Relaxed Wyner's Common Information

Michael Gastpar and Erixhen Sula
School of Computer and Communication Sciences, EPFL, CH-1015 Lausanne, Switzerland
{michael.gastpar,erixhen.sula}@epfl.ch

***Abstract*—In the problem of coded caching for media delivery, two separate coding opportunities have been identified. The first opportunity is a multi-user advantage and crucially hinges on a public broadcast link in the delivery phase. This has been explored in a plethora of works. The second opportunity has received far less attention and concerns similarities between files in the database. Here, the paradigm is to cache "the similarity" between the files. Upon the request, the encoder refines this by providing the specific details for the requested files. Extending Gray and Wyner's work (1974), it follows that the right measure of file similarity is Wyner's Common Information and its generalizations. The present paper surveys and extends the role of Wyner's Common Information in caching. As a novel result, explicit solutions are found for the Gaussian case under mean-squared error, both for the caching problem as well as for the network considered by Gray and Wyner. Our solution leverages and extends the recent technique of factorization of convex envelopes.**

I. Introduction

The Gray-Wyner network [1] is composed of one sender and two receivers. In a nutshell, the sender compresses two underlying correlated sources X and Y (with fixed $p(x, y)$) into three descriptions. The central description, of rate R_c, is provided to both receivers. Additionally, each receiver also has access to a tailored private description. Let us denote the rates of the private descriptions by $R_{u,x}$ and $R_{u,y}$, respectively. The main result of [1], Theorem 4, is that the set of trade-offs amongst these rates is given by the closure of the union of the regions

$$\{R_c \geq I(X, Y; W), R_{u,x} \geq H(X|W), R_{u,y} \geq H(Y|W)\}, \tag{1}$$

where the union is over all probability distributions $p(w, x, y)$ with marginals $p(x, y)$.

An alternative interpretation of this setting is in terms of multimedia contents caching, illustrated in Figure 1: There are files X and Y to choose from, but before the user chooses, the encoder provides a partial description, supposedly during a time when communication cost is much smaller. This partial description is called the *cache contents* and is of rate R_c. Then, the user selects one of the two files and the encoder provides the rest of the description. That is, the two receivers of the Gray-Wyner network now represent the two possible user requests, and the common description of the Gray-Wyner network is precisely the cache contents. When the user selects uniformly at random, using Eqn. (1), the resulting optimization problem can thus be written as

$$\min I(X, Y; W) \text{ s.t. } \frac{1}{2}H(X|\ W) + \frac{1}{2}H(Y|\ W) \leq \beta, \tag{2}$$

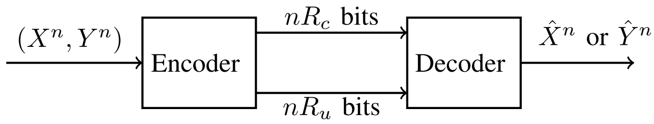

Fig. 1. A caching scenario with two correlated files, X^n and Y^n. The encoder produces two descriptions, one *before* knowing which file is requested, of rate R_c, and one *after finding out* which file is requested, of average rate R_u, where the average is taken over the file choice.

where the minimum is over all probability distributions $p(w, x, y)$ with marginals $p(x, y)$. That is, we minimize the cache rate R_c subject to a constraint on the expected delivery rate $\frac{1}{2}R_{u,x} + \frac{1}{2}R_{u,y}$. Connecting back to [1], we point out that this optimization problem is closely related to the quantity $T(\alpha)$ in Remark (3) of Theorem 4.

Alternatively (and equivalently), one can minimize the expected delivery rate subject to a constraint on the cache rate, which can be rewritten as

$$\min H(X, Y) - I(X, Y; W) + I(X; Y|W) \quad \text{s.t. } I(X, Y; W) \leq \alpha. \tag{3}$$

Since $H(X, Y)$ is fixed throughout, the essence of this optimization problem can thus be reduced to

$$\min I(X, Y; W) \text{ s.t. } I(X; Y|W) \leq \gamma. \tag{4}$$

We note that for $\gamma = 0$, this is precisely Wyner's Common Information [2]. For the case of discrete memoryless sources, a full account of this perspective on caching was developed in [3], [4]. An alternative connection between caching and the Gray-Wyner network, from a worst-case request perspective, is developed in [5].

II. Relaxed Wyner's Common Information

A. Lossless Case

The first key definition of the present paper is as follows.

Definition 1 (Relaxed Wyner's Common Information)**.** For a fixed probability distribution $p(x, y)$, we define

$$C_\gamma(X, Y) = \min I(X, Y; W) \tag{5}$$

such that $I(X; Y|W) \leq \gamma$, where the minimum is over all probability distributions $p(w, x, y)$ with marginals $p(x, y)$.

We note that $C_\gamma(X, Y)$ represents a relaxation of Wyner's common information in the sense that for the special case $\gamma = 0$, it recovers the latter. The quantity $C_\gamma(X, Y)$ should be expected to be of interest in its own right, but as we have detailed in the Introduction, it is directly related to the natural data compression question resulting in the optimization problem stated in Equation (2) and illustrated in Figure 1.

The optimization problem of Definition 1 does not easily admit explicit solutions. From [3, Thm.1], it follows that when X and Y are discrete memoryless sources, the cardinality of W need not be larger than the product of the cardinalities of X and Y plus 1. But even in the case of the binary symmetric source, the problem does not appear to admit a full and explicit solution for $\gamma > 0$ (see Example 1 in [3, p.6398]). By contrast, as we show here, for the case where X and Y are jointly Gaussian, an explicit solution can be given.

B. Lossy Case

As in the original work of Gray and Wyner [1] (Theorem 8), one may instead ask for *lossy* reconstructions of the original sources X and Y with respect to fidelity criteria. The optimization problem stated in Equation (2) can be directly extended to this scenario. This motivates the following definition:

Definition 2 (Relaxed lossy Wyner's Common Information)**.** For a fixed probability distribution $p(x, y)$, we define

$$C_{D,\beta}(X, Y) = \min I(X, Y; W) \tag{6}$$

such that $I(X; \hat{X}|W) + I(Y; \hat{Y}|W) \leq \beta$, where the minimum is over all probability distributions $p(\hat{x}, \hat{y}, w, x, y)$ with marginals $p(x, y)$ and satisfying

$$\mathbb{E}[d_x(X, \hat{X})] \leq D_x \text{ and } \mathbb{E}[d_y(Y, \hat{Y})] \leq D_y, \tag{7}$$

where $d_x(\cdot, \cdot)$ and $d_y(\cdot, \cdot)$ are arbitrary single-letter distortion measures (as in, e.g., [1, Eqn. (30) ff.]).

It does not appear possible to obtain a similarly canonical form of this optimization problem as was obtained for Definition 1. We note that a different (non-relaxed) notion of lossy Wyner's common information was defined in [6], imposing additional Markov chain constraints.

The operational significance of this definition for the caching problem described in the Introduction and illustrated in Figure 1 is to capture the minimum cache rate necessary to attain distortions D_x and D_y, respectively, when the average delivery rate is β.

III. Gaussian Sources

When the underlying distribution of X and Y is jointly Gaussian, it can be shown that both the (lossless) relaxed Wyner's common information as well as the lossy relaxed Wyner's common information (under mean-squared error) are attained by W that is jointly Gaussian with X and Y. This represents the main technical contribution of the present paper.

A. Relaxed Wyner's Common Information

In general, it does not appear feasible to obtain a closed-form explicit formula for the relaxed Wyner's common information. The case of jointly Gaussian sources presents a notable exception (though it does not have an immediate operational significance in the Gaussian (lossy) caching problem).

Theorem 1 (Gaussian Relaxed Wyner's Common Information)**.** *Let X and Y be jointly Gaussian with correlation coefficient ρ. Then,*

$$C_\gamma(X, Y) = \frac{1}{2}\log^+ \frac{(1+|\rho|)(1-\sqrt{1-e^{-2\gamma}})}{(1-|\rho|)(1+\sqrt{1-e^{-2\gamma}})}. \tag{8}$$

A proof outline is provided in Appendix A. The key part of the proof is to show that when X and Y are jointly Gaussian, the optimizing auxiliary W in Definition 1 is also Gaussian. For the special case of $\gamma = 0$, the result was established previously via a substantially different proof [7], [8].

B. Relaxed Lossy Wyner's Common Information

For Gaussian sources, the relaxed lossy Wyner's Common Information has a direct application both to the Gray-Wyner network as well as to the problem of caching. In the latter, it characterizes the minimum cache rate needed when the (average) delivery rate is β and the target distortion is D.

Theorem 2 (Gaussian Relaxed Lossy Wyner's Common Information)**.** *Let X and Y be jointly Gaussian with mean zero, equal variance σ^2, and with correlation coefficient ρ. Let $d_x(\cdot, \cdot)$ and $d_y(\cdot, \cdot)$ be the mean-squared error distortion measure. Let $D_x = D_y = D$. Then,*

$$C_{D,\beta}(X, Y) = \begin{cases} \frac{1}{2}\log^+ \frac{1+\rho}{2\frac{D}{\sigma^2}e^\beta+\rho-1}, & \textit{if } \sigma^2(1-\rho) \leq De^\beta \leq \sigma^2 \\ \frac{1}{2}\log^+ \frac{1-\rho^2}{\frac{D^2}{\sigma^4}e^{2\beta}}, & \textit{if } De^\beta \leq \sigma^2(1-\rho). \end{cases}$$

The proof proceeds along the same arguments as in Appendix A. We note that *assuming* the optimizing W in Definition 2 to be jointly Gaussian with the sources, the formula given in Theorem 2 was found in [9, Theorem 4.3, p.42].

IV. Gaussian Vector Sources

A. Relaxed Wyner's Common Information

Let us now turn to the special case where $\boldsymbol{X}$ and $\boldsymbol{Y}$ are jointly Gaussian random vectors. Most importantly, for this case, it can again be shown that the optimal auxiliary W in Definition 1 is jointly Gaussian with the sources (note that W is not generally scalar), using proof steps very similar to those presented in Appendix A. In the present paper, we provide the explicit result only for the case of two-dimensional vectors, although the following theorem directly extends to arbitrary dimensions.

Theorem 3. *Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be jointly Gaussian random vectors with mean zero and covariance matrix $K_{(\boldsymbol{X},\boldsymbol{Y})}$. Then,*

$$C_\gamma(\boldsymbol{X}, \boldsymbol{Y}) = \min_{\gamma_1,\gamma_2:\gamma_1+\gamma_2=\gamma} \frac{1}{2}\log^+ \frac{(1+\rho_1)(1-\sqrt{1-e^{-2\gamma_1}})}{(1-\rho_1)(1+\sqrt{1-e^{-2\gamma_1}})} + \frac{1}{2}\log^+ \frac{(1+\rho_2)(1-\sqrt{1-e^{-2\gamma_2}})}{(1-\rho_2)(1+\sqrt{1-e^{-2\gamma_2}})}, \tag{9}$$

where ρ_i (for $i = 1, 2$) are the singular values of $K_{\boldsymbol{X}}^{-1/2} K_{\boldsymbol{XY}} K_{\boldsymbol{Y}}^{-1/2}$.

The remaining optimization problem over γ_1 and γ_2 can be solved explicitly, as shown in the following corollary:

Corollary 4. *Assuming without loss of generality that $\rho_1 > \rho_2$, we have*

$$C_\gamma(\boldsymbol{X}, \boldsymbol{Y}) = \begin{cases} \frac{1}{2} \log \frac{(1+\rho_1)(1+\rho_2)(1-\sqrt{1-e^{-\gamma}})^2}{(1-\rho_1)(1-\rho_2)(1+\sqrt{1-e^{-\gamma}})^2}, & 0 \le \gamma < 2\overline{\gamma}_2, \\ \frac{1}{2} \log \frac{(1+\rho_1)\left(1-\sqrt{1-\frac{e^{-2\gamma}}{1-\rho_2^2}}\right)}{(1-\rho_1)\left(1+\sqrt{1-\frac{e^{-2\gamma}}{1-\rho_2^2}}\right)}, & 2\overline{\gamma}_2 \le \gamma < \overline{\gamma}_1 + \overline{\gamma}_2, \\ 0, & \overline{\gamma}_1 + \overline{\gamma}_2 \le \gamma, \end{cases} \tag{10}$$

where

$$\overline{\gamma}_i = \frac{1}{2} \log \frac{1}{1-\rho_i^2}. \tag{11}$$

V. Extension to More Than Two Sources

The Gray-Wyner network can be extended to more than two sources in several ways. For the context of caching, perhaps the most natural extension is as described in [3, Sec.III.C]: With respect to Figure 2 in [1], there are now N sources and there are N receivers. There is a common channel to all receivers, and there are private channels separately to each receiver. For the caching problem illustrated in Figure 1, this models a scenario where the user selects between N possible files rather than just two files, as in the figure. This perspective gives rise to the following extended definition.

Definition 3 (Relaxed Wyner's Common Information for multiple variables)**.** For a fixed probability distribution $p(x_1, x_2, \ldots, x_N)$, we define

$$C_\gamma(X_1, X_2, \ldots, X_N) = \min I(X_1, X_2, \ldots, X_N; W) \tag{12}$$

such that $\sum_{i=1}^N H(X_i|W) - H(X_1, X_2, \ldots, X_N|W) \le \gamma$, where the minimum is over all probability distributions $p(w, x_1, x_2, \ldots, x_N)$ with marginals $p(x_1, x_2, \ldots, x_N)$.

Using arguments similar to the ones outlined here, one can again show that for jointly Gaussian random variables $X_1, X_2, \ldots, X_N$, the optimizing W is jointly Gaussian.

Appendix

A. Proof Outline for Theorem 1

The proof of the converse for the Theorem involves two main steps. In this section, we prove that one optimal distribution is jointly Gaussian via a variant of the factorization of convex envelope. Then, we tackle the resulting (non-convex) optimization problem with Lagrange duality for the scalar and vector case respectively. We form the Lagrangian for our optimization problem as

$$s_\lambda(W|T) := I(XY; W|T) + \lambda I(X; Y|W, T) - \lambda\gamma, \tag{13}$$

and the two-letter version of the Lagrangian as

$$\begin{aligned} s_\lambda(W_1, W_2|T) &:= I(X_1X_2Y_1Y_2; W_1W_2|T) & (14) \\ &\quad + \lambda I(X_1X_2; Y_1Y_2|W_1W_2, T) - 2\lambda\gamma. & (15) \end{aligned}$$

Furthermore, we denote the *lower convex envelope* of $s_\lambda(W)$, (where $s_\lambda(W)$ is defined by dropping the random variable T in (13)) by

$$S_\lambda(W) = \inf_{p(t|x,y,w)} s_\lambda(W|T) \tag{16}$$

The dual function of our problem is

$$V_\lambda(\gamma) := \inf_{p(w|x,y)} S_\lambda(W). \tag{17}$$

Alternatively, we have

$$V_\lambda(\gamma) = \inf_{p(t,w|x,y)} s_\lambda(W|T) = \inf_{p(w|x,y)} \underbrace{\inf_{p(t|x,y,w)} s_\lambda(W|T)}_{S_\lambda(W)}. \tag{18}$$

Note that $S_\lambda(W)$ is a convex function of $p(w, x, y)$ as $S_\lambda(W)$ is the lower convex envelope of $s_\lambda(W)$. Thus, $S_\lambda(W)$ is a convex function of $p(w|x, y)$ since $p(x, y)$ is fixed and $p(w|x, y)$ is proportional to $p(w, x, y)$.

In addition, we define

$$S_\lambda(W|T) = \sum_t p(t) S_\lambda(W|T = t). \tag{19}$$

Lemma 5. *For any $\lambda > 0$ we have*

$$s_\lambda(W_{\theta_1}, W_{\theta_2}) \ge s_\lambda(W_{\theta_1}|W_{\theta_2}) + s_\lambda(W_{\theta_2}|W_{\theta_1}, X_{\theta_1}, Y_{\theta_1}) \tag{20}$$

with equality if and only if

- $I(X_{\theta_1}Y_{\theta_1}; W_{\theta_2}) = 0$
- $I(X_{\theta_2}Y_{\theta_2}; W_{\theta_1}|X_{\theta_1}Y_{\theta_1}) = 0$
- $I(X_{\theta_1}; Y_{\theta_2}|W_{\theta_1}W_{\theta_2}Y_{\theta_1}) = 0$
- $I(X_{\theta_2}; Y_{\theta_1}|W_{\theta_1}W_{\theta_2}X_{\theta_1}) = 0.$

Proof. Conditioned on a fixed $T = t$, we have

$$\begin{aligned} &s_\lambda(W_{\theta_1}, W_{\theta_2}) + 2\lambda\gamma = I(X_{\theta_1}X_{\theta_2}Y_{\theta_1}Y_{\theta_2}; W_{\theta_1}W_{\theta_2}) \\ &+ \lambda I(X_{\theta_1}X_{\theta_2}; Y_{\theta_1}Y_{\theta_2}|W_{\theta_1}W_{\theta_2}) \\ &\overset{(a)}{=} I(X_{\theta_1}Y_{\theta_1}; W_{\theta_1}W_{\theta_2}) + I(X_{\theta_2}Y_{\theta_2}; W_{\theta_1}, W_{\theta_2}|X_{\theta_1}Y_{\theta_1}) \\ &+\lambda I(X_{\theta_1}; Y_{\theta_1}Y_{\theta_2}|W_{\theta_1}W_{\theta_2}) + \lambda I(X_{\theta_2}; Y_{\theta_1}Y_{\theta_2}|W_{\theta_1}W_{\theta_2}X_{\theta_1}) \\ &\overset{(a)}{=} I(X_{\theta_1}Y_{\theta_1}; W_{\theta_1}|W_{\theta_2}) + \underline{I(X_{\theta_1}Y_{\theta_1}; W_{\theta_2})} \\ &+ I(X_{\theta_2}Y_{\theta_2}; W_{\theta_2}|X_{\theta_1}Y_{\theta_1}W_{\theta_1}) + \underline{I(X_{\theta_2}Y_{\theta_2}; W_{\theta_1}|X_{\theta_1}Y_{\theta_1})} \\ &+ \lambda I(X_{\theta_1}; Y_{\theta_1}|W_{\theta_1}W_{\theta_2}) + \lambda \underline{I(X_{\theta_1}; Y_{\theta_2}|W_{\theta_1}W_{\theta_2}Y_{\theta_1})} \\ &+\lambda I(X_{\theta_2}; Y_{\theta_2}|W_{\theta_1}W_{\theta_2}X_{\theta_1}Y_{\theta_1}) + \lambda \underline{I(X_{\theta_2}; Y_{\theta_1}|W_{\theta_1}W_{\theta_2}X_{\theta_1})} \\ &\overset{(b)}{\ge} s_\lambda(W_{\theta_1}|W_{\theta_2}) + s_\lambda(W_{\theta_2}|W_{\theta_1}, X_{\theta_1}, Y_{\theta_1}) + 2\lambda\gamma, \end{aligned}$$

where (a) follows from splitting the information terms and (b) follows from the non-negativity of the underlined terms. Thus, we have $\sum_t p(T = t) s_\lambda(W_{\theta_1}, W_{\theta_2}|T = t) \ge \sum_t p(T = t) \left(s_\lambda(W_{\theta_1}|W_{\theta_2}, T = t) + s_\lambda(W_{\theta_2}|W_{\theta_1}, X_{\theta_1}, Y_{\theta_1}, T = t) \right).$ $\square$

Proposition 6. *There is a pair of random variables $(T_*, W_*)|((X,Y)=(x,y))$ with $|\mathcal{T}_*| \leq 3$ such that*

$$V_\lambda(\gamma) = s_\lambda(W_*|T_*). \tag{21}$$

Proof. The proof comes along with Lemma 7 and Theorem 8. Hereafter we assume that the sequence $\{W_n\}|(X,Y)=(x,y)$ has finite variance.

Lemma 7. *If $\{W_n\}|(X,Y)=(x,y)$ has finite variance, then the sequence is tight and there exists a subsequence $\{W_{n_i}\}|(X,Y)=(x,y)$ and a limiting probability distribution $W_*|(X,Y)=(x,y)$ such that $W_{n_i}|(X,Y)=(x,y) \overset{w}{\Rightarrow} W_*|(X,Y)=(x,y)$ converges weakly in distribution.*

This lemma follows from [10, Proposition 17] and [10, Theorem 4]. To prove that the minimizer exists, it is enough to show that $s_\lambda(W)$ is lower semi-continuous. The critical terms in $s_\lambda(W)$ are $I(XY;W)$ and $I(X;Y|W)$. We will show by utilizing the following Theorem.

Theorem 8 ([11])**.** *If $X_n \overset{w}{\Rightarrow} X$ and $Y_n \overset{w}{\Rightarrow} Y$, then $D(p_X||q_Y) \leq \liminf_{n\to\infty} D(p_{X_n}||q_{Y_n})$.*

Note that $I(XY;W) = D(p_{XYW}||q_{XY}q_W)$, therefore

$$I(XY;W) \leq \lim \inf_{n\to\infty} I(X_nY_n;W_n), \tag{22}$$

since $(X_n,Y_n,W_n) \overset{w}{\Rightarrow} (X,Y,W)$, $(X_n,Y_n) \overset{w}{\Rightarrow} (X,Y)$ and $W_n \overset{w}{\Rightarrow} W$. On the other hand $I(X;Y|W) = D(p_{XYW}||q_{XYW})$, where q_{XYW} should satisfy the Markov chain $X \to W \to Y$, therefore

$$I(X;Y|W) \leq \lim \inf_{n\to\infty} I(X_n;Y_n|W_n). \tag{23}$$

The inequality holds since $(X_n,Y_n,W_n) \overset{w}{\Rightarrow} (X,Y,W)$ subject to the Markov chain $X_n \to W_n \to Y_n$. Then, we conclude that

$$s_\lambda(W) \leq \lim \inf_{n\to\infty} s_\lambda(W_n).$$

$\square$

Lemma 9. *Let $p_*(t,w|x,y)$ attain $V_\lambda(\gamma)$ where $p(x,y) \sim \mathcal{N}(0,K_{(X,Y)})$, and let $(\boldsymbol{T},\boldsymbol{W},\boldsymbol{X},\boldsymbol{Y}) \sim p_*(t_1,w_1,x_1,y_1)p_*(t_2,w_2,x_2,y_2)$. Let $(W,X,Y)_t$ denote the conditional distribution $p_*(w,x,y|t)$ and define*

$$(W_{\theta_1},X_{\theta_1},Y_{\theta_1})|((T_1,T_2)=(t_1,t_2)) \sim \frac{1}{\sqrt{2}}((W,X,Y)_{t_1}+(W,X,Y)_{t_2}),$$
$$(W_{\theta_2},X_{\theta_2},Y_{\theta_2})|((T_1,T_2)=(t_1,t_2)) \sim \frac{1}{\sqrt{2}}((W,X,Y)_{t_1}-(W,X,Y)_{t_2}).$$

Then:

1) *$(T,W_{\theta_1},X_{\theta_1},Y_{\theta_1})$ also attains $V_\lambda(\gamma)$.*
2) *$(T,W_{\theta_2},X_{\theta_2},Y_{\theta_2})$ also attains $V_\lambda(\gamma)$.*
3) *The joint distribution $(T,W_{\theta_1},W_{\theta_2},X_{\theta_1},X_{\theta_2},Y_{\theta_1},Y_{\theta_2})$ must satisfy*
   - $I(X_{\theta_1}Y_{\theta_1};W_{\theta_2}|T)=0$
   - $I(X_{\theta_2}Y_{\theta_2};W_{\theta_1}|X_{\theta_1}Y_{\theta_1},T)=0$
   - $I(X_{\theta_1};Y_{\theta_2}|W_{\theta_1}W_{\theta_2}Y_{\theta_1},T)=0$
   - $I(X_{\theta_2};Y_{\theta_1}|W_{\theta_1}W_{\theta_2}X_{\theta_1},T)=0.$

Proof. We have

$$\begin{aligned}
2V_\lambda(\gamma) &\overset{(c)}{=} s_\lambda(W_1|T_1)+s_\lambda(W_2|T_2)\\
&\overset{(d)}{=} s_\lambda(W_1,W_2|T_1,T_2)\\
&\overset{(e)}{=} s_\lambda(W_{\theta_1},W_{\theta_2}|T_1,T_2)\\
&\overset{(f)}{\geq} s_\lambda(W_{\theta_1}|W_{\theta_2},T_1,T_2)\\
&\quad + s_\lambda(W_{\theta_2}|W_{\theta_1},X_{\theta_1},Y_{\theta_1},T_1,T_2)\\
&\overset{(g)}{\geq} S_\lambda(W_{\theta_1}|W_{\theta_2})+S_\lambda(W_{\theta_2}|W_{\theta_1},X_{\theta_1},Y_{\theta_1})\\
&\overset{(h)}{\geq} S_\lambda(W_{\theta_1})+S_\lambda(W_{\theta_2})\\
&\overset{(i)}{\geq} 2V_\lambda(\gamma).
\end{aligned} \tag{24}$$

Here (c) holds for the distribution $p_*(t,w|x,y)p(x,y)$ that attains V_λ; (d) holds since (T_1,W_1,X_1,Y_1) and (T_2,W_2,X_2,Y_2) are independent by assumption; (e) follows by variable transformation since mutual information is preserved under bijective transformation; (f) follows by Lemma 5; (g) follows from

$$\begin{aligned}
s_\lambda(W_{\theta_1}|T,W_{\theta_2}) &= \sum_{w_{\theta_2}} p(w_{\theta_2})s_\lambda(W_{\theta_1}|T,W_{\theta_2}=w_{\theta_2})\\
&\overset{(j)}{\geq} \sum_{w_{\theta_2}} p(w_{\theta_2})S_\lambda(W_{\theta_1}|W_{\theta_2}=w_{\theta_2})\\
&\overset{(k)}{=} S_\lambda(W_{\theta_1}|W_{\theta_2})
\end{aligned} \tag{25}$$

where (j) holds because $S_\lambda(W_{\theta_1}|W_{\theta_2}=w_{\theta_2})$ is the lower convex envelope of $s_\lambda(W_{\theta_1}|W_{\theta_2}=w_{\theta_2})$ and (k) is the definition of $S_\lambda(.|.)$; (h) holds since $S_\lambda(W_{\theta_1})$ is convex in $p(w_{\theta_1}|x,y)$ and by Jensen's inequality $S_\lambda(W_{\theta_1}|W_{\theta_2}) \geq S_\lambda(W_{\theta_1})$; (i) follows from definition of $V_\lambda(\gamma)$. $\square$

Our approach only shows that Gaussian is a maximizer but not necessarily the unique maximizer. For simplicity let $\boldsymbol{Z}=(X,Y,W)$.

Corollary 10. *For every $\ell \in \mathbb{N}$, $n=2^\ell$, let $(T^n,\boldsymbol{Z}) \sim \prod_{i=1}^n p_*(t_i,\boldsymbol{z}_i)$. Then $(T^n,\tilde{\boldsymbol{Z}}_n)$ achieves $V_\lambda(\gamma)$ where $\tilde{\boldsymbol{Z}}_n|(T_n=(t_1,t_2,\ldots,t_n)) \sim \frac{1}{\sqrt{n}}(\boldsymbol{Z}_{t_1}+\boldsymbol{Z}_{t_2}+\cdots+\boldsymbol{Z}_{t_n})$. We take $\boldsymbol{Z}_{t_1},\boldsymbol{Z}_{t_2},\ldots,\boldsymbol{Z}_{t_n}$ to be independent random variables here.*

Proof. The proof follows by induction using Lemma 9. $\square$

Lemma 11. *For $\lambda>0$, there is a single Gaussian distribution (i.e. no mixture is required) that achieves $V_\lambda(\gamma)$.*

Proof. The proof is the same as in [10, Appendix IV]. $\square$

Remark 1. *Consider $\hat{\boldsymbol{X}} = K_{\boldsymbol{X}}^{-1/2}\boldsymbol{X}$, $\hat{\boldsymbol{Y}} = K_{\boldsymbol{Y}}^{-1/2}\boldsymbol{Y}$ and $\tilde{\boldsymbol{W}} = K_{\boldsymbol{W}}^{-1/2}\boldsymbol{W}$ (under the assumption that $K_{\boldsymbol{X}}$ and $K_{\boldsymbol{Y}}$ are invertible). We have $K_{\hat{\boldsymbol{X}}}=I_2$ and $K_{\hat{\boldsymbol{Y}}}=I_2$. In* *addition, we have* $K_{\hat{\boldsymbol{X}}\hat{\boldsymbol{Y}}} = K_{\boldsymbol{X}}^{-1/2} K_{\boldsymbol{X}\boldsymbol{Y}} K_{\boldsymbol{Y}}^{-1/2}$. *By singular value decomposition we have* $K_{\hat{\boldsymbol{X}}\hat{\boldsymbol{Y}}} = R_{\boldsymbol{X}} \Lambda R_{\boldsymbol{Y}}$, *then we define* $\tilde{\boldsymbol{X}} = R_{\boldsymbol{X}} \hat{\boldsymbol{X}}$ *and* $\tilde{\boldsymbol{Y}} = R_{\boldsymbol{Y}} \hat{\boldsymbol{Y}}$. *Thus, we have* $K_{\tilde{\boldsymbol{X}}} = K_{\hat{\boldsymbol{X}}}$, $K_{\tilde{\boldsymbol{Y}}} = K_{\hat{\boldsymbol{Y}}}$ *and* $K_{\tilde{\boldsymbol{X}}\tilde{\boldsymbol{Y}}} = \Lambda$. *Under the following transformation it holds that* $I(\boldsymbol{X}\boldsymbol{Y};\boldsymbol{W}) = I(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{Y}};\tilde{\boldsymbol{W}})$ *and* $I(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{W}) = I(\tilde{\boldsymbol{X}};\tilde{\boldsymbol{Y}}|\tilde{\boldsymbol{W}})$ *thus*

$$C_\gamma(\boldsymbol{X},\boldsymbol{Y}) = \min_{p(\boldsymbol{w}|\boldsymbol{x}\boldsymbol{y}):I(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{W})\leq\gamma} I(\boldsymbol{X}\boldsymbol{Y};\boldsymbol{W}) \tag{26}$$

$$= \min_{p(\tilde{\boldsymbol{w}}|\tilde{\boldsymbol{x}}\tilde{\boldsymbol{y}}):I(\tilde{\boldsymbol{X}};\tilde{\boldsymbol{Y}}|\tilde{\boldsymbol{W}})\leq\gamma} I(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{Y}};\tilde{\boldsymbol{W}}) \tag{27}$$

1) Lower Bound: By using remark 1 it suffices to consider an arbitrary covariance matrix for the triple (X,Y,W), which is of the form

$$K_{(X,Y,W)} = \begin{bmatrix} 1 & \rho & \rho_1 \\ \rho & 1 & \rho_2 \\ \rho_1 & \rho_2 & 1 \end{bmatrix}. \tag{28}$$

Thus, we have

$$\begin{aligned}
&C_\gamma(X,Y) = \min_{p(w|xy):I(X;Y|W)\leq\gamma} I(XY;W) \overset{(l)}{\geq} \max_\lambda V_\lambda(\gamma) \\
&\overset{(m)}{=} \max_\lambda \min_{\rho_1,\rho_2:K_{(X,Y,W)}\succeq 0} \frac{1}{2}\log\left(\frac{\det(K_W)\det(K_{(X,Y)})}{\det(K_{(X,Y,W)})}\right) \\
&+ \frac{\lambda}{2}\log\left(\frac{\det(K_{(X,W)})\det(K_{(Y,W)})}{\det(K_{(X,Y,W)})\det(K_W)}\right) - \lambda\gamma \\
&\overset{(n)}{=} \max_\lambda \min_{\rho_1,\rho_2:K_{(X,Y,W)}\succeq 0} \frac{1}{2}\log\left(\frac{1-\rho^2}{1-\rho^2-\rho_1^2-\rho_2^2+2\rho\rho_1\rho_2}\right) \\
&+ \frac{\lambda}{2}\log\left(1+\frac{(\rho-\rho_1\rho_2)^2}{1-\rho^2-\rho_1^2-\rho_2^2+2\rho\rho_1\rho_2}\right) - \lambda\gamma \\
&\overset{(o)}{\geq} \max_\lambda \min_{\rho_1,\rho_2:K_{(X,Y,W)}\succeq 0} \frac{1}{2}\log\left(\frac{1-\rho^2}{1-\rho^2-2\rho_1\rho_2+2\rho\rho_1\rho_2}\right) \\
&+ \frac{\lambda}{2}\log\left(1+\frac{(\rho-\rho_1\rho_2)^2}{1-\rho^2-2\rho_1\rho_2+2\rho\rho_1\rho_2}\right) - \lambda\gamma \\
&\overset{(p)}{=} \max_\lambda \min_{\eta:K_{(X,Y,W)}\succeq 0,\eta\geq 0} \frac{1}{2}\log\left(\frac{(1-\rho^2)(1-\eta)^{2\lambda}}{(1-\rho^2-2\eta+2\rho\eta)^{1+\lambda}}\right) - \lambda\gamma \\
&\overset{(q)}{=} \frac{1}{2}\log^+ \frac{(1+\rho)(1-\sqrt{1-e^{-2\gamma}})}{(1-\rho)(1+\sqrt{1-e^{-2\gamma}})}
\end{aligned}$$

where (l) comes from weak duality; (m) comes from Appendix A, which shows that one optimal distribution is Gaussian; (n) comes from simplifying the previous step; (o) comes from the bound $\rho_1^2 + \rho_2^2 \geq 2\rho_1\rho_2$ and equality is reached if and only if $\rho_1 = \rho_2$; (d) comes from plugging $\eta = \rho_1\rho_2$. We assume w.l.o.g that $\eta \geq 0$, meaning ρ_1 and ρ_2 have the same sign. Otherwise, for $\eta < 0$ we use the bound $\rho_1^2+\rho_2^2 \geq -2\rho_1\rho_2$ where ρ_1 and ρ_2 have different sign and (p) reduces to the same expression independent of the sign of η; (q) we find the optimal λ and η by finding the global maxima and minima respectively. The optimal λ and η are

$$\eta = \frac{\lambda\rho - 1}{\lambda - 1}, \quad \lambda = \sqrt{\frac{e^{2\gamma}}{e^{2\gamma}-1}}.$$

Combining the optimal λ and η we get

$$\eta = \frac{\rho - \sqrt{1-e^{-2\gamma}}}{1-\sqrt{1-e^{-2\gamma}}}. \tag{29}$$

Since $\eta \geq 0$, then $\rho \geq \sqrt{1-e^{-2\gamma}}$. If $\rho < \sqrt{1-e^{-2\gamma}}$, then generalized Wyner's common information becomes zero.

2) Upper Bound: Let us assume (without loss of generality) that X and Y have unit variance and are non-negatively correlated with correlation coefficient $\rho \geq 0$. Since they are jointly Gaussian, we can express them as

$$X = \sqrt{\beta}W + \sqrt{1-\beta}N_X \tag{30}$$

$$Y = \sqrt{\beta}W + \sqrt{1-\beta}N_Y, \tag{31}$$

where W, N_X, N_Y are jointly Gaussian, and where $W \sim \mathcal{N}(0,1)$ is independent of (N_X, N_Y). Letting the covariance of the vector (N_X, N_Y) be

$$K_{(N_X,N_Y)} = \begin{bmatrix} 1 & \alpha \\ \alpha & 1 \end{bmatrix} \tag{32}$$

for some $0 \leq \alpha \leq \rho$, we find that we need to choose $\beta = \frac{\rho-\alpha}{1-\alpha}$. Specifically, let us select $\alpha = \sqrt{1-e^{-2\gamma}}$, for some $0 \leq \gamma \leq \frac{1}{2}\log\frac{1}{1-\rho^2}$. For this choice, we find $I(X;Y|W) = \gamma$ and

$$I(XY;W) = \frac{1}{2}\log\frac{(1+\rho)(1-\alpha)}{(1-\rho)(1+\alpha)}. \tag{33}$$

ACKNOWLEDGMENT

This work was supported in part by the Swiss National Science Foundation under Grant 169294, Grant P2ELP2_165137.

REFERENCES

[1] R. M. Gray and A. D. Wyner, "Source coding for a simple network," *Bell Syst. Tech. J*, vol. 53, no. 9, pp. 1681–1721, 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] C. Y. Wang, S. H. Lim, and M. Gastpar, "Information-theoretic caching: Sequential coding for computing," *IEEE Transactions on Information Theory*, vol. 62, no. 11, pp. 6393–6406, Nov 2016.

[4] ——, "Information-theoretic caching: The multi-user case," *IEEE Transactions on Information Theory*, vol. 63, no. 11, pp. 7018–7037, Nov 2017.

[5] R. Timo, S. S. Bidokhti, M. A. Wigger, and B. C. Geiger, "A rate-distortion approach to caching," *IEEE Transactions on Information Theory*, vol. 64, no. 3, pp. 1957–1976, 2018.

[6] 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.

[7] G. Xu, W. Liu, and B. Chen, "A lossy source coding interpretation of Wyner's common information," *IEEE Transactions on Information Theory*, vol. 62, no. 2, pp. 754–768, 2016.

[8] 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.

[9] G. J. Op 't Veld, "A Gaussian source coding perspective on caching and total correlation," Ph.D. dissertation, EPFL (Lausanne), 2017.

[10] Y. Geng and C. Nair, "The capacity region of the two-receiver Gaussian vector broadcast channel with private and common messages," *IEEE Transactions on Information Theory*, vol. 60, no. 4, pp. 2087–2104, April 2014.

[11] E. Posner, "Random coding strategies for minimum entropy," *IEEE Transactions on Information Theory*, vol. 21, no. 4, pp. 388 – 391, 1975.