Caching (a pair of) Gaussians

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

Abstract

A source produces i.i.d. vector samples from a Gaussian distribution, but the user is interested in only one component. In the cache phase, not knowing which component the user is interested in, a first compressed description is produced. Upon learning the user's choice, a second message is provided in the update phase so as to attain the desired fidelity on that component. We aim to find the cache strategy that minimizes the average update rate. We show that for Gaussian codebooks, the optimal strategy depends on whether or not the cache is large enough to make the vector conditionally independent. If it is, infinitely many equally optimal strategies exist. If it is not, we show that the encoder should project the source onto some subspace prior to coding. For a pair of Gaussians, we exactly characterize this projection vector.

1 Introduction

Nowadays, streaming-services draw a huge chunk of the available bandwidth. The *on-demand* aspect of video-on-demand results in an overload of individual requests at slightly different times of the day, albeit concentrated during peak hours. Caching is a strategy to move part of that load to off-peak times. During the night, a service could pre-load data onto your hard drive, taking an estimated guess of the content you might ask for during the day. If a user has a limited cache budget on his drive, what should the server put there in order to minimize traffic during the day? In this paper, we study these applied questions in a theoretical context of Gaussian vector sources.

A source produces length K vector samples of a Gaussian distribution, but the user is only interested in one of the components. In the cache phase, the encoder can code a first message up to cache rate R_c, without knowing the user's desired component. In the update phase, the user chooses component k uniformly, i.e., $p_Y(Y = k) = 1/K$ and reveals it to the encoder, who then sends an update at a rate R_u. The decoder then uses the cache and this update to construct a lossy representation of the k'th component at the desired fidelity. A schematic of this is depicted in Figure 1. Our goal is to find the caching strategy that minimizes the average update rate.

We show in this paper that for Gaussian codebooks the optimal coding strategy depends on whether or not the cache is sufficiently large so as to make the source components conditionally independent when conditioned on the cache. If that is so, Section 3.1 explains that there are infinitely many coding strategies that are equally optimal. If not, we argue in Section 3.2 that the encoder must project the source vector to a shorter vector. For a pair of Gaussians, we find this projection exactly; it turns out to be solely defined by the source's covariance and it does not change for different values of R_c.

All that we discuss in this paper relies on the successive refinability of Gaussian sources to connect the cache and update phase. For a general discussion we refer the reader to [1, 2]. For the (Gaussian) vector case, one should read [3] and [4] as its precursor. It describes the refinability of $\mathbf{X}_1$ to $\mathbf{X}_2$ as being possible if and only if their

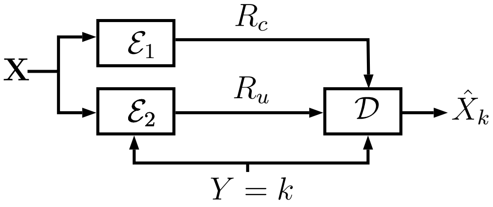

Figure 1: The caching scheme with cache rate R_c. After revealing $Y = k$ as the selection variable, $\mathcal{E}_2$ sends an update at rate R_u s.t. the decoder can retrieve $\hat{X}_k$.

covariances admit a semidefinite ordering $\Sigma_{X_1} \succeq \Sigma_{X_2}$. A discussion on the general rate-distortion function for Gaussian vectors was discussed in [5]. An attentive reader might also notice that our problem shows close resemblance to the Gray-Wyner system [6]. Namely, one could draw all the events of the user asking for one of the K components as K different decoders; the cache would then be their shared link and the required update their individual ones [7].

2 Definitions and Cache Rate-Distortion Function

Let $\mathbf{X}$ be an i.i.d. Gaussian vector source of dimension K, following the distribution $\mathcal{N}(0, \Sigma_{\mathbf{X}})$ with some potentially correlated covariance $\Sigma_{\mathbf{X}}$. That is, at each time instant, the source independently produces a vector sampled from this fixed distribution. We denote the source sample at time n by $\mathbf{X}(n)$, and we denote its kth component by $X_k(n)$, for $k = 1, 2, \dots, K$. Independently of $\mathbf{X}$, a single random variable Y is drawn from the set $\{1, 2, \dots K\}$ uniformly at random; we call it the selection variable.

We consider block coding of length N with two encoders. The first, referred to as the *cache encoder,* observes only $\{\mathbf{X}(n)\}_{n=1}^N$ and produces a description using NR_c bits, where R_c is called the *cache rate.* The second, referred to as the *update encoder,* gets to observe $\{\mathbf{X}(n)\}_{n=1}^N$ as well as the value of the random variable $Y = k$ and produces a description using $NR_u(k)$ bits, where $R_u(k)$ is called the *update rate* for the case $Y = k$. Hence, the average update rate of the encoder is given by $\overline{R}_u = \frac{1}{K}\sum_{k=1}^K R_u(k)$. Notation-wise, the sub- or superscript c stands for cache, while u stands for update.

Upon observing the realization y and both compressed descriptions, the decoder must output a sequence of estimates $\hat{X}_y(n)$ in such a way as to satisfy

$$\frac{1}{N}\sum_{n=1}^{N}\left(X_y(n) - \hat{X}_y(n)\right)^2 \leq D_u.$$

The question addressed in this paper is to characterize, for a fixed caching rate R_c, the smallest average update rate $\overline{R}_u$ for which the distortion constraint can be satisfied (irrespective of the value of Y).

For the cache phase, we allow the server to code any $\mathbf{Z}$ that is jointly Gaussian with the source. For large R_c, one can easily argue that Gaussian codebooks are optimal; for small R_c, it remains a difficult question that we unfortunately can not yet address in this article. In the update phase, one can compute $\hat{\mathbf{X}}^c = \mathbb{E}[\mathbf{X}|\mathbf{Z}]$ as the MSE-estimate of the source and subsequently the error as:

$$\mathbf{D}^c = \mathbb{E}[(\mathbf{X} - \hat{\mathbf{X}}^c)(\mathbf{X} - \hat{\mathbf{X}}^c)^T] \preceq \Sigma_{\mathbf{X}}. \tag{1}$$

The semidefinite ordering $\mathbf{D}^c \preceq \Sigma_{\mathbf{X}}$ means that $\Sigma_{\mathbf{X}} - \mathbf{D}^c$ is positive semidefinite. It yields an engineering perspective: any real symmetric matrix $\mathbf{D}^c$ that satisfies this ordering is an achievable Gaussian codebook.

At this point, there is no operational interest for a first estimate $\hat{\mathbf{X}}^c$ or its error $\mathbf{D}^c$. However, $\mathbf{D}^c$ has theoretical value. Namely, for any $\hat{\mathbf{X}}$ jointly (not necessarily Gaussian) distributed with $\mathbf{X}$, the mutual information satisfies

$$I(\mathbf{X}; \hat{\mathbf{X}}) \geq \frac{1}{2} \log \frac{|\Sigma_{\mathbf{X}}|}{|\mathbf{D}|} = R(\mathbf{D}). \tag{2}$$

The last step in (2) is met with equality if indeed we use Gaussians codebooks, i.e., $\hat{\mathbf{X}}^c = \mathbb{E}[\mathbf{X}|\mathbf{Z}]$ with $\mathbf{Z} = \mathbf{X} + \mathbf{W}$ where $\mathbf{W}$ is independent from $\mathbf{X}$ and Gaussian as well [5, Lemma 2]. Thus, we may not need $\mathbf{D}^c$, but we can use it to characterize the rate associated with the cache phase. Therefore, a cache strategy that yields a particular error covariance $\mathbf{D}^c$ must have had a rate satisfying

$$R_c \geq \frac{1}{2} \log \frac{|\Sigma_{\mathbf{X}}|}{|\mathbf{D}^c|}. \tag{3}$$

Since our goal is to minimize $\overline{R}_u$ for a *fixed* R_c we can reverse (3) and state the following:

Definition 1. *A (valid)* caching strategy *is any real symmetric matrix* $\mathbf{D}^c$ *that satisifies the following two conditions:*

1. $|\mathbf{D}^c| = |\Sigma_{\mathbf{X}}| e^{-2R_c}$*, the rate-constraint* (3)*.*
2. $0 \preceq \mathbf{D}^c \preceq \Sigma_{\mathbf{X}}$*, the semidefinite ordering constraint* (1)*.*

In the update phase, $Y = k$ is revealed and consequently only an interest for X_k remains. Both the encoder and decoder have access to the side information presented by the cache. The MSE-estimator $\mathbb{E}[X_k|\mathbf{Z}]$ forms the first step to an estimate $\hat{X}_k$ and since $p(X_k|\mathbf{Z})$ is also a normal distribution, the update rate is lower bounded by the Gaussian rate distortion function. Namely, Gaussians are successively refinable [1, 3], which allows to combine the messages from the first and second phase. The variance of $p(X_k|\mathbf{Z})$ is simply the k'th diagonal entry of $\mathbf{D}^c$ and thus we have:

$$R_u(k) \geq \frac{1}{2} \log^+ \frac{D^c_{kk}}{D_u},$$

which yields an *average* update rate for this construction:

$$\overline{R}_{u,\mathbf{D}^c}(D_u) = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{2} \log^+ \frac{D^c_{kk}}{D_u}. \tag{4}$$

The subscript $\mathbf{D}^c$ emphasizes that $\overline{R}_u$ depends on a particular cache strategy $\mathbf{D}^c$.

Definition 2. *The* cache rate-distortion function *is the average update rate needed to attain distortion* D_u *on any component, minimized over all caching strategies:*

$$\overline{R}_u(D_u) = \min_{\mathbf{D}^c} \quad \overline{R}_{u,\mathbf{D}^c}(D_u) \quad s.t. \begin{cases} 0 \preceq \mathbf{D}^c \preceq \Sigma_{\mathbf{X}} \\ |\mathbf{D}^c| = |\Sigma_{\mathbf{X}}| e^{-2R_c} \end{cases} \tag{5}$$

Our search for the best caching strategy thus translates to: What choice of $\mathbf{D}^c$ minimizes (4) given the search space of matrices set by Definition 1? What distortion profile for the cache phase minimizes that rate needed for the update phase?

Unfortunately, the cache rate-distortion function (5) is a minimization over a concave function. In many Gaussian source coding problems, the optimization variable $\mathbf{D}$ is found in the denominator, which is convex. It is now found in the numerator, which makes it concave and thus hard to solve. In the next section, we will argue on the different optimal caching strategies for small and large R_c.

3 Optimal Caching Strategies

A different way of writing (5) is to pull the sum of (4) inside the log:

$$\overline{R}_u(D_u) = \min_{\mathbf{D}^c} \frac{1}{2K} \log^+ \frac{\prod_k D^c_{kk}}{D_u^K} \quad \text{s.t.} \quad \begin{cases} 0 \preceq \mathbf{D}^c \preceq \Sigma_{\mathbf{X}} \\ |\mathbf{D}^c| = |\Sigma_{\mathbf{X}}| e^{-2R_c} \end{cases} \tag{6}$$

which leads to the insight that the numerator is lower bounded by the Hadamard inequality $\prod_k \mathbf{D}^c_{k,k} \geq |\mathbf{D}^c|$, hence

$$\overline{R}_u(D_u) \geq \frac{1}{2K} \log^+ \frac{|\mathbf{D}^c|}{D_u^K}, \tag{7}$$

and in turn $|\mathbf{D}^c|$ is bounded (or fixed even) by R_c, see again Definition 1. An interesting read on the relationship between the Hadamard inequality and Gaussians was presented in for example [8, Chapter 17]. The difference between a product of the diagonal entries of a covariance and its determinant stems from $h(\mathbf{X}) \leq \sum_k h(X_k)$. The mutual exclusiveness of the update phase, where the encoder only refines the one component the decoder asked for, combined with an objective to minimize the *average* update rate is the reason for why this product $\prod_k D^c_{kk}$ popped up instead of $|\mathbf{D}^c|$.

Interestingly, there are two distinct coding strategies depending on whether the lower bound (7) can be met or not. The Hadamard Inequality is met with equality if and only if the matrix $\mathbf{D}^c$ is diagonal. Algebraically, this is not trivial as one cannot have a diagonal $\mathbf{D}^c$ and satisfy $\mathbf{D}^c \preceq \Sigma_{\mathbf{X}}$ at the same time if $|\mathbf{D}^c|$ is too large. In terms of information theory, a diagonal cache distortion implies that the components of $\mathbf{X}$ become independent when conditioned on the cache. This is impossible if R_c is too small. These algebraic and information theoretic arguments are the same. In the next subsection, we elaborate on a threshold R^* on R_c and show that there are infinitely many equally optimal cache strategies if the rate is larger than R^*. In the subsection thereafter, we show that for smaller rates, the optimal strategy requires a dimensionality reduction. The cache should be a particular projection of the source components to some space. For a pair of Gaussians, we derive this projection exactly.

3.1 Large cache rates

The Hadamard inequality that was the lower bound in (7) hits equality if and only if a matrix is diagonal. Hence there must exist a decomposition $\Sigma_{\mathbf{X}} = \mathbf{D}^c + \Sigma_{\hat{\mathbf{X}}}$ where $\Sigma_{\hat{\mathbf{X}}}$ and $\mathbf{D}^c$ are both positive semidefinite* and $\mathbf{D}$ is diagonal. For this we derive:

Theorem 1. *For any cache rate $R_c \geq R^*$, there exists a caching strategy $\mathbf{D}^c$ that achieves the lower bound on the average update rate* (7)*, where R^* is the solution to*

$$\min_{\mathbf{D}^c} \frac{1}{2} \log \frac{|\Sigma_{\mathbf{X}}|}{|\mathbf{D}^c|} \quad s.t. \quad \begin{cases} 0 \preceq \mathbf{D}^c \preceq \Sigma_{\mathbf{X}}, \\ \mathbf{D}^c \textit{ is diagonal.} \end{cases} \tag{8}$$

Proof. Recall that $R_c = \frac{1}{2} \log \frac{|\Sigma_{\mathbf{X}}|}{|\mathbf{D}^c|}$ implies the reverse relation on the determinant, $|\mathbf{D}^c| = |\Sigma_{\mathbf{X}}| e^{-2R_c}$. Suppose that $\mathbf{D}^*$ is the distortion matrix that minimizes (8) and let R^* be the cache rate associated to this point. Evidently, there cannot be a $\mathbf{D}'$ that is

*Demanding that $\Sigma_{\mathbf{X}}$ decomposes into a sum of two positive semidefinite matrices is equivalent to demanding $\mathbf{D}^c \preceq \Sigma_{\mathbf{X}}$ like we did before.

diagonal and has a determinant larger than that of $\mathbf{D}^*$ (or equivalenty, an R_c smaller than R^*), otherwise $\mathbf{D}'$ would have been the minimizer of (8). On the other end, for all $R_c \geq R^*$ there does exist a diagonal candidate caching strategy $\mathbf{D}'$. Namely, for diagonal matrices $\mathbf{D}' \preceq \mathbf{D}^*$ holds if and only if $D'_{i,i} \leq D^*_{i,i}$ $\forall i$. So one can construct another distortion matrix $\mathbf{D}'$ by decreasing the values on some arbitrary subset of the diagonal entries of $\mathbf{D}^*$. In doing so, any determinant $|\mathbf{D}'| \leq |\mathbf{D}^*|$ (and thus any $R_c \geq R^*$) can be achieved by a matrix that satisfies the chain $\mathbf{D}' \preceq \mathbf{D}^* \preceq \Sigma_{\mathbf{X}}$ and is thus both diagonal and achievable. □

In the proof we constructed matrices $\mathbf{D}'$ with any particular determinant in the region $R_c > R^*$ by decreasing some diagonal entries of $\mathbf{D}^*$, the solution to (8). It does not matter which entries we use for this or by what amount we decrease them, as long as the resulting determinant has the value we are after. Hence, in this high-R_c regime, there exists infinitely many diagonal $\mathbf{D}'$ with the same determinant that thus all achieve the same lower bound on $\overline{R}_u$ (7); they are equally optimal.

The minimization of (8) is simply a MaxDet problem, which can be solved efficiently numerically. The constraint that $\mathbf{D}^c$ must be diagonal is also a simple linear constraint, namely one can replace it by $\mathbf{D}^c - \text{diag}(\mathbf{D}^c) \preceq 0$ and we already had $\mathbf{D}^c \succeq 0$. This brings about an interesting contrast with the original problem: Finding the general optimal distortion profile for our problem was a hard-to-solve concave minimization. The high-rate regime, however, now appears to be characterizable by a convex problem which is easily solvable. To our knowledge, we do not know of any analytical expression for $\mathbf{D}^c$ that minimizes (8), except for some special cases, one of which we will explain now.

3.1.1 Example: a Pair of Gaussians

Theorem 2. *For a pair of Gaussians the minimizer of* (8) *is* $R^* = \frac{1}{2}\log\frac{1+|\rho|}{1-|\rho|}$, *which is achieved by a distortion matrix*

$$\mathbf{D}^* = \begin{bmatrix} \sigma_1^2(1-|\rho|) & 0 \\ 0 & \sigma_2^2(1-|\rho|) \end{bmatrix}. \tag{9}$$

Proof. Let us find a decomposition of $\Sigma_{\mathbf{X}} = \mathbf{D}^c + \Sigma_{\hat{\mathbf{X}}}$ of a diagonal $\mathbf{D}^c$ by setting $\mathbf{D}^c = \text{diag}(\alpha^2, \beta^2)$ and work out:

$$\begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} \alpha^2 & 0 \\ 0 & \beta^2 \end{bmatrix} + \begin{bmatrix} \sigma_1^2 - \alpha^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 - \beta^2 \end{bmatrix}.$$

Such a decomposition yields positive semidefinite matrices (which is equivalent to $0 \preceq \mathbf{D}^c \preceq \Sigma_{\mathbf{X}}$) and is a valid caching strategy if and only if the following conditions are met:

1. $0 \leq \alpha^2 \leq \sigma_1^2$ and $0 \leq \beta^2 \leq \sigma_2^2$ ($\mathbf{D}^c$ is PSD).
2. $\frac{\rho^2\sigma_1^2\sigma_2^2}{(\sigma_1^2-\alpha^2)(\sigma_2^2-\beta^2)} \leq 1$ ($\Sigma_{\hat{\mathbf{X}}}$ is PSD).
3. $\alpha^2\beta^2 = |\mathbf{D}^c| =$ fixed (cache rate constraint).

Let us start by evaluating the 2nd condition:

$$\begin{aligned} 0 &\leq (\sigma_1^2 - \alpha^2)(\sigma_2^2 - \beta^2) - \rho^2\sigma_1^2\sigma_2^2 \\ &= -\alpha^2\sigma_2^2 - \frac{|\mathbf{D}^c|}{\alpha^2}\sigma_1^2 + |\Sigma_{\mathbf{X}}| + |\mathbf{D}^c|, \end{aligned} \tag{10}$$

where we replaced $\alpha^2\beta^2 = |\mathbf{D}^c|$ and $\beta^2 = \frac{|\mathbf{D}^c|}{\alpha^2}$ by condition 3. Moreover, note that for 2×2 matrices we have $|\Sigma_{\mathbf{X}}| = \sigma_1^2\sigma_2^2(1-\rho^2)$. The right hand side is convex. If it has two roots α_-^2 and α_+^2, all $\alpha^2 \in [\alpha_-^2, \alpha_+^2]$ are valid solutions, given that condition 1 is satisfied. Hence, if there are two roots then there exist infinitely many $\mathbf{D}^c$ that are equally optimal.

Equation (10) has only one root -and thus yields only one optimal $\mathbf{D}^c$- if the minimum of the right hand side is exactly at zero. By setting its derivative to zero, one finds that $\alpha_{min}^2 = \sqrt{|\mathbf{D}^c|\frac{\sigma_1^2}{\sigma_2^2}}$. Substituting α_{min}^2 into (10) and demanding equality:

$$\begin{aligned} 0 &= |\Sigma_{\mathbf{X}}| + |\mathbf{D}^c| - 2\sqrt{|\mathbf{D}^c|\sigma_1^2\sigma_2^2} \\ &= |\mathbf{D}^c|^2 - 2|\mathbf{D}^c|\sigma_1^2\sigma_2^2(1+\rho^2) + (\sigma_1^2\sigma_2^2(1-\rho^2))^2. \end{aligned}$$

This follows from pulling $-2\sqrt{|\mathbf{D}^c|\sigma_1^2\sigma_2^2}$ to the left hand side, squaring both sides and then pulling it back, while at the same time filling in $|\Sigma_{\mathbf{X}}| = \sigma_1^2\sigma_2^2(1-\rho^2)$. This is a new quadratic equation, which now revolves around $|\mathbf{D}^c|$ instead of α^2. Its roots are

$$\begin{aligned} |\mathbf{D}^c|_\pm^* &= \sigma_1^2\sigma_2^2(1+\rho^2) \pm \sqrt{\sigma_1^4\sigma_2^4(1+\rho^2)^2 - \sigma_1^4\sigma_2^4(1-\rho^2)^2} \\ &= \begin{cases} \sigma_1^2\sigma_2^2(1-|\rho|)^2 & \text{valid,} \\ \sigma_1^2\sigma_2^2(1+|\rho|)^2 & \text{invalid (since } |\mathbf{D}^c| > |\Sigma_{\mathbf{X}}| \text{ cannot be).} \end{cases} \end{aligned}$$

This bifurcation point $|\mathbf{D}^c|^*$ corresponds to a cache rate

$$R^* = \frac{1}{2}\log\frac{|\Sigma_{\mathbf{X}}|}{|\mathbf{D}^c|^*} = \frac{1}{2}\log\frac{\sigma_1^2\sigma_2^2(1-\rho^2)}{\sigma_1^2\sigma_2^2(1-|\rho|)^2} = \frac{1}{2}\log\frac{1+|\rho|}{1-|\rho|},$$

and marks the transition from having no to one and then to infinitely many $\mathbf{D}^c$ that have no correlation. We denote the actual distortion profile that achieves this rate $\mathbf{D}^*$ (9) and find it by filling $|\mathbf{D}^c|^* = \sigma_1^2\sigma_2^2(1-|\rho|)^2$ into $\alpha_{min}^2 = \sqrt{|\mathbf{D}^c|\frac{\sigma_1^2}{\sigma_2^2}}$ and $\beta^2 = \frac{|\mathbf{D}^c|}{\alpha^2}$. □

The value $R^* = \frac{1}{2}\log\frac{1+|\rho|}{1-|\rho|}$ also came forward in [7] as Wyner's Common Information for a pair of Gaussians.

3.2 Small Cache Rates for a Pair of Gaussians

For R_c smaller than the R^* of Theorem 1, no $\mathbf{D}^c$ can close the Hadamard inequality, but perhaps we can find another achievable lower bound. Here, we find this optimal strategy for a pair of Gaussians, which shows a strong connection to Theorem 2. For general dimensions, the problem remains open. One thing that is clear is the following:

Lemma 1. *If $R_c \leq R^*$, the $\mathbf{D}^c$ that minimizes* (6) *yields $\mathbf{D}^c \preceq \Sigma_{\mathbf{X}}$, but not $\mathbf{D}^c \prec \Sigma_{\mathbf{X}}$.*

We will not fully prove this here, but imagine that $\bar{\mathbf{D}}$ is some candidate strategy that yields $\bar{\mathbf{D}} \prec \Sigma_{\mathbf{X}}$. Since the ordering is not strict, we have room to rotate the matrix. Determinants are rotation-invariant, hence the R_c required for this rotated distortion profile is the same (3). The key insight is that rotation can always further minimize the product of the main diagonal, e.g., by bringing the matrix closer to eigendecomposition.

The difference between $\mathbf{D}^c \prec \Sigma_{\mathbf{X}}$ and $\mathbf{D}^c \preceq \Sigma_{\mathbf{X}}$ is that the latter implies $\exists v$ such that $v^T(\Sigma_{\mathbf{X}} - \mathbf{D})v = 0$; there exists a direction of which one learns nothing by observing the cache. In other words, the cache encoder must have projected $\mathbf{X}$ onto some subspace prior to coding. For a *pair* of Gaussians, a lower-dimensional coding strategy simply means one codes a representation of $v^T\mathbf{X}$ for any (normalized) vector v in the cache, rather than $\mathbf{X}$ itself. The code in the cache can be represented as $v^T\mathbf{X}+W$ where W is a Gaussian noise and independent of $\mathbf{X}$. Then the error is found as $\mathbf{D}^c = [||\mathbf{X}-\mathbb{E}[\mathbf{X}|v^T\mathbf{X}+W]||^2]$, which can be worked out completely using channel models for lossy representations, e.g., [8, Chapter 10]. In short, any caching strategy featuring a projection to a vector v leads to a Schur complement:

$$\mathbf{D}^c(R_c) = \Sigma_{\mathbf{X}} - (1-e^{-2R_c})\frac{1}{v^T\Sigma_{\mathbf{X}}v}\Sigma_{\mathbf{X}}vv^T\Sigma_{\mathbf{X}}. \tag{11}$$

We specifically express this matrix as a function of R_c. Even the optimal choice of a vector v could in principle be different for different R_c, but for a pair of Gaussians we will prove that this is actually not the case. Note that (11) always satisfies both conditions of Definition 1 by construction. The border case $\mathbf{D}^*$ is also still a one-dimensional coding operation. We derived $\mathbf{D}^*$ algebraically, we can plug it into (11) and solve for the vector v that could have led us to it. The particular vector associated to $\mathbf{D}^*$ turns out to be of more importance than simply the border case:

Theorem 3. *If for a pair of Gaussians $R_c \leq \frac{1}{2}\log\frac{1+|\rho|}{1-|\rho|}$, then the caching strategy that uniquely minimizes* (6) *requires one to code $v^{*T}\mathbf{X}$ with*

$$v^* = \frac{1}{\sqrt{\mathrm{tr}\left(\Sigma_{\mathbf{X}}\right)}}\begin{bmatrix}\sigma_2\\ \mathrm{sign}(\rho)\cdot\sigma_1\end{bmatrix}. \tag{12}$$

Proof. By Lemma 1 we know it suffices to constrain the search space of $\mathbf{D}^c$ to those we can describe by means of (11). Hence, we can plug (11) into (6) and minimize over all v such that $v^Tv=1$. To find the optimal v it suffices to look at $\arg\min\prod_{k=1,2}D^c_{kk}$:

$$\arg\min_{\substack{v\\ :v^Tv=1}}\left(\sigma_1^2 - \frac{1-e^{-2R_c}}{v^T\Sigma_{\mathbf{X}}v}\left(\Sigma_{\mathbf{X}}vv^T\Sigma_{\mathbf{X}}\right)_{1,1}\right)\cdot\left(\sigma_2^2 - \frac{1-e^{-2R_c}}{v^T\Sigma_{\mathbf{X}}v}\left(\Sigma_{\mathbf{X}}vv^T\Sigma_{\mathbf{X}}\right)_{2,2}\right)$$

For a 2×2 matrix, one can work out the expression above by hand; it is not hard, but for length constraints we choose to omit this from this paper. Its derivative with respect to v has a clear root at (12), regardless of R_c. The $\mathrm{sign}(\rho)$ then ensures one picks the minimum rather than a maximum. □

As a closing comment, let us briefly explain where v^* comes from and what it entails. The vector can be found at the border case of $R_c = R^*$ by setting (11) equal to (9) and solve for v. As for the intuition, every positive semidefinite matrix can be uniquely represented by the ellipsoid $\mathcal{E}_{\mathbf{A}} = \{v : v^T\mathbf{A}^{-1}v = 1\}$. Its semiprincipal axes match the eigenvectors of $\mathbf{A}$ and have lengths equal to $\sqrt{\lambda_i}$. In Figure 2 we plot both $\mathcal{E}_{\Sigma_{\mathbf{X}}}$ and $\mathcal{E}_{\mathbf{D}^*}$. Recall that $\mathbf{D}^*$ is the covariance matrix with the largest possible determinant that still satisfies $\mathbf{D}^c \preceq \Sigma_{\mathbf{X}}$ without having any correlation. Since it is the border case, $\mathcal{E}_{\Sigma_{\mathbf{X}}}$ and $\mathcal{E}_{\mathbf{D}^*}$ touch (that is the impact of having $\mathbf{D}^c \preceq \Sigma_{\mathbf{X}}$ rather than $\mathbf{D}^c \prec \Sigma_{\mathbf{X}}$). Even more so, the vector where these ellipses touch is the orthogonal complement to our coding vector v^*; the cache provides information on all directions spanned by the source, except the one orthogonal to the one we coded.

A second consequence is that, since one should use the same vector to code $v^{*T}\mathbf{X}$ for all $R_c \leq R^*$, all resulting $\mathcal{E}_{\mathbf{D}(R_c)}$ touch $\mathcal{E}_{\Sigma_{\mathbf{X}}}$ at this same orthogonal complement. In other words, $\mathcal{E}_{\mathbf{D}^c(R_c)}$ is sandwiched between $\mathcal{E}_{\mathbf{D}^*}$ and $\mathcal{E}_{\Sigma_{\mathbf{X}}}$. The result is that for a

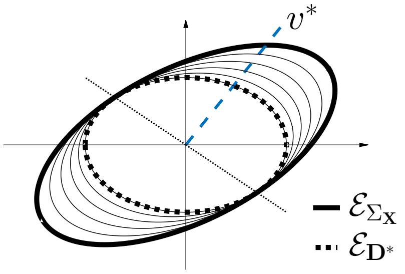
v^*
— $\mathcal{E}_{\Sigma_\mathbf{X}}$
··· $\mathcal{E}_{\mathbf{D}^*}$

Figure 2: Ellipse of $\Sigma_\mathbf{X}$ and $\mathbf{D}^*$, together with the coding transform vector v^* (dashed) and its orthogonal complement ${}^1/_{\sqrt{\mathrm{tr}\,(\Sigma_\mathbf{X})}}[\sigma_1,\ -\mathrm{sign}(\rho)\sigma_2]^T$ (dotted) that intersects the points where both ellipses touch. The thinner ellipses in between are the optimal $\mathbf{D}^c(R_c)$ for increasing R_c, coded with the same v^*, showing the ordering of (13).

sequence of cache rates $0 \leq R_1 \leq R_2 \leq \cdots \leq R_\ell \leq R^*$, the caching strategies that minimize (6) admit a semidefinite ordering:

$$\Sigma_\mathbf{X} \succeq \mathbf{D}^c(R_1) \succeq \mathbf{D}^c(R_2) \succeq \cdots \succeq \mathbf{D}^c(R_\ell) \succeq \mathbf{D}^*. \tag{13}$$

Hence, as a conclusion that stands apart from the goal of this paper, the Gaussian coding strategies that minimize the gap on the Hadamard Inequality for increasing rates form a Markov chain and are because of this successively refinable.

References

[1] W. Equitz and T. Cover, "Successive refinement of information," *Information Theory, IEEE Transactions on*, vol. 37, no. 2, pp. 269–275, 1991.

[2] B. Rimoldi, "Successive refinement of information: characterization of the achievable rates," *Information Theory, IEEE Transactions on*, vol. 40, no. 1, pp. 253–259, 1994.

[3] J. Nayak, E. Tuncel, D. Gunduz, and E. Erkip, "Successive refinement of vector sources under individual distortion criteria," *Information Theory, IEEE Transactions on*, vol. 56, no. 4, pp. 1769–1781, April 2010.

[4] H. Wang and P. Viswanath, "Vector gaussian multiple description with individual and central receivers," *Information Theory, IEEE Transactions on*, vol. 53, no. 6, pp. 2133–2153, June 2007.

[5] 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.

[6] 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.

[7] G. Xu, W. Liu, and B. Chen, "Wyner's common information: Generalizations and A new lossy source coding interpretation," *CoRR*, vol. abs/1301.2237, 2013. [Online]. Available: http://arxiv.org/abs/1301.2237

[8] T. M. Cover and J. A. Thomas, *Elements of information theory (2. ed.)*. Wiley, 2006.