

Caching Gaussians: Minimizing Total Correlation on the Gray–Wyner Network

Giel J. Opt’ 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 caching problem that resembles a lossy Gray–Wyner network: A source produces vector samples from a Gaussian distribution, but the user is interested in the samples of only one component. The encoder first sends a cache message without any knowledge of the user’s preference. Upon learning her request, a second message is provided in the update phase so as to attain the desired fidelity on that component.

The cache is efficient if it exploits as much of the correlation in the source as possible, which connects to the notions of Wyner’s common information (for high cache rates) and Watanabe’s total correlation (for low cache rates). For the former, we extend known results for 2 Gaussians to multivariates by showing that common information is a simple linear program, which can be solved analytically for circulant correlation matrices.

Total correlation in a Gaussian setting is less well-studied. We show that for bivariate and using Gaussian auxiliaries it is captured in the dominant eigenvalue of the correlation matrix. For multivariates the problem is a more difficult optimization over a non-convex domain, but we conjecture that circulant matrices may again be analytically solvable.

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

I. INTRODUCTION

Caching revolves around coding information you expect to be useful in the near future. Imagine a source continuously produces two samples from two correlated distributions. The user, though, is only interested in the samples produced by one of the two. Before revealing which one has its interest, the encoder can send a *cache* that is hopefully useful when the actual request takes place. Once the user reveals its choice, the encoder sends an update to complement the cache such that either X_1^n or X_2^n can be recovered at the required fidelity, as is pictured in Figure 1. By sketching both possible requests as two separate decoders, our idea of caching resembles a Gray–Wyner network [1], [2], a popular source coding network in information theory. In that analogy, an efficient caching strategy is a common message R_0 that minimizes the average of the individual updates R_1 and R_2 in Figure 2.

In this paper, we study this interpretation of caching applied to Gaussians, similar to the approach of [3], [4] for discrete sources. As a start, we assume the requests are uniformly distributed; the focus of our study is to achieve efficient caching through exploiting correlation in the data streams. This viewpoint will reveal that Caching Gaussian sources revolves around the two notions of Wyner’s common information [2]

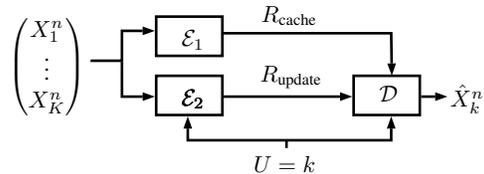


Fig. 1. The caching problem for a length- K vector source, with a caching and subsequent update phase. The side information $U = k$ models the request.

for high cache rates and Watanabe’s total correlation [5] for low cache rates; to show this will be our first task.

Wyner used the Gray–Wyner network to find and define

$$C_W(X_1, X_2) = \inf_{X_1 - Y - X_2} I(X_1, X_2; Y). \quad (1)$$

It is the minimum rate needed on the common branch of Figure 2 so as to make X_1 and X_2 conditionally independent. Recent work has already sought to extend this to a lossy setting. The lossy notion of Viswanatha, Akyol and Rose [6] is denoted $C_W(X_1, X_2; D_1, D_2)$. This extension takes the end distortion constraints of the decoders of Figure 2 in mind. Formally, it is the minimum rate needed on the common branch such that the sum-rate on all branches does not exceed the joint rate-distortion function $R_{X_1, X_2}(D_1, D_2)$. Consequently, if D_1 and D_2 are too large, the common branch must code $R_{X_1, X_2}(D_1, D_2)$ entirely by itself. Constraining the sum-rate would imply for our caching analogy that the cache rate is necessarily large, which covers only half of our interest.

Xu, Liu and Chen [7], [8] also studied Gaussians in particular. By treating lossy common information as if it were lossless, i.e., by removing the end distortion constraint like $C_W(X_1, X_2; 0, 0)$, they found an analytic expression for $C_W(X_1, X_2)$. They then extended this result to K Gaussians who are all pairwise correlated with the same coefficient ρ . Also their work specializes on high rates on the common branch only. We do, however, continue on the perspective of assuming that the end distortion constraints are small: that is the region where the cache-update-rate trade-off is all about how to use correlation.

Our aim is to characterize our caching problem and to study how the concepts of common information for a high cache rate, and total correlation for a low rate pan out for Gaussians. The former is the topic of Section III, where we follow up on some results from [6], [8]: We generalize the

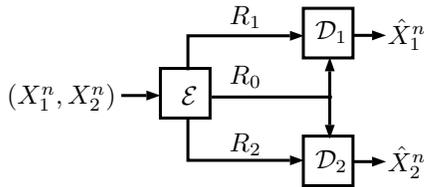


Fig. 2. A Gray–Wyner network for a vector source of length 2.

notion of common information to K sources and show that the auxiliary is necessarily a Gaussian found by a simple convex optimization problem. Furthermore, we show that the $K = 2$ closed-form expression of Xu et al. [8] corresponds to a much wider set of sources: Gaussians with a circulant correlation matrix.

If the cache rate is too small to capture *all* common information, finding a full characterization is both challenging and unclear. Most notably, we do not know if Gaussian auxiliaries are optimal. Nonetheless, we restrict our attention to Gaussian quantizers and the resulting inner bounds are smooth between the high and low cache rate region, where—as stated—we know that in the former the bound is tight. We show that the optimal *Gaussian* auxiliary is one that minimizes the total correlation that is left when you condition \mathbf{X} on the cache message; one should pick a \mathbf{Y} that minimizes $\sum_i h(X_i|\mathbf{Y}) - h(\mathbf{X}|\mathbf{Y})$.

Finding this auxiliary is a challenging non-convex optimization problem. For the case of a bivariate Gaussian source, we show that total correlation is captured in the dominant eigenvalue of the correlation matrix. For an arbitrary number of sources, though, these eigenvalues are generally *not* the solution. What is instead remains an open problem. This is the topic of Section IV, which we end with a conjecture that a closed-form expression might again exist for circulant correlation matrices.

II. PRELIMINARIES AND PROBLEM STATEMENT

Let us refresh our idea of caching: The source is a vector \mathbf{X} of length K , but the user is only interested in the samples of one vector element. In the cache phase, the encoder sends a message without knowing the user’s preference. In the second phase, the update, the user shows an interest in X_k^n uniformly at random. The encoder then sends an update such that both messages combined will provide the user with an estimate \hat{X}_k^n that satisfies some required fidelity criterion.

A. Caching as an interpretation of the Gray–Wyner Network

At each time instant, the source independently produces a vector of length K , following the Gaussian distribution $\mathcal{N}(0, \Sigma_{\mathbf{X}})$ with some correlated covariance $\Sigma_{\mathbf{X}}$. That is, the samples are drawn iid in time, while the vector elements among themselves are correlated. We denote the sample at time n by $\mathbf{X}(n)$, and its k th component by $X_k(n)$, for $k = 1, 2, \dots, K$. Independently of \mathbf{X} , a single random variable U is drawn from the set $\{1, 2, \dots, K\}$, the selection variable. As a starting point, we discuss the uniform case $P_U(u) = \frac{1}{K}$.

We consider block coding of length N with two encoding phases. The first, referred to as the *cache*, observes only $\{\mathbf{X}(n)\}_{n=1}^N$ and produces a description using NR_{cache} bits, where R_{cache} is called the *cache rate*. The second, referred to as the *update*, gets to observe $\{\mathbf{X}(n)\}_{n=1}^N$ as well as the value of the random variable $U = k$ and produces a description using $NR_{u,k}$ bits, where $R_{u,k}$ is called the *update rate* for the case $U = k$. Upon observing both compressed descriptions, the decoder must output estimates $\hat{X}_k(n)$. A rate-distortion $(K + 2)$ -tuple $(R_{\text{cache}}, R_{u,1}, \dots, R_{u,k}, D_{\text{final}})$ is said to be achievable if for every k there exists a sequence of $(2^{nR_{\text{cache}}}, 2^{nR_{u,k}}, n)$ codes with

$$\limsup_{N \rightarrow \infty} \mathbb{E} \left(\frac{1}{N} \sum_{n=1}^N (X_k(n) - \hat{X}_k(n))^2 \right) \leq D_{\text{final}}.$$

Our interest is in how correlation in \mathbf{X} affects the cache-update-rate trade-off, not so much how these rates behave as a function of this D_{final} . One could therefore assume $D_{\text{final}} = 0$. However, if it is sufficiently small, it will have no effect on the trade-off; how small will be discussed in Corollary 2.

If we sketch the events of the user asking for some X_k as K separate, mutually exclusive events, our concept of caching can be drawn as a Gray–Wyner network with K end terminals; there is one common communication link to all decoders (the cache) and K individual ones (the possible updates). The case of $K = 2$ is drawn in Figure 2.

Our goal is to find a good caching strategy that minimizes the average update rate. In the Figure this corresponds to a good common message that minimizes the average (or sum) of all the individual branches. To that end we also define the average update rate:

$$\bar{R}_{\text{update}} = \frac{1}{K} \sum_{i=1}^K R_{u,i}.$$

A shorthand notation is to say the triple $(R_{\text{cache}}, \bar{R}_{\text{update}}, D_{\text{final}})$ is achievable to mean that the entire $(K + 2)$ -tuple is. The convex closure of all these tuples is called the achievable cache-rate-distortion region $\mathcal{R}_{\text{cache}}(D_{\text{final}})$.

Theorem 1. *The cache-rate-distortion region $\mathcal{R}_{\text{cache}}(D_{\text{final}})$ is characterized by the following inequalities*

$$R_{\text{cache}} \geq I(\mathbf{X}; \mathbf{Y}), \quad (2)$$

$$\bar{R}_{\text{update}} \geq \frac{1}{K} \sum_{i=1}^K I(X_i; \hat{X}_i|\mathbf{Y}), \quad (3)$$

$$D_{\text{final}} \geq \mathbb{E}[(X_i - \hat{X}_i)^2] \quad i = 1, \dots, K, \quad (4)$$

over all auxiliaries \mathbf{Y} and $\hat{\mathbf{X}}$.

The outline of the proof is the analogy with the Gray–Wyner network, whose lossy extension to K users was discussed in [8]. The difference is that we take the average over all possible individual links, since this is the average update rate we want to take as an objective. The choice of the user through the selection variable $U = k$ can be considered as

side information for the update-encoder. However, U follows a uniform distribution and its statistics cannot be leveraged in any of the rates.

B. Applying Gaussian Auxiliaries leads to Total Correlation

In general, it is difficult to find a closed-form characterization of the boundary (2)-(4), because the minimization over all auxiliaries \mathbf{Y} is far from trivial. The distribution of \mathbf{X} does not imply that \mathbf{Y} should also be Gaussian. This remains an open problem. In this work, we are able to show that for large R_{cache} , \mathbf{Y} should indeed be jointly Gaussian with \mathbf{X} . For small R_{cache} , we restrict our attention to Gaussian quantizers nevertheless. By doing so, the boundary of the rate region (2)-(4) can be captured into one function by fixing two parameters and minimizing the other. We shall call this the cache-rate function $R_{\text{cache}}(d)$, where d will be one parameter to capture the end distortion constraint and average update rate as is outlined in the following corollary, again for small D_{final} :

Corollary 1. *For Gaussian auxiliaries, the boundary of achievable rates (2)-(4) can be characterized by the cache-rate function:*

$$R_{\text{cache}}(d) = \min_{\mathbf{D}} \frac{1}{2} \log \frac{|\Sigma_{\mathbf{X}}|}{|\mathbf{D}|} \quad \text{s.t.} \quad \begin{cases} \mathbf{D} \preceq \Sigma_{\mathbf{X}}, \\ \prod_i D_{i,i} = d, \end{cases} \quad (5)$$

where $d = D_{\text{final}}^K \cdot 2^{2K\bar{R}_{\text{update}}}$ and $D_{i,i} \geq D_{\text{final}} \forall i$.

Proof. First, for \mathbf{Y} jointly Gaussian with \mathbf{X} , fill in (2)-(4):

$$R_{\text{cache}} \geq \frac{1}{2} \log \frac{|\Sigma_{\mathbf{X}}|}{|\mathbf{D}|}, \quad (6)$$

$$\bar{R}_{\text{update}} \geq \frac{1}{K} \sum_{i=1}^K \frac{1}{2} \log \frac{D_{i,i}}{D_{\text{final}}} = \frac{1}{2K} \log \frac{\prod_i D_{i,i}}{D_{\text{final}}^K}, \quad (7)$$

where \mathbf{D} is the MSE distortion after receiving \mathbf{Y} :

$$\mathbf{D} = \Sigma_{\mathbf{X}|\mathbf{Y}} = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}|\mathbf{Y}])(\mathbf{X} - \mathbb{E}[\mathbf{X}|\mathbf{Y}])^T]. \quad (8)$$

Hence, we can describe a Gaussian auxiliary \mathbf{Y} equivalently by its consequent distortion \mathbf{D} w.r.t. \mathbf{X} . This \mathbf{D} is the intermediate distortion *between* the cache and update phase.

The boundary of these inequalities can be captured in one function by fixing one rate and minimizing the other; w.l.o.g. we fix \bar{R}_{update} and minimize R_{cache} . Hence, we look for a distortion profile that maximizes the determinant in (6) for a fixed product of its marginal distortions in (7), i.e., we fix

$$\prod_{i=1}^K D_{i,i} = D_{\text{final}}^K \cdot 2^{2K\bar{R}_{\text{update}}} \stackrel{\text{def}}{=} d.$$

The search space is then further constrained to $\mathbf{D} \preceq \Sigma_{\mathbf{X}}$, which corresponds to achievable Gaussian codebooks, the proof of which can be found in [9]. Lastly, $D_{i,i} \geq D_{\text{final}}$, ensures the update rates in (7) are non-negative¹. \square

¹The condition $D_{i,i} \geq D_{\text{final}}$ may falsely suggest that the distortion constraint will not be satisfied. It only says, though, it may not *yet* be satisfied by the cache. If it is not, the update phase will complement any shortcoming (see (7)). Otherwise, the inequality simply says the cache should not waste rate on components that already meet the final distortion constraint.

The auxiliary \mathbf{Y} thus influences R_{cache} and \bar{R}_{update} through the MSE distortion matrix \mathbf{D} that is left after observing \mathbf{Y} . For a small cache rate, one must code \mathbf{X} up to some distortion matrix with a large determinant. Yet, for a small average update rate, the encoder will want to minimize the product of the marginal distortions found on the diagonal of \mathbf{D} . These objectives are contradictory and are linked through total (conditional) correlation as defined by Watanabe [5], i.e.,

$$T(\mathbf{X}|\mathbf{Y}) \stackrel{\text{def}}{=} \sum_{i=1}^K h(X_i|\mathbf{Y}) - h(\mathbf{X}|\mathbf{Y}) \geq 0. \quad (9)$$

Namely if \mathbf{Y} is Gaussian, then the above equals

$$\frac{1}{2} \log(2\pi e)^K \prod_i D_{i,i} - \frac{1}{2} \log(2\pi e)^K |\mathbf{D}| \geq 0,$$

which is the algebraic Hadamard inequality [10, Section 17.9]:

$$\prod_i D_{ii} \geq |\mathbf{D}|.$$

Hence, the cache-rate function $R_{\text{cache}}(d)$ is a search over Gaussian auxiliaries \mathbf{Y} with an MSE \mathbf{D} that has the smallest possible gap on the Hadamard inequality, given that one side of it is fixed. That gap corresponds to the total correlation present in the conditional distribution $\mathbf{X}|\mathbf{Y}$. In other words: a good caching strategy is one that exploits as much of the correlation as possible in the cache phase, such that as little as possible goes to waste in the individual update. Although not surprising, this insight is of most importance and now has a measure in the form of $T(\mathbf{X}|\mathbf{Y})$.

C. Variance is irrelevant for minimizing Total Correlation

The issue is that although $R_{\text{cache}}(d)$ is a minimization over a convex function, it is one over a non-convex domain. This problem is not efficiently solvable by numeric methods. It requires insight into the correlation structure of matrices and an analytic solution might not exist.

An important observation is that total correlation can be minimized by only operating on the correlation matrix of the source; variance is irrelevant. To illustrate this, let us for the following lemma define $R_{\text{cache},\mathbf{X}}(d)$ as $R_{\text{cache}}(d)$ with respect to the source distribution \mathbf{X} .

Lemma 1. *Let $\mathbf{A} = \text{diag}(a_1, a_2, \dots, a_K)$, with $a_i > 0$. Then*

$$R_{\text{cache},\mathbf{X}}(d) = R_{\text{cache},\mathbf{A}^T\mathbf{X}}(d \cdot \prod_i a_i^2).$$

Proof. The key is that determinant (objective value) and the product of the main diagonal (constraint) of a covariance matrix are equally affected by scaling X_i by some positive factor a_i . Consider the scaled random variable $\mathbf{A}^T\mathbf{X}$ with covariance $\mathbf{A}^T\Sigma_{\mathbf{X}}\mathbf{A}$ and let us solve $R_{\text{cache},\mathbf{A}^T\mathbf{X}}(d)$:

$$\max |\mathbf{D}| \quad \text{s.t.} \quad \mathbf{D} \preceq \mathbf{A}^T\Sigma_{\mathbf{X}}\mathbf{A} \quad \text{and} \quad \prod_i D_{i,i} = d.$$

Rewrite $\mathbf{D} = \mathbf{A}^T \bar{\mathbf{D}} \mathbf{A}$ and its constraint-function as $\prod_i D_{i,i} = \prod_i a_i^2 \bar{D}_{i,i} = d$. Hence, one can equivalently solve:

$$\max |\bar{\mathbf{D}}| \quad \text{s.t.} \quad \bar{\mathbf{D}} \preceq \Sigma_{\mathbf{X}} \quad \text{and} \quad \prod_i \bar{D}_{i,i} = \frac{d}{\prod_i a_i^2}.$$

For completeness, we note that $I(\mathbf{X}; \mathbf{Y}) = I(\mathbf{A}^T \mathbf{X}; \mathbf{A}^T \mathbf{Y})$ for any diagonal \mathbf{A} , which in our algebra expresses itself as:

$$\frac{1}{2} \log \frac{|\Sigma_{\mathbf{X}}|}{|\bar{\mathbf{D}}|} = \frac{1}{2} \log \frac{|\mathbf{A}^T \Sigma_{\mathbf{X}} \mathbf{A}|}{|\mathbf{A} \bar{\mathbf{D}} \mathbf{A}|}.$$

By doing these steps with $R_{\text{cache}, \mathbf{A}^T \mathbf{X}}(\prod_i a_i^2 d)$ instead of $R_{\text{cache}, \mathbf{A}^T \mathbf{X}}(d)$, we find the nicer expression as in the Lemma. \square

The Lemma says that if one scales the variance of \mathbf{X} and the end distortion constraint together, the \mathbf{D} that minimizes (5) scales accordingly. Thus, w.l.o.g. we can scale \mathbf{X} to unit variance, find the optimal cache auxiliary \mathbf{Y} and scale that \mathbf{Y} back. This seems redundant for now, but it will help us solve $R_{\text{cache}}(d)$ for $K = 2$, as well as derive some analytic results for high R_{cache} for arbitrary K in the following sections.

III. ZERO TOTAL CONDITIONAL CORRELATION

A good caching strategy is thus an auxiliary \mathbf{Y} that minimizes the total correlation in the distribution $\mathbf{X}|\mathbf{Y}$. Unfortunately, it required a minimization over a non-convex domain. As it turns out, the problem is much easier if the cache can reduce that total correlation all the way to zero. Evidently, R_{cache} needs to be sufficiently large to do so. This section discusses the high cache-rate region.

In fact, $T(\mathbf{X}|\mathbf{Y}) = 0$ corresponds to the work on $K = 2$ of [8] and [6]. Those authors sought rate-tuples on the Gray-Wyner network whose sum equaled the joint rate-distortion function. This is possible if and only if the common auxiliary \mathbf{Y} makes the components of \mathbf{X} conditionally independent. The minimum rate needed to achieve this independence is the notion of Wyner's common information (1), which we here generalize to K sources. To that end, define:

$$C_W(\mathbf{X}) \stackrel{\text{def}}{=} \inf_{\mathbf{Y}: T(\mathbf{X}|\mathbf{Y})=0} I(\mathbf{X}; \mathbf{Y}),$$

where $T(\mathbf{X}|\mathbf{Y})$ is the total conditional correlation as defined in (9). Then for Gaussians we can show the following:

Theorem 2. *For Gaussian \mathbf{X} the common information equals:*

$$C_W(\mathbf{X}) = \min_{\mathbf{D}} \frac{1}{2} \log \frac{|\Sigma_{\mathbf{X}}|}{|\mathbf{D}|} \quad \text{s.t.} \quad \begin{cases} \mathbf{D} \preceq \Sigma_{\mathbf{X}}, \\ \mathbf{D} \text{ is diagonal.} \end{cases} \quad (10)$$

Proof. The proof uses standard arguments:

$$\begin{aligned} I(\mathbf{X}; \mathbf{Y}) &\geq I(\mathbf{X}; \mathbb{E}[\mathbf{X}|\mathbf{Y}]) \\ &\geq \frac{1}{2} \log \frac{|\Sigma_{\mathbf{X}}|}{|\mathbf{D}|} \\ &\geq \frac{1}{2} \log \frac{|\Sigma_{\mathbf{X}}|}{\prod_i D_{i,i}}. \end{aligned}$$

The first line follows from the data processing inequality and the second follows from the Gaussian rate distortion function being the lower bound to any $\mathbf{D} \preceq \Sigma_{\mathbf{X}}$ (this inequality is in turn implied by using $\mathbb{E}[\mathbf{X}|\mathbf{Y}]$, see [9, Lemma 2 and 3]). The last line is due to the Hadamard inequality, which is met with equality if and only if \mathbf{D} is diagonal. A diagonal \mathbf{D} stands for zero correlation, which may not guarantee independence. but for Gaussians zero correlation and independence do have an if and only if relationship. Since all $\mathbf{D} \preceq \Sigma_{\mathbf{X}}$ are achievable by Gaussian distributions (also shown in [9]), the lower bound is also achievable; hence the theorem. \square

Note the similarity between (5) and (10). The objective functions are equal. However, total correlation is captured in the non-linear constraint on $\prod_i D_{i,i}$, whereas common information lies in the linear constraint $(\mathbf{D} - \text{diag}(\mathbf{D})) = 0$. Consequently, the latter is efficiently solvable by linear programming. Identifying *all* the common information therefore poses no challenge, but when we can only code parts of it, it is unclear how to identify which part is most important.

Theorem 2 has the following consequence: If the cache can make the components of \mathbf{X} conditionally independent, the trade-off between R_{cache} and \bar{R}_{update} will be a straight line; there will be no more correlation the cache could have benefited from. To illustrate this, define $\mathbf{D}_{C_W(\mathbf{X})}$ to be the distortion profile that achieves $C_W(\mathbf{X})$, then:

Corollary 2. $R_{\text{cache}}(d) \geq \frac{1}{2} \log \frac{|\Sigma_{\mathbf{X}}|}{d}$. If $D_{\text{final}} \cdot \mathbf{I} \preceq \mathbf{D}_{C_W(\mathbf{X})}$, then we have equality for $d \leq |\Sigma_{\mathbf{X}}| \cdot 2^{-2C_W(\mathbf{X})}$.

Proof. The lower bound on $R_{\text{cache}}(d)$ follows from plugging the Hadamard inequality, $|\mathbf{D}| \leq \prod_i D_{i,i} = d$, into (5). Since $C_W(\mathbf{X})$ is the minimum rate needed to have an achievable diagonal \mathbf{D} , we cannot have equality for $R_{\text{cache}} < C_W(\mathbf{X})$.

Let $\mathbf{D}_{C_W(\mathbf{X})}$ be the distortion profile that achieves $C_W(\mathbf{X})$. If D_{final} is smaller than any of the diagonal entries of $\mathbf{D}_{C_W(\mathbf{X})}$, then there exist infinitely many diagonal distortions \mathbf{D}' that satisfy $D_{\text{final}} \cdot \mathbf{I} \preceq \mathbf{D}' \preceq \mathbf{D}_{C_W(\mathbf{X})}$. Since they are all diagonal, they all hit equality on the Hadamard inequality and span all $R_{\text{cache}} \in [C_W(\mathbf{X}), \frac{1}{2} \log \frac{|\Sigma_{\mathbf{X}}|}{D_{\text{final}}}]$. We find the threshold on d by observing that at $R_{\text{cache}}(d) = C_W(\mathbf{X})$ we have

$$|\Sigma_{\mathbf{X}}| \cdot 2^{-2C_W(\mathbf{X})} = |\mathbf{D}| = \prod_i D_{i,i} = d. \quad \square$$

They above simply says that if $R_{\text{cache}} \geq C_W(\mathbf{X})$ then we have that $R_{\text{cache}} + \sum_k R_{u,k} = R_{\mathbf{X}}(D_{\text{final}}, \dots, D_{\text{final}})$; all rates sum up to the joint rate-distortion function. This requires that D_{final} is smaller than the smallest non-zero entry in $\mathbf{D}_{C_W(\mathbf{X})}$. When we said we wanted D_{final} to be small, this is it.

Xu et al. [8] analytically solved (10) for $K = 2$ which is $C_W(X_1, X_2) = \frac{1}{2} \log \frac{1+|\rho|}{1-|\rho|}$ and found that their result was extendable to K Gaussians whose pairwise correlations all equal ρ . In that special case, $C_W(\mathbf{X}) = \frac{1}{2} \log \left(1 + \frac{K|\rho|}{1-|\rho|} \right)$. In fact, their results are part of a much wider class of sources, i.e., Gaussians whose correlation matrix is circulant:

Corollary 3. For any Gaussian source \mathbf{X} of dimension K and whose covariance $\Sigma_{\mathbf{X}}$ is circulant, we have that

$$C_W(\mathbf{X}) = \frac{1}{2} \log \frac{|\Sigma_{\mathbf{X}}|}{\lambda_{\min}(\Sigma_{\mathbf{X}})^K},$$

where $\lambda_{\min}(\Sigma_{\mathbf{X}})$ is the smallest eigenvalue of $\Sigma_{\mathbf{X}}$.

Proof. First, the solution to (10) yields a unique distortion profile \mathbf{D} . In [11], among others, the strict convexity of the Gaussian channel was discussed. Our formulation, though, more closely resembles a geometric problem: $\min -\log |\mathbf{D}|$ under linear constraints corresponds to finding a maximum volume ellipsoid inscribed inside a convex body. It is called the (Löwner–)John ellipsoid and was shown in [12] to be unique.

Secondly, uniqueness implies that a circulant $\Sigma_{\mathbf{X}}$ will yield a circulant \mathbf{D} as the minimizer of (10). Consider the search space of all diagonal matrices \mathbf{D} that satisfy $\mathbf{D} \preceq \Sigma_{\mathbf{X}}$. If it holds that $D_{i,i} \neq D_{j,j}$ for some $i \neq j$, then there must exist other feasible distortion profiles with the same determinant (and hence objective value). Namely, $\Sigma_{\mathbf{X}}$ is circulant and thus one could swap components $D_{i,i}$ and $D_{j,j}$ and find a profile with the same determinant that also satisfies $\mathbf{D} \preceq \Sigma_{\mathbf{X}}$.

Any \mathbf{D} whose diagonal entries are not equal cannot be unique and can thus not be the solution to (10). Concluding that all non-zero entries must be equal, one can verify that $\mathbf{D} = \lambda_{\min}(\Sigma_{\mathbf{X}})\mathbf{I}$ has the largest determinant among all scaled identity matrices \mathbf{I} subject to $\mathbf{D} \preceq \Sigma_{\mathbf{X}}$. As a side note, by Lemma 1 it suffices that the correlation matrix is circulant. \square

IV. NON-ZERO TOTAL CONDITIONAL CORRELATION

If R_{cache} is too small to cache all the common information, one should code most of it. Recall that the general notion of total conditional correlation $T(\mathbf{X}|\mathbf{Y})$ was associated to the non-convex $R_{\text{cache}}(d)$ (5). One thing is for sure for low cache rates, $\mathbf{D} \preceq \Sigma_{\mathbf{X}}$ is an active constraint. Consequently, if \mathbf{D}^* is the minimizer of (5) then it must be that $\text{null}(\Sigma_{\mathbf{X}} - \mathbf{D}^*) \neq 0$; the optimal distortion profile takes on the form

$$\mathbf{D}^* = \Sigma_{\mathbf{X}} - \mathbf{V}\mathbf{V}^T,$$

where $\mathbf{V} \in \mathbb{R}^{K \times M}$ with $M < K$; correlation is minimized by a rank deficient correction along some subspace of $\Sigma_{\mathbf{X}}$.

In this section, we show that for bivariate Gaussians this subspace is simply the eigenspace of the correlation matrix. Afterwards, we show by counterexample that this cannot be the case in general for $K > 2$.

A. Bivariates

Assume $\Sigma_{\mathbf{X}}$ is normalized, otherwise apply Lemma 1.

Theorem 3. For bivariate (unit-variance) Gaussians, we have

$$R_{\text{cache}}(d) = \begin{cases} \frac{1}{2} \log \frac{1-\rho^2}{d} & \text{if } d \leq (1-|\rho|)^2, \\ \frac{1}{2} \log \frac{1-\rho^2}{d-(|\rho|-(1-\sqrt{d}))^2} & \text{otherwise,} \end{cases}$$

where $d = D_{\text{final}}^2 \cdot 2^{4\bar{R}_{\text{update}}} \in [D_{\text{final}}^2, 1]$.

Proof. The case of $d \leq (1-|\rho|)^2$ corresponds to all $R_{\text{cache}}(d) \geq C_W(\mathbf{X})$ (Corollary 2 and 3).

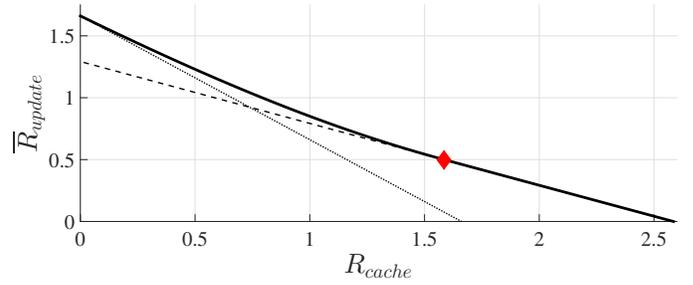


Fig. 3. Example with $\sigma_1^2 = \sigma_2^2 = 1$, $\rho = 0.8$ and $D_{\text{final}} = 0.1$. The thick line represents the cache-update rate trade-off of Theorem 3 and (6)-(7). The diamond corresponds to $R_{\text{cache}} = C_W(\mathbf{X})$. The dashed line is the lower bound of Corollary 2. The dotted straight line goes between $(0, \frac{1}{2} \log \frac{\sigma_1^2}{D_{\text{final}}^2})$ and $(\frac{1}{2} \log \frac{\sigma_1^2}{D_{\text{final}}^2}, 0)$, assuming the cache knows the request upfront.

For the low-rate regime, $\mathbf{D} \preceq \Sigma_{\mathbf{X}}$ must be an active constraint, i.e., the equality is not strict and $\Sigma_{\mathbf{X}} - \mathbf{D}$ is singular. The latter implies that we can model the optimal distortion profile as a rank-one correction:

$$\mathbf{D} = \Sigma_{\mathbf{X}} - \alpha \mathbf{v}\mathbf{v}^T,$$

for some normalized vector \mathbf{v} and scalar α . Let us maximize the determinant of this expression, as that is the core of (5):

$$\begin{aligned} R_{\text{cache}}(d) &\rightarrow \max_{\mathbf{D}} |\mathbf{D}| \\ &= \max_{\alpha, \mathbf{v}} (1 - \alpha v_1^2)(1 - \alpha v_2^2) - (\rho - \alpha v_1 v_2)^2 \\ &\leq \max_{\alpha, \mathbf{v}} d - (\rho - \alpha v_1 v_2)^2. \end{aligned} \quad (11)$$

The last step follows from the constraint $\prod D_{i,i} = d$. Let us continue by finding the argument that maximizes the above:

$$\arg \min_{\alpha, \mathbf{v}} (\rho - \alpha v_1 v_2)^2 = \arg \max_{\alpha, \mathbf{v}} \alpha v_1 v_2.$$

Assume w.l.o.g. that $\{v_1, v_2\}$ are normalized (α can take care of proper scaling). Then we can add the constraint that $v_1^2 + v_2^2 = 1$. We thus end up at maximizing a product of numbers under a sum constraint, which is known to be solved by taking all numbers equal. Thus,

$$v_1 = v_2 = 1/\sqrt{2}.$$

Plugging in $(1 - \alpha v_1^2)(1 - \alpha v_2^2) = d$ from (11) tells us that $\alpha = 2(1 - \sqrt{d})$, and thus

$$\mathbf{D} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} - (1 - \sqrt{d})\mathbf{1}\mathbf{1}^T. \quad (12)$$

The above holds for $\rho > 0$. Otherwise, one can verify we should take $v_1 = -v_2$. Both cases combined gives us

$$|\mathbf{D}| = d - (|\rho| - (1 - \sqrt{d}))^2.$$

The two other conditions in (5), $D_{i,i} \geq D_{\text{final}}$ and $\mathbf{D} \preceq \Sigma_{\mathbf{X}}$, pose no problem if $d = D_{\text{final}}^2 \cdot 2^{4\bar{R}_{\text{update}}} \in [D_{\text{final}}^2, 1]$. Namely, $D_{1,1} = D_{2,2} = \sqrt{d}$ by (12) and there will be no negative update-rates or negative cache-rate. \square

The insight of (12) is that $\frac{1}{\sqrt{2}}\mathbf{1}$ (resp. $\frac{1}{\sqrt{2}}[1, -1]^T$) is the dominant eigenvector of *any* bivariate correlation matrix with $\rho > 0$ (resp. $\rho < 0$). For bivariate, the optimal caching strategy is thus a reversed water filling procedure on the eigenvalues of the correlation matrix: For low cache rates one should only code λ_{\max} until some distortion. If $R_{\text{cache}} = C_W(\mathbf{X})$ the above results in a diagonal distortion matrix. Then for $R_{\text{cache}} > C_W(\mathbf{X})$ one should keep X_1 and X_2 conditionally independent. The resulting trade-off between R_{cache} and \bar{R}_{update} is plotted in Figure 3.

B. Multivariates of dimension larger than 2

The bivariate result of Theorem 3 does not extend to $K > 2$: the eigenvalues of the correlation matrix do *not* capture total correlation. Consider this counterexample: For small D_{final} , the threshold between the characterizable high- R_{cache} region and the low rate region lies where total conditional correlation can be made zero, i.e., at $R_{\text{cache}} = C_W(\mathbf{X})$. Let $\mathbf{D}_{C_W(\mathbf{X})}$ denote the distortion profile that achieves $C_W(\mathbf{X})$. Then for $K = 3$ we have the following example:

$$\Sigma_{\mathbf{X}} = \begin{bmatrix} 1 & 2/3 & 1/3 \\ 2/3 & 1 & 1/3 \\ 1/3 & 1/3 & 1 \end{bmatrix} \Rightarrow \mathbf{D}_{C_W(\mathbf{X})} = \begin{bmatrix} 1/3 & 0 & 0 \\ 0 & 1/3 & 0 \\ 0 & 0 & 5/6 \end{bmatrix}.$$

If the bivariate result did extend to higher dimensions, the implication would have been that $\mathbf{D}_{C_W(\mathbf{X})}$ is the result of an operation on the dominant eigenvalues of $\Sigma_{\mathbf{X}}$ only. Contrarily,

$$\mathbf{D}_{C_W(\mathbf{X})} = \Sigma_{\mathbf{X}} - \frac{3}{2} \cdot \begin{bmatrix} 2/3 & 2/3 & 1/3 \\ 2/3 & 2/3 & 1/3 \\ 1/3 & 1/3 & 1 \end{bmatrix}.$$

The entire common information is captured in an elegant and structured subspace, whereas the eigenvectors can be verified to be different and not nearly as nice. This single rate point disproves that total correlation is in general minimized by corrections that commute with the eigenspace of $\Sigma_{\mathbf{X}}$.

Instead, we conjecture the following: $T(\mathbf{X}|\mathbf{Y})$ of K Gaussians is minimized by a reversed water filling procedure [10, Section 10.3.3] on the eigenvalues of the correlation matrix if and only if this matrix is circulant. We have not been able to disprove our claim by exhaustive search over all possible distortion profiles. Recall that the special case of common information $C_W(\mathbf{X})$ showed similar behavior: it was an optimization problem in general, with an analytic solution for circulants.

If our conjecture is true, Theorem 3 is a logical consequence of the fact that all *bivariate* correlation matrices are circulant and it explains why eigenvalue-operations are insufficient in general for $K > 2$. In addition, our conjecture would again for circulants smoothly connect the low- R_{cache} case to the known high- R_{cache} region at Corollary 3.

If one can prove that the solution to (5) is unique, an argument like the one of Corollary 3 proves the conjecture. In case one cannot prove uniqueness, there are some indications in favor of the hypothesis: (5) tries to find a \mathbf{D} with $\max |\mathbf{D}|$, whose eigenvectors are as close to the identity matrix (as this

means independence) as possible under the constraint $\mathbf{D} \preceq \Sigma_{\mathbf{X}}$. To find a \mathbf{D} with a larger determinant whose eigenvectors are closer to \mathbf{I} , it is necessary that the dominant eigenvectors of $\Sigma_{\mathbf{X}}$ are leaning towards some unit basis vector(s). Circulant $\Sigma_{\mathbf{X}}$, however, have no such bias: $\lambda_0(\Sigma_{\mathbf{X}})$ is associated to the eigenvector $\mathbf{1}$, $\lambda_{K-1}(\Sigma_{\mathbf{X}})$ to $[+1, -1, +1, -1 \dots]^T$ (if K even) and all other λ come in pairs of two and have no uniquely defined eigenvectors. Consequently, one can observe in simulations that the \mathbf{D} that minimizes (5) keeps the same eigensystem as $\Sigma_{\mathbf{X}}$. Any non-circulant $\Sigma_{\mathbf{X}}$ has eigenvectors that show at least some bias towards a certain direction and we observe that the eigenbasis of the optimal \mathbf{D} follows this bias to approach \mathbf{I} as R_{cache} increases.

V. CONCLUSIONS

We introduced a caching problem for Gaussian multivariates and discussed the trade-off between cache and average update rate. A good caching strategy codes as much of the common information as possible. If the cache rate is sufficiently large, characterizing all common information of K Gaussians is a simple linear program. For small rates and assuming Gaussian auxiliaries, coding *most* of the common information was captured in the general notion of total (conditional) correlation; this was unfortunately an optimization problem over a non-convex domain. Analytic expressions we found for common information for any K and total correlation for $K = 2$ both revolved around circulant correlation matrices. Future efforts should either try to tackle the non-convexity of total conditional correlation, or leverage matrix structure further.

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] 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.
- [4] —, "Information-theoretic caching: Sequential coding for computing," 2015. [Online]. Available: <http://arxiv.org/abs/1504.00553>
- [5] S. Watanabe, "Information theoretical analysis of multivariate correlation," *IBM Journal of Research and Development*, vol. 4, no. 1, pp. 66–82, Jan 1960.
- [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, "Wyner's 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.
- [8] —, "Wyner's common information: Generalizations and a new lossy source coding interpretation," 2013. [Online]. Available: <http://arxiv.org/abs/1301.2237>
- [9] 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.
- [10] T. M. Cover and J. A. Thomas, *Elements of information theory (2. ed.)*. Wiley, 2006.
- [11] Y.-H. Kim and S.-J. Kim, "On the convexity of $\log \det (\mathbf{I} + K \mathbf{x} \mathbf{x}^T)$," 2006. [Online]. Available: <http://arxiv.org/abs/cs/0611043>
- [12] O. Güler and F. Gürtuna, "Symmetry of convex sets and its applications to the extremal ellipsoids of convex bodies," *Optimization Methods and Software*, vol. 27, no. 4-5, pp. 735–759, 2012.