Total Correlation of Gaussian Vector Sources on the Gray–Wyner Network

Giel J. Op 't Veld and Michael C. Gastpar
School of Computer and Communication Sciences, EPFL
Lausanne, Switzerland
Email: {giel.optveld, michael.gastpar}@epfl.ch

***Abstract*—We study a generalization of Wyner's Common Information to Watanabe's Total Correlation. The first minimizes the description size required for a variable that can make two other random variables conditionally independent. If independence is unattainable, Watanabe's total (conditional) correlation is measure to check just how independent they have become. Following up on earlier work for scalar Gaussians, we discuss the minimization of total correlation for Gaussian vector sources. Using Gaussian auxiliaries, we show one should transform two vectors of length d into d independent pairs, after which a reverse water filling procedure distributes the minimization over all these pairs. Lastly, we show how this minimization of total conditional correlation fits a lossy coding problem by using the Gray–Wyner network as a model for a caching problem.**

***Index Terms*—Source Coding, Gray–Wyner network, Common Information, Total Correlation, Caching**

I. Introduction

The Gray–Wyner network, depicted in Figure 1, is a coding problem that quite elegantly offers a discussion on the trade-off between jointly and separately coding information [1], [2]. The encoder has access to two sources, X and Y, and is in contact with two decoders, each of which is only interested in one of the two sources. The encoder sends one joint message to both decoders and then two individual messages. Consequently, only on the common branch does the encoder benefit from the correlation between X and Y to reduce communication rates.

This network motivated Wyner to introduce his notion of common information between two random variables as

$$C_W(X,Y) = \inf_{X-V-Y} I(X,Y;V), \tag{1}$$

and he proved that this entity (for discrete sources) equals the minimum rate required on the common branch such that the sum-rate does not exceed the joint entropy of the source [1], [2]. Work in the last five years extended this notion to lossy sources. Viswanatha, Akyol and Rose [3], as well as Xu, Liu and Chen [4], [5] have introduced *lossy* common information:

$$C_W(X,Y,D_x,D_y) = \inf R_0 \in \mathcal{R}_W(D_x,D_y), \tag{2}$$

where $\mathcal{R}_W(D_x,D_y)$ is the set of all common rates R_0 such that $R_0 + R_1 + R_2 \leq R_{X,Y}(D_x,D_y) + \epsilon \quad \forall \epsilon > 0$ and (R_0,R_1,R_2) is an achievable tuple in the rate-distortion region of the Gray–Wyner network. In words, what is the minimum amount of rate required on the common branch such that the sum-rate does not exceed the joint rate-distortion function?

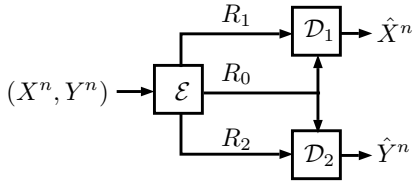

Fig. 1. A Gray–Wyner network for a source of two random variables.

For two scalar Gaussian sources, a full and closed-form characterization of $C_W(X,Y,D_x,D_y)$ was presented in [3].

The interpretation of all these results is that one can distribute the rate required by the joint rate-distortion function over a common and two individual branches if one spends at least enough rate on the common branch so as to make the samples of X and Y conditionally independent. But what if the common rate is limited and one *cannot* make these streams conditionally independent? How does the rate trade-off on the Gray–Wyner network behave then?

In our previous work [6], we stumbled over this question when we explored a caching problem for Gaussian sources for which we used the Gray–Wyner network as a model. The model first appeared for discrete sources in [7], [8]. We viewed upon the common branch as a cache message, and the individual branches as update messages in case the samples of either X or Y were requested. Wyner's common information provided only one point on the boundary of the cache-update rate trade-off. If $R_0 > C_W(X,Y)$, we knew the sum-rate could not exceed the joint rate distortion function. The difficulty was found in the low cache rate regime. It required an auxiliary for the common branch that could make X and Y conditionally "as independent as possible". The latter is measured by Watanabe's total (conditional) correlation [9]:

$$TC(X,Y|V) = h(X|V) + h(Y|V) - h(X,Y|V). \tag{3}$$

Whereas finding the common information is surprisingly a convex problem for Gaussians, minimizing total conditional correlation is not. Therefore we focused on auxiliaries V that are jointly Gaussian with the source. For two Gaussian sources, it turned out one can minimize the total conditional correlation by coding the contribution of X,Y along the dominant eigenvector of their correlation (not covariance) matrix and send that on the common branch [6]. By increasing R_0 one can capture more of this correlation until $R_0 = C_W(X,Y)$

and X, Y become conditionally independent. Surprisingly, for more than two sources the eigenvalue decomposition of the correlation matrix does not seem to be related to this measure of total correlation. What is instead, remains an open problem.

In [10], Satpathy and Cuff found *en passant* a closed-form expression for Wyner's common information of two Gaussian *vectors* by a transformation that turns two vectors $(\mathbf{X}, \mathbf{Y})$ of length d into d independent Gaussian pairs $(\tilde{X}_i, \tilde{Y}_i)$. The common information equals the sum of that of all the pairs:

$$C_W(\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{d} C_W(\tilde{X}_i, \tilde{Y}_i).$$

In this paper, we ask ourselves if this transform of Satpathy and Cuff is also useful for the Gray–Wyner network on points on the rate trade-off other than $R_0 = C_W(\mathbf{X}, \mathbf{Y})$. The answer is affirmative. We first introduce a notion similar to common information to capture this sense of "as independent as possible", i.e., to minimize total conditional correlation:

$$T_{\mathbf{X},\mathbf{Y}}(\gamma) = \inf_{\mathbf{V}} \; I(\mathbf{X}, \mathbf{Y}; \mathbf{V}) \;\text{ s.t. }\; h(\mathbf{X}|\mathbf{V}) + h(\mathbf{Y}|\mathbf{V}) \leq \gamma.$$

Even though the objective function is convex, its constraint is not; this is also part of the reason why the trade-off on the Gray–Wyner network becomes hard when the common branch has a small communication rate. We therefore focus our attention to auxiliaries $\mathbf{V}$ which are jointly Gaussian with $\mathbf{X}, \mathbf{Y}$.

The assumption of Gaussianity does not solve the trade-off being non-convex, but we can show that indeed it suffices to transform the vectors $(\mathbf{X}, \mathbf{Y})$ into d pairs. The problem then splits into two: a convex part to distribute the minimization of total correlation over all the pairs, and a non-convex part of the minimization within a pair. The latter was thus solved by hand before [6]. The most efficient way to minimize total conditional correlation is by minimizing the correlation of each pair separately in a *reverse water filling* fashion. This is the topic of Section III. Afterwards, we fit this building block into our caching coding problem on the Gray–Wyner network and show it fits if the end distortion constraints on $\hat{\mathbf{X}}$ and $\hat{\mathbf{Y}}$ are not too large. The latter will be discussed in Section IV.

II. Preliminaries

Considering notation, capital letters will indicate random variables, e.g. X, while boldface characters are for multivariate vectors, e.g., $\mathbf{X}$. Also matrices will be denoted by boldface letters, while the normal typesetting will refer to an element of the matrix, e.g., $\mathbf{D}$ versus D_{11}. To avoid confusion with the vectors, the end of the alphabet is reserved for random variables.

The bulk of this paper will be about two Gaussian vectors $\mathbf{X}, \mathbf{Y}$ of equal length d and covariance

$$\Sigma = \begin{bmatrix} \Sigma_{\mathbf{X}} & \Sigma_{\mathbf{XY}} \\ \Sigma_{\mathbf{YX}}^T & \Sigma_{\mathbf{Y}} \end{bmatrix}.$$

No subscript indicates the $2d \times 2d$ matrix corresponding to the joint random vector $(\mathbf{X}, \mathbf{Y})$, whereas a subscript is to refer to a corner of that matrix. Straight bars, $|\cdot|$, denote the determinant of a matrix, or the absolute value if the argument is a scalar.

$R_{X,Y}(D_x, D_y)$ denotes the joint rate-distortion function. For two unit-variance scalar Gaussians and symmetric mean squared distortions $D_x = D_y$ this function equals

$$R_{X,Y}(D, D) = \begin{cases} \frac{1}{2}\log\left(\frac{1-\rho^2}{D^2}\right) & \text{if } 0 < D \leq 1-\rho, \\ \frac{1}{2}\log\left(\frac{1+\rho}{2D-1+\rho}\right) & \text{if } 1-\rho \leq D \leq 1. \end{cases}$$

For two *vectors* $\mathbf{X}, \mathbf{Y}$ with joint covariance Σ under trace-constraints one has

$$R_{\mathbf{X},\mathbf{Y}}(D_x, D_y) = \min_{\mathbf{D}} \; \frac{1}{2}\log\frac{|\Sigma|}{|\mathbf{D}|} \quad \text{s.t. } \begin{cases} 0 \preceq \mathbf{D} \preceq \Sigma, \\ \operatorname{tr}(\mathbf{D}_{\mathbf{X}}) \leq D_x, \\ \operatorname{tr}(\mathbf{D}_{\mathbf{Y}}) \leq D_y. \end{cases}$$

The above is a direct consequence from [11]; their proof for the rate-distortion function of scalar Gaussians under individual distortion criteria generalizes to Gaussians with any distortion metric $d(\cdot)$ that respects semidefinite ordering, i.e., $\mathbf{D}_1 \preceq \mathbf{D}_2 \Rightarrow d(\mathbf{D}_1) \leq d(\mathbf{D}_2)$.

A. Vectors to Pairs Decomposition

The key operation of this paper is a set of operations to turn two vectors $(\mathbf{X}, \mathbf{Y})$ into a set of d independent unit-variance pairs $(\tilde{X}_i, \tilde{Y}_i)$. Satpathy and Cuff [10] used these operations as a *transform*, whereas we -for convenience later on- will rather *decompose* $(\mathbf{X}, \mathbf{Y})$. To start, one can pull out the variance:

$$\mathbf{X} = \Sigma_{\mathbf{X}}^{1/2} \bar{\mathbf{X}},$$
$$\mathbf{Y} = \Sigma_{\mathbf{Y}}^{1/2} \bar{\mathbf{Y}}.$$

The random vectors $\bar{\mathbf{X}}, \bar{\mathbf{Y}}$ are unit-variance independent Gaussians, but their cross-correlation has not disappeared:

$$\bar{\Sigma} = \begin{bmatrix} \mathbf{I} & \Sigma_{\mathbf{X}}^{-1/2}\Sigma_{\mathbf{XY}}\Sigma_{\mathbf{Y}}^{-1/2} \\ \Sigma_{\mathbf{X}}^{-1/2}\Sigma_{\mathbf{XY}}\Sigma_{\mathbf{Y}}^{-1/2} & \mathbf{I} \end{bmatrix}.$$

The next step is to note that also this cross-correlation can be diagonalized by a singular value decomposition:

$$\Sigma_{\bar{\mathbf{X}}\bar{\mathbf{Y}}} = \Sigma_{\mathbf{X}}^{-1/2}\Sigma_{\mathbf{XY}}\Sigma_{\mathbf{Y}}^{-1/2} = \mathbf{B}_{\mathbf{X}}\Lambda\mathbf{B}_{\mathbf{Y}},$$

which gives

$$\mathbf{X} = \Sigma_{\mathbf{X}}^{1/2}\mathbf{B}_{\mathbf{X}}\tilde{\mathbf{X}},$$
$$\mathbf{Y} = \Sigma_{\mathbf{Y}}^{1/2}\mathbf{B}_{\mathbf{Y}}\tilde{\mathbf{Y}}.$$

The elements of both $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$ are also independent and of unit-variance, because $\mathbf{B}_{\mathbf{X}}, \mathbf{B}_{\mathbf{Y}}$ are orthonormal matrices. The covariance of $(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}})$ equals

$$\tilde{\Sigma} = \begin{bmatrix} \mathbf{I} & \Lambda \\ \Lambda & \mathbf{I} \end{bmatrix}, \tag{4}$$

and features diagonal matrices in all its four corners. Thus $(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}) = (\tilde{X}_1, \tilde{Y}_1), \cdots, (\tilde{X}_d, \tilde{Y}_d)$; two Gaussian vectors of length d have been decomposed into d independent pairs. Throughout this paper, a tilde over a random variable implies the above decomposition.

Lastly, we attend the reader that even though mutual information is invariant to one-to-one transformations, entropy is not. Therefore:

$$\begin{aligned} h(\mathbf{X},\mathbf{Y}) &= \frac{1}{2}\log(2\pi e)^{2d}|\Sigma| \\ &= \frac{1}{2}\log(2\pi e)^{2d}|\Sigma_{\mathbf{X}}||\Sigma_{\mathbf{Y}}||\tilde{\Sigma}| \\ &= \sum_{i=1}^{d} h(\tilde{X}_i,\tilde{Y}_i) + \frac{1}{2}\log|\Sigma_{\mathbf{X}}||\Sigma_{\mathbf{Y}}|. \end{aligned} \tag{5}$$

B. Gray–Wyner Network

The inspiration of Wyner's common information came from an operational perspective, the Gray–Wyner network of Figure 1 [1]. The encoder observes two sequences $(\mathbf{X}^n,\mathbf{Y}^n)$ drawn in an iid fashion; in our case each sample is a length-d vector. The encoder then maps these sequences to three messages M_0, M_1, M_2 drawn from an alphabet of size $2^{nR_0}, 2^{nR_1}$ and 2^{nR_2}, respectively. Decoder 1 reconstructs a lossy $\hat{\mathbf{X}}^n$ using only (M_0,M_1), whereas decoder 2 tries to do the same for $\hat{\mathbf{Y}}^n$ with (M_0,M_2). A code is defined by an encoder f to map $(\mathbf{X}^n,\mathbf{Y}^n)$ into these three messages, two decoders g_X and g_Y to produce the lossy estimates $(\hat{\mathbf{X}}^n,\hat{\mathbf{Y}}^n)$ and two distortion metrics $d_{\mathbf{X}}(\cdot,\cdot), d_{\mathbf{Y}}(\cdot,\cdot)$.

A tuple (R_0,R_1,R_2,D_x,D_y) is said to be achievable if there exist such encoders and decoders and if furthermore

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} d_{\mathbf{X}}(\mathbf{X}_i,\hat{\mathbf{X}}_i) &\le D_x, \\ \frac{1}{n}\sum_{i=1}^{n} d_{\mathbf{Y}}(\mathbf{Y}_i,\hat{\mathbf{Y}}_i) &\le D_y. \end{aligned}$$

The region of achievable rate-distortion tuples on the Gray–Wyner network is the union of all (R_0,R_1,R_2,D_x,D_y) satisfying

$$\begin{aligned} R_0 &\ge I(\mathbf{X},\mathbf{Y};\mathbf{V}) \\ R_1 &\ge I(\mathbf{X};\hat{\mathbf{X}}|\mathbf{V}) \\ R_2 &\ge I(\mathbf{Y};\hat{\mathbf{Y}}|\mathbf{V}) \\ D_x &\ge \mathbb{E}[d_{\mathbf{X}}(\mathbf{X},\hat{\mathbf{X}})] \\ D_y &\ge \mathbb{E}[d_{\mathbf{Y}}(\mathbf{Y},\hat{\mathbf{Y}})], \end{aligned}$$

over joint densities $p(\mathbf{X},\mathbf{Y},\mathbf{V},\hat{\mathbf{X}},\hat{\mathbf{Y}})$ [5]. Working with Gaussian vector sources, let us choose for the distortion metric a trace-constraint, i.e.,

$$d_{\mathbf{X}}(\mathbf{X},\hat{\mathbf{X}}) = \operatorname{tr}\left(\mathbb{E}[(\mathbf{X}-\hat{\mathbf{X}})(\mathbf{X}-\hat{\mathbf{X}})^T]\right) \le D_x,$$

the same for $\mathbf{Y}$.

III. Common Information and Total Correlation

We begin by a *lossless* discussion on independence and total correlation.

A. Common Information

First, we would like to add one comment to the work of Satpathy and Cuff on the common information for Gaussian vectors [10]. For scalars, a closed-form solution of $C_W(X,Y)$ is known, but there is no analytic expression of how to make *three* or more random variables conditionally independent. The problem is, however, surprisingly convex and can be solved efficiently by linear programming numerically [6]. For Gaussian *vectors*, it turns out to be the same.

For two scalar jointly Gaussian random variables, the common information equals

$$C_W(X,Y) = \min_{X-V-Y} I(X,Y;V) = \frac{1}{2}\log\frac{1+|\rho|}{1-|\rho|},$$

where ρ is the correlation between X and Y [4], [5]. For two Gaussian vectors, it is the following:

Lemma 1 (Satpathy and Cuff [10])**.** *For jointly Gaussian $\mathbf{X},\mathbf{Y}\in\mathbb{R}^d$, Wyner's common information is given by*

$$C_W(\mathbf{X},\mathbf{Y}) = \min_{\mathbf{X}-\mathbf{V}-\mathbf{Y}} I(\mathbf{X},\mathbf{Y};\mathbf{V}) = \frac{1}{2}\sum_{i=1}^{d}\log\frac{1+|\rho_i|}{1-|\rho_i|},$$

where $\{\rho_i\}$ are the singular values of $\Sigma_{\mathbf{X}}^{-1/2}\Sigma_{\mathbf{XY}}\Sigma_{\mathbf{Y}}^{-1/2}$.

In short, the vector common information equals the sum of the common information of all $(\tilde{X}_i,\tilde{Y}_i)$-pairs obtained by the vector-to-pairs decomposition described in Section II-A. Unfortunately, such a transformation becomes harder to develop when there are more than two vectors to make independent. However, at least numerically finding the common information is not a hard problem.

To that end, define:

$$\begin{aligned} &C_W(\mathbf{X}^{(1)},\cdots,\mathbf{X}^{(M)}) \stackrel{\text{def}}{=} \inf_{\mathbf{V}}\ I(\mathbf{X}^{(1)},\cdots,\mathbf{X}^{(M)};\mathbf{V}), \\ &\text{such that } h(\mathbf{X}^{(1)},\cdots,\mathbf{X}^{(M)}|\mathbf{V}) = \sum_{i=1}^{M} h(\mathbf{X}^{(i)}|\mathbf{V}). \end{aligned}$$

Theorem 1. *For jointly Gaussian $\mathbf{X}^{(1)},\cdots,\mathbf{X}^{(M)}\in\mathbb{R}^d$ with joint covariance Σ, the common information is given by*

$$C_W(\mathbf{X}^{(1)},\cdots,\mathbf{X}^{(M)}) = \min_{\mathbf{K}} \frac{1}{2}\log\frac{|\Sigma|}{|\mathbf{K}|} \tag{6}$$

$$\textit{such that } \begin{cases} 0 \preceq \mathbf{K} \preceq \Sigma, \\ \mathbf{K} \textit{ is block-diagonal.} \end{cases}$$

Proof. Consider any $\mathbf{V}$ that is jointly distributed with $\mathbf{X}^{(1)},\cdots,\mathbf{X}^{(M)}$ and that makes all $\mathbf{X}^{(i)}$ conditionally inde-

pendent. Then,

$$\begin{aligned}
&I(\mathbf{X}^{(1)},\cdots,\mathbf{X}^{(M)};\mathbf{V})\\
&\quad= h(\mathbf{X}^{(1)},\cdots,\mathbf{X}^{(M)}) - h(\mathbf{X}^{(1)},\cdots,\mathbf{X}^{(M)}|\mathbf{V})\\
&\quad= h(\mathbf{X}^{(1)},\cdots,\mathbf{X}^{(M)}) - \sum_{i=1}^{M} h(\mathbf{X}^{(i)}|\mathbf{V})\\
&\quad= \sum_{i=1}^{M}\Big(h(\mathbf{X}^{(i)}) - h(\mathbf{X}^{(i)}|\mathbf{V})\Big) +\\
&\qquad\quad h(\mathbf{X}^{(1)},\cdots,\mathbf{X}^{(M)}) - \sum_{i=1}^{M} h(\mathbf{X}^{(i)})\\
&\quad\geq \frac{1}{2}\log\frac{\prod_{i=1}^{M}|\Sigma_{\mathbf{X}^{(i)}}|}{\prod_{i=1}^{M}|\mathbf{K}^{(i)}|} + \frac{1}{2}\log\frac{|\Sigma|}{\prod_{i=1}^{M}|\Sigma_{\mathbf{X}^{(i)}}|}\\
&\quad= \frac{1}{2}\log\frac{|\Sigma|}{\prod_{i=1}^{M}|\mathbf{K}^{(i)}|},
\end{aligned}$$

where $\mathbf{K}^{(i)} = \mathbb{E}[(\mathbf{X}^{(i)} - \mathbb{E}[\mathbf{X}^{(i)}|\mathbf{V}])(\mathbf{X}^{(i)} - \mathbb{E}[\mathbf{X}^{(i)}|\mathbf{V}])^T]$, is the MMSE matrix of $\mathbf{X}^{(i)}$ based on $\mathbf{V}$. Let $\mathbf{K}$ be the block-diagonal matrix formed like $\mathbf{K} = \operatorname{diag}\left(\left\{\mathbf{K}^{(i)}\right\}_{i=1}^{M}\right)$. The one inequality follows from Lemma 2 from [11] and is -in fact- just the Gaussian rate-distortion function. The lower bound is met with equality for any jointly Gaussian $\mathbf{V}$ yielding a block-diagonal distortion matrix $\mathbf{K}$ that satisfies $\mathbf{K} \preceq \Sigma_{\mathbf{X}}$. The block-diagonal structure emphasizes that zero correlation between any $(\mathbf{X}^{(i)}, \mathbf{X}^{(j)})$-pair is necessary for conditional independence, but for Gaussians also happens to be sufficient. □

The Theorem in itself is not so much a revelation as is the insight that the problem of common information is strictly convex and constrained by only linear constraints; it is a so-called MaxDet problem [12]. Consequently, as we argued in [6] for scalar Gaussians, the common information and the distortion matrix that attains it can be found efficiently by linear programming.

Using the popular CVX package for Matlab [13], one can easily see for oneself that for $M = 2$ the above optimization problem leads to the analytically found result by Satpathy and Cuff. For $M > 2$ we know no analytic expression for the optimal distortion matrix $\mathbf{K}$, though numerically the problem remains tractable.

B. Total Conditional Correlation

We concentrate ourselves again on only two Gaussian vectors and on the problem we started off on: If one cannot make $\mathbf{X}$ and $\mathbf{Y}$ conditionally independent, how can they be made "as independent as possible"? This problem is, unlike finding Gaussian common information, non-convex. The hypothesis is that one should still apply the vectors-to-pair decomposition (Section II-A) and treat the vectors $(\mathbf{X}, \mathbf{Y})$ as d pairs $(\tilde{X}_i, \tilde{Y}_i)$. This claim turns out to be true if one restricts his attention to jointly Gaussian auxiliaries. Moreover, Theorem 2 will show that a reversed water filling procedure minimizes the correlation of these $(\tilde{X}_i, \tilde{Y}_i)$ in the most efficient and distributed manner.

First, let us define this sense of "as independent as possible", which is motivated by Watanabe's notion of total (conditional) correlation [9]:

$$TC(\mathbf{X},\mathbf{Y}|\mathbf{V}) = h(\mathbf{X}|\mathbf{V}) + h(\mathbf{Y}|\mathbf{V}) - h(\mathbf{X},\mathbf{Y}|\mathbf{V}).$$

Regarding the Gray–Wyner network (Figure 1), this gap between the sum of marginal entropies and the joint entropy characterizes the loss incurred by first sending an (insufficient) common message followed by two individual updates. On the individual branches, the encoder does not benefit from jointly coding $(\mathbf{X}, \mathbf{Y})$. Hence, any correlation left unused by the common branch is essentially a wasted opportunity to reduce communication rates. At best, $TC(\mathbf{X},\mathbf{Y}|\mathbf{V}) = 0$, meaning that (at least) all the common information was captured by the auxiliary $\mathbf{V}$.

With the Gray–Wyner network in mind, the goal is to find a formulation similar to that of common information (1): A minimization over a cost-function to make two random variables not completely independent, but to bring down their total correlation to an acceptable level. To that end, let us rewrite the notion of total correlation into a minimization:

$$T_{\mathbf{X},\mathbf{Y}}(\gamma) \stackrel{\text{def}}{=} \inf_{\mathbf{V}} \; I(\mathbf{X},\mathbf{Y};\mathbf{V}) \;\text{ s.t. }\; h(\mathbf{X}|\mathbf{V}) + h(\mathbf{Y}|\mathbf{V}) \leq \gamma.$$

$T_{\mathbf{X},\mathbf{Y}}(\gamma)$ is convex and non-increasing in γ, however, the constraint-function is concave in $\mathbf{V}$. So far, it is still unclear whether $\mathbf{X}, \mathbf{Y}$ being Gaussian implies it suffices to also take $\mathbf{V}$ Gaussian. We shall therefore take Gaussianity as an assumption, and show an attainable smooth bound on the problem.

Let $\mathbf{K}$ be a conditional covariance matrix associated to the distribution $p(\mathbf{X},\mathbf{Y}|\mathbf{V})$ and let $\mathbf{K}_{\mathbf{X}}$, $\mathbf{K}_{\mathbf{Y}}$ be the top-right and bottom-left corner of that matrix. Then:

$$h(\mathbf{X}|\mathbf{V}) + h(\mathbf{Y}|\mathbf{V}) = \frac{1}{2}\log(2\pi e)^{2d}|\mathbf{K}_{\mathbf{X}}||\mathbf{K}_{\mathbf{Y}}|.$$

For convenience, one can redefine the Gaussian $T_{\mathbf{X},\mathbf{Y}}(\gamma)$ to not worry about the constants, and focus on this conditional covariance $\mathbf{K}$:

$$T_{\mathbf{X},\mathbf{Y}}(\gamma) = \min_{\mathbf{K}} \frac{1}{2}\log\frac{|\Sigma|}{|\mathbf{K}|} \quad \text{s.t.} \quad \begin{cases} 0 \preceq \mathbf{K} \preceq \Sigma, \\ |\mathbf{K}_{\mathbf{X}}||\mathbf{K}_{\mathbf{Y}}| \leq \gamma, \end{cases} \tag{7}$$

for which now $\gamma \in [0, |\Sigma_{\mathbf{X}}| \cdot |\Sigma_{\mathbf{Y}}|]$.

Since $I(\mathbf{X},\mathbf{Y};\mathbf{V}) = h(\mathbf{X},\mathbf{Y}) - h(\mathbf{X},\mathbf{Y}|\mathbf{V})$, the minimization of $T_{\mathbf{X},\mathbf{Y}}(\gamma)$ is actually a maximization of the joint conditional entropy. This objective and the constraint are bounds to each other,

$$h(\mathbf{X},\mathbf{Y}|\mathbf{V}) \leq h(\mathbf{X}|\mathbf{V}) + h(\mathbf{Y}|\mathbf{V}), \tag{8}$$

and $T_{\mathbf{X},\mathbf{Y}}(\gamma)$ tries to close this inequality by maximizing the left-hand side, while bounding the right. For Gaussians, the same bound is expressed by the Hadamard inequality for block matrices:

$$|\mathbf{K}| \leq |\mathbf{K}_{\mathbf{X}}||\mathbf{K}_{\mathbf{Y}}|. \tag{9}$$

At best, the inequality is met with equality, which happens if and only if $\mathbf{X}$ and $\mathbf{Y}$ become conditionally independent. Consequently, there is a close relationship between our objective and Wyner's common information:

Lemma 2. *For jointly Gaussian $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^d$,*

$$T_{\mathbf{X},\mathbf{Y}}(\gamma') = C_W(\mathbf{X}, \mathbf{Y}),$$

for $\gamma' = |\Sigma_{\mathbf{X}}||\Sigma_{\mathbf{Y}}| \prod_{i=1}^{d}(1-|\rho_i|)^2$ and where $\{\rho_i\}$ are the singular values of $\Sigma_{\mathbf{X}}^{-1/2}\Sigma_{\mathbf{XY}}\Sigma_{\mathbf{Y}}^{-1/2}$.

Proof. Conditional independence is equivalent to the condition $h(\mathbf{X}|\mathbf{V})+h(\mathbf{Y}|\mathbf{V}) = h(\mathbf{X},\mathbf{Y}|\mathbf{V})$. For Gaussian distributions, this equality means that the covariance matrix associated to $p(\mathbf{X},\mathbf{Y}|\mathbf{V})$ satisfies $|\mathbf{K}| = |\mathbf{K}_{\mathbf{X}}||\mathbf{K}_{\mathbf{Y}}|$. For minimizing total conditional correlation, this equality is the best one can achieve, as can be seen in (7). Filling equality in (9) into (7) gives:

$$T_{\mathbf{X},\mathbf{Y}}(\gamma) = \frac{1}{2}\log\frac{|\Sigma|}{\gamma}.$$

The $\mathbf{V}$ that achieves common information corresponds to a matrix $\mathbf{K}$ that is diagonal after the transformation of Section II-A, and is of the form $\tilde{\mathbf{K}}_{\mathbf{X}} = \tilde{\mathbf{K}}_{\mathbf{Y}} = \mathrm{diag}(\{1-|\rho_i|\}_{i=1}^{d})$. So the $\mathbf{V}$ that achieves $C_W(\mathbf{X},\mathbf{Y})$ yields:

$$\begin{aligned} h(\mathbf{X}|\mathbf{V}) + h(\mathbf{Y}|\mathbf{V}) &= \frac{1}{2}\log(2\pi e)^{2d}|\mathbf{K}_{\mathbf{X}}||\mathbf{K}_{\mathbf{Y}}| \\ &= \frac{1}{2}\log(2\pi e)^{2d}|\Sigma_{\mathbf{X}}||\Sigma_{\mathbf{Y}}|\prod_{i=1}^{d}(1-|\rho_i|)^2. \end{aligned}$$

Hence, choosing γ equal to the argument of the expression above gives $T_{\mathbf{X},\mathbf{Y}}(\gamma) = C_W(\mathbf{X},\mathbf{Y})$. □

For $\gamma < |\Sigma_{\mathbf{X}}||\Sigma_{\mathbf{Y}}|\prod_{i=1}^{d}(1-|\rho_i|)^2$ there is still equality in both (9) and (8), which implies that the choice of picking $\mathbf{V}$ jointly Gaussian with $\mathbf{X},\mathbf{Y}$ is not just an assumption, it is optimal in general. Note, however, that for small γ one does have that $T_{\mathbf{X},\mathbf{Y}}(\gamma) > C_W(\mathbf{X},\mathbf{Y})$. Another implication is that this regime of small γ and conditional independence is not as challenging as large γ, where the total conditional correlation remains strictly positive. As shown later, minimizing the total conditional correlation between two Gaussian vectors will require one to decompose those vectors into d pairs and then minimize the correlation of those pairs separately, in a cleverly distributed way. To that end, one must understand how to most effectively minimize total correlation between two scalar Gaussians, for which we cite an older result from us:

Lemma 3 (Op 't Veld and Gastpar [6])**.** *For Gaussian X, Y of unit variance and correlation ρ we have*

$$T_{X,Y}(\gamma) = R_{X,Y}\left(\sqrt{\gamma},\sqrt{\gamma}\right).$$

Relying on a Gaussian auxiliary V, we showed in [6] that for two scalars $T_{X,Y}(\gamma)$ (7) is solved by choosing a $\mathbf{K}$ that is a rank-1 correction along the dominant eigenvector of the *correlation* matrix until X and Y become conditionally independent. This dominant eigenvector is either $\frac{1}{\sqrt{2}}[1\ 1]^T$ if $\rho > 0$ or $\frac{1}{\sqrt{2}}[1\ \ -1]^T$ if $\rho < 0$. Either how, consequently $K_X = K_Y$ (though we note this choice is no longer unique if $T_{\mathbf{X},\mathbf{Y}}(\gamma) > C_W(X,Y)$).[1]

This brings us to our main result: The total conditional correlation of two Gaussian vectors is minimized by a reverse water filling procedure on the common information of each pair found by the decomposition of Section II-A:

Theorem 2. *For jointly Gaussian $\mathbf{X},\mathbf{Y} \in \mathbb{R}^d$, the total conditional correlation is minimized by a reverse water filling procedure until $\mathbf{X},\mathbf{Y}$ become conditionally independent, i.e.*

$$T_{\mathbf{X},\mathbf{Y}}(\gamma) = \begin{cases} \frac{1}{2}\log\frac{|\Sigma|}{\gamma} & T_{\mathbf{X},\mathbf{Y}}(\gamma) \geq C_W(\mathbf{X},\mathbf{Y}) \\ \sum_{i=1}^{d} R_i & T_{\mathbf{X},\mathbf{Y}}(\gamma) \leq C_W(\mathbf{X},\mathbf{Y}) \end{cases}$$

where

$$R_i = \max(C_W(\tilde{X}_i,\tilde{Y}_i) - \theta, 0),$$

and θ is a positive constant chosen such that

$$|\Sigma_{\mathbf{X}}||\Sigma_{\mathbf{Y}}|\prod_{i-1}^{d}\left(2^{-2R_i}(1+|\rho_i|)+(1-|\rho_i|)\right)^2 = \gamma.$$

Proof. First, for γ small such that $T_{\mathbf{X},\mathbf{Y}}(\gamma) \geq C_W(\mathbf{X},\mathbf{Y})$, conditional independence and hence equality in (9) (and, as a matter of fact, also (8)) is attainable, see Lemma 2. Hence,

$$\begin{aligned} \frac{1}{2}\log\frac{|\Sigma|}{|\mathbf{K}|} &\geq \frac{1}{2}\log\frac{|\Sigma|}{|\mathbf{K}_{\mathbf{X}}||\mathbf{K}_{\mathbf{Y}}|} \\ &\geq \frac{1}{2}\log\frac{|\Sigma|}{\gamma}, \end{aligned}$$

can be met with equality.

For large γ (and thus small $T_{\mathbf{X},\mathbf{Y}}(\gamma)$), conditional independence is not attainable. In principle, the problem is this optimization:

$$\begin{aligned} \max_{\mathbf{K}\preceq\Sigma} \quad & |\mathbf{K}|, \\ \text{s.t.} \quad & |\mathbf{K}_{\mathbf{X}}||\mathbf{K}_{\mathbf{Y}}| \leq \gamma. \end{aligned}$$

Without loss of generality, one can do a change of variable by applying the vectors-to-pairs decomposition of Section II-A to the source and the *same* transformation to the variable $\mathbf{K}$. Since this decomposition scales out the variance, which are not orthonormal matrices, the objective and constraint are affected:

$$\begin{aligned} \max_{\tilde{\mathbf{K}}\preceq\tilde{\Sigma}} \quad & |\Sigma_{\mathbf{X}}||\Sigma_{\mathbf{Y}}||\tilde{\mathbf{K}}|, \\ \text{s.t.} \quad & |\Sigma_{\mathbf{X}}||\Sigma_{\mathbf{Y}}||\tilde{\mathbf{K}}_{\mathbf{X}}||\tilde{\mathbf{K}}_{\mathbf{Y}}| \leq \gamma. \end{aligned}$$

Both are, however, affected *equally* and we can restrict our attention to finding a suitable $\tilde{\mathbf{K}} \preceq \tilde{\Sigma}$.

Note the very special structure of $\tilde{\Sigma}$ (4): there only exists correlation between $(\tilde{X}_i,\tilde{Y}_i)$-pairs. Our hypothesis is that the optimal $\tilde{\mathbf{K}}$ has the same eigenbasis that generates a

[1] For non-unit variance scalar Gaussians, the optimal $\mathbf{K}$ is a rank-1 correction along a scaled version of the dominant eigenvector of the correlation matrix, e.g., $1/\sqrt{\mathrm{tr}(\Sigma)}[\sigma_1\ \sigma_2]^T$ instead of $1/\sqrt{2}[1\ 1]^T$.

block-matrix with diagional matrices in all its four corners. If $\tilde{\mathbf{K}}$ would introduce correlation between an $(\tilde{X}_i, \tilde{Y}_j)_{i\neq j}$-pair, then $|\tilde{\mathbf{K}}|$ would decrease, while $|\tilde{\mathbf{K}}_{\mathbf{X}}||\tilde{\mathbf{K}}_{\mathbf{Y}}|$ observes no improvement. If $\tilde{\mathbf{K}}$ would introduce correlation between a $(\tilde{X}_i, \tilde{X}_j)_{i\neq j}$-pair (or $\tilde{Y}$ respectively), the gain on the constraint can never exceed the drop on the objective, because of the Hadamard inequality $|\mathbf{K}| \leq |\tilde{\mathbf{K}}_{\mathbf{X}}||\tilde{\mathbf{K}}_{\mathbf{Y}}|$. Hence, there is no added value in a $\tilde{\mathbf{K}}$ with a different eigenbasis than $\tilde{\Sigma}$.

Hence, one arrives at the modular approach of looking for a 2×2 distortion matrix $\tilde{\mathbf{K}}^{(i)}$ for each $(\tilde{X}_i, \tilde{Y}_i)$-pair, i.e.,

$$\tilde{\mathbf{K}}^{(i)} \preceq \tilde{\Sigma}^{(i)} = \begin{bmatrix} 1 & \rho_i \\ \rho_i & 1 \end{bmatrix},$$

where again $\{\rho_i\}$ are the singular values of $\Sigma_{\mathbf{X}}^{-1/2}\Sigma_{\mathbf{XY}}\Sigma_{\mathbf{Y}}^{-1/2}$. For a pair of Gaussians, Lemma 3 states that $\tilde{\mathbf{K}}^{(i)}$ should be a rank-one correction along the dominant eigenvector of the correlation matrix of $(\tilde{X}_i, \tilde{Y}_i)$. That leaves the question of which pairs $(\tilde{X}_i, \tilde{Y}_i)$ have the biggest impact on minimizing the total conditional correlation of the vectors $\mathbf{X}$ and $\mathbf{Y}$. This problem turns out to be convex, but it requires the proper variable to expose so. To that end, let

$$R_i \stackrel{\text{def}}{=} I(\tilde{X}_i, \tilde{Y}_i; V_i) = \frac{1}{2}\log\frac{1-\rho_i^2}{|\tilde{\mathbf{K}}^{(i)}|}.$$

Then if $\tilde{\mathbf{K}}^{(i)}$ is indeed only an update along the dominant eigenvector, then equivalently:

$$\begin{aligned} |\tilde{\mathbf{K}}_{\mathbf{X}}||\tilde{\mathbf{K}}_{\mathbf{Y}}| &= \prod_{i=1}^{d}(\tilde{K}_{X_i}^{(i)}\tilde{K}_{Y_i}^{(i)}) \\ &= \prod_{i=1}^{d}\left(2^{-2R_i}(1+|\rho_i|) + (1-|\rho_i|)\right)^2. \end{aligned}$$

Applying logarithms, also the constraint-function is convex and one can use Lagrangian multipliers to construct:

$$J = \sum_{i=1}^{d} R_i + \lambda \sum_{i=1}^{d} \frac{1}{2}\log\left(\left(2^{-2R_i}(1+|\rho_i|) + (1-|\rho_i|)\right)^2\right),$$

giving

$$\frac{\partial J}{\partial R_i} = 1 - 2\lambda\frac{(1+|\rho_i|)2^{-2R_i}}{2^{-2R_i}(1+|\rho_i|) + (1-|\rho_i|)} = 0.$$

Rewriting the above expression leads to

$$\begin{aligned} R_i &= \frac{1}{2}\log\frac{1+|\rho_i|}{1-|\rho_i|} - \frac{1}{2}\log(2\lambda - 1) \\ &= C_W(\tilde{X}_i, \tilde{Y}_i) - \theta. \end{aligned}$$

Each R_i is ideally the common information of its $(\tilde{X}_i, \tilde{Y}_i)$-pair minus a constant θ. However, R_i must be non-negative. Incorporating also this extra constraint leads to the reverse water filling procedure as stated in the Theorem. □

The rate–distortion function of a Gaussian vector $\mathbf{X}$ subject to a trace distortion constraint, i.e. $\operatorname{tr}\left(\mathbb{E}[(\mathbf{X}-\hat{\mathbf{X}})(\mathbf{X}-\hat{\mathbf{X}})^T]\right) \leq D$, is a classic result that also admits a reverse water filling procedure [14, Theorem 10.3.3]. We attend the reader to a subtle difference: the Gaussian vector rate-distortion function applies reverse water filling to the *eigenvalues* of the covariance matrix Σ, whereas the minimization of total conditional correlation uses $R_i = I(\tilde{X}_i, \tilde{Y}_i; V_i)$ as the variable. Consequently, one will not observe similar thresholding behavior by plotting the evolution of the eigenvalues. The right way to plot the water filling of total conditional correlation is by plotting the common information of each $(\tilde{X}_i, \tilde{Y}_i)$-pair as a bar graph. An example is shown in Figure 2.

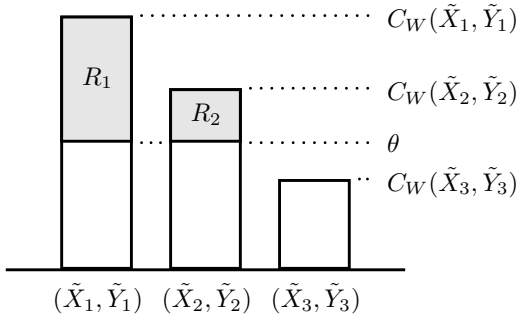

Fig. 2. Example of the reverse water-filling procedure of Theorem 2. The bars represent the common information of each $(\tilde{X}_i, \tilde{Y}_i)$-pair and the shaded area equals R_i. In this example, $d = 3$ and θ is such that $R_3 = 0$.

IV. Caching: an application of Total Correlation on the Gray–Wyner Network

This section is to serve as an example of how the essentially lossless definition of $T_{\mathbf{X},\mathbf{Y}}(\gamma)$ applies to a lossy coding problem on the Gray–Wyner network. Viswanatha *et al.* proved in [3] that their lossy interpretation of common information (2) matched Wyner's lossless one (1) if and only if the end distortion constraints on $\hat{X}$ and $\hat{Y}$ were sufficiently small, i.e.:

$$C_W(X, Y, D_x, D_y) = C_W(X, Y) \text{ if } \max(D_x, D_y) \leq 1 - |\rho|.$$

For larger distortion constraints, the lossy common information is different and in some regimes even equals the entire rate-distortion function.

In the introduction we stated our motivation came from a caching model we studied earlier, based on the Gray–Wyner network [6]: Imagine a user is interested in the samples produced by either $\mathbf{X}$ or $\mathbf{Y}$. Before revealing her preference, the encoder sends a cache message. The user then announces her request after which the encoder sends an update tailored to either $\mathbf{X}$ or $\mathbf{Y}$ to complement the cache message such that both messages together provide the user with an acceptable lossy representation of the samples requested. The cache message can be viewed upon as the common message in the Gray–Wyner network, whereas the individual branches stand for the events of the user asking for either $\mathbf{X}$ or $\mathbf{Y}$.

If the user makes her choice uniformly at random, the statistics of her choice cannot be leveraged to reduce communication rates. The update rate needed on average would simply be the sum-rate of the individual branches in the Gray–

Wyner network divided by two. So referring to Section II-B, one finds:

$$R_{\text{cache}} \geq I(\mathbf{X}, \mathbf{Y}; \mathbf{V}),$$
$$\overline{R}_{\text{update}} \geq \frac{1}{2}\Big(I(\mathbf{X}, \hat{\mathbf{X}}|\mathbf{V}) + I(\mathbf{Y}, \hat{\mathbf{Y}}|\mathbf{V})\Big).$$

Applying Gaussian distributions to these equations, the *Gaussian* achievable cache-rate-distortion region is the union of $(R_{\text{cache}}, \overline{R}_{\text{update}}, D_x, D_y)$ satisfying

$$R_{\text{cache}} \geq \frac{1}{2}\log\frac{|\Sigma|}{|\mathbf{K}|} \tag{10}$$
$$\overline{R}_{\text{update}} \geq \frac{1}{4}\log\frac{|\mathbf{K_X}||\mathbf{K_Y}|}{|\mathbf{D_X}||\mathbf{D_Y}|} \tag{11}$$
$$D_x \geq \text{tr}(\mathbf{D_X})$$
$$D_y \geq \text{tr}(\mathbf{D_Y}),$$

over positive semidefinite matrices $\mathbf{D_X}, \mathbf{D_Y} \in \mathbb{R}^{d\times d}$ and $\mathbf{K} \in \mathbb{R}^{2d\times 2d}$ satisfying $\mathbf{D_X} \preceq \mathbf{K_X}, \mathbf{D_Y} \preceq \mathbf{K_Y}$, and $\mathbf{K} \preceq \Sigma$.

One can observe the same trade-off as in the lossless discussion of Section III: $|\mathbf{K}| \leftrightarrow |\mathbf{K_X}||\mathbf{K_Y}|$. However, if the end distortion constraints become too large, the $\mathbf{K}$ that optimizes the trade-off between R_{cache} and $\overline{R}_{\text{update}}$ necessarily depends on D_x, D_y. Hence, like for common information, the distortion constraints cannot be large for the minimization of total conditional correlation $T_{\mathbf{X},\mathbf{Y}}(\gamma)$ to be directly applicable to a coding problem on the Gray–Wyner network.

Corollary 1. *Let $\mathbf{K}^{C_W(\mathbf{X},\mathbf{Y})}$ be the distortion matrix that attains the common information $C_W(\mathbf{X}, \mathbf{Y})$. Then, the Gaussian trade-off between R_{cache} (10) and $\overline{R}_{update}$ (11) can be controlled by a parameter γ such that*

$$R_{cache} \geq T_{\mathbf{X},\mathbf{Y}}(\gamma),$$
$$\overline{R}_{update} \geq \frac{1}{4}\log\frac{\gamma}{|\mathbf{D_X}||\mathbf{D_Y}|},$$

for the regime of end distortion constraints satisfying:

$$D_x \leq d \cdot \lambda_{min}(\mathbf{K}_{\mathbf{X}}^{C_W(\mathbf{X},\mathbf{Y})}),$$
$$D_y \leq d \cdot \lambda_{min}(\mathbf{K}_{\mathbf{Y}}^{C_W(\mathbf{X},\mathbf{Y})}).$$

Proof. The Gaussian rate region (10)-(11) features the same trade-off as $T_{\mathbf{X},\mathbf{Y}}(\gamma)$, i.e., $|\mathbf{K}| \leftrightarrow |\mathbf{K_X}||\mathbf{K_Y}|$. The rate-distortion theorem for Gaussian multivariates under a trace-constraint dictates the update phase is most efficiently coded via a reverse water filling procedure on the eigenvalues of $\mathbf{K_X}$ and $\mathbf{K_Y}$ [14, Theorem 10.3.3]. Hence, for large D_x, D_y the optimal choice of a $\mathbf{K}$ not only depends on the determinants, but also the specific spectra of the submatrices in its top-left and bottom-right corner, $\mathbf{K_X}$ and $\mathbf{K_Y}$. If $D_x \leq d \cdot \lambda_{\min}(\mathbf{K_X})$, the distortion matrix $\mathbf{D_X}$ that minimizes the update rate does not depend on the spectrum of $\mathbf{K_X}$, but equals $\mathbf{D_X} = (\frac{D_x}{d})\cdot\mathbf{I}$.

If, specifically, $D_x \leq d \cdot \lambda_{\min}(\mathbf{K}_{\mathbf{X}}^{C_W(\mathbf{X},\mathbf{Y})})$ then the choice of $\mathbf{K}$ and $\mathbf{D_X}, \mathbf{D_Y}$ that minimize R_{cache} and $\overline{R}_{\text{update}}$ decouple. Namely, in the regime $R_{\text{cache}} \in [0,\ C_W(\mathbf{X}, \mathbf{Y})]$ the trade-off $|\mathbf{K}| \leftrightarrow |\mathbf{K_X}||\mathbf{K_Y}|$ solved by $T_{\mathbf{X},\mathbf{Y}}(\gamma)$ produces a $\mathbf{K}'$ that satisfies $\mathbf{K}' \succeq \mathbf{K}^{C_W(\mathbf{X},\mathbf{Y})}$, a consequence of Theorem 2. Consequently, the optimal choice of $\mathbf{D_X}$ remains $\mathbf{D_X} = (\frac{D_x}{d})\cdot\mathbf{I}$. The same for $\mathbf{Y}$. In the regime $R_{\text{cache}} \geq C_W(\mathbf{X}, \mathbf{Y})$, $\mathbf{X}$ and $\mathbf{Y}$ can become conditionally independent and the trade-off between cache and update rate comes without rate loss. □

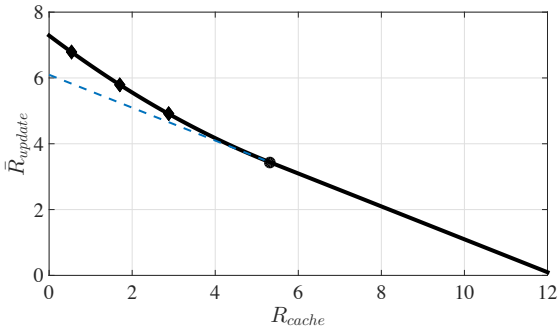

Fig. 3. Example of the caching trade-off with $d = 4$ and $\rho_1 = 0.9$, $\rho_2 = 0.8$, $\rho_3 = 0.6$ and $\rho_4 = 0.4$. The diamonds correspond, from left to right, to the points were respectively R_1, R_2 and R_3 become positive, following the waterfilling of Theorem 2. The circle corresponds to $R_{\text{cache}} = C_W(\mathbf{X}, \mathbf{Y})$. The dotted line is the straight line connecting $R_{\text{cache}} = R_{\mathbf{X},\mathbf{Y}}(D_x, D_y)$ and $\overline{R}_{\text{update}} = \frac{1}{2}R_{\mathbf{X},\mathbf{Y}}(D_x, D_y)$.

The Corollary implies that also in the coding problem on the Gray–Wyner network, for a small common rate one should apply the reverse waterfilling of Theorem 2, given the end distortion constraints are not too large. An example of this trade-off is plotted in Figure 3 for two vectors of length $d = 4$. The diamonds mark the points where the waterfilling procedure hits a new threshold and starts including another $(\tilde{X}_i, \tilde{Y}_i)-$pair into the coding process. Once $R_{\text{cache}} \geq C_W(\mathbf{X}, \mathbf{Y})$, it is large enough to make $\mathbf{X}, \mathbf{Y}$ conditionally independent. Consequently, the trade-off between R_{cache} and $\overline{R}_{\text{update}}$ coincides with the straight line connecting the points of $\big(R_{\text{cache}}, \overline{R}_{\text{update}}\big) = (R_{\mathbf{X},\mathbf{Y}}(D_x, D_y), 0)$ and $\big(0, \frac{1}{2}R_{\mathbf{X},\mathbf{Y}}(D_x, D_y)\big)$.

REFERENCES

[1] R. Gray and A. Wyner, "Source coding for a simple network," *Bell System Technical Journal, The*, vol. 53, no. 9, pp. 1681–1721, Nov 1974.

[2] A. Wyner, "The common information of two dependent random variables," *IEEE Transactions on Information Theory*, vol. 21, no. 2, pp. 163–179, Mar 1975.

[3] K. Viswanatha, E. Akyol, and K. Rose, "The lossy common information of correlated sources," *IEEE Transactions on Information Theory*, vol. 60, no. 6, pp. 3238–3253, June 2014.

[4] G. Xu, W. Liu, and B. Chen, "Wyners common information for continuous random variables - a lossy source coding interpretation," in *45th Annual Conference on Information Sciences and Systems (CISS)*, March 2011, pp. 1–6.

[5] ——, "Wyner's common information: Generalizations and a new lossy source coding interpretation," 2013. [Online]. Available: http://arxiv.org/abs/1301.2237

[6] G. J. Op 't Veld and M. C. Gastpar, "Caching gaussians: Minimizing total correlation on the gray-wyner network," in *2016 Annual Conference on Information Science and Systems (CISS)*, March 2016, pp. 478–483.

[7] C.-Y. Wang, S. H. Lim, and M. Gastpar, "Information-theoretic caching," in *IEEE International Symposium on Information Theory (ISIT)*, June 2015, pp. 1776–1780.

[8] ——, "Information-theoretic caching: Sequential coding for computing," 2015. [Online]. Available: http://arxiv.org/abs/1504.00553
[9] S. Watanabe, "Information theoretical analysis of multivariate correlation," *IBM Journal of Research and Development*, vol. 4, no. 1, pp. 66–82, Jan 1960.
[10] S. Satpathy and P. Cuff, "Gaussian secure source coding and wyner's common information," in *2015 IEEE International Symposium on Information Theory (ISIT)*, June 2015, pp. 116–120.
[11] J. Xiao and Q. Luo, "Compression of correlated gaussian sources under individual distortion criteria," in *43rd Allerton Conference on Communication, Control, and Computing*, 2005, pp. 438–447.
[12] L. Vandenberghe, S. Boyd, and S. P. Wu, "Determinant maximization with linear matrix inequality constraints," *SIAM Journal on Matrix Analysis and Applications*, vol. 19, no. 2, pp. 499–533, 1998. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.46.1561
[13] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," http://cvxr.com/cvx, Mar. 2014.
[14] T. M. Cover and J. A. Thomas, *Elements of information theory (2. ed.)*. Wiley, 2006.