin{array}{ll} \max_{p(x_1,x_2)} & h(X_1+X_2+Z) - \mu h(X_2+Z) \\ \text{subject to} & \mathrm{Var}(X_1) \leq a_1, \end{array} \quad (168)$$

where $\mu \geq 1$, and the maximization is over all jointly distributed random variables (X_1, X_2) independent of Z. Let (X_{1G}^*, X_{2G}^*) be the Gaussian optimal solution of the (168). Then

$$h(X_1 + X_2 + Z) - \mu h(X_2 + Z) \leq h(X_{1G}^* + X_{2G}^* + Z) - \mu h(X_{2G}^* + Z) \qquad (169)$$

for any jointly distributed random variables (X_1, X_2) such that $\mathrm{Var}(X_1) \leq a_1$.

It is easy to verify that $h(X_{2G}^* + Z)$ is a continuous function of μ. When $\mu = \infty$, $h(X_{2G}^* + Z) = h(Z)$; when $\mu = 1$, $h(X_{2G}^* + Z) = a_2^*$ where a_2^* was defined in (30). By the intermediate value theorem, for any $h(Z) \leq a_2 \leq a_2^*$ there is a μ for which $h(X_{2G}^* + Z) = a_2$. Hence for any jointly distributed random variables (X_1, X_2) such that $\mathrm{Var}(X_1) \leq a_1$ and $h(X_2 + Z) \leq a_2$, we have by (169) that

$$
\begin{aligned}
h(X_1 + X_2 + Z) &\leq h(X_{1G}^* + X_{2G}^* + Z) + \mu\left(h(X_2 + Z) - h(X_{2G}^* + Z)\right) &\qquad (170)\\
&\leq h(X_{1G}^* + X_{2G}^* + Z). &\qquad (171)
\end{aligned}
$$

We conclude that a Gaussian solution is an optimal solution of (29) for any $a_1 \geq 0$ and any $h(Z) \leq a_2 \leq a_2^*$. This completes the proof.

F   Proof of the Outer Bound (50)

Let W_1 and W_2 be the encoded messages for $\{\mathbf{Y}_1[m]\}$ and $\{\mathbf{Y}_2[m]\}$, respectively. Let $\underline{\mathbf{Y}_i^m} := (\mathbf{Y}_i[1], \cdots, \mathbf{Y}_i[m])$ and $U[m] := (W_2, \underline{\mathbf{Y}_1^{m-1}})$. We have

$$
\begin{aligned}
NR_2 &= H(W_2) &\qquad (172)\\
&\geq H(W_2) - H(W_2|\underline{\mathbf{Y}_2^N}) &\qquad (173)\\
&= I(W_2; \underline{\mathbf{Y}_2^N}) &\qquad (174)\\
&= \sum_{m=1}^{N} I(W_2; \mathbf{Y}_2[m] | \underline{\mathbf{Y}_2^{m-1}}) &\qquad (175)\\
&= \sum_{m=1}^{N} \left(h(\mathbf{Y}_2[m]|\underline{\mathbf{Y}_2^{m-1}}) - h(\mathbf{Y}_2[m]|W_2, \underline{\mathbf{Y}_2^{m-1}}) \right) &\qquad (176)\\
&= \sum_{m=1}^{N} \left(h(\mathbf{Y}_2[m]) - h(\mathbf{Y}_2[m]|W_2, \underline{\mathbf{Y}_2^{m-1}}, \underline{\mathbf{Y}_1^{m-1}}) \right) &\qquad (177)\\
&\geq \sum_{m=1}^{N} \left(h(\mathbf{Y}_2[m]) - h(\mathbf{Y}_2[m]|W_2, \underline{\mathbf{Y}_1^{m-1}}) \right) &\qquad (178)\\
&= \sum_{m=1}^{N} I(U[m]; \mathbf{Y}_2[m]), &\qquad (179)
\end{aligned}
$$

29

where (177) follows from the fact that $\underline{\mathbf{Y}_1^{m-1}}$ is a degraded version of $\underline{\mathbf{Y}_2^{m-1}}$ for $m = 1, \cdots, N$. Furthermore,

$$
\begin{align}
NR_1 &= H(W_1) \tag{180} \\
&\geq H(W_1) - H(W_1|W_2, \underline{\mathbf{Y}_1^N}) \tag{181} \\
&= I(W_1; W_2, \underline{\mathbf{Y}_1^N}) \tag{182} \\
&= I(W_1; \underline{\mathbf{Y}_1^N}|W_2) + I(W_1; W_2) \tag{183} \\
&\geq I(W_1; \underline{\mathbf{Y}_1^N}|W_2) \tag{184} \\
&= \sum_{m=1}^{N} I(W_1; \mathbf{Y}_1[m]|W_2, \underline{\mathbf{Y}_1^{m-1}}) \tag{185} \\
&= \sum_{m=1}^{N} I(W_1, \widehat{\mathbf{Y}}_1[m]; \mathbf{Y}_1[m]|W_2, \underline{\mathbf{Y}_1^{m-1}}) \tag{186} \\
&\geq \sum_{m=1}^{N} I(\widehat{\mathbf{Y}}_1[m]; \mathbf{Y}_1[m]|W_2, \underline{\mathbf{Y}_1^{m-1}}) \tag{187} \\
&= \sum_{m=1}^{N} I(\widehat{\mathbf{Y}}_1[m]; \mathbf{Y}_1[m]|U[m]) \tag{188}
\end{align}
$$

where (186) follows from the Markov chain $\mathbf{Y}_1[m] \to (W_1, W_2) \to \widehat{\mathbf{Y}}_1[m]$ for $m = 1, \cdots, N$. Finally, let Q be a random variable uniformly distributed over $\{1, \cdots, N\}$ and independent of any other random variables/vectors. We have from (179) and (188) that

$$
R_1 \geq I(U[Q]; \mathbf{Y}_2[Q]|Q) = I(U[Q], Q; \mathbf{Y}_2[Q]) - I(Q; \mathbf{Y}_2[Q]) = I(U[Q], Q; \mathbf{Y}_2[Q]) = I(U; \mathbf{Y}_2) \tag{189}
$$

and that

$$
R_2 \geq I(\widehat{\mathbf{Y}}_1[Q]; \mathbf{Y}_1[Q]|U[Q], Q) = I(\widehat{\mathbf{Y}}_1; \mathbf{Y}_1|U) \tag{190}
$$

by defining

$$
U := (Q, U[Q]), \quad \widehat{\mathbf{Y}}_1 := \widehat{\mathbf{Y}}_1[Q], \quad \mathbf{Y}_1 := \mathbf{Y}_1[Q], \quad \mathbf{Y}_2 := \mathbf{Y}_2[Q]. \tag{191}
$$

For each $m = 1, \cdots, N$,

$$
\mathbf{Y}_1[m] = \mathbf{Y}_2[m] + \mathbf{Z}[m] \to \mathbf{Y}_2[m] \to U[m] = (W_2, \underline{\mathbf{Y}_1^{m-1}}) \tag{192}
$$

forms a Markov chain because $\mathbf{Z}[m]$ is independent of $(W_2, \mathbf{Y}_1^{m-1})$. Therefore,

$$
\mathbf{Y}_1 \to \mathbf{Y}_2 \to U \tag{193}
$$

also forms a Markov chain. This completes the proof.

Acknowledgment

References

[1] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 623-656, Oct. 1948.

[2] A. J. Stam, "Some inequalities satisfied by the quantities of information of Fisher and Shannon," *Info. Ctrl.*, vol. 2, no. 2, pp. 101-112, Jun. 1959.

[3] N. M. Blachman, "The convolution inequality for entropy powers", *IEEE Trans. Info. Theory*, vol. 11, no, 2, pp. 267-271, Apr. 1965.

[4] P. P. Bergman, "A simple converse for broadcast channels with additive white Gaussian noise," *IEEE Trans. Info. Theory*, vol. 20, no. 2, pp. 279-280, Mar. 1974.

[5] Y. Oohama, "The rate-distortion function for the quadratic Gaussian CEO problem," *IEEE Trans. Info. Theory*, vol. 44, vol. 3, pp. 1057-1070, May 1998.

[6] L. Ozarow, "On a source coding problem with two channels and three receivers," *Bell Syst. Tech. J.*, vol. 59, no. 10, pp. 1909-1921, Dec. 1980.

[7] H. Weingarten, Y. Steinberg, and S. Shamai (Shitz), "The capacity region of the Gaussian MIMO broadcast channel," *IEEE Trans. Info. Theory*, vol. 52, no. 9, pp. 3936-3964, Sept. 2006.

[8] S. Ihara, "On the capacity of channels with additive non-Gaussian noise," *Info. Ctrl.*, vol. 37, no. 1, pp. 34-39, Apr. 1978.

[9] S. N. Diggavi and T. M. Cover, "The worst additive noise under a covariance constraint," *IEEE Trans. Info. Theory*, vol. 47, no. 7, pp. 3072-3081, Nov. 2001.

[10] A. Dembo, T. M. Cover, and J. A. Thomas, "Information theoretic inequalities," *IEEE Trans. Info. Theory*, vol. 37, no. 6, pp. 1501-1518, Nov. 1991.

[11] R. Zamir, "A proof of the Fisher information inequality via a data processing argument," *IEEE Trans. Info. Theory*, vol. 44, no. 3, pp. 1246-1250, May 1998.

[12] D. Guo, S. Shamai (Shitz), and S. Verdú, "Additive non-Gaussian noise channels: mutual information and conditional mean estimation," in *Proc. IEEE Int. Symp. Info. Theory*, Adelaide, Australia, Sept. 2005.

[13] T. M. Cover and Z. Zhang, "On the maximum entropy of the sum of two dependent random variables," *IEEE Trans. Info. Theory*, vol. 40, no. 4, pp. 1244-1246, Jul. 1994.

[14] K. Marton, "A coding theorem for the discrete memoryless broadcast channel," *IEEE Trans. Info. Theory*, vol. 44, vol. 25, pp. 306-311, May 1979.

[15] O. Johnson, *Information Theory and The Central Limit Theorem*, London: Imperial College Press, 2004.

[16] G. Caire and S. Shamai, "On the achievable throughput of a multi-antenna Gaussian broadcast channel," *IEEE Trans. Info. Theory*, vol. 49, no. 7, pp. 1691-1706, Jul. 2003.

[17] W. Yu, A. Sutivong, D. Juian, T. M. Cover, and M. Chiang, "Writing on colored paper," in *Proc. IEEE. Int. Symp. Inform. Theory*, Washington DC, Jun. 2001.

[18] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Info. Theory*, vol. 19, no. 4, pp. 471-480, Jul. 1973.

[19] R. Gallager, "Capacity and coding for degraded broadcast channels," *Prob. Pered. Info.*, vol. 10, no. 3, pp. 3-14, Jul.-Sept. 1974.

[20] H. Weingarten, T. Liu, Y. Steinberg, S. Shamai (Shitz), and P. Viswanath, "The capacity region of the degraded multiple input multiple output broadcast compound channel," in prepration.