A Gaussian Source Coding Perspective on Caching and Total Correlation

THÈSE N$^{\text{O}}$ 8001 (2017)

PRÉSENTÉE LE 15 SEPTEMBRE 2017
À LA FACULTÉ INFORMATIQUE ET COMMUNICATIONS
LABORATOIRE D'INFORMATION DANS LES SYSTÈMES EN RÉSEAUX
PROGRAMME DOCTORAL EN INFORMATIQUE ET COMMUNICATIONS

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Guillaume Jean OP 'T VELD

acceptée sur proposition du jury:

Prof. B. Rimoldi, président du jury
Prof. M. C. Gastpar, directeur de thèse
Prof. M. Wigger, rapporteuse
Dr J. Goseling, rapporteur
Dr O. Lévêque, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2017

HOUDT HET DAN NOOIT OP?
— Carice van Houten, “*Zwartboek*”

Aan Jeaphianne van Rijn,
je had volkomen gelijk,
sorry.

Acknowledgments

My favorite thesis chapter is the Acknowledgments. Formulas and theorems reflect academic competence, but they tell you nothing about the journey. When I graduated from Eindhoven University of Technology, I sat down in a room alone and actually shed a tear. University felt like a playground of endless opportunities, where students could let their ambitions burst. I had the time of my life. The emotion I feel now, though, as I close the door on EPFL, is relief.

In 2012, Frans Willems tapped me on the back saying “oh by the way, if you want to go to Lausanne, you can”, after storming off to an appointment. That one-line conversation is how I ended up at EPFL. His trust in me opened the door and I am forever in his debt. At the time, I left Eindhoven with enormous momentum. At EPFL, I was never able to attain the same level of success or energy. I struggled to see the practical value in much of what was going on in information theory and I missed the opportunity use my soft skills. Michael’s guidance is probably a big reason in why I still succeeded in the end. Michael excels as a supervisor as he puts the student’s interest first and scientific output of the lab second. This may seem trivial, but is sadly too often not the case in academia.

I also wish to acknowledge my thesis jury: Jasper Goseling, Olivier Lévêque, Michèle Wigger and president Bixio Rimoldi were all exceptionally nice and I appreciate they genuinely read my work. I met Jasper through the *Werkgemeenschap voor Informatie- en Communicatietheorie*, a society that promotes networking among people in the field around the Benelux. Michael has always encouraged me to visit their events and I greatly enjoyed doing so.

In the EPFL-professional context I thank my labmates Chien-Yi, Saeid, Chen, Sung Hoon and Adriano, who were there for most of my time at LINX. I would like to mention our ‘newest recruits’ Su and Erixhen separately; they made the lab significantly more social in my last year. Un merci tout particulier à France, as without her I would have probably been homeless, uninsured and bored. However, saving the best for last: Jingge and Ye. Jingge and I shared an office for years and he might not know, but I felt genuinely sad over his departure. We were always very silent and polite at work, but clicked well on a personal level. Ye has been a great asset as well in making sure we would also *enjoy* our conferences, as opposed to attending too many dreadful talks. Probably thanks to her, I acquired an obsession over spotting sea turtles on holidays. In that light, I also realized that Jingge and Ye are the only persons in my life whom I met on three continents. I challenge them to bring that number up to four.

Daniel ridiculed me sometimes in front of others for describing a PhD like ‘the black hole that slowly eats away your resume’. I did say those words and as I am writing this section I still stand by them. As a disclaimer, I always like to phrase my thoughts fiercely and slightly exaggerated; so a grain of salt is advised. Yet, I do believe that the opportunity cost of doing a PhD is exceptionally high. Imagine yourself sitting in front of a recruiter; how do you pitch the value you either gained or generated in those extra years of academia? It is the struggle I am going through now and I am therefore grateful for having been an active member in the student association ShARE. It has saved my resume from that ‘black hole’, as well as provided me with some greatly needed social interaction outside of the PhD-bubble. Whatever job I will land in next, Galyna and Daniel will have had a great hand in helping me secure it.

I greatly enjoyed having Thijs and Laura around; they were the link to my past here in Switzerland. Back in the Netherlands my special thanks goes out to Gerald and Lieneke. I often feel bad over the amount of hospitality they have shown me, but their open door has allowed me to keep my social circle in Eindhoven alive. The value of that is too big to quantify. Here in Lausanne I met many great people, but a few have been exceptional: Catherine, Victor and Jonas were the driving forces of my social life at the very start. Damian and Katarina (strong competitors alongside Julia and Zygie for the award for most stellar couple) are friends I truly felt I could always rely on. But above all: Adrian and Julia have been there on `days(1:end)` and that is pretty amazing. Julia and I in particular have bonded so strongly over the last two years that our clocks even synced to simultaneous graduation. Often I felt like her secretary or coach, but in the end, everybody needs a person whose office they can enter without knocking. Julia’s office was mine and there is no doubt I will miss her the most.

Years ago, I told a good friend that giving ‘special thanks’ to too many people devaluates its meaning. This honor is best saved for someone who truly stood out among all the people who supported you during these years. After contemplating tirelessly for a full two minutes, I concluded that my special thanks could only go to Jeaphianne van Rijn. Together we cursed, complained and let out our frustrations as rawly as we felt them. We are very much the same: ambitious, craving something meaningful, but instead we found ourselves stuck being insignificant, staring at what felt meaningless. Both of us were caught in academia for all the wrong reasons. I do not know how I would have fared without her echo, as hers was the only one. As one might have noticed, I dedicated this thesis to her. After all, I am the culprit who persuaded her *not* to quit her thesis. For that, Jeaphianne, I am truly sorry.

Lausanne, August 2017

Abstract

Communication technology has advanced up to a point where children are getting unfamiliar with the most iconic symbol in IT: the loading icon. We no longer wait for something to come on TV, nor for a download to complete. All the content we desire is available in instantaneous and personalized streams. Whereas users benefit tremendously from the increased freedom, the network suffers. Not only do personalized data streams increase the load overall, the instantaneous aspect concentrates traffic around peak hours. The heaviest (mostly video) applications are used predominantly during the evening hours.

Caching is a tool to balance traffic without compromising the 'on-demand' aspect of content delivery; by sending data *in advance* a server can avoid peak traffic. The challenge is, of course, that in advance the server has no clue *what* data the user might be interested in. We study this problem in a lossy source coding setting with Gaussian sources specifically, using a model based on the Gray–Wyner network. Ultimately caching is a trade-off between anticipating the precise demand through user habits versus 'more bang for buck' by exploiting correlation among the files in the database.

For *two* Gaussian sources and using Gaussian codebooks we derive this trade-off completely. Particularly interesting is the case when the user has no preference for some content a-priori, caching then becomes an application of the concepts of Wyner's common information and Watanabe's total correlation. We study these concepts in databases of more than two sources where we derive that caching *all* of the information shared by multiple Gaussians is easy, whereas caching *some* is hard. We characterize the former, provide an inner bound for the latter and conjecture for which class of Gaussians it is tight. Later we also study how to most efficiently capture the total correlation that exists between two *sets* of Gaussians.

As a final chapter, we study the applicability of caching of discrete information sources by actually building such algorithms, using convolutional codes to 'cache and compress'. We provide a proof of concept of the practicality for doubly symmetric and circularly symmetric binary sources. Lastly we provide a discussion on challenges to be overcome for generalizing such algorithms.

Keywords: coded caching, source coding, Gaussian distributions, common information, total correlation, Gray–Wyner network, convolutional codes

Résumé

Les technologies de l'information et des communications ont avancé jusqu'à un niveau où les enfants ne connaissent plus le symbole le plus iconique dans l'informatique : l'icône de chargement. Nous ne devons plus attendre la diffusion d'un programme à la télé, ni la fin d'un téléchargement. Tout le contenu que nous désirons est disponible instantanément. Alors que les utilisateurs en bénéficient, le réseau souffre. Les flux de données personnalisés non seulement augmentent le trafic en général, mais l'aspect instantané le concentre aux heures de pointes. Les applications les plus lourdes sont surtout populaires le soir.

'La mise en cache' est une technique qui permet de repartir le trafic sans compromettre la liberté d'une 'diffusion sur demande' ; un serveur peut éviter les heures de pointes en transmettant les données à l'avance. Pourtant, le serveur ne sait évidemment pas à l'avance quelles données seront demandées par l'utilisateur dans le futur. Nous étudions ce problème par un modèle fondé sur le système de Gray–Wyner. Finalement, la mise en cache est un compromis de l'anticipation de la demande exacte fondé sur les habitudes des utilisateurs versus la corrélation entre les fichiers enregistrés dans la base de données.

Nous décrivons complètement ce compromis pour deux sources Gaussiennes en utilisant les codes Gaussiens. Un cas intéressant est lorsque l'utilisateur n'a aucune préférence pour un fichier spécifique. La mise en cache devient alors une application des concepts de l'information commune de Wyner et de la corrélation totale de Watanabe. Nous étudions ensuite ces concepts dans des bases de données de plus de deux sources Gausiennes. Nous découvrons que la mise en cache de toute l'information commune entre divers Gaussiens est facile, alors que faire de même pour seulement une partie est un processus difficile. Le premier est caractérisé et nous proposons une limite pour le dernier. Ensuite, nous étudions comment capturer la corrélation totale entre deux ensembles des sources Gaussiennes le plus efficacement possible.

Le denier chapitre discute l'applicabilité des algorithmes pour la mise en cache des sources discrètes par l'utilisation des codes convolutionnels. Nous apportons une preuve du concept pratique des sources binaire symétrique. Finalement, nous discutons les obstacles à surmonter pour généraliser ces algorithmes.

Mots clefs : cache, Gaussian, l'information commune, la corrélation totale, le système de Gray–Wyner, les codes convolutionnels

Contents

List of Figures

List of Tables

1 Introduction

For every advancement made in communication technologies users become even more demanding. The privilege of being perfectly connected everywhere to endless amount of content becomes so natural that people become increasingly unable to handle setbacks in that connectivity. Speed, responsiveness, storage space, etc. are all expected to only increase monotonically. This makes the work of engineers difficult: for every step forward there is no way back. Hence, we must make sure the gains offset any new problems these advancements might introduce.

The widespread adoption of streaming video services in the last few years has been the biggest gamechanger in online content. Sure, Youtube has been around for longer, but the HD services of Netflix, Hulu and Amazon Prime truly set the standard of what quality users nowadays expect. The 'on-demand' aspect in particular has been the biggest leap in user experience and the biggest headache for communication engineers. The instantaneous and personalized properties of video-on-demand causes network traffic to be concentrated in 'internet rush hour', the evening peak when users all log in simultaneously.

The engineering challenge ahead is to design communication technology that can provide users the freedom of 'on demand' streaming with a balanced network load that providers can handle. Guaranteeing both forces servers to anticipate demand and to communicate data before it is even requested. Of course, correlation does not imply causation, but one cannot help but notice that in this same timeframe *caching* became a hot topic in the information theory community. The most popular caching paper by Maddah-Ali and Niessen [1] was first published on arXiv in the same year in which Netflix tripled her stock value [2].

In this thesis we consider the same idea of coding information in two phases: to first anticipate demand by writing something in the *cache* of the user, followed by responding to an actual request for data by that same person. This second phase is most often called *delivery*, while we prefer the term *update*. The impatience of the encoder makes it almost inevitable that some data in the cache will turn out to be redundant; after all, at first he is blind to what he actually needs to transmit. This loss is accepted in the knowledge that otherwise waiting for a user to make a request would put a heavy burden on the network during those ‘peak hours’. The grand caching question is whether *despite* this loss there are still smart strategies to reduce overall communication rates.

We study these problems in a lossy source coding setting and even more specific: using Gaussian databases. The model is based on the (single user) work of Wang, Lim and Gastpar who ventured in this direction using discrete information sources [3]. Recently this model was extended to a lossy setting as well by Timo, Bidokhti, Wigger and Geiger [4]. Their focus was on designing caching strategies that optimize the worst-case update scenario. Our objective, however, will be to design caching codes that do well on average and to -when possible- also model asymmetric user preference for certain data. In our model, the cache encoder weighs two parameters in its strategy: the preference a user has for some files in the database versus the correlation between those files to be able to transmit data that is useful no matter the user’s request.

For related but a little more distant work one may also consider reading the following: Hassanzadeh, Erkip, Llorca and Tulino also study Gaussian caching in a broadcast scenario to multiple users in a very hands-on, practical work [5]. Yang and Gündüz discuss the multi-user case and distinguish their model by letting the distortion constraints vary per user using a worst-case metric to evaluate update communication rates [6].

The Wang-Lim-Gastpar caching model is closely related to the classic Gray–Wyner network, in which an encoder communicates two files to two decoders (one for each) via one common and two individual communication links [7]. The common link stands analogous to the cache that gets transmitted in any case, whereas the individual ones can be viewed upon as the update messages that get send if either one file or the other is requested. If the user has no preference for any data, the encoder must try to cache as much of the information that is shared by the files as possible. After all, that data is useful no matter what, whereas individual information might go to waste if the user makes a different choice. This perspective opens up a whole range of questions on how to capture the information that is shared between Gaussians. As this thesis progresses, this scenario receives most attention.

Caching without a bias in user preference is therefore closely related to the notions of Wyner’s common information [8] and Watanabe’s total correlation [9]. It is through this analogy with the Gray–Wyner network that these definitions are a right fit for this application rather than, e.g., the Gács-Körner common information [10]. Research that is closely related to this thesis therefore also includes the work of Viswanatha, Akyol and Rose [11] and Xu, Liu and Chen [12]. These authors have introduced and studied *lossy* common information, as well as provided some important characterizations of these properties in Gaussian distributions. These tools help us to identify what common information an encoder needs to cache in order to -for now- speed up research, followed by hopefully one day Netflix.

1.1 Outline and Contributions

Our caching model is motivated by real world applications, yet simultaneously touches upon more fundamental questions on the correlation structure of Gaussian multivariates. As the chapters progress the emphasis slowly moves from being application-driven to focusing more on these fundamental statistical questions. The final chapter is an exception to this trend as it is actually the most applied of all.

- **Chapters 2 and 3** serve as preliminary chapters. The latter addresses some basics on Gaussian source coding, rate-distortion functions and successive refinability. The first chapter covers a novel unification of well-known relationships between positive definite matrices and ellipsoids and aims to better understand information theoretic concepts through geometry. Many results in this thesis will be accompanied by such geometrical pictures to appeal to intuition; those special sections will be clearly marked and Chapter 2 will thus explain how to read them.
- **Chapter 4** starts with the most basic ánd most developed model of caching *bivariate* Gaussian sources. We present a model based on the Gray–Wyner network [7], similar to [3, 4]. Our model model distinguishes itself by focusing on average performance and also including user preference. We provide a full characterization of optimal caching strategies when using Gaussian codebooks. Specifically, we determine two key drivers: user preference and the correlation among the information sources. We address each separately and together:
  - When there is user preference, but the sources are independent: we show that the encoder should cache the most popular source exclusively.
  - When the user's choice is uniform, but the sources are dependent: we show that the encoder should perform caching via a reverse water-filling procedure on the correlation matrix (not the covariance!).
  - When we take both drivers into account we derive the optimal caching strategy, which does not follow a formula of simple intuition like the other cases. In addition, we discuss how caching strategies depend on the size of the cache, how caching is not a successively refinable process and we bound rate loss due to absence of knowledge on user preference.
- **Chapter 5** extends the caching model to databases of more than two information sources, while we take user preference out of the equation. Contributions include:
  - A characterization of high cache rates by deriving Wyner's common information as a convex optimization problem. We also derive this entity analytically for Gaussians whose covariance is circulant.
  - An inner bound for low cache rates using the result of Chapter 4.
  - A conjecture that this inner bound is tight only for Gaussians whose correlation matrix is circulant, accompanied by a discussion how this connects to known results from the same chapter as well as the previous one.

- **Chapter 6** looks at databases of *sets* of Gaussians and studies the concept of total correlation in particular. Contributions include:
  - A derivation of the Wyner's common information that exists between K Gaussian vectors as a convex optimization problem.
  - For 2 Gaussian vectors, we show that the optimal way to capture their total correlation is by first transforming these 2 vectors into independent sets of Gaussian pairs, followed by an optimization over these pairs which resembles reverse water-filling.
- **Chapter 7** stands as the odd one out as it is the only chapter on discrete information sources. We discuss the practicality of caching by building an actual caching algorithm. We use convolutional codes to translate random coding arguments into systems of acceptable block length and running time. Presented are experimental results for the caching of doubly and circularly symmetric binary sources, as well as a discussion on what barriers are to be overcome for implementing more universal caching systems.

1.2 Notation

The star of this thesis is the Gaussian multivariate $\mathbf{X}$, distributed as $\sim \mathcal{N}(\mathbf{0}, \Sigma)$. Unless explicitly stated otherwise, covariance matrices Σ decompose as follows,

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1\sigma_3 & \cdots \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \rho_{23}\sigma_2\sigma_3 & \\ \rho_{13}\sigma_1\sigma_3 & \rho_{23}\sigma_2\sigma_3 & \sigma_3^2 & \\ \vdots & & & \ddots \end{bmatrix}$$

where σ_i is the variance of X_i, and ρ_{ij} the correlation between X_i and X_j. More so than not will unit variance be assumed for simplicity.

Boldface letters are reserved for either vectors (like $\mathbf{v}$) or matrices (like $\mathbf{A}$). Capital letters denote either matrices, random variables or random vectors (like $\mathbf{A}$, X and $\mathbf{X}$ respectively). To avoid confusion between matrices and random vectors, the end of the alphabet is reserved for random variables. $A_{i,j}$ is a single element of the matrix $\mathbf{A}$. As working with Gaussians requires a hefty amount of linear algebra, Table 1.1 summarizes the most used operators in this thesis. The semidefinite ordering symbol $\mathbf{A} \preceq \mathbf{B}$ means $\mathbf{B} - \mathbf{A}$ is positive semidefinite. The strict inequality $\prec$ implies the difference of both matrices is positive definite. The unit basis vectors are denoted $\mathbf{e}_i$, who thus are all zero except for a 1 in the i'th position.

Logarithms are all base-2, and $\log^+(x) = \max(0, \log(x))$; the superscript $^+$ is also used on other occasions to enforce non-negativity. The set $[a\ b]$ with square brackets indicates the set of all elements between a and b, whereas $\{a, b\}$ contains only a and b. $H(X)$ denotes the classic discrete entropy, whereas $h(X)$ stands for the (in this thesis more frequently used) differential entropy and in both the continuous and discrete domain $I(X;Y)$ stands for mutual information. Lastly, $X - Y - Z$ signifies a Markov chain.

Notation	**Meaning**
$\lambda_i(\mathbf{A})$	the i'th eigenvalue of **A**.
$\lambda_{\min}, \lambda_{\max}$	the smallest and respectively largest eigenvalue.
diag(**A**)	a vector containing the diagonal elements of the matrix **A**.
diag(**v**)	a diagonal matrix with non-zero entries equal to the vector **v**.
$\lvert\mathbf{A}\rvert$	the determinant of **A**.
$\lvert\lvert\mathbf{v}\rvert\rvert$	the Euclidean norm of the vector **v**.
ker(**A**)	the kernel (or nullspace) of **A**.
$\dim(\cdot)$	the dimension of a vector space.

Table 1.1 – Most used matrix and vector functions.

2 Preliminaries I: the Geometry to visualize Gaussian Relationships

This chapter complements the thesis; a reader with little time can comfortably skip to Chapter 3 without compromising his or her understanding of any of the material.

Many information theoretic results for Gaussians can also be told in pictures, which may not always help to write proofs, but certainly aids in grasping the intuition. This new graphical language stems from two properties:

- Operations to Gaussian distributions can often be expressed by algebraic operations on covariance matrices.
- Every $K-$dimensional covariance matrix, being real and symmetric, can be drawn as an ellipsoid in $K-$dimensional space.

Consequently, concepts like correlation, independence, estimators as well as information-theoretic properties like mutual information or Wyner's common information can be explained and understood by geometry. Many of the results in this thesis will be accompanied by a plot that illustrates the intuition behind the formula. Look out for the boxed paragraphs at the end of a section:

Theorems in Pictures

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis at eleifend lacus, et efficitur elit. Quisque efficitur auctor ipsum, vel rutrum mauris tincidunt ut. Morbi a ante eget lacus viverra iaculis ut sed metus. Integer iaculis consequat elit eget fringilla. Aliquam erat volutpat. Vestibulum rutrum tortor quis velit pellentesque commodo. Proin mattis arcu quis mauris feugiat aliquam.

2.1 Ellipsoids and Joint Covariance

Every $K \times K$ positive definite matrix corresponds uniquely to an ellipsoid in K-dimensional space, and vice versa. In this thesis, we stick to the following definition:

Definition 2.1. *The ellipsoid corresponding to a positive definite matrix* $\mathbf{A}$ *is the set:*

$$\mathcal{E}_{\mathbf{A}} = \{\mathbf{u} : \mathbf{u}^T \mathbf{A}^{-1} \mathbf{u} = 1\},$$

or, equivalently:

$$\mathcal{E}_{\mathbf{A}} = \{\mathbf{A}^{1/2}\mathbf{v} : ||\mathbf{v}|| = 1\}.$$

When we take a look at the Gaussian distribution

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^K |\Sigma_{\mathbf{X}}|}} e^{-\frac{1}{2}\mathbf{x}^T \Sigma_{\mathbf{X}}^{-1} \mathbf{x}} \tag{2.1}$$

we see that $\mathcal{E}_{\Sigma_{\mathbf{X}}}$ is an equipotential line of this distribution. An example of such an ellipsoid is plotted in Figure 2.1.

Property 2.1. *The semi-principal axes of* $\mathcal{E}_{\mathbf{A}}$ *are the eigenvectors of* $\mathbf{A}$ *with lengths equal to the square root of their respective eigenvalue.*

It is easy to see why. Namely, take the second definition of the ellipsoid: $\mathcal{E}_{\mathbf{A}} = \{\mathbf{A}^{1/2}\mathbf{u} : ||\mathbf{u}|| = 1\}$ and recall that the square root of a symmetric matrix has the same set of eigenvectors and only the eigenvalues have their square root taken. Thus, if $\mathbf{v}_i$ is the i'th eigenvector then

$$\mathbf{A}^{1/2}\mathbf{v}_i = \sqrt{\lambda_i(\mathbf{A})}\mathbf{v}_i. \tag{2.2}$$

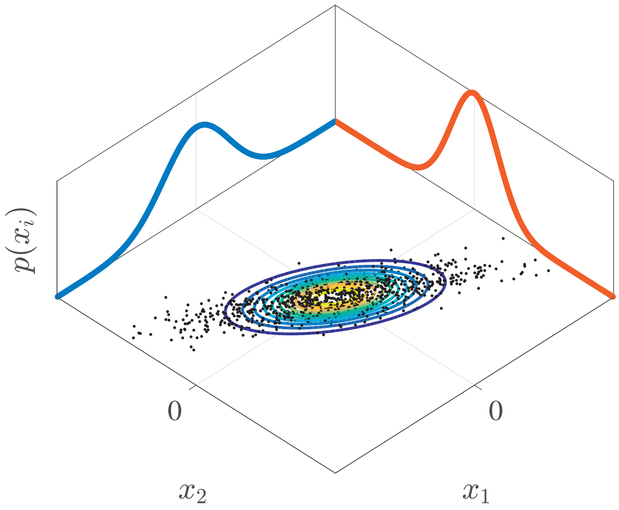
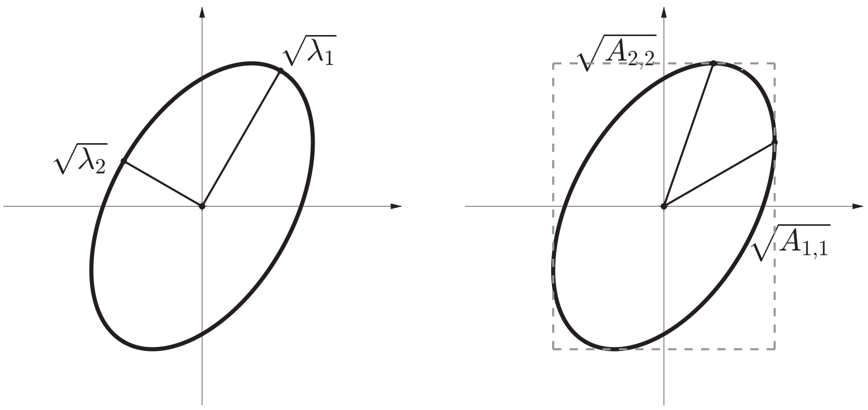


Figure 2.1 – Ellipses are the equipotential lines of a bivariate Gaussian distribution, as illustrated here via a scatter plot.



(a) The eigenvectors of **A** form the semi-principal axes of the ellipsoid.

(b) The 'bounding box' correspond to the diagonal entries of **A**.

Figure 2.2 – The ellipsoid $\mathcal{E}_\mathbf{A}$ corresponding to the matrix **A**.

Property 2.2. *The volume of* $\mathcal{E}_\mathbf{A}$ *relates to the determinant of* **A***:*

$$Vol(\mathcal{E}_\mathbf{A}) = Vol(\mathcal{E}_\mathbf{I}) \cdot \sqrt{|\mathbf{A}|}. \tag{2.3}$$

Proof. For simplicity, let us stick to 3 dimensions (the extension to arbitrary dimensions is identical). Any ellipsoid

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1,$$

can be viewed upon as a linear transformation $(x, y, z) \to (ax, by, cz)$ of a unit-sphere. In turn, that transformation can also be expressed in matrix form as $\mathrm{diag}(a, b, c)$. Consequently, the volume of an ellipsoid equals the volume of a unit-sphere multiplied by abc. In our definition of an ellipsoid $\mathcal{E}_\mathbf{A}$, we have

$$a = \sqrt{\lambda_1(\mathbf{A})}, \quad b = \sqrt{\lambda_2(\mathbf{A})}, \quad c = \sqrt{\lambda_3(\mathbf{A})},$$

which proves the property. In 3-dimensions, we have $Vol(\mathcal{E}_\mathbf{A}) = \frac{4}{3}\pi\sqrt{|\mathbf{A}|}$. □

The volume of the unit sphere is computed via the so-called Gamma-function, which we will not elaborate on here, nor actually use in this thesis.

2.2 Hyperrectangles and Marginal Variance

While most undergrad students have seen a Gaussian elliptical scatter plot and the role of eigenvalues, a less well known property is that also the variance can be clearly seen in the same plot. To be precise, for any $\mathcal{E}_{\mathbf{A}}$ the height/width in any dimension corresponds to the diagonal entries of **A**:

Definition 2.2. *Let $\mathcal{B}_{a_1,a_2,\cdots,a_K}$ be a K-dimensional hyperrectangle spanned by corner points $(\pm\sqrt{a_1},\pm\sqrt{a_2},\cdots,\pm\sqrt{a_K})$. In addition, for matrices we use the shorthand notation $\mathcal{B}_{\mathbf{A}} = \mathcal{B}_{\text{diag}(\mathbf{A})}$.*

We refer to this hyperrectangle as a box. The corner points are taken to be the square root of the set of numbers given so as to match our other geometrical figure, the ellipsoid of Definition 2.1. Namely, $\mathcal{B}_A$ and $\mathcal{E}_{\mathbf{A}}$ are connected as follows:

Property 2.3 (The Bounding Box)**.** *Let $\mathbf{e_i}^T$ be the unit vector with a 1 in the i'th position as its only non-zero component, then*

$$\max_{\mathbf{v}\in\mathcal{E}_{\mathbf{A}}} \mathbf{e}_i^T\mathbf{v} = \sqrt{A_{i,i}}.$$

In other words, $\mathcal{B}_{\mathbf{A}}$ fits snugly around $\mathcal{E}_{\mathbf{A}}$ in all dimensions.

Proof. We make use of the second definition of the ellipsoid, i.e., $\mathcal{E}_{\mathbf{A}} = \{\mathbf{A}^{1/2}\mathbf{u} : ||\mathbf{u}||^2 = 1\}$. Then the Cauchy-Schwarz inequality shows us that:

$$\begin{aligned}\max_{\mathbf{v}\in\mathcal{E}_{\mathbf{A}}} \mathbf{e}_i^T\mathbf{v} &= \max_{\mathbf{u}:||\mathbf{u}||=1} \mathbf{e}_i^T\mathbf{A}^{1/2}\mathbf{u}\\ &\leq ||\mathbf{e}_i^T\mathbf{A}^{1/2}|| \cdot ||\mathbf{u}||\\ &= \sqrt{A_{i,i}}.\end{aligned}$$

The last step follows from $||\mathbf{u}|| = 1$ and that the norm of the i'th row of $\mathbf{A}^{1/2}$ equals $\sqrt{A_{i,i}}$. □

Back to Gaussians, the property says that the height and width of $\mathcal{E}_{\Sigma_{\mathbf{X}}}$ correspond to the standard deviation σ_i. Figure 2.2b shows an example. This property is useful when in source coding we put distortion constraints on these marginal variances. Searching for a Gaussian distribution with individual constraints on variance is a search for an ellipsoid that fits inside a particular bounding box.

Corollary 2.1. *The ellipsoid $\mathcal{E}_{\mathbf{A}}$ is at its most wide with respect to the i'th basis vector $\mathbf{e}_i$ at the i'th column vector of* **A***, scaled accordingly:*

$$\arg\max_{\mathbf{v}\in\mathcal{E}_{\mathbf{A}}} \mathbf{e}_i^T\mathbf{v} = \frac{\sqrt{A_{i,i}}}{||\mathbf{A}\mathbf{e}_i||}\left(\mathbf{A}\mathbf{e}_i\right).$$

2.3 Inscribed Ellipsoids and Conditional Covariance

Property 2.4. $\mathbf{A} \preceq \mathbf{B}$ *if and only if* $\mathcal{E}_\mathbf{A}$ *lies inside* $\mathcal{E}_\mathbf{B}$.

Property 2.5. $\mathbf{A} \prec \mathbf{B}$ *if and only if* $\mathcal{E}_\mathbf{A}$ *lies inside* $\mathcal{E}_\mathbf{B}$ and they do not touch.

Namely, we have $\mathbf{A} \preceq \mathbf{B}$ if and only if $\mathbf{B}^{-1} \preceq \mathbf{A}^{-1}$. Therefore, for all vectors $\mathbf{v}$ it holds that

$$\mathbf{v}^T\mathbf{B}^{-1}\mathbf{v} \leq \mathbf{v}^T\mathbf{A}^{-1}\mathbf{v}. \tag{2.4}$$

Combining this with the definition of $\mathcal{E}_\mathbf{A}$, we conclude that $\mathcal{E}_\mathbf{A}$ must be smaller than $\mathcal{E}_\mathbf{B}$ in all directions. If the inequality is non-strict, then there exists at least one $\mathbf{v}$ (that is not the all-zeroes vector) for which $\mathbf{v}^T\mathbf{A}\mathbf{v} = \mathbf{v}^T\mathbf{B}\mathbf{v}$. Consequently, the ellipsoids $\mathcal{E}_\mathbf{A}$ and $\mathcal{E}_\mathbf{B}$ are tangential at this point (or these points). They are necessarily tangential and cannot cross, because that would contradict the ordering $\mathbf{A} \preceq \mathbf{B}$.

Consider $(\mathbf{X}, \mathbf{Y})$ that are jointly Gaussian with the following covariance:

$$\Sigma = \begin{bmatrix} \Sigma_\mathbf{X} & \Sigma_\mathbf{XY} \\ \Sigma_\mathbf{YX} & \Sigma_\mathbf{Y} \end{bmatrix}, \tag{2.5}$$

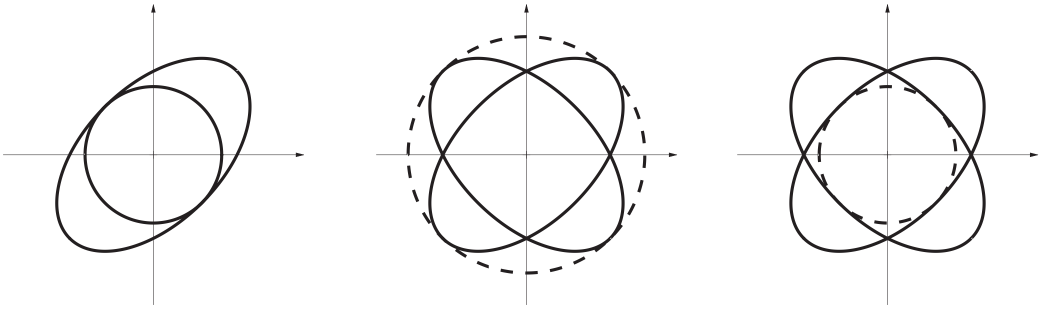
where $\Sigma_\mathbf{YX} = \Sigma_\mathbf{XY}^T$. Then the distribution $p(\mathbf{x}|\mathbf{y})$ is Gaussian with a conditional covariance that is found via the Schur-complement:

$$\Sigma_{\mathbf{X}|\mathbf{Y}} = \Sigma_\mathbf{X} - \Sigma_\mathbf{XY}\Sigma_\mathbf{Y}^{-1}\Sigma_\mathbf{YX}. \tag{2.6}$$

The rightmost term is positive semidefinite. Furthermore, it is necessarily so that

$$\Sigma_{\mathbf{X}|\mathbf{Y}} \preceq \Sigma_\mathbf{X} \tag{2.7}$$

and by the just established properties we have that $\mathcal{E}_{\Sigma_{\mathbf{X}|\mathbf{Y}}}$ is inscribed inside $\mathcal{E}_{\Sigma_\mathbf{X}}$. Examples and derivatives of the ordering property are plotted in Figure 2.3.



(a) $\mathbf{A} \preceq \mathbf{B}$ implies $\mathcal{E}_\mathbf{A}$ lies inside $\mathcal{E}_\mathbf{B}$.

(b) $\mathbf{A} \preceq \mathbf{C}$ and $\mathbf{B} \preceq \mathbf{C}$ imply $\mathcal{E}_\mathbf{C}$ encircles both $\mathcal{E}_\mathbf{A}$ and $\mathcal{E}_\mathbf{B}$.

(c) $\mathbf{A} \succeq \mathbf{C}$ and $\mathbf{B} \succeq \mathbf{C}$ imply $\mathcal{E}_\mathbf{C}$ lies in the intersection of $\mathcal{E}_\mathbf{A}$ and $\mathcal{E}_\mathbf{B}$.

Figure 2.3 – Some consequences from the ordering properties of Section 2.3.

2.4 Touching Ellipsoids and Rank-Deficient Schur Complements

This section discusses the significance of the equality mark underneath $\mathbf{A} \preceq \mathbf{B}$, especially when applied to the Gaussian relationship between conditional covariance and the algebraic Schur-complement (2.6)–(2.7). If a strict matrix ordering does not hold, then $\exists \mathbf{v}$ s.t. $\mathbf{v}^T \Sigma_{\mathbf{X}} \mathbf{v} = \mathbf{v}^T \Sigma_{\mathbf{X}|\mathbf{Y}} \mathbf{v}$; there exists a direction in which $\mathbf{Y}$ does not reduce the variance and hence contains no information on $\mathbf{X}$. This is particularly useful in information theory.

Imagine the following setting: Let us say that an encoder possesses information produced by K dependent Gaussian sources $\mathbf{X} \in \mathbb{R}^K$. He wishes to communicate this data, but only has a limited budget. So he decides not to transmit all his sources, but only a clever *mixture* and to apply compression as well by the following test channel:

$$\mathbf{Y} = \mathbf{A}^T \mathbf{X} + \mathbf{W}, \tag{2.8}$$

where $\mathbf{A} \in \mathbb{R}^{K \times L}$ is an orthonormal projection matrix, i.e., $L < K$ and $\mathbf{A}^T \mathbf{A} = \mathbf{I}$, and $\mathbf{W}$ is Gaussian noise independent of $\mathbf{X}$.

The receiver only learns partial information on $\mathbf{X}$ and still has the following uncertainty:

$$\begin{aligned} \Sigma_{\mathbf{X}|\mathbf{Y}} &= \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}|\mathbf{Y}])(\mathbf{X} - \mathbb{E}[\mathbf{X}|\mathbf{Y}])^T] \\ &= \Sigma_{\mathbf{X}} - \Sigma_{\mathbf{X}} \mathbf{A} \left(\mathbf{A}^T \Sigma_{\mathbf{X}} \mathbf{A} + \Sigma_{\mathbf{W}} \right)^{-1} \mathbf{A}^T \Sigma_{\mathbf{X}}. \end{aligned} \tag{2.9}$$

The crux is that the rightmost term is necessarily rank-deficient due to $\mathbf{A}$ being tall. Consequently, $\dim\left(\ker\left(\Sigma_{\mathbf{X}} - \Sigma_{\mathbf{X}|\mathbf{Y}}\right)\right) \neq 0$. The result is that there exists a vector (or subspace) at which $\mathcal{E}_{\Sigma_{\mathbf{X}|\mathbf{Y}}}$ and $\mathcal{E}_{\Sigma_{\mathbf{X}}}$ 'touch'; this is the little 'or equal' mark under the semidefinite ordering $\Sigma_{\mathbf{X}|\mathbf{Y}} \preceq \Sigma_{\mathbf{X}}$ as said earlier in Property 2.4.

The vectors/space where $\mathcal{E}_{\Sigma_{\mathbf{X}}}$ and $\mathcal{E}_{\Sigma_{\mathbf{X}|\mathbf{Y}}}$ touch stands orthogonal to the set of projection vectors $\mathbf{A}$. We will prove this fact and use it (in the next Section) to argue that any point on the contour of an ellipsoid corresponds to the variance of $\mathbf{X}$ conditioned on the space that stands orthogonal to it. A hint of this could have already been seen in Definition 2.1, as the ellipsoid is spanned by Σ^{-1}, also known as the *precision* matrix of a set of random variables.

Property 2.6. *Let* $\mathbf{A}^{\perp} \in \mathbb{R}^{K \times (K-L)}$ *be the orthogonal complement to* $\mathbf{A}$*, which in turn was used to compute* $\mathbf{Y} = \mathbf{A}^T \mathbf{X} + \mathbf{W}$ *for any independently drawn Gaussian noise* $\mathbf{W}$*. Then* $\mathcal{E}_{\Sigma_{\mathbf{X}}}$ *and* $\mathcal{E}_{\Sigma_{\mathbf{X}|\mathbf{Y}}}$ *touch in the direction of* $\mathbf{A}^{\perp}$*, i.e.,*

$$\mathbf{A}^{\perp^T} \Sigma_{\mathbf{X}}^{-1} \mathbf{A}^{\perp} = \mathbf{A}^{\perp^T} \Sigma_{\mathbf{X}|\mathbf{Y}}^{-1} \mathbf{A}^{\perp}. \tag{2.10}$$

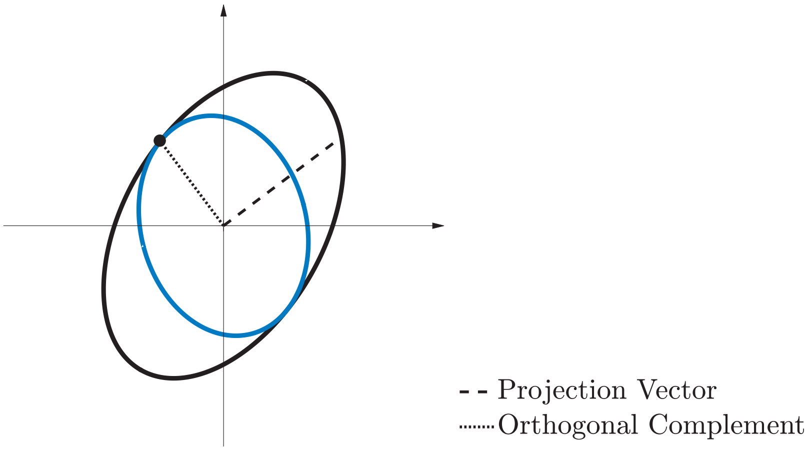


Figure 2.4 – If $\mathcal{E}_{\Sigma_\mathbf{X}}$ and $\mathcal{E}_{\Sigma_{\mathbf{X}|\mathbf{Y}}}$ touch, they do so at the space orthogonal to a set of projection vectors $\mathbf{A}$ belonging to the test channel $\mathbf{Y} = \mathbf{A}^T\mathbf{X} + \mathbf{W}$.

Proof. The proof is a simple application of the Woodbury identity for inverses of a sum of matrices:

$$\Sigma_{\mathbf{X}|\mathbf{Y}}^{-1} = \left(\Sigma_\mathbf{X} - \Sigma_\mathbf{X}\mathbf{A}\left(\mathbf{A}^T\Sigma_\mathbf{X}\mathbf{A} + \Sigma_\mathbf{W}\right)^{-1}\mathbf{A}^T\Sigma_\mathbf{X}\right)^{-1} \tag{2.11}$$

$$= \Sigma_\mathbf{X}^{-1} - \Sigma_\mathbf{X}^{-1}\Sigma_\mathbf{X}\mathbf{A}\left(\mathbf{A}^T\Sigma_\mathbf{X}\mathbf{A} + \Sigma_\mathbf{W} - \mathbf{A}^T\Sigma_\mathbf{X}\Sigma_\mathbf{X}^{-1}\Sigma_\mathbf{X}\mathbf{A}\right)^{-1}\mathbf{A}^T\Sigma_\mathbf{X}\Sigma_\mathbf{X}^{-1} \tag{2.12}$$

$$= \Sigma_\mathbf{X}^{-1} - \mathbf{A}\Sigma_\mathbf{W}^{-1}\mathbf{A}^T. \tag{2.13}$$

Since ${\mathbf{A}^\perp}^T\mathbf{A} = \mathbf{0}$, we observe that the rightmost part drops out when we multiply left and right by $\mathbf{A}^\perp$, which proves the theorem. □

A 2-dimensional example is shown in Figure 2.4.

Property 2.6 can be phrased the other way around: draw two nested ellipsoids and if they touch, reason about which Gaussian test channel could attain it by recognizing that the coding projection matrix $\mathbf{A}$ stands orthogonal to $\ker\left(\Sigma_{\mathbf{X}|\mathbf{Y}}^{-1} - \Sigma_\mathbf{X}^{-1}\right)$ (mind the inverses). This reverse phrasing is useful when in information theoretic applications one wants to 'shape' $\Sigma_{\mathbf{X}|\mathbf{Y}}$, e.g., by distortion constraints. If $\dim\left(\ker\left(\Sigma_\mathbf{X} - \Sigma_{\mathbf{X}|\mathbf{Y}}\right)\right) = 0$ (and thus $\Sigma_{\mathbf{X}|\mathbf{Y}} \prec \Sigma_\mathbf{X}$ is strict) then the orthogonal complement spans the entire space, hence one can take $\mathbf{A} = \mathbf{I}$.

2.5 Ellipsoidal Subspaces

In the previous two subsections we conditioned $\mathbf{X}$ on another jointly Gaussian $\mathbf{Y}$, but a different property arises when one partitions the elements of $\mathbf{X}$ and conditions some of them on all the others.

Property 2.7. *A two-dimensional ellipse based on covariance matrix $\Sigma_{\mathbf{X}}$ cuts the axis beloning to X_i at* $\pm\sqrt{\sigma_i^2(1-\rho^2)}$.

Proof. The proof is straightforward:

$$\Sigma_{\mathbf{X}}^{-1} = \frac{1}{\sigma_1^2\sigma_2^2(1-\rho^2)}\begin{bmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{bmatrix}. \tag{2.14}$$

Hence, $\left(\Sigma_{\mathbf{X}}^{-1}\right)_{i,i} = \frac{1}{\sigma_i^2(1-\rho^2)}$. Combined with the definition of $\mathcal{E}_{\Sigma_{\mathbf{X}}}$ this proves the statement. □

See Figure 2.5 for an illustration of this fact.

This property stems from a more general and far more abstract property: any $\mathbf{v} \in \mathcal{E}_{\Sigma_{\mathbf{X}}}$ has a length equal to the (square root of the) variance of $\mathbf{X}$ conditioned on $\mathbf{X}$ projected on the space that stands orthogonal to that $\mathbf{v}$. In higher dimensions every lower-dimensional ellipsoid on the contour of $\mathcal{E}_{\Sigma_{\mathbf{X}}}$ that is centered around $\mathbf{0}$ corresponds to the covariance of $\Sigma_{\mathbf{X}|\mathbf{A}^{\mathsf{T}}\mathbf{X}}$, where $\mathbf{A}$ is the space orthogonal to that subdimensional ellipsoid. The meaning of this statement is visualized in Figure 2.6 and more precisely formulated as follows:

Property 2.8. *Partition any $K \times K$ orthonormal matrix as $\mathbf{Q} = [\mathbf{A}\,\mathbf{A}^{\perp}]$, where $\mathbf{A} \in \mathbb{R}^{K\times L}$ with $L < K$. Then, we have*

$$\mathbf{A}^{\perp^T}\Sigma_{\mathbf{X}}^{-1}\mathbf{A}^{\perp} = \mathbf{A}^{\perp^T}\Sigma_{\mathbf{X}|\mathbf{A}^{\mathsf{T}}\mathbf{X}}^{-1}\mathbf{A}^{\perp}.$$

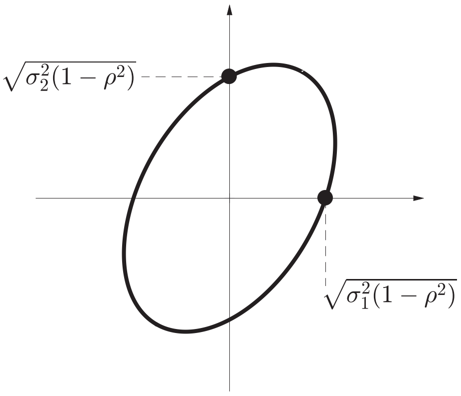


Figure 2.5 – A two-dimensional ellipse corresponding to a covariance cuts the X_1−axis at $\sqrt{\mathrm{var}(X_1|X_2)}$ and vice-versa.

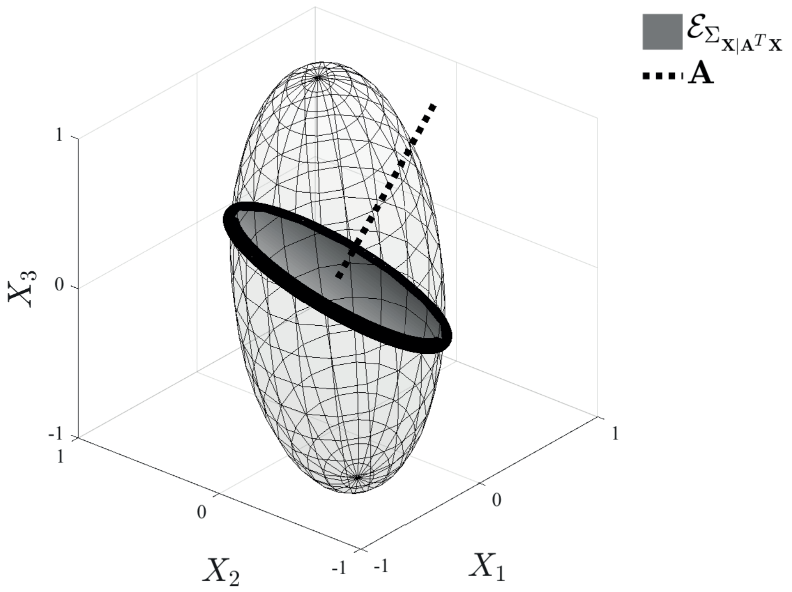


Figure 2.6 – Any lower dimensional ellipsoid along the contour of $\Sigma_{\mathbf{X}}$ corresponds to the covariance of $\Sigma_{\mathbf{X}|\mathbf{A}^{\mathbf{T}}\mathbf{X}}$, where $\mathbf{A}$ is the space orthogonal to that subdimensional ellipsoid.

Consequently, any $L-$ dimensional ellipsoid along the surface of $\mathcal{E}_{\Sigma_{\mathbf{X}}}$ and centered around $\mathbf{0}$ *characterizes $\mathcal{E}_{\Sigma_{\mathbf{X}|\mathbf{A}^{\mathbf{T}}\mathbf{X}}}$ where $\mathbf{A} \in \mathbb{R}^{K\times(K-L)}$ stands orthogonal to that lower-dimensional ellipsoid.*

Proof. First, consider the Gaussian test channel of Section 2.4, $\mathbf{Y} = \mathbf{A}^T\mathbf{X} + \mathbf{W}$. Evaluate the conditional covariance (2.9) and let the noise power of $\mathbf{W}$ go to zero:

$$\lim_{\Sigma_{\mathbf{W}}\to\mathbf{0}} \Sigma_{\mathbf{X}|\mathbf{Y}} = \lim_{\Sigma_{\mathbf{W}}\to\mathbf{0}} \Sigma_{\mathbf{X}} - \Sigma_{\mathbf{X}}\mathbf{A}\left(\mathbf{A}^T\Sigma_{\mathbf{X}}\mathbf{A} + \Sigma_{\mathbf{W}}\right)^{-1}\mathbf{A}^T\Sigma_{\mathbf{X}} \tag{2.15}$$

$$= \Sigma_{\mathbf{X}|\mathbf{A}^T\mathbf{X}}. \tag{2.16}$$

Then apply Property 2.6. Taking the limit seems redundant, but is necessary as the proof of Property 2.6 relies on the Woodbury Identity, which in turn requires invertibility. □

From an information theoretic perspective, the difference between Section 2.4 and this one can be seen as providing someone with perfect or imperfect side-information on parts of $\mathbf{X}$. For this reason, we explicitly included the proof technique using the Gaussian test channel and a limit of noise power going to 0. One could have also gone a different way by remarking that $\Lambda \triangleq \Sigma_{\mathbf{X}}^{-1}$ is the *precision* matrix. Any principal submatrix of Λ is the conditional covariance of all X_i indiced by that submatrix, conditioned on the others. Combining this property with an orthonormal projection is another, but notation-wise more cumbersome route to prove Property 2.8. A nice discussion on this characteristic of the precision matrix can be found in [13, Section 2.3.1].

2.6 Information Theoretic Properties

Property 2.9. *The mutual information between two jointly Gaussian random variables* $\mathbf{X}$ *and* $\mathbf{Y}$ *is measured by the ratio of volume of* $\mathcal{E}_{\Sigma_\mathbf{X}}$ *and* $\mathcal{E}_{\Sigma_{\mathbf{X}|\mathbf{Y}}}$*:*

$$I(\mathbf{X};\mathbf{Y}) = \log \frac{\mathrm{Vol}(\mathcal{E}_{\Sigma_\mathbf{X}})}{\mathrm{Vol}(\mathcal{E}_{\Sigma_{\mathbf{X}|\mathbf{Y}}})}. \tag{2.17}$$

The above is the logical consequence of Property 2.2 and the fact that $I(\mathbf{X};\mathbf{Y}) = \frac{1}{2}\log\frac{|\Sigma_\mathbf{X}|}{|\Sigma_{\mathbf{X}|\mathbf{Y}}|}$. Information theoretic applications in which one wants to minimize $I(\mathbf{X};\mathbf{Y})$ over Gaussian test channels thus have an intuitive geometric interpretation: it corresponds to maximizing the volume of any ellipsoid inside $\mathcal{E}_{\Sigma_\mathbf{X}}$. Subjecting the minimization to some constraints corresponds to desiring a particular shape for $\mathcal{E}_{\Sigma_{\mathbf{X}|\mathbf{Y}}}$. For example the following:

Property 2.10. $X_1, X_2, \cdots, X_K$ *are independent if an only if their covariance* $\Sigma_\mathbf{X}$ *spans an ellipsoid that is straight, i.e., whose semiprincipal axes align with the basis vectors of the system.*

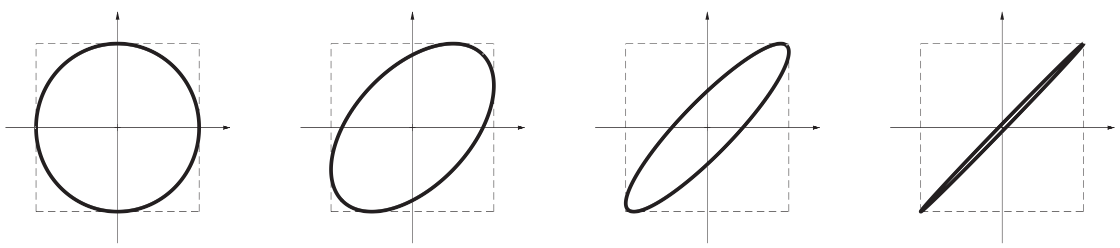
Gaussian random variables are independent if and only if their covariance matrix is diagonal. Evidently, such a matrix has eigenvectors that are the unit basis vectors $\mathbf{e}_i$. Therefore, the semiprincipal axis of $\mathcal{E}_{\Sigma_\mathbf{X}}$ must align with the coordinate system. The impact of correlation on the shape of $\mathcal{E}_{\Sigma_\mathbf{X}}$ is demonstrated in Figure 2.7.

Property 2.11 (Nested Ellipsoids)**.** *For jointly Gaussian* $(\mathbf{X},\mathbf{Y},\mathbf{Z})$ *the following two statements are equivalent:*

$$\mathbf{X}-\mathbf{Y}-\mathbf{Z} \quad \overset{\textit{implies}}{\longleftrightarrow} \quad \Sigma_\mathbf{X} \succeq \Sigma_{\mathbf{X}|\mathbf{Z}} \succeq \Sigma_{\mathbf{X}|\mathbf{Y}}$$

and therefore $\mathcal{E}_{\Sigma_{\mathbf{X}|\mathbf{Z}}}$ *is necessarily 'sandwiched' between* $\mathcal{E}_{\Sigma_\mathbf{X}}$ *and* $\mathcal{E}_{\Sigma_{\mathbf{X}|\mathbf{Y}}}$.

This property follows from -for example- Nayak, Tuncel, Gündüz and Erkip proving the necessity of the semidefinite ordering for the successive refinability of Gaussian sources [14] and the general assertion of Equitz and Cover that a Markov chain is required [15]. A definition of and discussion on successive refinability will follow in Section 3.4; we will refer back to here once needed. A more direct proof of the property can be read in the work by Ando and Petz on Gaussian Markov triplets, to be precise Theorem 1 (property (d)) and/or Corollary 1 in [16].



ρ =0.000 ρ =0.500 ρ =0.900 ρ =0.999

Figure 2.7 – Increasing the correlation ρ while keeping $\sigma_1^2 = \sigma_2^2 = 1$ constant.

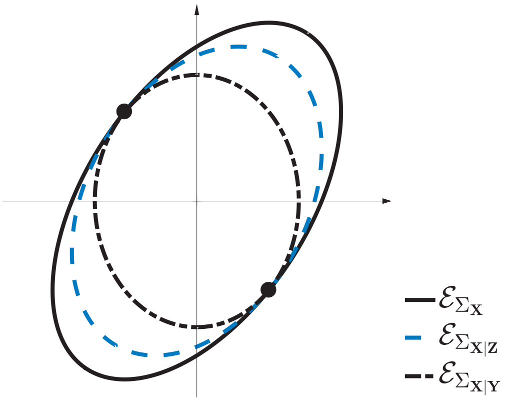


Figure 2.8 – Property 2.11 states that $(\mathbf{X},\mathbf{Y},\mathbf{Z})$ form a Markov chain if and only if the ellipsoids $\mathcal{E}_{\Sigma_{\mathbf{X}}}, \mathcal{E}_{\Sigma_{\mathbf{X|Z}}}$ and $\mathcal{E}_{\Sigma_{\mathbf{X|Y}}}$ are nested (in that order). In this specific example, $\mathcal{E}_{\Sigma_{\mathbf{X}}}$ and $\mathcal{E}_{\Sigma_{\mathbf{X|Y}}}$ touch and therefore $\mathcal{E}_{\Sigma_{\mathbf{X|Z}}}$ is necessarily sandwiched at these points of contact.

The Markov property is most interesting when $\mathcal{E}_{\Sigma_{\mathbf{X|Y}}}$ and $\mathcal{E}_{\Sigma_{\mathbf{X}}}$ touch, i.e., when the outer ends form a non-strict inequality $\Sigma_{\mathbf{X}} \succeq \Sigma_{\mathbf{X|Y}}$; if one desires a Markov chain $\mathbf{X}-\mathbf{Y}-\mathbf{Z}$, then by the sandwiching property it must be so that $\mathcal{E}_{\Sigma_{\mathbf{X|Z}}}$ touches $\mathcal{E}_{\Sigma_{\mathbf{X}}}$ in the same point(s). This connects to Property 2.6: **Z** and **Y** can apparently be constructed via a test channel based on the same projection matrix (though technically, **Y** could contain more projection vectors). Figure 2.8 shows an example of this ‘tight’ version of the Markovian sandwich.

2.7 Closing Remarks

In the following chapters more information theoretic properties will follow when appropriate, including Wyner’s Common Information and Gaussian rate-distortion functions. The main take-away of this Chapter are two properties in particular:

- Property 2.9: minimizing mutual information over Gaussian test channels equals maximizing the volume of an ellipsoid inside $\mathcal{E}_{\Sigma_{\mathbf{X}}}$.
- Property 2.6: If an inner ellipsoid (associated $\mathcal{E}_{\Sigma_{\mathbf{X|Y}}}$) touches the outer ellipsoid (associated to $\mathcal{E}_{\Sigma_{\mathbf{X}}}$), then this **Y** is a Gaussian test channel based on projecting **X** onto the space orthogonal to the space where the ellipsoids touch.

All other properties aid in thinking *what shape* that inscribed ellipsoid should have to ensure statistical properties that are desirable with respect to the application.

An interesting, but very different take on elliptical geometry is the work of Friendly, Monette and Fox [17]. After asserting similar basics, they use ellipsoids to reason about regression and learning applications.

2.8 Overview of Properties

1 The semi-principal axis of $\mathcal{E}_{\mathbf{A}}$ are the eigenvectors of $\mathbf{A}$ with lengths equal to $\sqrt{\lambda_i(\mathbf{A})}$.

2 Volume relates to determinant, i.e., $Vol(\mathcal{E}_{\mathbf{A}}) = Vol(\mathcal{E}_{\mathbf{I}}) \cdot \sqrt{|\mathbf{A}|}$.

3 Height and width relate to diagonal entries, i.e., $\max_{\mathbf{v}\in\mathcal{E}_{\mathbf{A}}} \mathbf{e}_i^T \mathbf{v} = \sqrt{A_{i,i}}$.

4 $\mathbf{A} \preceq \mathbf{B}$ if and only if $\mathcal{E}_{\mathbf{A}}$ lies inside $\mathcal{E}_{\mathbf{B}}$.

5 $\mathbf{A} \prec \mathbf{B}$ if and only if $\mathcal{E}_{\mathbf{A}}$ lies inside $\mathcal{E}_{\mathbf{B}}$ and they do not touch.

6 $\mathbf{A} \preceq \mathbf{C}$ and $\mathbf{B} \preceq \mathbf{C}$ hold if and only if $\mathcal{E}_{\mathbf{C}}$ circumferes $\mathcal{E}_{\mathbf{A}}$ and $\mathcal{E}_{\mathbf{B}}$.

7 $\mathbf{A} \succeq \mathbf{C}$ and $\mathbf{B} \succeq \mathbf{C}$ hold if and only if $\mathcal{E}_{\mathbf{C}}$ lies in the intersection of $\mathcal{E}_{\mathbf{A}}$ and $\mathcal{E}_{\mathbf{B}}$.

8 The ellipse of a 2-dimensional covariance matrix cuts the axis of X_i in $\sqrt{\sigma_i^2(1-\rho^2)}$.

9 Any lower-dimensional ellipsoid along the contour of $\mathcal{E}_{\Sigma_{\mathbf{X}}}$ characterizes the covariance $\Sigma_{\mathbf{X}|\mathbf{A}^T\mathbf{X}}$ where $\mathbf{A}$ is the space orthogonal to that subdimensional ellipsoid.

10 $I(\mathbf{X};\mathbf{Y}) = \log \frac{\mathrm{Vol}(\mathcal{E}_{\Sigma_{\mathbf{X}}})}{\mathrm{Vol}(\mathcal{E}_{\Sigma_{\mathbf{X}|\mathbf{Y}}})}$.

11 $X_1, X_2, \cdots, X_K$ are independent if and only if $\mathcal{E}_{\Sigma_{\mathbf{X}}}$ is straight.

12 $\mathbf{X}-\mathbf{Y}-\mathbf{Z}$ is a valid Markov chain if and only if $\mathcal{E}_{\Sigma_{\mathbf{X}}}, \mathcal{E}_{\Sigma_{\mathbf{X}|\mathbf{Z}}}$ and $\mathcal{E}_{\Sigma_{\mathbf{X}|\mathbf{Y}}}$ are nested (in that order).

Table 2.1 – List of properties of ellipsoids. All random variables are Gaussian.

3 Preliminaries II: Gaussian Rate-Distortion Functions

The single most important tool in this thesis is the (Gaussian) rate distortion function with respect to a variety of distortion criteria. With the exception of Chapter 7 all material concerns the encoding of Gaussian sources. Hence, we are operating with information produced in a continuous domain and the teachings of *lossy* source coding are required.

In addition, this entire thesis is on *vector* sources in particular ($\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \Sigma_{\mathbf{X}})$), not scalars ($X \sim \mathcal{N}(0, \sigma_x^2)$). The 'classic' Gaussian vector rate-distortion function as often taught in undergraduate courses is computed with respect to a sum (or mean) squared error [18, Section 10.3.3]. We emphasize that such a distortion function is a choice best fitted for a particular application, but that in general one can consider many other classes of criteria as well. To that end, we review some of these functions.

The source sequence is denoted is $\mathbf{X}^N$, which consists of N samples that are drawn in an i.i.d. fashion from a $K-$dimensional Gaussian distribution. The encoder maps this sequence to an index m by a function $f(\mathbf{X}^N) \in \{1, 2, \cdots, 2^{NR}\}$. The decoder, in turn, maps this index back to an estimate $\hat{\mathbf{X}}^N$ via its own function $g(m) = \hat{\mathbf{X}}^N$. A $(2^{NR}, N)$ rate-distortion code consist of such encoder and decoder functions. The distortion resulting from such a code is

$$\mathbb{E}[d(\mathbf{X}^N, g(f(\mathbf{X}^N)))], \tag{3.1}$$

where $d(\cdot, \cdot)$ is the distortion measure, which can output a scalar, vector or a matrix.

A rate-distortion pair (R, D) is achievable if there exists a sequence of $(2^{NR}, N)$ codes such that

$$\lim_{N \to \infty} \mathbb{E}[d(\mathbf{X}^N, \hat{\mathbf{X}}^N)] \leq D. \tag{3.2}$$

By the same style of notation, we say that a $(R, [D_1, \cdots, D_K])$ pair is achievable if the distortion measure is a vector constraint, or replace the second argument by a matrix and the inequality by $\preceq$.

The keystone of lossy source coding is that the infimum of achievable rates over particular distortion levels D is the information rate-distortion function:

$$R_{\mathbf{X}}(D) \triangleq \min_{\substack{p(\hat{\mathbf{x}}|\mathbf{x}) \\ :\mathbb{E}[d(\mathbf{X},\hat{\mathbf{X}})] \leq D}} I(\mathbf{X};\hat{\mathbf{X}}). \tag{3.3}$$

We draw special attention to the *conditional* rate-distortion function

$$R_{\mathbf{X}|\mathbf{W}}(D) \triangleq \min_{\substack{p(\hat{\mathbf{x}}|\mathbf{x},\mathbf{w}) \\ :\mathbb{E}[d(\mathbf{X},\hat{\mathbf{X}})] \leq D}} I(\mathbf{X};\hat{\mathbf{X}}|W). \tag{3.4}$$

Again we stress that depending on the distortion measure, the constraint can also be a set of inequalities or a matrix ordering.

For a scalar Gaussian source $X \sim \mathcal{N}(0, \sigma_X^2)$ subject to a squared error constraint $(X - \hat{X})^2$, the rate-distortion function equals

$$R_X(D) = \frac{1}{2} \log \frac{\sigma_X^2}{D}. \tag{3.5}$$

In general we simply denote $R(D)$. The subscript is reserved for drawing explicit attention to which distribution constitutes the rate-distortion function, if required so by the context.

3.1 Maxdet Optimization and the Optimality of Gaussian Codebooks

This thesis follows most closely the point of view of Jin-Jun Xiao and Zhi-Quan Luo [19]: for any 'nice' (to be defined shortly) squared-error distortion criterion, the Gaussian rate-distortion function reduces to semidefinite programming: a minimization over distortion *matrices* **D**. Sometimes, these problems can be solved by hand, but always by well-studied tools in convex optimization. Even though Xiao and Luo wrote a paper specifically on individual distortion criteria for each X_i, their proof extends to many classes of distortion functions by just a tiny generalization. Here, we briefly review the core of their argument, followed by a brief overview of some of those 'nice' distortion functions.

First, let **D** denote the mean squared error matrix between **X** and the lossy representation $\hat{\mathbf{X}}$:

$$\mathbf{D} \triangleq \mathbb{E}[(\mathbf{X}-\hat{\mathbf{X}})(\mathbf{X}-\hat{\mathbf{X}})^T]. \tag{3.6}$$

Theorem 3.1. *Let the distortion function only depend on the mean squared error matrix* (3.6)*, i.e.,* $d(\mathbf{X},\hat{\mathbf{X}}) = d(\mathbf{D})$*. Then, the Xiao-Luo general form of a Gaussian rate-distortion function is the following:*

$$R_{\mathbf{X}}(\gamma) = \min_{\mathbf{D}} \frac{1}{2}\log\frac{|\Sigma_{\mathbf{X}}|}{|\mathbf{D}|} \quad s.t. \begin{cases} \mathbf{0} \preceq \mathbf{D} \preceq \Sigma_{\mathbf{X}} \\ d(\mathbf{D}) \le \gamma, \end{cases} \tag{3.7}$$

which is the solution to (3.3) *for all functions* $d(\mathbf{D})$ *that preserve semidefinite ordering, i.e.:*

$$\mathbf{D}_1 \preceq \mathbf{D}_2 \Rightarrow d(\mathbf{D}_1) \le d(\mathbf{D}_2). \tag{3.8}$$

Proof. We briefly review the main steps of the argument; the details of each step and lemma can be read in [19].

Achievability:
Consider the Gaussian channel $\mathbf{Z} = \mathbf{X} + \mathbf{W}$, where $\mathbf{W} \sim \mathcal{N}(0, \Sigma_{\mathbf{W}})$ is independent of **X**. Then, by forming an estimator $\hat{\mathbf{X}} = \mathbb{E}[\mathbf{X}|\mathbf{Z}]$ one can obtain *any* distortion (3.6) in the range $\mathbf{0} \preceq \mathbf{D} \preceq \Sigma_{\mathbf{X}}$ by generating the noise **W** along the following covariance matrix:

$$\Sigma_{\mathbf{W}} = \Sigma_{\mathbf{X}}(\Sigma_{\mathbf{X}} - \mathbf{D})^{-1}\Sigma_{\mathbf{X}} - \Sigma_{\mathbf{X}}. \tag{3.9}$$

This can easily be verified by computing the Schur complement based on $\hat{\mathbf{X}} = \mathbb{E}[\mathbf{X}|\mathbf{Z}]$:

$$\mathbb{E}[(\mathbf{X}-\hat{\mathbf{X}})(\mathbf{X}-\hat{\mathbf{X}})^T] = \Sigma_{\mathbf{X}} - \Sigma_{\mathbf{X}}(\Sigma_{\mathbf{X}} + \Sigma_{\mathbf{W}})^{-1}\Sigma_{\mathbf{X}} = \mathbf{D}. \tag{3.10}$$

Note that $\mathbf{D} \preceq \Sigma_{\mathbf{X}}$ implies that $\Sigma_{\mathbf{W}} \succeq \mathbf{0}$.

Problem Type	**Constraint**	**Featured in e.g.**
1: Trace (sum-squared error)	$\mathrm{tr}(\mathbf{D}) \leq \gamma$	[18, Section 10.3.3],[20]
2: Individual squared error	$D_{i,i} \leq \gamma_i$	[14, 19]
3: Matrix ordering	$\mathbf{D} \preceq \mathbf{K}$	[21]

Table 3.1 – Examples of distortion functions $d(\mathbf{X}, \hat{\mathbf{X}}) = d(\mathbf{D})$ that depend only on the squared error matrix and who preserve semidefinite ordering.

Converse:

The converse starts with two lemmas:

Lemma 3.1. *For any random vector* $\hat{\mathbf{X}}$ *that is jointly distributed with* $\mathbf{X}$*, we have*

$$I(\mathbf{X}; \hat{\mathbf{X}}) \geq \frac{1}{2} \log \frac{|\Sigma_{\mathbf{X}}|}{|\mathbf{D}|}. \tag{3.11}$$

Equality holds for $\hat{\mathbf{X}} = \mathbb{E}[\mathbf{X}|\mathbf{Z}]$ *where* $\mathbf{Z}$ *is a Gaussian test channel of the form* $\mathbf{Z} = \mathbf{X} + \mathbf{W}$*, where is* $\mathbf{W}$ *is Gaussian distributed and independent of* $\mathbf{X}$*.*

Lemma 3.2. *Suppose* $\mathbf{Z}$ *is a random variable jointly distributed with* $\mathbf{X}$*. Then for* $\hat{\mathbf{X}} = \mathbb{E}[\mathbf{X}|\mathbf{Z}]$ *it holds that* $\mathbf{D} \preceq \Sigma_{\mathbf{X}}$*.*

Then, a few logical steps assess the sufficiency of Gaussian codebooks.
If $\hat{\mathbf{X}} \neq \bar{\mathbf{X}} = \mathbb{E}[\mathbf{X}|\hat{\mathbf{X}}]$ then one can consider $\bar{\mathbf{X}}$ instead, because

1. $I(\mathbf{X}; \hat{\mathbf{X}}) \geq I(\mathbf{X}; \bar{\mathbf{X}})$ (data processing inequality); $\bar{\mathbf{X}}$ has a lower objective value.
2. Let $\bar{\mathbf{D}}$ and $\hat{\mathbf{D}}$ be the distortion matrices (3.6) based on $\bar{\mathbf{X}}$ and $\hat{\mathbf{X}}$ respectively. Then, $\bar{\mathbf{D}} \preceq \hat{\mathbf{D}}$, since $\bar{\mathbf{X}}$ is the MMSE estimator based on $\hat{\mathbf{X}}$. Therefore, if the distortion measure preserves semidefinite ordering (as described in the theorem), then the constraint will not be violated by picking $\hat{\mathbf{X}}$ instead[1].

Since one can restrict one's attention to $\bar{\mathbf{X}} = \mathbb{E}[\mathbf{X}|\hat{\mathbf{X}}]$, then by Lemma 3.2 this is equivalent to considering only auxiliary random variables for which $\mathbf{D} \preceq \Sigma_{\mathbf{X}}$ holds. For this class of random variables, it holds by Lemma 3.1 that Gaussian test channels give the lowest objective value for each possible distortion matrix in that specified range. □

Table 3.1 lists a few examples of used distortion constraints that preserve semidefinite ordering and thus for which (3.7) is the rate-distortion function.

[1] This step was only assessed for individual squared error constraints for each X_i when originally written in [19]: $\bar{\mathbf{D}} \preceq \hat{\mathbf{D}}$ implies $\mathrm{diag}(\bar{\mathbf{D}}) \leq \mathrm{diag}(\hat{\mathbf{D}})$, which generalizes to all ordering-preserving distortion functions.

Theorems in Pictures

The Gaussian rate-distortion function subject to *linear* constraints tries to find an ellipsoid of maximum volume inside a convex body. Take for example, the second entry of Table 3.1: individual squared error constraints. The convex search space is the intersection of the following two spaces:

1. $\mathcal{E}_{\Sigma_{\mathbf{X}}}$, from $\mathbf{D} \preceq \Sigma_{\mathbf{X}}$,
2. $\mathcal{B}_{D_1,D_2}$, from $\mathrm{diag}(\mathbf{D}) \leq [D_1\ D_2]$.

As both spaces are convex, so is their intersection. The plot below depicts the geometry of this optimization problem. The rate-distortion function subject to a matrix-ordering constraint ($\mathbf{D} \preceq \mathbf{K}$) follows a similar pattern, but optimizes inside the convex intersection of *two ellipses* (the ones belonging to $\Sigma_{\mathbf{X}}$ and $\mathbf{K}$).

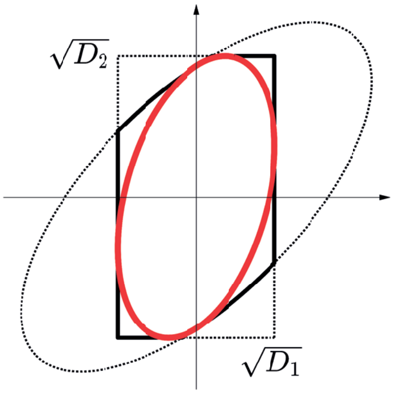


Figure 3.1 – The convex geometry of the Gaussian bivariate rate-distortion function. The thick black line marks the intersection of the two convex spaces spanned by both constraints. The red line is then the ellipse of maximum volume inside this space.

We specifically attend the reader to the link between these optimization problems and Section 2.4. Since we *maximize* the volume of an ellipsoid inside a convex space, the solution will touch the borders of the space spanned by the contraints. If the inner ellipsoid touches $\mathcal{E}_{\Sigma_{\mathbf{X}}}$ in particular, then Section 2.4 teaches us about the Gaussian test channel that stands at its foundation.

3.2 The Rate-Distortion Function under a Sum Squared Error

The classic textbook example of Gaussian vector coding is the so-called ‘reverse water-filling’ procedure on the eigenvalues of $\Sigma_{\mathbf{X}}$ [18, Section 10.3.3]. This technique is the solution to (3.7) under a sum squared error constraint:

$$\sum_{i=1}^{K} \mathbb{E}[(X_i - \hat{X}_i)^2] = \mathrm{tr}(\mathbb{E}[(\mathbf{X}-\hat{\mathbf{X}})(\mathbf{X}-\hat{\mathbf{X}})^T]) \leq D. \tag{3.12}$$

Theorem 3.2. *Let* **X** *be a Gaussian random vector of length K, then the rate-distortion function subject to a sum-squared error equals*

$$R_{\mathbf{X}}(D) \triangleq \min_{\substack{p(\hat{\mathbf{x}}|\mathbf{x})\\ :\mathrm{tr}(\mathbf{D})\leq D}} I(\mathbf{X};\hat{\mathbf{X}}) \tag{3.13}$$

$$= \frac{1}{2}\sum_{i=1}^{K} \log \frac{\lambda_i(\Sigma_{\mathbf{X}})}{D_i}, \tag{3.14}$$

where D_i is chosen as follows using a parameter θ s.t. $\sum_{i=1}^{K} D_i = D$:

$$D_i = \begin{cases} \theta & \theta < \lambda_i(\Sigma_{\mathbf{X}}), \\ \lambda_i(\Sigma_{\mathbf{X}}) & \theta \geq \lambda_i(\Sigma_{\mathbf{X}}). \end{cases} \tag{3.15}$$

A crucial insight is that **D** adheres to the eigenbasis of $\Sigma_{\mathbf{X}}$. One should not underestimate how special this property is. Namely, *both* the objective $|\mathbf{D}|$ (3.7) and the constraint $\mathrm{tr}(\mathbf{D})$ are rotation-invariant, meaning that a multiplication with an orthonormal matrix changes neither the objective value nor the constraint. This is not the case for most other distortion functions, for example those in Table 3.1.

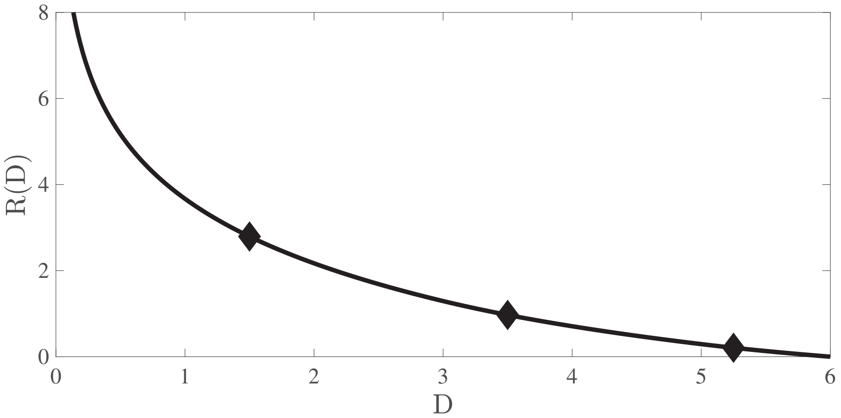


Figure 3.2 – Example of $R_{\mathbf{X}}(D)$ for a $\Sigma_{\mathbf{X}}$ with eigenvalues equal to 3, 2 and 1. The diamonds correspond f.l.t.r. to the example points of Figures 3.3–3.5 on the next page.

Theorems in Pictures

The geometric proof of Theorem 3.2 is as follows: Having proved that Gaussian codebooks are optimal in Theorem 3.1, the problem reduces to choosing a **D** that maximizes its determinant under a trace-constraint. In other words:

$$R_{\mathbf{X}}(D): \quad \max \prod_i \lambda_i(\mathbf{D}) \quad \text{s.t.} \sum_i \lambda_i(\mathbf{D}) \le D, \text{ and } \mathbf{0} \le \mathbf{D} \le \Sigma_{\mathbf{X}}.$$

Maximizing a product of numbers under a sum-constraint is a well-known optimization problem whose solution is to choose all λ_i to be equal. The optimal **D** is therefore a scaled identity matrix, or geometrically, $\mathcal{E}_{\mathbf{D}}$ is ideally a sphere.

Tracing Figure 3.2 from left to right (thus increasing D) corresponds to inflating a sphere inside $\mathcal{E}_{\Sigma_{\mathbf{X}}}$. The largest sphere that fits inside the source ellipsoid is $\mathbf{D} = \lambda_{\min}(\Sigma_{\mathbf{X}})\mathbf{I}$. Afterwards, the sphere can still inflate *equally* in the direction of the other eigenvectors until it hits the wall of the second smallest eigenvalue. This process continues iteratively and is another way of looking at water filling.

The plots below depict this process at the three diamonds marked in Figure 3.2. The transparent mesh corresponds to $\mathcal{E}_{\Sigma_{\mathbf{X}}}$ and the colored ellipsoid on the inside to $\mathcal{E}_{\mathbf{D}}$.

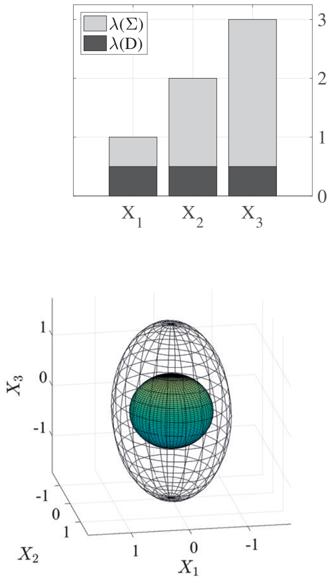


Figure 3.3 – $\theta = 0.5$.

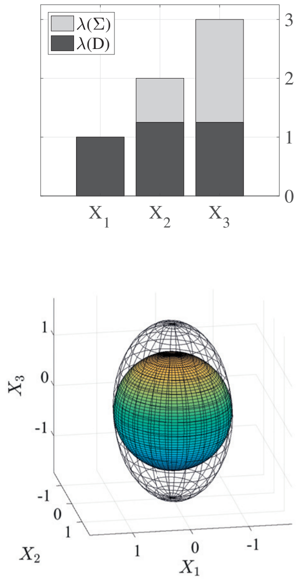


Figure 3.4 – $\theta = 1.25$.

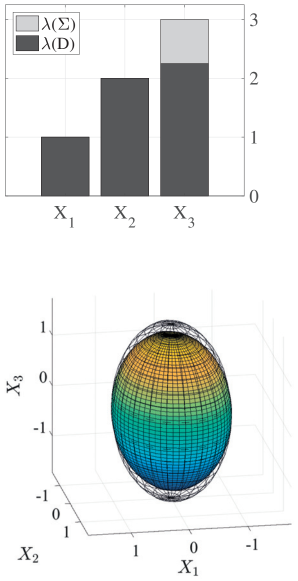


Figure 3.5 – $\theta = 2.25$.

3.3 The Rate-Distortion Function under Individual Criteria

Corollary 3.1. *The Gaussian rate-distortion function subject to individual squared error constraints is the following:*

$$R(D_1, D_2, \cdots, D_K) = \min_{\mathbf{D}} \frac{1}{2} \log \frac{|\Sigma_{\mathbf{X}}|}{|\mathbf{D}|} \quad s.t. \begin{cases} \mathbf{0} \preceq \mathbf{D} \preceq \Sigma_{\mathbf{X}}, \\ \operatorname{diag}(\mathbf{D}) \leq [D_1 \; D_2 \; \cdots \; D_K]. \end{cases} \tag{3.16}$$

The above follows directly from plugging individual criteria into (3.7), as was the original work of Xiao and Luo [19]. This function has not been reduced to a closed-form analytic solution, but can be solved efficiently via interior point methods [22]. Namely, observe that the objective function is strictly convex and the set of feasible $\mathbf{D}$ is convex as well.

A crucial observation is the role of the Hadamard inequality in this problem:

$$|\mathbf{D}| \leq \prod_{i=1}^{K} D_{i,i} \leq \prod_{i=1}^{K} D_i, \tag{3.17}$$

resulting into the following lower bound:

$$R(D_1, D_2, \cdots, D_K) \geq \frac{1}{2} \log \frac{|\Sigma_{\mathbf{X}}|}{\prod_{i=1}^{K} D_i}, \tag{3.18}$$

which is met with equality if and only if $\operatorname{diag}(D_1, D_2, \cdots, D_K) \preceq \Sigma_{\mathbf{X}}$.

Thus far, the case of bivariate Gaussians $\mathbf{X} \in \mathbb{R}^2$ is the only instance that is solved into a closed-form expression. This function $R(D_1, D_2)$ will be widely used throughout Chapter 4:

Corollary 3.2. *Let $K = 2$ and assume w.l.o.g. that $\sigma_1^2 = \sigma_2^2 = 1$. Then for distortions $0 \leq D_1, D_2 \leq 1$, the Gaussian rate-distortion function with respect to individual squared error criteria equals*

$$R(D_1, D_2) = \begin{cases} \frac{1}{2} \log\left(\frac{1-\rho^2}{D_1 D_2}\right) & \text{if } (D_1, D_2) \in \mathcal{D}_1, \\ \frac{1}{2} \log\left(\frac{1-\rho^2}{D_1 D_2 - \left(|\rho| - \sqrt{(1-D_1)(1-D_2)}\right)^2}\right) & \text{if } (D_1, D_2) \in \mathcal{D}_2, \\ \frac{1}{2} \log\left(\frac{1}{\min(D_1, D_2)}\right) & \text{if } (D_1, D_2) \in \mathcal{D}_3, \end{cases} \tag{3.19}$$

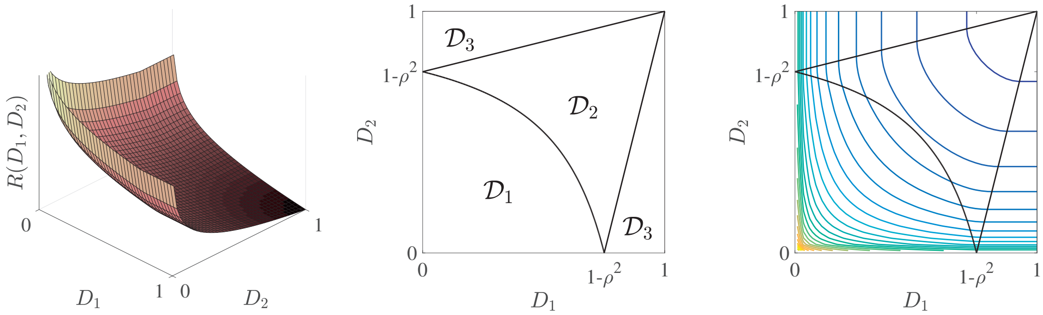
where

$$\mathcal{D}_1 = \{D_1, D_2 : (1 - D_1)(1 - D_2) \geq \rho^2\}, \tag{3.20}$$

$$\mathcal{D}_2 = \{D_1, D_2 : (1 - D_1)(1 - D_2) \leq \rho^2 \leq \min\left(\frac{1 - D_1}{1 - D_2}, \frac{1 - D_2}{1 - D_1}\right)\}, \tag{3.21}$$

$$\mathcal{D}_3 = \mathcal{D}_1^c \cap \mathcal{D}_2^c. \tag{3.22}$$

For any other variance one has to normalize each X_i and scale the respective D_i accordingly.



(a) $R(D_1, D_2)$.

(b) The distortion plane as defined by (3.20)-(3.22).

(c) The distortion plane with contour lines of $R(D_1, D_2)$.

Figure 3.6 – Visualization of the Gaussian joint rate-distortion function and the regions of (D_1, D_2) in which it exhibits different behavior.

Figure 3.6 makes $R(D_1, D_2)$ tangible through visualization. The plane of distortion levels for $\hat{X}_1$ and $\hat{X}_2$ is cut into different regions in which the rate-distortion function exhibits different behavior. At each coordinate, the 2×2 error matrix $\mathbf{D}$ takes on a different shape (assuming w.lo.g. $\rho > 0$):

$$\mathcal{D}_1 \to \mathbf{D} = \begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix} \tag{3.23}$$

$$\mathcal{D}_2 \to \mathbf{D} = \begin{bmatrix} D_1 & \rho - \sqrt{(1-D_1)(1-D_2))} \\ \rho - \sqrt{(1-D_1)(1-D_2)} & D_2 \end{bmatrix}. \tag{3.24}$$

In $\mathcal{D}_1$ this matrix is diagonal, whereas in $\mathcal{D}_2$ it is correlated. $\mathcal{D}_3$ is degenerate: the distortion D_i on one X_i is so small (in comparison to the other) that the best strategy is to only code that X_i. One can then achieve any distortion in $\mathcal{D}_3$ on the other component by an estimator. We call Figure 3.6b in its entirety the $\mathcal{D}$−plane. On the next page we show some further examples of the impact of this plane.

Theorems in Pictures

The following plots match the $\mathcal{D}-$plane to actual distortion matrices $\mathbf{D}$ (3.23)–(3.24) and the role of the inequality $\mathbf{D} \preceq \Sigma_{\mathbf{X}}$. The black outer ellipse corresponds to $\Sigma_{\mathbf{X}}$, the inner ones to distortion matrices that stem from the same-color coordinates in the left plots.

The first 2 examples illustrate that the difference between $\mathcal{D}_1$ and $\mathcal{D}_2$ stems from whether or not $\mathbf{D} \preceq \Sigma_{\mathbf{X}}$ is a strict inequality. In $\mathcal{D}_1$ the constraints are so small that consequently it becomes strict. This also means that the bound of (3.18) can be met with equality and $\mathbf{D}$ is consequently diagonal. The third example illustrates $\mathcal{D}_3$.

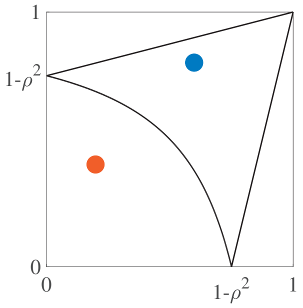


Figure 3.7 – The difference between $\mathcal{D}_1$ and $\mathcal{D}_2$ is that the inequality $\mathbf{D} \prec \Sigma_{\mathbf{X}}$ is...

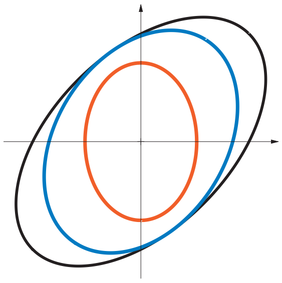


Figure 3.8 – ... only strict in $\mathcal{D}_1$. In $\mathcal{D}_2$ Section 2.4 applies.

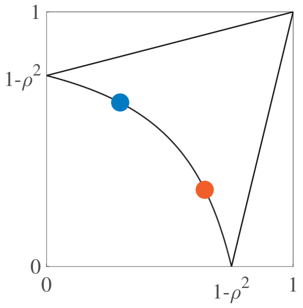


Figure 3.9 – On the border of $\mathcal{D}_1$ and $\mathcal{D}_2$ the inequality $\mathbf{D} \preceq \Sigma_{\mathbf{X}}$ is not strict...

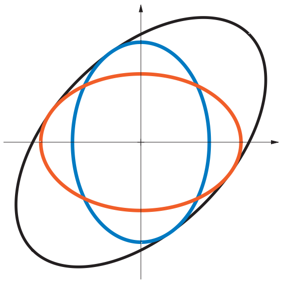


Figure 3.10 – ... but $\mathbf{D}$ is diagonal, hence $\mathcal{E}_{\mathbf{D}}$ still aligns with the system axes.

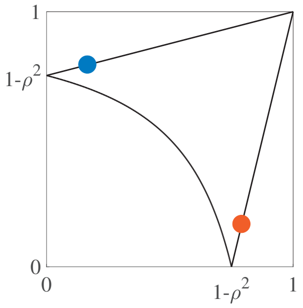


Figure 3.11 – On the border of $\mathcal{D}_2$ and $\mathcal{D}_3$ the encoder only codes one...

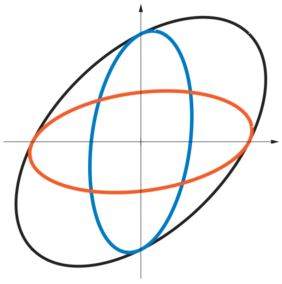


Figure 3.12 – ... X_i. Therefore, $\mathcal{E}_{\mathbf{D}}$ touches $\mathcal{E}_{\Sigma_{\mathbf{X}}}$ at the axis of the other $X_{j \neq i}$.

3.4 Successive Refinability

In the early 90s, Equitz and Cover [15] and later Rimoldi [23] puzzled over iterative source coding: Can one split the bits of $R(D)$ into steps of which each is decodable and rate-distortion optimal by itself? The answer turned out to be affirmative, but also a property that is not common to all sources of information or all ranges of distortions.

The technicality of the matter is depicted in Figure 3.13: Two lossy representations of a source X are to be decoded. There are two encoders of rate R_1 and R_2 and the second decoder gets both messages to produce a lossy description $\hat{X}_2$ that is of better quality than $\hat{X}_1$ with respect to the distortion measure used. Such communication was shown to be achievable for a rate-distortion quadruple (R_1, R_2, D_1, D_2) if and only if there exists a distribution $p(\hat{x}_1, \hat{x}_2|x)$ such that [15, 23]:

$$\begin{aligned} I(X; \hat{X}_1) &\le R_1 \\ I(X; \hat{X}_1 \hat{X}_2) &\le R_2 \\ \mathbb{E}[d(X, \hat{X}_1)] &\le D_1 \\ \mathbb{E}[d(X, \hat{X}_1)] &\le D_2. \end{aligned}$$

Definition 3.1. *We say* $\mathbf{X}$ *is successively refinable from* D_1 *to* $D_2 \le D_1$ *with respect to a particular rate-distortion function if communication is achievable at the following rates:*

$$R_1 = R(D_1) \tag{3.25}$$

$$R_1 + R_2 = R(D_2). \tag{3.26}$$

Equitz and Cover proved that successive refinability is attainable if and only if the individual rate-distortion solutions $p(\hat{x}_1|x)$ and $p(\hat{x}_2|x)$ are such that $X - \hat{X}_2 - \hat{X}_1$ holds as a Markov chain [15]. For Gaussians, Markovity translates to a semidefinite ordering of conditional covariances, as was mentioned in Property 2.11 in Chapter 2. Nayak, Tuncel, Gündüz and Erkip studied explicitly the refinability with respect to the rate-distortion function subject to individual squared error criteria [14]. Their result extends to the successive refinability of any rate-distortion function whose solution is Gaussian distributed as follows:

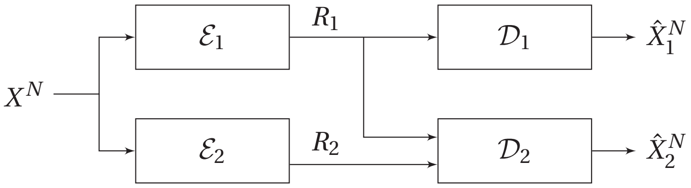


Figure 3.13 – Successive Refinability in two stages.

Theorem 3.3 (Nayak *et al.* [14])**.** *A Gaussian random vector* **X** *is successively refinable from* D_1 *to* $D_2 \leq D_1$ *if the respective distortion matrices resulting from* $R(D_1)$ *and* $R(D_2)$ *result in a semidefinite ordering:*

$$\mathbf{D}_1 \succeq \mathbf{D}_2.$$

It is irrelevant whether the distortion constraint is a scalar, vector or matrix inequality.

A reference that is particularly relevant for this thesis is earlier work of Nayak and Tuncel on the successive coding of two correlated sources [24]. The setting is equivalent, but the distortion measures in stage 1 and 2 are different in the sense that they measure the precision with respect to only some and different elements of **X**, rather than **X** as a whole. This case is encapsulated in their aforementioned general and later result.

3.4.1 Successive Refinability of the Bivariate Rate-Distortion Function under Individual Distortion Criteria

We specifically wish to connect this discussion of refinability to Section 3.3. Later in Chapter 4 when we develop a model for caching we will ask ourselves whether the caching of bivariate Gaussians is a successively refinable process or not. The bivariate Gaussian rate-distortion function subject to individual squared error criteria was an explicit example in the work of Nayak *et al.* [14]. Their result: such a situation is successively refinable, but not everywhere. We review the important insight, but coming from a different point-of-view and notation. To ensure successive refinability, we must assess whether $R(D_1, D_2)$ produces distortion matrices **D** that adhere to a semidefinite ordering as D_1, D_2 decrease.

Assume w.l.o.g. that **X** is of unit variance. Otherwise, normalize variance and scale the respective distortion matrix **D** accordingly. Then, the bivariate rate-distortion function $R(D_1, D_2)$ will result in the distortion matrices mentioned in (3.23) and (3.24), depending on (D_1, D_2) being in $\mathcal{D}_1$ or $\mathcal{D}_2$[2] [19]. Now we discuss three possible refinement moves, all of which behave differently:

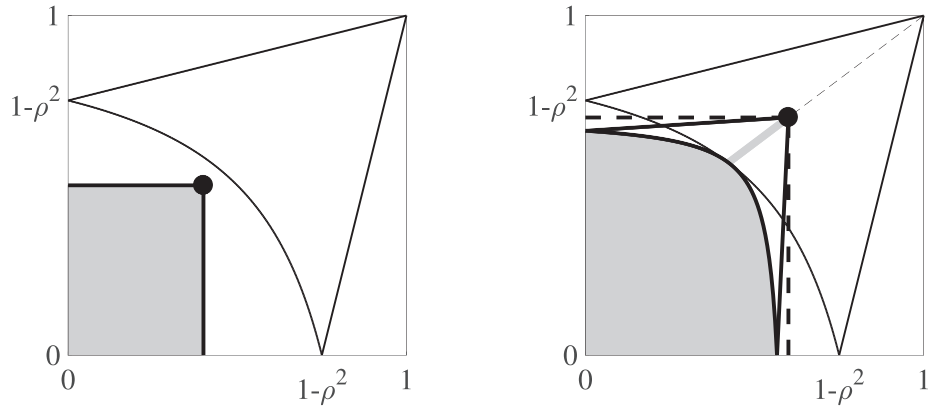
1: $\mathcal{D}_1 \rightarrow \mathcal{D}_1$
Successive refinability from any two coordinates (D_1^1, D_2^1) to $(D_1^2, D_2^2) \leq (D_1^1, D_2^1)$ is evidently possible, as the distortion matrices will necessarily be diagonal therefore $\mathbf{D}^1 \succeq \mathbf{D}^2$ is ensured.

2: $\mathcal{D}_2 \rightarrow \mathcal{D}_2$
Observe that the distortion matrix (3.24) can be constructed as follows:

$$\mathbf{D} = \Sigma_{\mathbf{X}} - \begin{bmatrix} \sqrt{1-D_1} \\ \sqrt{1-D_2} \end{bmatrix} \begin{bmatrix} \sqrt{1-D_1} & \sqrt{1-D_2} \end{bmatrix}; \tag{3.27}$$

it is a rank-one correction of the original covariance matrix. Two distortion matrices generated from different $(D_1, D_2)-$pairs in $\mathcal{D}_2$ cannot respect a semidefinite ordering if they are rank-1

[2] We omit $\mathcal{D}_3$ since it is degenerate; for the same rate one can code up to distortions on the border of $\mathcal{D}_2$ and $\mathcal{D}_3$.



(a) $\mathcal{D}_1 \to \mathcal{D}_1$, one can successively refine to all $(D_1, D_2) \leq (D_1^1, D_2^1)$.

(b) $\mathcal{D}_2 \to \mathcal{D}_2$ or $\mathcal{D}_1$, a new $\mathcal{D}$−plane is drawn from condition (3.28).

Figure 3.14 – Successive Refinability of $R(D_1, D_2)$. The black dot represents example distortions (D_1^1, D_2^1) after the first coding phase, the gray line and area are all coordinates up to which one can successively refine the source.

corrections along different spaces. The outer product on the right-hand side must be the same up to scaling. This implies that $\mathbf{X}$ is successively refinable from (D_1^1, D_2^1) to (D_1^2, D_2^2) in $\mathcal{D}_2$ *if and only if those two coordinates lie on a straight line originating from* $(1,1)$.

3: $\mathcal{D}_2 \to \mathcal{D}_1$

After coding up to a first distortion matrix $\mathbf{D}^1$ in $\mathcal{D}_2$, one *cannot* refine to all coordinates in $\mathcal{D}_1$. Namely, all diagonal $\mathbf{D}$ in $\mathcal{D}_1$ respect the ordering $\mathbf{D} \preceq \Sigma_{\mathbf{X}}$ but may not respect $\mathbf{D} \preceq \mathbf{D}^1$. Consider a refinement from $(D_1^1, D_2^1) \in \mathcal{D}_2$ to $(D_1^2, D_2^2) \in \mathcal{D}_1$. We will verify whether the matrices generated by $R(D_1, D_2)$ at these coordinates satisfy $\mathbf{D}^1 \succeq \mathbf{D}^2$ by checking whether $|\mathbf{D}^1 - \mathbf{D}^2| \geq 0$:

$$\mathbf{D}^1 - \mathbf{D}^2 = \begin{bmatrix} D_1^1 & \rho - \sqrt{(1-D_1^1)(1-D_2^1)} \\ \rho - \sqrt{(1-D_1^1)(1-D_2^1)} & D_2^1 \end{bmatrix} - \begin{bmatrix} D_1^2 & 0 \\ 0 & D_2^2 \end{bmatrix},$$

hence successive refinability is achievable if and only if

$$(D_1^1 - D_1^2)(D_2^1 - D_2^2) \geq (\rho - \sqrt{(1-D_1^1)(1-D_2^1)})^2. \tag{3.28}$$

Note the resemblance to the original condition for $\mathcal{D}_1$ (3.20). They are essentially the same: $\mathcal{D}_1$ characterizes all diagonal matrices $\mathbf{D}$ satisfying $\mathbf{D} \preceq \Sigma_{\mathbf{X}}$. We derived the same condition not for $\Sigma_{\mathbf{X}}$, but for a non-diagonal distortion $\mathbf{D}^1$. In other words one has to redraw a *new* $\mathcal{D}$−plane based on this smaller matrix $\mathbf{D}^1$. Coding from $\mathcal{D}_2$ to either $\mathcal{D}_2$ or $\mathcal{D}_1$ is drawn in Figure 3.14b.

Theorems in Pictures

Successively refining $\mathbf{X}$ from one set of distortion levels to another in $\mathcal{D}_2$ requires that both set of coordinates lie on a straight line originating from $(1,1)$. While the condition can be drawn on the $\mathcal{D}-$plane, it is impossible to see why from the same plot. The crux is the following: if two matrices $\mathbf{D}^1$ and $\mathbf{D}^2$ are both rank-one corrections of $\Sigma_{\mathbf{X}}$ but along *different* subspaces, then neither $\mathbf{D}^1 \succeq \mathbf{D}^2$ nor the reverse can hold.

The geometry is a contradiction of Properties 2.6 and 2.11 in Chapter 2.

1. For successive refinability, we need a Markov chain. Hence $R(D_1, D_2)$ should code up to distortion matrices that are ordered as $\Sigma_{\mathbf{X}} \succeq \mathbf{D}^1 \succeq \mathbf{D}^2$. This ordering is equivalent to their respective ellipses needing to be nested.

2. Since any coordinate $(D_1, D_2) \in \mathcal{D}_2$ is achieved by a $\mathbf{D}$ that is a rank-one correction of $\Sigma_{\mathbf{X}}$, we have Property 2.6: $\Sigma_{\mathbf{X}}$ and $\mathbf{D}$ must touch at a pair of symmetric points.

If by the first argument the ellipses need to be nested, then all three $\mathcal{E}_{\Sigma_{\mathbf{X}}}, \mathcal{E}_{\mathbf{D}^1}$ and $\mathcal{E}_{\mathbf{D}^2}$ must touch at the *same* points, otherwise $\mathcal{E}_{\mathbf{D}^1}$ and $\mathcal{E}_{\mathbf{D}^2}$ intersect. Consequently, the encoder must code information about the same subspace of $\Sigma_{\mathbf{X}}$ in both stages. By the construction of $\mathbf{D}$ in (3.27) it is clear this requires the ratio $\frac{1-D_2}{1-D_1}$ to be a constant, which in turn constitutes the aforementioned line originating from $(1,1)$.

An example of a confirmation and a contradiction are plotted below.

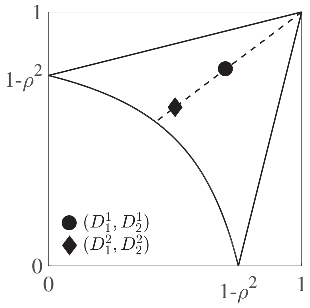


Figure 3.15 – Refinement on a line.

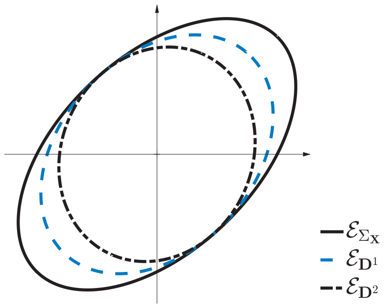

$\mathcal{E}_{\Sigma_{\mathbf{X}}}$
$\mathcal{E}_{\mathbf{D}^1}$
$\mathcal{E}_{\mathbf{D}^2}$

Figure 3.16 – The ellipses are nested and all touch at the same pair of points.

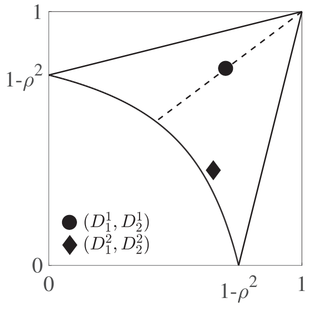


Figure 3.17 – 'Attempt' to refine to distortions not on the line.

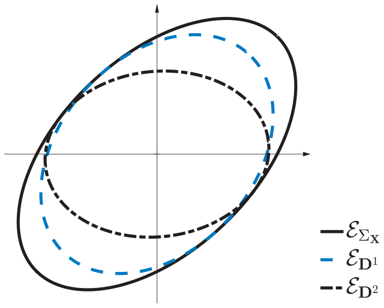

$\mathcal{E}_{\Sigma_{\mathbf{X}}}$
$\mathcal{E}_{\mathbf{D}^1}$
$\mathcal{E}_{\mathbf{D}^2}$

Figure 3.18 – $\mathcal{E}_{\mathbf{D}^1}$ and $\mathcal{E}_{\mathbf{D}^2}$ touch $\mathcal{E}_{\Sigma_{\mathbf{X}}}$ at different points and hence intersect.

4 Caching of Bivariate Gaussian Sources

‘On-demand’ is the keyword for communication technology in this decade. Advancements in information theory have brought us to a point of immense freedom for the end user: First, downloading videos of one’s own choice replaced traditional broadcast of TV and radio. Then, instantaneous streaming replaced downloading. While providing great flexibility, this demand for personal and instantaneous data streams also increased the load on the network.

A second cost is often overlooked: on-demand services increase the *imbalance* of network load. Notoriously data-heavy applications like Netflix and Amazon Prime are hardly popular during the day; almost all users use the service sometime between dinner and their bedtime. Network and server capacity suffer from this imbalance; they are installed to withstand *peak* traffic and not average. Clearly, the trend of on-demand streaming is both costly and inefficient.

A challenge for information theorists is to combine the user experience of on-demand streaming with a balanced network load; caching can be the tool to break that impasse. The key of caching is that a server does not wait for a user to make a request for data. Instead the server tries to *anticipate* what data will be requested and sends it in advance. Netflix, for example, could already transmit parts of the next episode of your favorite series assuming that you will continue your viewing habits. Imperfect prediction of the user’s request will increase the overall need for data, but a well designed system will reduce the average network load during the peak hours. This is the trade-off we intend to study.

Caching has been looked at from different angles, with the work from Maddah-Ali and Niesen being the most popular [1]. In this thesis[1], we take a lossy source coding perspective and model the problem in a way that resembles the Gray–Wyner network [7]. This model was introduced in a lossless discrete setting before [3]. Also Timo, Bidokhti, Wigger, and Geiger studied the lossy case in a similar setting, but took a worst-case metric to design good caching strategies, whereas we look at average performance [4]. Moreover, our focus is on Gaussian sources in particular. An effective caching strategy carefully weighs two parameters: the correlation between the elements of the database, and the user’s preference for one.

[1] The material of this chapter appeared in [25–27]

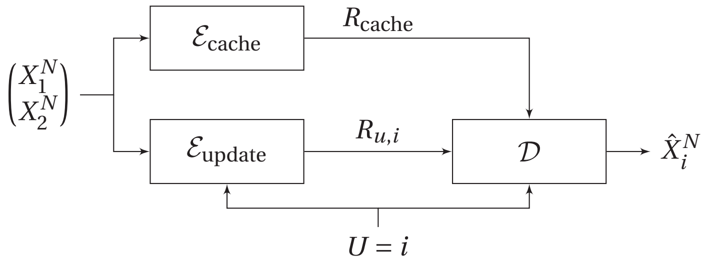

$\begin{pmatrix} X_1^N \\ X_2^N \end{pmatrix}$
$\mathcal{E}_{\text{cache}}$
R_{cache}
$\mathcal{E}_{\text{update}}$
$R_{u,i}$
$\mathcal{D}$
$\hat{X}_i^N$
$U = i$

Figure 4.1 – The Caching Network

4.1 Problem Statement

Figure 4.1 serves as our model to capture the essence of caching as earlier described. For discrete sources, this model was also studied in [3]. The database consists of two 'files', X_1^N and X_2^N and at some point the user will submit a request for either of the two. However, like explained in the introduction the encoder does not wish to wait for this to happen, but instead wants to already transmit some data ahead of time.

The model is accompanied by a timeline of three events:

1. A *cache* message is sent while the encoder is still unaware of the user's choice.
2. The user submits a request for either X_i^N.
3. A second encoder sends an *update* message tailored towards X_i^N to complement the cache. Both messages combined need to create a final lossy description $\hat{X}_i^N$ at a precision that is acceptable for the user.

The fundamental question is: What data should the first encoder write in the cache if it does not know what will be requested?

4.1.1 The Caching Network

Let $\mathbf{X}^N$ be a sequence of two-dimensional random vectors which fulfills the role of the database in our model. Each sample of $\mathbf{X}$ is drawn in an i.i.d. fashion, but within one sample the vector elements X_1 and X_2 can be correlated. This sequence is to be encoded into three messages, m_c for the 'cache' and $m_{u,1}, m_{u,2}$ for the 'update'. The cache message m_c is transmitted in any case, while for the update the decoder will only receive $m_{u,1}$ if it requests X_1^N (and similar for $m_{u,2}$ and X_2^N). Each message m is an integer in the set $I_M = \{1, 2, \cdots, M\}$, where the set size for each message is denoted by $M_c, M_{u,1}$ and $M_{u,2}$.

A code consists of encoder mappings

$$f: \quad \mathbb{R}^{2\times N} \to I_{M_c} \times I_{M_{u,1}} \times I_{M_{u,2}}$$

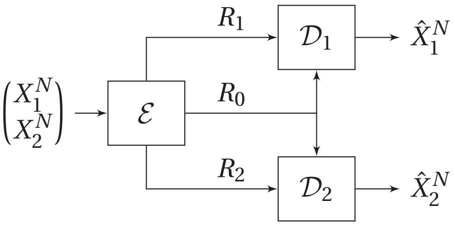


Figure 4.2 – The Gray–Wyner network.

and a decoder

$$g_i : \quad I_{M_c} \times I_{M_{u,i}} \to \mathbb{R}^{1\times N} \qquad \text{for } i = 1,2$$

of which the latter is meant to reconstruct $\hat{X}_i^N = g_i(m_c, m_{u,i})$ where i stands for the specific file requested.

A caching rate-distortion tuple $(R_{\text{cache}}, R_{u,1}, R_{u,2}, D_F)$ is said to be achievable if for arbitrary $\epsilon > 0$ there exist such encoders and decoders that for both $i = 1,2$ we have

$$\begin{aligned} M_c &\leq 2^{N(R_{\text{cache}}+\epsilon)} \\ M_{u,i} &\leq 2^{N(R_{u,i}+\epsilon)} && \text{for } i = 1,2 \\ \frac{1}{N}\sum_{n=1}^{N} d(X_i(n), \hat{X}_i(n)) &\leq D_F + \epsilon && \text{for } i = 1,2 \end{aligned}$$

where $d_X(\cdot,\cdot)$ is some single-letter distortion measure. As is clear from the definition, we apply symmetric end distortion criteria to X_1 and X_2. This is a matter of notational convenience and it will later become clear that (a)symmetry is neither important nor interesting. A far more interesting thing to study in the Gaussian case are the *intermediate* distortion levels after caching, these will be introduced in Section 4.1.4.

After the caching phase, the user submits a requests for either X_1^N or X_2^N which is modeled by the Bernoulli random variable $U \in \{1,2\}$, distributed as $P(U = 1) = p$. The main question we pose is: What does one need to cache in order to minimize the update rate that is still needed on average? To that end define also the *average* update rate:

$$\overline{R}_{\text{update}} \triangleq pR_{u,1} + (1-p)R_{u,2}. \tag{4.1}$$

A shorthand notation will be to say that $(R_{\text{cache}}, \overline{R}_{\text{update}}, D_F)$ is achievable to indicate that (at least one tuple) $(R_{\text{cache}}, R_{u,1}, R_{u,2}, D_F)$ is, of which the average update rate equals $\overline{R}_{\text{update}}$.

4.1.2 Analogy to the Gray–Wyner Network

From an operational perspective and the existence of codes, there is a complete equivalence with the Gray–Wyner network [7], as depicted in Figure 4.2. Namely, even though the decoder only needs X_1 or X_2, any code should be capable of doing both as the user could request any of the two. By taking Figure 4.1 and drawing the events of the user asking for X_1 or X_2 as two separate decoders one obtains 4.2. The equivalence is thus as follows:

$$R_0 \leftrightarrow R_{\text{cache}}, \quad R_1 \leftrightarrow R_{u,1}, \quad R_2 \leftrightarrow R_{u,2}.$$

The Gray–Wyner network was introduced originally in [7], but also featured more recently in an explicit lossy source coding setting [11, 12]. The region of achievable rate-distortion tuples on the Gray–Wyner network is the union of all $(R_0, R_1, R_2, D_1, D_2)$ satisfying

$$R_0 \geq I(\mathbf{X}; V) \tag{4.2}$$

$$R_1 \geq I(X_1; \hat{X}_1|V) \tag{4.3}$$

$$R_2 \geq I(X_2; \hat{X}_2|V) \tag{4.4}$$

$$D_1 \geq \mathbb{E}[d_{X_1}(X_1, \hat{X}_1)] \tag{4.5}$$

$$D_2 \geq \mathbb{E}[d_{X_2}(X_2, \hat{X}_2)] \tag{4.6}$$

over joint densities $p(\mathbf{x}, v, \hat{\mathbf{x}})$, for some distortion measures $d_X(\cdot,\cdot)$. Notation-wise, V may be a single random variable, or also a vector, but for consistency (and with the end result in mind) we stick to denoting V by a non-bold character.

Hence, by the equivalence with the Gray–Wyner network one knows which caching strategies are achievable:

Theorem 4.1. *Given a joint density $p(\mathbf{x}, v, \hat{\mathbf{x}})$, all caching rate-distortion tuples $(R_{cache}, \overline{R}_{update}, D_F)$ satisfying the following inequalities*

$$\begin{aligned} R_{cache} &\geq I(\mathbf{X}; V) \\ \overline{R}_{update} &\geq pI(X_1; \hat{X}_1|V) + (1-p)I(X_2; \hat{X}_2|V) \\ D_F &\geq \mathbb{E}[d_{X_i}(X_i, \hat{X}_i)] \qquad \textit{for } i = 1, 2 \end{aligned}$$

are achievable. The closure of such achievable tuples over joint densities $p(\mathbf{x}, v, \hat{\mathbf{x}})$ is denoted $\mathcal{R}_{caching}$.

The goal is now to better understand the boundary of the $(R_{\text{cache}}, \overline{R}_{\text{update}})$ trade-off, to understand which strategies are not only achievable, but are also *good*.

4.1.3 Characteristics and (Non-)Attainable Limits

For a fixed D_F, the boundary of the caching rate-distortion region of Theorem 4.1 is a curve that is convex in R_{cache} and lies inside the triangle depicted in Figure 4.3. That shape is the intersection of the following three bounds. Recall for these equations that $R(D)$ is the single rate-distortion function and $R(D_1, D_2)$ is the bivariate rate-distortion function subject to individual distortion constraints as defined in Chapter 3, specifically equations (3.3) and

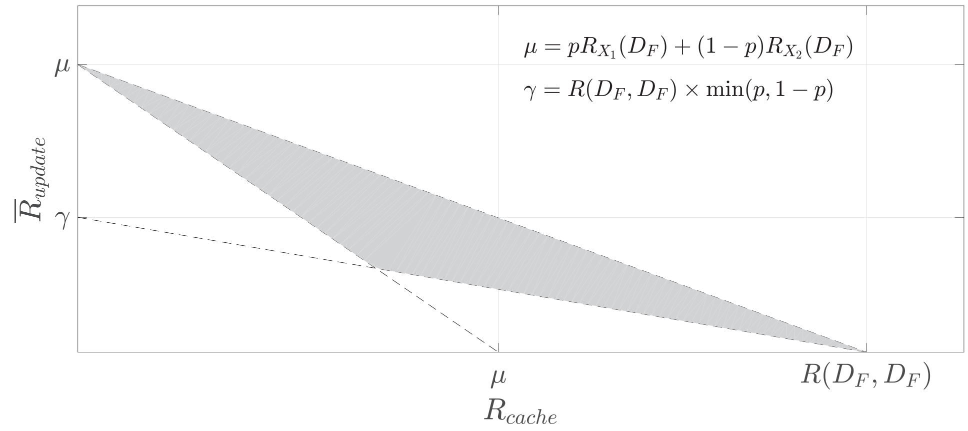


Figure 4.3 – The boundary of achievable $(R_{\text{cache}}, \overline{R}_{\text{update}})-$pairs lies inside this gray triangle.

(3.19). Figure 4.3 is then built as follows:

1. An achievable inner bound based on time-sharing the extremal strategies of caching both files completely in advance or sending only one file completely in the update phase by waiting for the user to make a request:

$$\begin{aligned} \text{update everything} \quad & (R_{\text{cache}}, \overline{R}_{\text{update}}) = (0, pR_{X_1}(D_F) + (1-p)R_{X_2}(D_F)) \\ \text{or cache everything} \quad & (R_{\text{cache}}, \overline{R}_{\text{update}}) = (R(D_F, D_F), 0). \end{aligned}$$

2. An outer bound connecting the two points:

$$\begin{aligned} (R_{\text{cache}}, \overline{R}_{\text{update}}) &= (0, pR_{X_1}(D_F) + (1-p)R_{X_2}(D_F)) \\ (R_{\text{cache}}, \overline{R}_{\text{update}}) &= (pR_{X_1}(D_F) + (1-p)R_{X_2}(D_F), 0), \end{aligned}$$

of which the first point is achievable with certainty. The bound stems from the following inequality:

$$\begin{aligned} R_{\text{cache}} + \overline{R}_{\text{update}} &= p(R_{\text{cache}} + R_{u,1}) + (1-p)(R_{\text{cache}} + R_{u,2}) \\ &\geq pR_{X_1}(D_F) + (1-p)R_{X_2}(D_F). \end{aligned}$$

If $R_{\text{cache}} = 0$, then all communication happens in the update phase when the encoder is aware of the user's request. Therefore indeed the leftmost point of this outer bound coincides with the start of the boundary of achievable $(R_{\text{cache}}, \overline{R}_{\text{update}})-$pairs.

3. An outer bound connecting the two points:

$$\begin{aligned} (R_{\text{cache}}, \overline{R}_{\text{update}}) &= (0, R(D_F, D_F)) \\ (R_{\text{cache}}, \overline{R}_{\text{update}}) &= (\min(p, 1-p) \times R(D_F, D_F), 0), \end{aligned}$$

of which the second is achievable with certainty.
This bound is the consequence of $R_{\text{cache}} + R_{u,1} + R_{u,2} \geq R(D_F, D_F)$. Equality implies that the joint rate-distortion function can be completely distributed over all branches of the Gray–Wyner network. This is known to be possible in some cases (R_{cache} needs to be large) and definitely not in others (see, e.g., [11]). The caching of Gaussian sources will also have equality for some $(R_{\text{cache}}, \overline{R}_{\text{update}})$, as will be detailed later.

4.1.4 The Gaussian Case

Now let us zoom in further: the database $\mathbf{X}^N$ are IID samples from a Gaussian distribution $\sim \mathcal{N}(0, \Sigma_{\mathbf{X}})$ with covariance

$$\Sigma_{\mathbf{X}} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

Reflecting on the equations of Theorem 4.1, one must note that the source $\mathbf{X}$ being Gaussian distributed does not imply that V and $\hat{\mathbf{X}}$ are necessarily also Gaussian on the boundary of this region. In Corollaries 4.2 and 4.5 we will show that using Gaussian codebooks it is possible to attain communication at rates for which it holds that $R_{\text{cache}} + R_{u,1} + R_{u,2} = R(D_F, D_F)$. Since one cannot do better than the joint rate-distortion function, Gaussian auxiliaries are thus sufficient for optimality in these cases.

However, in general it holds that $R_{\text{cache}} + R_{u,1} + R_{u,2} \geq R(D_F, D_F)$ and it is clear a priori that this condition cannot be met with equality everywhere. For starters, we have in Figure 4.3 that the outer bound associated to this inequality crosses another outer bound. Whenever the inequality is strict, to the best of our knowledge it is not known whether Gaussian auxiliaries are sufficient for optimality; this remains an open problem. In this work, we restrict ourselves to all variables being Gaussian. From here onwards we therefore speak of the *Gaussian* achievable caching rate-distortion region.

Corollary 4.1. *The Gaussian boundary of the caching rate-distortion region is characterized by*

$$R_{cache}(d, D_F) = \min_{D_F \leq D_1, D_2 \leq 1} R(D_1, D_2) \quad s.t. \quad D_1^p D_2^{1-p} \leq d, \tag{4.7}$$

for a normalized parameter $d \in [D_F, 1]$ that relates back to $\overline{R}_{update}$ by picking $d \leq D_F 2^{2\overline{R}_{update}}$.

Proof. A shorthand but equally correct characterization of $\mathcal{R}_{\text{caching}}$ is by means of the conditional rate-distortion function, i.e. to take the union over all $p(\mathbf{x}, v)$ of

$$\begin{cases} R_{\text{cache}} & \geq I(\mathbf{X}; V) \\ \overline{R}_{\text{update}} & \geq pR_{X_1|V}(D_F) + (1-p)R_{X_2|V}(D_F). \end{cases}$$

By taking $p(\mathbf{x}|v)$ to be Gaussian, also $R_{X_i|V}(D_F)$ is solved by Gaussian distributions since it is a regular rate-distortion function. Then, the cache rate condition translates to:

$$I(\mathbf{X}; V) = \frac{1}{2}\log\frac{|\Sigma_{\mathbf{X}}|}{|\mathbf{D}|}, \tag{4.8}$$

and the update

$$\begin{aligned} pI(X_1; \hat{X}_1|V) + (1-p)I(X_2; \hat{X}_2|V) &= \frac{p}{2}\log^+\frac{D_{1,1}}{D_F} + \frac{1-p}{2}\log^+\frac{D_{2,2}}{D_F} && (4.9)\\ &= \frac{1}{2}\log^+\frac{D_{1,1}^{p}D_{2,2}^{1-p}}{D_F}, && (4.10)\end{aligned}$$

where

$$\mathbf{D} = \mathbb{E}[(\mathbf{X}-\mathbb{E}[\mathbf{X}|V])(\mathbf{X}-\mathbb{E}[\mathbf{X}|V])^T] = \Sigma_{\mathbf{X}|V}. \tag{4.11}$$

Any positive semidefinite matrix $\mathbf{D}$ that satisfies $\mathbf{D} \preceq \Sigma_{\mathbf{X}}$ can be associated to a random variable V that is jointly Gaussian with $\mathbf{X}$, and vice versa (see Section 3.1). Such a matrix can be interpreted as a mean-squared error distortion after caching, but before the update phase. In that light, one can equivalently optimize over all $\mathbf{D}$ rather than over all Gaussian distributions $p(\mathbf{x}, v)$.

To characterize the boundary, we wish to fix one rate and minimize the other:

$$R_{\text{cache}}(\gamma) = \min R_{\text{cache}} \quad \text{s.t.} \quad \overline{R}_{\text{update}} \le \gamma.$$

As a matter of definition, instead of fixing $\overline{R}_{\text{update}}$ one can equivalently fix $D_{1,1}^{p}D_{2,2}^{1-p}$ to emphasize that the distortions up to which one caches the sources are truly the intrinsic variables of this problem, they are both objective *and* constraint. Observe also that it serves no purpose to cache either X_1 or X_2 beyond the final distortion constraint D_F (4.9). In other words, it is futile to pick a caching distortion profile $\mathbf{D}$ of which $D_{1,1} < D_F$ or $D_{2,2} < D_F$; that rate is better spent on caching a component that does not yet satisfy the end criterion.

Combining all, define and simplify the following *caching rate-distortion function*:

$$\begin{aligned} R_{\text{cache}}(d, D_F) &= \min_{\mathbf{D}} \quad \frac{1}{2}\log\frac{|\Sigma_{\mathbf{X}}|}{|\mathbf{D}|} \quad \text{s.t.} \begin{cases} 0 \preceq \mathbf{D} \preceq \Sigma_{\mathbf{X}} \\ D_{1,1}^{p}D_{2,2}^{1-p} \le d \\ D_{1,1}, D_{2,2} \ge D_F \end{cases} && (4.12)\\ &= \min_{D_1, D_2 \ge D_F} \quad \min_{\mathbf{D} : \operatorname{diag}(\mathbf{D}) = (D_1, D_2)} \quad \frac{1}{2}\log\frac{|\Sigma_{\mathbf{X}}|}{|\mathbf{D}|} \quad \text{s.t.} \begin{cases} 0 \preceq \mathbf{D} \preceq \Sigma_{\mathbf{X}} \\ D_1^{p}D_2^{1-p} \le d \end{cases} && (4.13)\\ &= \min_{D_F \le D_1, D_2 \le 1} \quad R(D_1, D_2) \quad \text{s.t.} \quad D_1^{p}D_2^{1-p} \le d. && (4.14)\end{aligned}$$

□

Note that the caching rate-distortion function is a minimization of a convex function over a non-convex domain. Furthermore, also note that even though we constrain $D_1^p D_2^{1-p} \leq d$, the minimizer to (4.7) will result in equality by construction.

Definition 4.1. *A* caching strategy *refers to a pair of intermediate distortions* (D_1, D_2) *after the caching phase. Such a strategy is said to be optimal if it is the minimizer to* (4.7) *w.r.t. an instance of* d *and* D_F.

The extreme ends of (4.7) like depicted in Figure 4.3 simplify thanks to the symmetry assumptions (assuming $p \notin \{0,1\}$):

$$\begin{aligned} R_{\text{cache}}(D_F, D_F) &\rightarrow (R_{\text{cache}}, \overline{R}_{\text{update}}) = (R(D_F, D_F), 0) \qquad \textit{all cache,} \\ R_{\text{cache}}(1, D_F) &\rightarrow (R_{\text{cache}}, \overline{R}_{\text{update}}) = (0, R(D_F)) \qquad \textit{all update.} \end{aligned}$$

In the following sections we aim to understand $R_{\text{cache}}(d, D_F)$ while playing with both the correlation ρ and the file preference p. As the bivariate rate-distortion function plays a key role, we urge the reader to take note of its construction in equation (3.19) and in particular the $\mathcal{D}-$plane that distinguishes its behavior in (3.20)–(3.22) and Figure 3.6.

4.2 Preference, but no dependence

First, we establish that in a database in which files have no information in common, the server should cache the most preferred file exclusively. Only when the distortion constraint on that data is met, should he continue caching the other. This remark is not specific to Gaussians.

Theorem 4.2. *Assuming without loss of generality that $p > 1-p$, if X_1 and X_2 are independent then the boundary of $\mathcal{R}_{caching}$ is the following connection of two straight lines:*

$$\begin{aligned} R_{cache} &= \delta \\ \overline{R}_{update} &= p\left(R_{X_1}(D_F)-\delta_1\right)^+ + (1-p)\left(R_{X_2}(D_F)-\delta_2\right)^+ \end{aligned}$$

for $\delta \in [0, R_{X_1}(D_F)+R_{X_2}(D_F)]$ and

$$(\delta_1,\delta_2) = \begin{cases} (\delta, 0) & \text{if } \delta \le R_{X_1}(D_F) \\ (R_{X_1}(D_F), \delta - R_{X_1}(D_F)) & \text{if } R_{X_1}(D_F) \le \delta \le R_{X_1}(D_F)+R_{X_2}(D_F). \end{cases}$$

If $p < 1-p$, X_1 and X_2 switch roles.

Proof. Recall an important lower bound from the Gray–Wyner network:

$$\begin{aligned} R_{\text{cache}} + R_{u,1} + R_{u,2} &\ge R(D_F, D_F) \\ &\ge R_{X_1}(D_F) + R_{X_2}(D_F). \end{aligned} \tag{4.15}$$

When X_1 and X_2 are independent equality holds in the last step by definition. Equality in the first line is attainable for all $R_{\text{cache}} \in [0, R_{X_1}(D_F)+R_{X_2}(D_F)]$ by splitting the cache message into two parts, one for each decoder. Each decoder then has a personal cache and update link over which the encoder can split the bits of $R_{X_i}(D_F)$. Call the size of those parts δ_1 and δ_2 for decoders 1 and 2. Let $\delta \in [0, R_{X_1}(D_F)+R_{X_2}(D_F)]$. Then $\overline{R}_{\text{update}}$ is minimized by:

$$\underset{\delta_1+\delta_2=\delta}{\arg\min} \quad p\left(R_{X_1}(D_F)-\delta_1\right)^+ + (1-p)\left(R_{X_2}(D_F)-\delta_2\right)^+, \tag{4.16}$$

which is solved by distributing rate in a greedy fashion as mentioned in the theorem: the encoder should cache X_1 exclusively until it satisfies the desired end distortion constraint. □

It is important to realize that the cache message may very well be meaningless on its own (which is part of the reason why it is hard to prove that Gaussian strategies are optimal for Gaussian sources). However, if X_1, X_2 are independent *and* they are successively refinable, then the caching could also be implemented as an application of successive refinability [15]. Instead of splitting the bits of $R(D_F)$ over both phases, one caches the most popular X_i up to some distortion D_{cache} for $R(D_{\text{cache}})$ bits and then refines this to $D_F \le D_{\text{cache}}$ using $R(D_F)-R(D_{\text{cache}})$ bits as explained in Section 3.4. The cache message then serves as a 'thumbnail' of what is to come. For Gaussians, this is already intrinsic to $R_{\text{cache}}(d, D_F)$ (4.7).

4.3 Dependence, but no preference

In the absence of user preference, the caching problem revolves completely around correlation. On a high level, if it is equally likely that the user picks X_1 or X_2 then there is no bias to be leveraged. The best strategy for the encoder is then to cache the information that is *shared* by both files; after all, the shared information is useful no matter the choice the user makes, whereas information that is unique to one file might go to waste. In that light, we will argue that if $p = \frac{1}{2}$ the caching problem becomes an application of the concepts of Wyner's common information [8] and Watanabe's total correlation [9]. In a Gaussian setting, these concepts are encapsulated in the Hadamard inequality. To be precise, the gap on that inequality; we will show that a good caching strategy is a distortion matrix of which that gap is as small as possible. The consequence is that symmetry in user preference results in symmetry in the caching strategy:

Theorem 4.3. *If* $p = 1 - p = \frac{1}{2}$ *then*

$$\begin{aligned} R_{cache}(d, D_F) &= R(d,d) \\ &= \begin{cases} \frac{1}{2}\log\frac{1-\rho^2}{d^2} & \textit{for } d \in [D_F,\ 1-|\rho|], \\ \frac{1}{2}\log\frac{1-\rho^2}{d^2-(d-(1-|\rho|))^2} & \textit{for } d \in [1-|\rho|,\ 1]. \end{cases} \end{aligned}$$

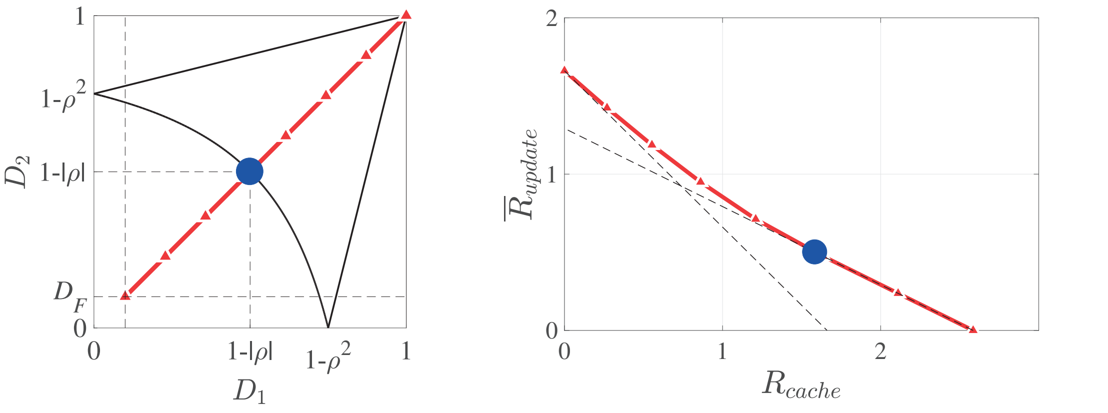
The proof will be divided into two key steps, given by upcoming Lemmas 4.2 and 4.3.

An example is plotted in Figure 4.4. The result may not appear as surprising; we set both end distortions to be equal to one criterion D_F, hence one could argue that it is logical that the optimal caching strategy would be to cache X_1 and X_2 equally. However, the symmetric caching strategy stems from $p = 1 - p$ and not from the symmetric end distortion constraints. It would also hold if the end criteria were asymmetric (assuming R_{cache} is not enough to fully code either X_i). To see this, we develop the proof of Theorem 4.3 in the next subsections at perhaps a slower pace than necessary.

4.3.1 Efficient Gaussian Caching is Closing the Hadamard Inequality

In Corollary 4.1 we established that a Gaussian caching strategy can be picked by means of an MSE distortion matrix $\mathbf{D}$ and that without loss of optimality one can always pick a matrix that is rate-distortion optimal with respect to $R(D_1, D_2)$. Instead of optimizing over all matrices $\mathbf{D}$, one would then only have to optimize over the marginal distortions D_1 and D_2. Let us, however, take one step back and evaluate $(R_{\text{cache}}, \overline{R}_{\text{update}})$ of a general Gaussian strategy:

$$\begin{cases} R_{\text{cache}} &= \frac{1}{2}\log\frac{|\Sigma_{\mathbf{X}}|}{|\mathbf{D}|}, \\ \overline{R}_{\text{update}} &= \frac{1}{2}\log\frac{D_{1,1}^{1/2}D_{2,2}^{1/2}}{D_F} = \frac{1}{4}\log\frac{D_{1,1}D_{2,2}}{D_F^2}. \end{cases}$$



(a) The optimal caching strategy requires $D_1 = D_2$ at all time.

(b) An example of the consequent trade-off between R_{cache} and $\overline{R}_{\text{update}}$.

Figure 4.4 – If $p = \frac{1}{2}$, optimal caching strategies must lie on the diagonal line in the $\mathcal{D}-$plane. The blue dot corresponds to $R_{\text{cache}} = C_W(X_1, X_2)$, the best strategy then moves from $\mathcal{D}_2$ into $\mathcal{D}_1$. For illustration purposes, the left is drawn with $\rho = 0.5$ and the right with $\rho = 0.8$.

Small R_{cache} requires the determinant to be large, whereas a small $\overline{R}_{\text{update}}$ forces us to minimize the product of the diagonal entries of $\mathbf{D}$. These two are related in the Hadamard inequality

$$D_{1,1}D_{2,2} \geq \mathbf{D}, \tag{4.17}$$

which relates to a classic information theoretic inequality

$$h(X_1|V) + h(X_2|V) \geq h(\mathbf{X}|V), \tag{4.18}$$

by realizing the following relationship:

$$\begin{cases} h(X_1|V) + h(X_2|V) & = \frac{1}{2}\log(2\pi e)^2 D_{1,1}D_{2,2}, \\ h(\mathbf{X}|V) & = \frac{1}{2}\log(2\pi e)^2|\mathbf{D}|. \end{cases} \tag{4.19}$$

A nice discussion of this one-to-one correspondence can be read in the classic book by Cover and Thomas [18, Section 17.9]. These inequalities put the following limit on performance:

Lemma 4.1. *The caching rate-distortion function is lower bounded as:*

$$R_{cache}(d, D_F) \geq \frac{1}{2}\log\frac{|\Sigma_{\mathbf{X}}|}{d^2}.$$

Proof. The determinant is bounded by the product of the diagonal, which in turn is bounded by the constraint on $\overline{R}_{\text{update}}$ (by the definition of $R_{\text{cache}}(d, D_F)$):

$$d^2 \geq D_{1,1}D_{2,2} \geq |\mathbf{D}|. \tag{4.20}$$

□

Realizing that R_{cache} is minimized by maximizing $|\mathbf{D}|$, the encoder must find a distortion profile $\mathbf{D}$ that closes the gap on this inequality. Reflecting on the information theoretic version of the inequality: 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.

It turns out that whether equality is attainable or not separates the caching problem into two distinct regions that exhibit different behavior.

4.3.2 High Cache Rate Region & Common Information

Equality on the Hadamard inequality requires the caching distortion profile $\mathbf{D}$ to be diagonal, In a Gaussian universe this implies

$$h(X_1|V) + h(X_2|V) = h(\mathbf{X}|V),$$

X_1 and X_2 have to become conditionally independent given V. This is not trivially attainable. Namely, if $\mathbf{D}$ must be diagonal, then $|\mathbf{D}|$ cannot be arbitrarily large in order for $\mathbf{D} \preceq \Sigma_{\mathbf{X}}$ to hold. From an information theoretic perspective, to make X_1 and X_2 conditionally independent, $I(\mathbf{X}; V)$ cannot be arbitrarily small. The latter formulation led to the concept of *Wyner's Common Information*:

Definition 4.2 ([8]). *Wyner's Common Information is defined as*

$$C_W(X_1, X_2) \triangleq \min_{X_1 - V - X_2} I(\mathbf{X}; V). \tag{4.21}$$

Thanks to [28] we know that for two Gaussians we have

$$C_W(X_1, X_2) = \frac{1}{2} \log \frac{1 - |\rho|}{1 + |\rho|}, \tag{4.22}$$

which corresponds to a jointly Gaussian V which results in the following distortion

$$\mathbf{D}_{C_W} = \begin{bmatrix} 1 - |\rho| & 0 \\ 0 & 1 - |\rho| \end{bmatrix}. \tag{4.23}$$

$C_W(X_1, X_2)$ being a minimum has consequences on the $(R_{\text{cache}}, \overline{R}_{\text{update}})$ trade-off as follows:

Lemma 4.2. *Equality in Lemma 4.1 is achievable if and only if $R_{cache} \geq C_W(X_1, X_2)$, which is in the following range of d:*

$$d \in \begin{cases} [D_F,\ 1 - |\rho|] & \textit{if } D_F \leq 1 - |\rho|, \\ \emptyset & \textit{if } D_F > 1 - |\rho|. \end{cases}$$

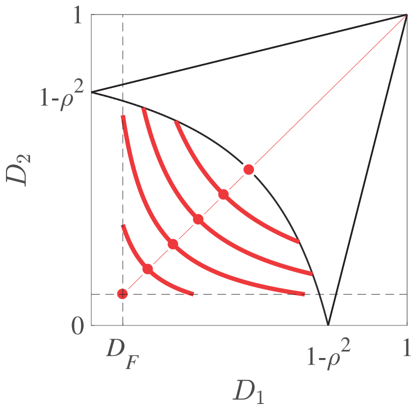


Figure 4.5 – Once $R_{\text{cache}} \geq C_W(X_1, X_2)$, the optimal caching strategies lie inside $\mathcal{D}_1$ where the joint rate-distortion function is separable. Therefore, all caching strategies on the hyperbola $\frac{1}{D_1 D_2} = \frac{1}{d^2}$ achieve $R_{\text{cache}}(d, D_F)$.

Proof. Tying this and the previous section together: Equality in Lemma 4.1 requires equality in (4.20), which for Gaussians means that X_1 and X_2 become conditionally independent. By the definition of Wyner's Common Information, this can only be done if $R_{\text{cache}} \geq C_W(X_1, X_2)$.

For $D_F > 1 - |\rho|$ we have, however, that $R(D_F, D_F) < C_W(X_1, X_2)$, i.e., even if the encoder would cache X_1 and X_2 together and completely it would not need more rate than the common information. It is the regime where $(D_F, D_F) \in \mathcal{D}_2$ (as defined in (3.21) in Section 3.3) and therefore the rate-distortion optimal encoding of X_1 and X_2 is associated to a distortion matrix that cannot be diagonal. For $D_F \leq 1 - |\rho|$, the lower bound is achievable for all $R_{\text{cache}} \in [C_W(X_1, X_2),\ R(D_F, D_F)]$. For example, let $\alpha \in [\frac{D_F}{1-|\rho|}, 1]$ and construct $\mathbf{D}' = \alpha \mathbf{D}_{C_W}$. Then it holds that $\mathbf{D}' \preceq \mathbf{D}_{C_W} \preceq \Sigma_{\mathbf{X}}$ (it is an achievable Gaussian distortion matrix) and all $\mathbf{D}'$ are diagonal, achieving the optimum and spanning all $R_{\text{cache}} \in [C_W(X_1, X_2),\ R(D_F, D_F)]$. □

Corollary 4.2. *If $R_{cache} \geq C_W(X_1, X_2)$ then $R_{cache}(d, D_F)$ also characterizes the boundary of $\mathcal{R}_{caching}$ in general.*

Proof. This optimality of Gaussians is a consequence of the rate-distortion function being separable over all links of the Gray–Wyner network:

$$\begin{aligned} R_{\text{cache}} + R_{u,1} + R_{u,2} &= \frac{1}{2}\log\frac{|\Sigma_{\mathbf{X}}|}{|\mathbf{D}|} + \frac{1}{2}\log\frac{D_{1,1}}{D_F} + \frac{1}{2}\log\frac{D_{2,2}}{D_F} \\ &= \frac{1}{2}\log\frac{1-\rho^2}{D_{1,1}D_{2,2}} + \frac{1}{2}\log\frac{D_{1,1}D_{2,2}}{D_F^2} \\ &= \frac{1}{2}\log\frac{1-\rho^2}{D_F^2} \\ &= R(D_F, D_F). \end{aligned}$$

Evidently, one cannot beat the joint rate-distortion function. In this regime caching achieves the third outer bound described in Section 4.1.3. □

The full story of the separability of $R(D_1, D_2)$ on the Gray–Wyner network and the role of Wyner's Common Information can be found in the work of Viswanatha, Akyol and Rose [11].

Corollary 4.3. *If $R_{cache} \geq C_W(X_1, X_2)$ then there exist infinitely many distortion profiles $\mathbf{D}$ that optimize $R_{cache}(d, D_F)$.*

Proof. For $(D_1, D_2) \in \mathcal{D}_1$ (3.20) the joint rate-distortion function behaves as $R(D_1, D_2) = \frac{1}{2}\log\frac{1-\rho^2}{D_1 D_2}$. Consequently the cache and update phase fit together seamlessly for caching strategies inside $\mathcal{D}_1$, since $\overline{R}_{\text{update}} = \frac{1}{4}\log\frac{D_1 D_2}{D_F^2}$ (also mentioned in the previous corollary). Hence, if $d \leq 1 - |\rho|$ then any $\mathbf{D} = \text{diag}(D_1, D_2)$ of which $D_1 D_2 = d^2$ and $(D_1, D_2) \in \mathcal{D}_1$ constitute the same performance of $(R_{\text{cache}}, \overline{R}_{\text{update}})$. In other words, for $R_{\text{cache}} \geq C_W(X_1, X_2)$ infinitely many optimal caching strategies lie in $\mathcal{D}_1$ on hyperbola $\frac{1}{D_1, D_2} = \frac{1}{d^2}$ as depicted in Figure 4.5. □

The high cache rate regime is plotted on the right-hand side of Figure 4.4b. Observe how the slope of the $(R_{\text{cache}}, \overline{R}_{\text{update}})$ trade-off becomes constant. This is the result of the caching strategy capturing all the information that is shared between X_1 and X_2. For $R_{\text{cache}} \leq C_W(X_1, X_2)$ all that remains to cache is information that is individual to either X_i. Therefore, should one decide to continue caching then only a per-file gain remains.

Theorems in Pictures

Wyner's Common Information also has a special geometric meaning. Namely, (4.22) can be found in two steps: first by proving that $C_W(X_1, X_2)$ is solved by a V that is jointly Gaussian with $\mathbf{X}$, then by finding the best possible V. For Gaussians independence means diagonal covariance/distortion. Therefore:

$$\min_{X_1 - V - X_2} I(\mathbf{X}; V) = \max_{\mathbf{D}} \frac{1}{2} \log \frac{|\Sigma_{\mathbf{X}}|}{|\mathbf{D}|} \quad \text{s.t.} \begin{cases} \mathbf{0} \preceq \mathbf{D} \preceq \Sigma_{\mathbf{X}}, \\ \mathbf{D} \text{ is diagonal.} \end{cases}$$

Geometrically, this optimization looks for the ellipse with the largest possible volume (the objective) that lies inside $\mathcal{E}_{\Sigma_{\mathbf{X}}}$ (constraint 1) and is straight, i.e., whose semiprincipal axes align with the axes of the system (constraint 2). For any general 2×2 covariance matrix, this is solved by $\mathbf{D} = \operatorname{diag}(\sigma_1^2(1 - |\rho|), \sigma_2^2(1 - |\rho|))$. Consequently, if X_1, X_2 are of unit variance then $\mathbf{D}_{C_W}$ is a scaled identity matrix: $\mathcal{E}_{\mathbf{D}_{\mathbf{C_W}}}$ is thus a circle and it touches $\mathcal{E}_{\Sigma_{\mathbf{X}}}$ at its semi-minor axis.

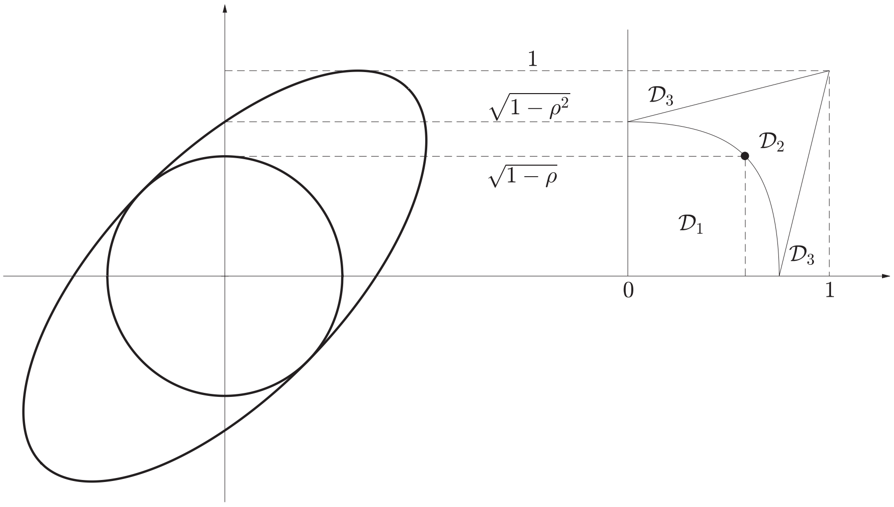


Figure 4.6 – The correspondence between the ellipses (left) and the $\mathcal{D}$−plane (right). To match points, the $\mathcal{D}$−plane is drawn with $\sqrt{D_1}$ instead of D_1 (same for D_2).

4.3.3 Low Cache Rate Region & Total Conditional Correlation

The previous section introduced the notion of Wyner's Common Information as a threshold of where the caching-update trade-off reaches its maximum performance: It was associated to a R_{cache} large enough to carry all the information that is shared between X_1 and X_2. What about the other half of the $(R_{\text{cache}}, \overline{R}_{\text{update}})$ trade-off, the region $R_{\text{cache}} < C_W(X_1, X_2)$? In any case, it must be so that for any caching auxiliary V we might consider it holds that (4.18) is strict, that

$$I(X_1; X_2|V) = h(X_1|V) + h(X_2|V) - h(\mathbf{X}|V) > 0. \tag{4.24}$$

This brings us to a notion that is closely related to common information:

Definition 4.3. *[9] Let* $\mathbf{X}$ *be a* K*-dimensional random variable. Then Watanabe's total correlation equals*

$$TC(\mathbf{X}) = \sum_{i=1}^{K} h(X_i) - h(\mathbf{X}),$$

which extends to total conditional correlation as:

$$TC(\mathbf{X}|V) = \sum_{i=1}^{K} h(X_i|V) - h(\mathbf{X}|V).$$

In our bivariate case this expression simplifies to $I(X_1; X_2|V)$, but we will refer to it as total conditional correlation nonetheless. In this terminology, the insights so far can be restated as saying that when $R_{\text{cache}} \geq C_W(X_1, X_2)$ we can apply caching strategies for which $TC(\mathbf{X}|V) = 0$ - no shared information is left after caching.

For $R_{\text{cache}} < C_W(X_1, X_2)$, whatever cache auxiliary V one picks it must be so that $TC(\mathbf{X}|V) > 0$, but the encoder should try to get as close to 0 as possible. One thing is certain for low cache rates: $\mathbf{D} \preceq \Sigma_{\mathbf{X}}$ is an active constraint (it is not a *strict* inequality). Namely, if it were not active then there would be opportunity to get closer to the bound of Lemma 4.1 until it is[2]. Consequently, if $\mathbf{D}^*$ is the minimizer of (4.7) then it must be that

$$\dim\left(\ker\left(\Sigma_{\mathbf{X}} - \mathbf{D}^*\right)\right) \neq 0. \tag{4.25}$$

Combining this with the condition $\mathbf{D} \preceq \Sigma_{\mathbf{X}}$, the optimal distortion profile takes on the form

$$\mathbf{D}^* = \Sigma_{\mathbf{X}} - \mathbf{v}\mathbf{v}^T \succeq 0; \tag{4.26}$$

correlation is minimized by a rank-one correction along some subspace of $\Sigma_{\mathbf{X}}$.
The question is: which?

[2] A simple trick would be rotation: Given a Gaussian caching strategy $\mathbf{D}$, multiply this distortion with a rotation matrix. The determinant (and hence R_{cache}) will remain constant, while the product of the diagonal entries (hence $\overline{R}_{\text{update}}$) can improve. If $\mathbf{D} \prec \Sigma_{\mathbf{X}}$, one can rotate until this inequality changes to $\mathbf{D} \preceq \Sigma_{\mathbf{X}}$.

Lemma 4.3. *For all $d \geq 1-|\rho|$, $R_{cache}(d, D_F)$ is optimized for the following distortion matrix:*

$$\mathbf{D} = \Sigma_{\mathbf{X}} - (1-d)\mathbf{1}\mathbf{1}^T.$$

Proof. We established $\mathbf{D} \preceq \Sigma_{\mathbf{X}}$ 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 $R_{\text{cache}}(d, D_F)$:

$$\begin{aligned} R_{\text{cache}}(d, D_F) \rightarrow \quad & \max_{\mathbf{D}} |\mathbf{D}| \\ & = \max_{\alpha, \mathbf{v}} (1-\alpha v_1^2)(1-\alpha v_2^2) - (\rho - \alpha v_1 v_2)^2 \\ & \leq \max_{\alpha, \mathbf{v}} d^2 - (\rho - \alpha v_1 v_2)^2. \end{aligned} \tag{4.27}$$

The last step follows from the constraint $D_{1,1}^{1/2} D_{2,2}^{1/2} = d$. 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, $v_1 = v_2 = \frac{1}{\sqrt{2}}$. Plugging in $(1-\alpha v_1^2)(1-\alpha v_2^2) = d$ from (4.27) tells us that $\alpha = 2(1-d)$, and thus

$$\mathbf{D} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} - (1-d)\mathbf{1}\mathbf{1}^T. \tag{4.28}$$

The above holds for $\rho > 0$. Otherwise, one can verify that $v_1 = -v_2$ is the right solution. Both cases combined gives us $|\mathbf{D}| = d - (|\rho| - (1-d))^2$. □

On a side note, this distortion matrix (4.28) is also the optimizer of the joint-rate distortion function subject to symmetric distortions $R(d, d)$; the insight of the proof is thus more that indeed these marginal distortions need to be equal. Observe as well that the all-ones vector in the rank-one correction is the dominant eigenvector of the correlation matrix $\Sigma_{\mathbf{X}}$ (for $\rho > 0$).

Lemma 4.2 and 4.3 together constitute Theorem 4.3 and Figure 4.4. We emphasize the nuance that the caching strategy in this low cache rate regime is unique, whereas it is not in the high rate regime (by Corollary 4.3).

Theorems in Pictures

The series of plots below show the evolution of distortion matrices that attain the optimal caching-updating trade-off as described by Theorem 4.3. From left to right d decreases, meaning that $R_{\text{cache}}(d, D_F)$ increases; the encoder caches more information and hence the distortion on X_1 and X_2 after caching decreases.

The middle plot depicts $\mathcal{E}_{\mathbf{D}_{\mathbf{C_W}}}$, the ellipse corresponding to $R_{\text{cache}} = C_W(X_1, X_2)$, the common information. The left two plots are two instances of the low cache-rate regime of Lemma 4.3, where the encoder caches the contribution along the dominant eigenvector **1**. The right two plots are two instances of the high cache-rate regime of Lemma 4.2, where the cache encoder spends so much rate that it can make X_1 and X_2 conditionally independent. The consequence is that the distortion matrices correspond to straight ellipses. Observe how $\mathcal{E}_{\mathbf{D}}$ no longer touches $\mathcal{E}_{\Sigma_{\mathbf{X}}}$. This indicates that $\mathbf{D} \preceq \Sigma_{\mathbf{X}}$ becomes a strict inequality.

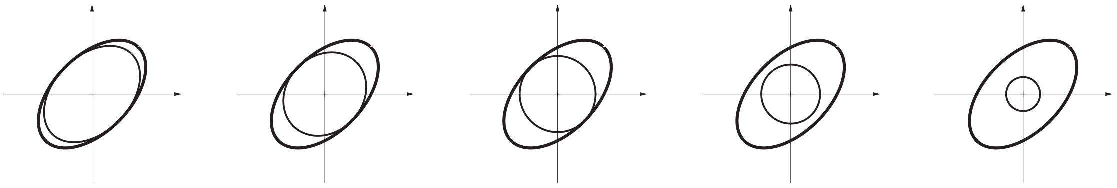


Figure 4.7 – Evolution of optimal caching distortion *matrices* that optimize $R_{\text{cache}}(d, D_F)$ for (f.l.t.r.) $d \in [0.8\ 0.6\ 0.5\ 0.3\ 0.1]$ and $\rho = \frac{1}{2}$.

4.3.4 Variance is irrelevant - Reverse Water-Filling the Correlation Matrix

Asymmetric distortion constraints

The choice for setting symmetric end distortion constraints was motivated solely by simplifying notation; the optimal caching strategy when $p = 1 - p$ results in X_1 and X_2 reaching the end distortion criteria simultaneously when $R_{\text{cache}} = R(D_F, D_F)$. The impact of choosing individual criteria for both sources leaves Theorem 4.3 intact for all R_{cache} up to the point where either $\hat{X}_i$ reaches its end criterion first. Let $D_{F,1}$ (resp. $D_{F,2}$) be the end distortion constraint for $\hat{X}_1$ (resp. $\hat{X}_2$). Define $R_{\text{cache}}(d, [D_{F,1}\ D_{F,2}])$ to be the cache rate-distortion function exactly as (4.7) but for asymmetric end distortion criteria, valid over $d \in [D_{F,1}^{1/2} D_{F,2}^{1/2},\ 1]$.

Corollary 4.4. *If $D_{F,1} > D_{F,2}$, then*

$$R_{cache}(d, [D_{F,1}\ D_{F,2}]) = R(\bar{D}_1, d^2/\bar{D}_1),$$

where $\bar{D}_1 = \max(d, D_{F,1})$. If $D_{F,1} < D_{F,2}$, switch X_1 and X_2.

Variance is irrelevant

Now consider asymmetric X_1, X_2, i.e., they are Gaussian distributed with a general covariance:

$$\Sigma_{\mathbf{X}} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}.$$

The main message of this section is that the variance is irrelevant; one can scale X_1 and X_2 to unit variance, apply the symmetric caching strategy as developed in this Section and then scale the end result back. This is possible because caching is a problem in which objective and constraint are equally affected by scaling. Namely consider

$$\max_{\mathbf{0} \preceq \mathbf{D} \preceq \Sigma_{\mathbf{X}}} |\mathbf{D}| \quad \text{s.t.} \quad D_{1,1} D_{2,2} \leq d^2.$$

Also define the matrix $\mathbf{S} = \text{diag}(\sigma_1\ \sigma_2)$. Then one can decompose $\mathbf{X} = \mathbf{S}\bar{\mathbf{X}}$, of which $\bar{\mathbf{X}}$ are two unit variance Gaussians. Then the optimization above can be equivalently written down as:

$$\max_{\mathbf{0} \preceq \mathbf{S}\bar{\mathbf{D}}\mathbf{S}^T \preceq \mathbf{S}\Sigma_{\bar{\mathbf{X}}}\mathbf{S}^T} |\mathbf{S}\bar{\mathbf{D}}\mathbf{S}^T| \quad \text{s.t.} \quad \sigma_1^2 \bar{D}_{1,1} \sigma_2^2 \bar{D}_{2,2} \leq d^2,$$

which is solved by the same $\bar{\mathbf{D}}$ as the following normalized version of the problem:

$$\max_{\mathbf{0} \preceq \bar{\mathbf{D}} \preceq \Sigma_{\bar{\mathbf{X}}}} |\bar{\mathbf{D}}| \quad \text{s.t.} \quad \bar{D}_{1,1} \bar{D}_{2,2} \leq \frac{d^2}{\sigma_1^2 \sigma_2^2}.$$

One could even redo Lemma 4.3 again for general covariances and find that the optimal caching distortion profile for $R_{\text{cache}} \leq C_W(X_1, X_2)$ takes on the form $\mathbf{D} = \Sigma_{\mathbf{X}} - (1-d)\mathbf{v}\mathbf{v}^t$, where

$\mathbf{v} = \frac{1}{\sqrt{\operatorname{tr}(\Sigma_{\mathbf{X}})}}[\sigma_1\ \sigma_2]^T$. Decomposing with the scaling matrix **S** one finds the solution we already had for caching of unit variance Gaussians.

This brings us to the main insight: The caching of bivariate Gaussians under $p = 1 - p = \frac{1}{2}$ is a reverse water-filling procedure on the eigenvectors of the *correlation* matrix. The encoder should first scale the covariance matrix to unit variance, hence turning it into a correlation matrix. Then it should cache the contribution along the dominant eigenvector of that correlation matrix according to Lemma 4.3 until R_{cache} is so large that this construction results in $\lambda_1(\mathbf{D}) = \lambda_2(\mathbf{D}) = 1 - |\rho|$. After this point the cache encoder should keep **D** diagonal and one way to do that is by reducing both eigenvalues of **D** equally.

Connecting everything to the start of this Section, we stated that an optimal strategy needs to cache **X** up to a distortion matrix with the smallest possible gap on the Hadamard inequality $|\mathbf{D}| \leq D_{1,1}D_{2,2}$. We noted the information theoretic meaning of this, which was that the encoder needs to cache as much of the shared information as possible. In this last subsection we observed that one can cache as much of the correlation as possible by ignoring variance and applying a reverse water-filling procedure on the correlation matrix. This is similar to $R_{\mathbf{X}}(D)$, the multivariate rate-distortion function subject to a sum squared error (Theorem 3.2), but different in the sense that this normalization of variance is a crucial step that comes first.

Theorems in Pictures

Recall Property 2.6 in Section 2.4: If an encoder codes a *projection* of $\mathbf{X}$ by constructing:

$$Y = \mathbf{v}^T\mathbf{X} + W,$$

with W independent Gaussian noise and $\mathbf{v}$ a projection vector, then $\mathcal{E}_{\Sigma_\mathbf{X}}$ and $\mathcal{E}_{\Sigma_{\mathbf{X}|\mathbf{Y}}}$ 'touch' at the vector that stands orthogonal to $\mathbf{v}$. Conversely, if the ellipses of two matrices $\mathbf{D} \preceq \Sigma_\mathbf{X}$ touch at a particular vector, then this distortion matrix $\mathbf{D}$ can be attained by coding a projection that again stands orthogonal to that vector.

The two plots below depict several caching-optimal distortion matrices in the range of $R_{\text{cache}} \in [0, \ C_W(X_1, X_2)]$. The left shows the caching of two Gaussians of arbitrary variance, the right of the same process after normalization. It can be seen that the ellipses of all caching profiles pass through and thus touch at the same symmetric pair of points. The arrow indicates the vector orthogonal to the vector on which the ellipses touch. While on the left this coding vector appears to be arbitrary, one can observe that after normalization they correspond to the eigenvectors of the correlation matrix.

One can derive that the arrow in the left plot constitutes the following construction:

$$Y = \frac{1}{\sqrt{\operatorname{tr}(\Sigma_\mathbf{X})}} \begin{bmatrix} \sigma_2 & \sigma_1 \end{bmatrix} \mathbf{X} + W$$

The encoder can do caching (for rates $R_{\text{cache}} \leq C_W(X_1, X_2)$) by mixing X_1 and X_2 with coefficients equal to the variance of the other. Consequently, $\sigma_2 X_1$ and $\sigma_1 X_2$ will have equal variance, thus rendering variance irrelevant. Caching (under uniform preference probabilities) is about correlation only.

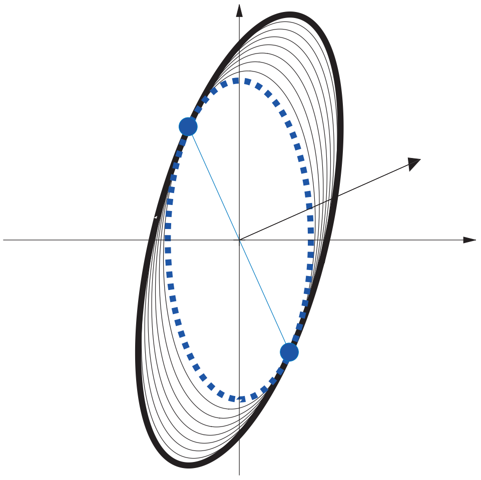


Figure 4.8 – $\sigma_1^2 = \frac{1}{2}$ and $\sigma_2^2 = \frac{5}{2}$.

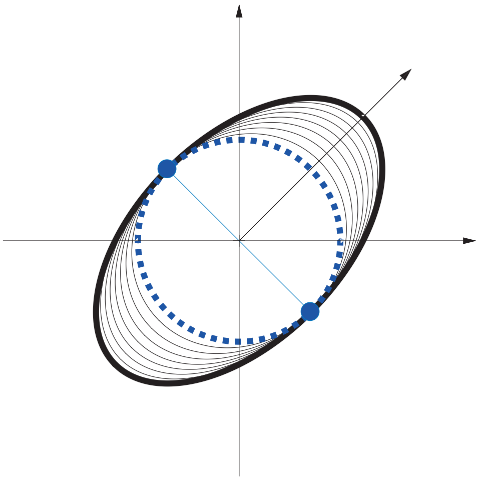


Figure 4.9 – Normalized, $\sigma_1^2 = \sigma_2^2 = 1$.

In these plots, the blue dotted lines represent the ellipses of $\mathbf{D}_{C_W}$, the distortion matrices corresponding to Wyner's Common Information.

4.4 The Full Picture: Dependent Sources with Non-Uniform Preference Probabilities

This section discusses the full $(R_{\text{cache}}, \overline{R}_{\text{update}})$ trade-off by considering $0 < |\rho| < 1$ and arbitrary p. So far we understand the following cases:

Section 4.2	$p \in \{0,1\}$ or $\rho = 0$	Cache only the most popular X_i, i.e., (D_1, D_2) must lie on the border of $\mathcal{D}_2$ and $\mathcal{D}_3$.
Section 4.3	$p = \frac{1}{2}$	Cache X_1 and X_2 equally, i.e., (D_1, D_2) must lie on the diagonal of the $\mathcal{D}-$plane.

In this section we argue that for any p between those extremes, the optimal caching strategy is a distortion pair (D_1, D_2) that lies between the diagonal and the border of $\mathcal{D}_2$ and $\mathcal{D}_3$. Moreover, we show at the end of this section that the caching problem turns out to be very sensitive to knowledge of this user preference p.

The optimal $(R_{\text{cache}}, \overline{R}_{\text{update}})$ trade-off and the strategies that attain it are shown by example in Figure 4.11. These curves are what we will derive the next few pages. To the best of our knowledge, no closed-form analytic expression exists for expressing these optimal cache-distortions in terms of p and d. Numerically, however, the problem is not hard.

First of all, it should be clear that if $p < 1 - p$ then a good caching strategy should mostly cache information on X_2 and must thus result in $D_1 > D_2$. To that end, let us cut up the $\mathcal{D}-$plane in an upper and lower triangle:

$$\mathcal{D}_{i,1} = \mathcal{D}_i \cap \{D_1, D_2 : \ D_2 \geq D_1\},$$
$$\mathcal{D}_{i,2} = \mathcal{D}_i \cap \{D_1, D_2 : \ D_2 \leq D_1\}.$$

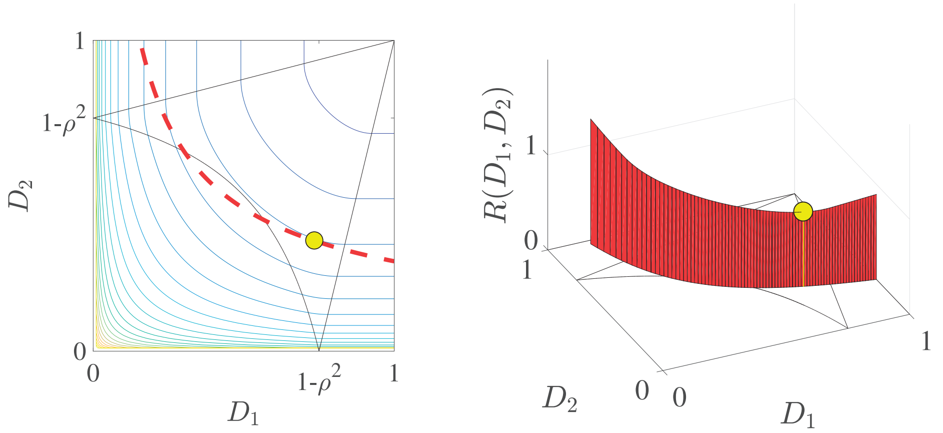
Lemma 4.4. *If $D_F = 0$, then the cache-update trade-off has one unique minimum on the $\mathcal{D}-$plane, which is the solution to*

$$R_{cache}(d, 0) = \min_{D_1} R(D_1, d^{\frac{1}{1-p}} D_1^{-\frac{p}{1-p}}).$$

Proof. Neglect the constraint involving D_F. In (4.7) one does not evaluate $R(D_1, D_2)$ over *all* $(D_1, D_2) \in \mathcal{D}$, but only along a 'slice' defined by the constraint:

$$D_1^p D_2^{1-p} = d \quad \longrightarrow \quad D_2 = d^{\frac{1}{1-p}} D_1^{-\frac{p}{1-p}}. \tag{4.29}$$

This slice is strictly convex with respect to (D_1, D_2). The contour lines (or isolines) of $R(D_1, D_2)$ are also convex (and continuous!) on the $\mathcal{D}-$plane. More importantly, though, these contour lines end straight ($\frac{dD_2}{dD_1} = 0$ in $\mathcal{D}_{3,2}$ and $\frac{dD_1}{dD_2} = 0$ in $\mathcal{D}_{3,1}$). Consequently, the minimum of $R(D_1, D_2)$ evaluated on a *strictly* convex curve is where that curve is tangential with a contour line; it cannot be at a simple crossing.



(a) The optimal caching strategy is where a contour line of $R(D_1, D_2)$ is tangential to the (dashed) line $D_1^p D_2^{1-p} = d$.

(b) A slice of $R(D_1, D_2)$ that is strictly convex in (D_1, D_2) has one unique minimum.

Figure 4.10 – Example of Lemma 4.4 for $p = \frac{2}{5}$ and $d = 0.475$.

This tangential part is either a unique point or a (set of) closed interval(s), the latter if and only if ∃ a contour line that is described by the same curve as (4.29) for some interval(s). This happens in $\mathcal{D}_1$ when $p = \frac{1}{2}$, but not for other p. Hence, there can only be one minimum. □

An illustration of the 'slicing' of Lemma 4.4 is depicted in Figure 4.10. Using this as a building block, we end with the following theorem:

Theorem 4.4. *Without loss of generality, assume $p < \frac{1}{2}$. Then,*

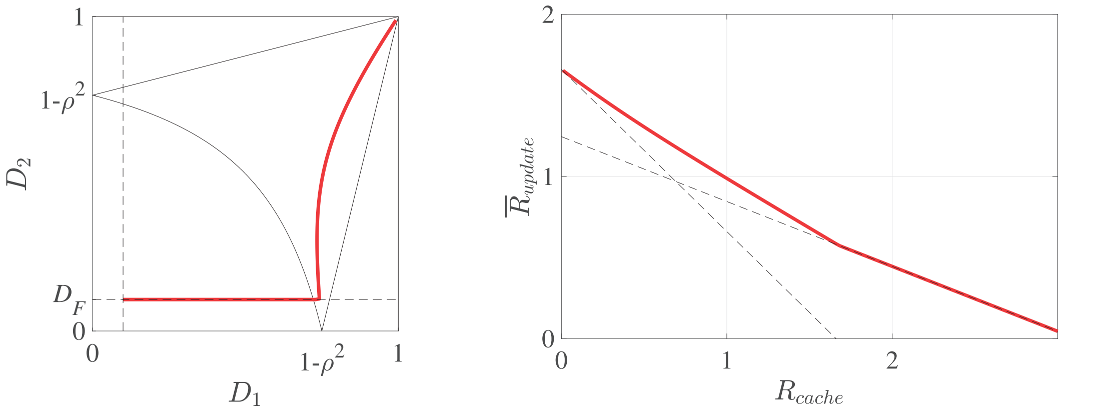
$$R_{cache}(d, D_F) = R(\bar{D}_1, d^{\frac{1}{1-p}} \bar{D}_1^{-\frac{p}{1-p}})$$

where $\bar{D}_1 = \max(D_F, D_1^*)$ and D_1^* is the solution to

$$\frac{d}{dD_1}\left(-1 + D_1 + d^{\frac{1}{1-p}} D_1^{\frac{-p}{1-p}} + 2\rho\sqrt{(1-D_1)(1-d^{\frac{1}{1-p}} D_1^{\frac{-p}{1-p}})}\right) = 0, \tag{4.30}$$

over $(D_1, d^{\frac{1}{1-p}} D_1^{-\frac{p}{1-p}}) \in \mathcal{D}_{2,2}$.

Proof. First, assume D_F plays no restricting role. Then, the minimum of Lemma 4.4 lies necessarily in $D_{2,2}$. Namely, it cannot lie in $\mathcal{D}_{3,2}$ since its boundary is strictly superior. Second, in $\mathcal{D}_{1,2}$ the equipotential lines of $R(D_1, D_2)$ behave as $D_1 D_2 =$ constant. Hence, they cannot be tangential to the curve $D_1^p D_2^{1-p} = d$, whose derivative is 'less steep' everywhere. That leaves $\mathcal{D}_{2,2}$, where $R(D_1, D_2)$ is minimized by maximizing $D_1 D_2 - \left(\rho - \sqrt{(1-D_1)(1-D_2)}\right)^2$ (3.19). By restricting $(D_1, D_2) = (D_1, d^{\frac{1}{1-p}} D_1^{-\frac{p}{1-p}}) \in \mathcal{D}_{2,2}$ first and only then setting this derivative w.r.t. D_1 to 0, one finds the optimum.



(a) As long as no file is cached completely, optimal strategies lie in $\mathcal{D}_2$.

(b) The resulting trade-off between R_{cache} and $\overline{R}_{\text{update}}$ with bounds.

Figure 4.11 – Example of optimal caching strategies and the resulting performance in terms of $(R_{\text{cache}}, \overline{R}_{\text{update}})$ for $p = 0.4$ and correlation $\rho = 0.5$.

Finally, should any D_i drop to D_F, then $R_{\text{cache}}(d, D_F)$ is minimized at the intersection of $D_i = D_F$ and $D_1^p D_2^{1-p} = d$, since $R(D_1, D_2)$ is monotonic along that curve on one side of the aforementioned unconstrained minimum. □

Figure 4.11 plots an example of Theorem 4.4. The qualitative interpretation of this result is the following: for $p = \frac{1}{2}$ the optimal strategy was to use all R_{cache} to make X_1 and X_2 conditionally independent as quickly as possible. The opposite is the case if $p \neq 1 - p$; after caching there should always remain *some* dependency between X_1 and X_2. The only exception is when one file is cached completely, otherwise the optimal strategy lies necessarily in $\mathcal{D}_2$.

Corollary 4.5. *Assume w.l.o.g. that $p < 1 - p$, then describes the boundary of $\mathcal{R}_{caching}$ for $R_{cache} \geq R(1 - \frac{\rho^2}{1-D_F}, D_F)$, provided $D_F \leq 1 - |\rho|$.*

Proof. For $R_{\text{cache}} \geq R(1 - \frac{\rho^2}{1-D_F}, D_F)$, the optimal caching strategies as described by Theorem 4.4 lie inside $\mathcal{D}_1$. In this region $R(D_1, D_2) = \frac{1}{2} \log \frac{1-\rho^2}{D_1 D_2}$, X_1 and X_2 become conditionally independent and the same argument as for Corollary 4.2 holds. In this high cache rate regime, the Gaussian $(R_{\text{cache}}, \overline{R}_{\text{update}})$ trade-off attains the outer bound described as number 3 in Subsection 4.1.3. □

4.5 Insights and Discussion

In Figure 4.12, we plot the evolution of optimal caching strategies and the $(R_{\text{cache}}, \overline{R}_{\text{update}})$ trade-off as we vary p. As expected, a stronger bias in preference results in more leverage on the encoder side and hence a more efficient rate trade-off. Furthermore, as $p \to 0$ or 1, the distortions that achieve the optimal trade-off converge to the border of $\mathcal{D}_2$ and $\mathcal{D}_3$, indicating the encoding of only the most popular X_i.

4.5.1 Size Matters

The presence or absence of preference determines whether or not the optimal caching strategy also depends on the size of the cache, in addition to it -obviously- depending on ρ and p.

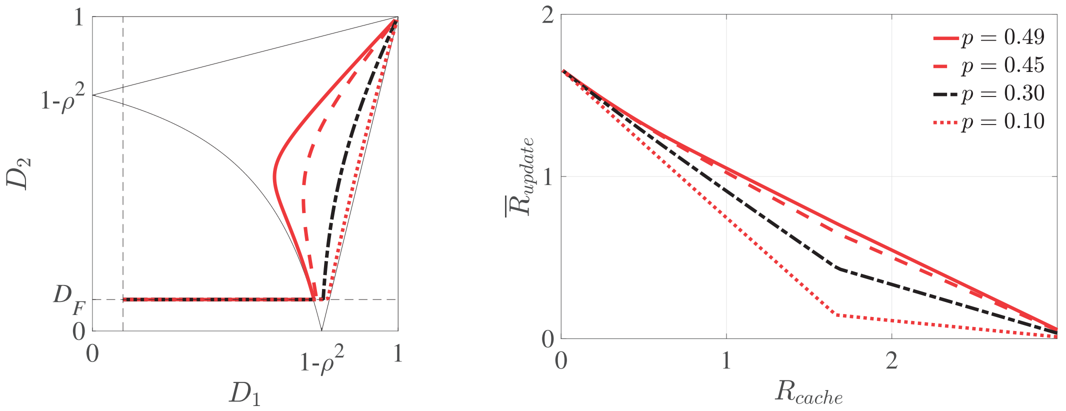
Corollary 4.6. *The optimal caching strategy depends on the size of the cache.*

Proof. Theorem 4.4 tells us that the optimal caching strategy necessarily lies inside $\mathcal{D}_2$ (except for extremely large R_{cache}). Every point in $\mathcal{D}_2$ is obtained by coding a *single* Gaussian random variable that is a mixture of the two sources (see again Chapter 3, specifically equation (3.27)). In others words, the encoder constructs $\alpha X_1 + \beta X_2 + W$, with α, β some constants and W independent Gaussian noise (not mentioned explicitly, but a direct consequence of [19]). These α and β are constant regardless of R_{cache} for two different pairs of distortions (D_1, D_2) if and only if those two pairs lie on a straight line originating from $(1,1)$. Due to the semi-triangular shape of $\mathcal{D}_2$, however, it is impossible for *all* optimal caching distortion pairs to lie on such a straight line, as can also be seen in Figure 4.12a. Therefore, the encoder must code different mixtures of X_1 and X_2 as R_{cache} increases. □

A direct follow-up is the observation that Gaussian bivariates are not successively refinable with respect to $R_{\text{cache}}(d, D_F)$ from some d_1 to a $d_2 < d_1$ according to Definition 3.1:

Corollary 4.7. *Caching cannot be split into two steps that are both optimal w.r.t. $R_{cache}(d, D_F)$. In other words, the caching rate-distortion function is not successively refinable from d_1 to $d_2 < d_1$. The only exceptions are when $p \in \{0, \frac{1}{2}, 1\}$ or $\rho = 0$.*

Proof. Gaussian sources subject to individual mean squared error constraints are successively refinable from (D_1, D_2) to (D'_1, D'_2) if the *matrices* associated to the rate-distortion optimal encoding satisfy $\mathbf{D} \succeq \mathbf{D}'$ (see Theorem 3.3). In $\mathcal{D}_2$, this again requires for (D_1, D_2) and (D'_1, D'_2) to lie on a straight line originating from $(1,1)$. Since optimal caching strategies generally lie on a curve instead of a line, it is impossible for the caching phase to be split into two of which both phases are optimal w.r.t. $R_{\text{cache}}(d, D_F)$. The only exceptions are when $p = \frac{1}{2}$ (when optimal strategies lie on the diagonal) or when either $\rho = 0$ or $p \in \{0, 1\}$ (when the encoder codes only X_1 or X_2). □



(a) Optimal strategies move away from the diagonal as p decreases.

(b) The resulting trade-off between R_{cache} and $\overline{R}_{\text{update}}$.

Figure 4.12 – Progression of strategies for $p = 0.49, 0.45, 0.3$ and 0.1 for correlation $\rho = 0.5$.

A last observation from Figure 4.12a is that encoder strategies diverge, but eventually converge:

Corollary 4.8. *For $D_F \to 0$ and $R_{cache} \to \infty$, the optimal strategy is to only cache the most popular X_i irrespective of the exact value p.*

Proof. $\mathcal{D}_2$ ends in two corners, i.e., $(0, 1-\rho^2)$ or $(1-\rho^2, 0)$. if R_{cache} grows very large and D_F plays no restricting role, the optimal caching strategy is necessarily squeezed into these corners. These points are associated to a perfect description of one X_i and the resulting MSE-estimator of the other. In other words: for very large R_{cache} the best caching strategy cares more about the most popular component and less about the correlation between the two, irrespective of the value of p. □

4.5.2 Sensitivity to Exact Knowledge of User Preference

The strong change in optimal strategy as a function of preference begs the question exactly how sensitive caching is to accurate knowledge of the user's habits. Considering the two parameters correlation and preference, one can argue that in practice the encoder most likely has perfect knowledge of the former and only an estimate of the latter.

If the encoder does not know the value of p, his best strategy is to assume $p = \frac{1}{2}$. The loss of performance due to this imperfect knowledge is bounded. This is already evident from the geometry of Figure 4.3: the actual best Gaussian $(R_{\text{cache}}, \overline{R}_{\text{update}})$ trade-off lies inside the bounded triangle, as well as an encoding strategy of assuming $p = \frac{1}{2}$ (even though that is not the real value of p). In the following corollary we characterize the maximum loss explicitly. A surprising result is that the maximum loss depends on p, but the point *where* this loss is maximized does not; that only depends on D_F.

Theorem 4.5. *Assume $D_F \le 1-|\rho|$ and w.l.o.g. that $p<\frac{1}{2}$. Loss due to lack of knowledge of p is no larger than*

$$\Delta_{R_u} = \frac{1}{2}\log\left(\frac{1-\frac{\rho^2}{1-D_F}}{D_F}\right)^{\frac{1}{2}-p}$$

and attains this maximum at

$$R_{cache} = R(1-\frac{\rho^2}{1-D_F}, D_F).$$

Proof. Let us use the superscript $(R_{\text{cache}}, \overline{R}_{\text{update}})^{\text{actual}}$ to refer to the Gaussian optimal caching trade-off via Theorem 4.4 based on some fixed $p<\frac{1}{2}$. Introduce as well $(R_{\text{cache}}, \overline{R}_{\text{update}})^{\text{unif}}$ to indicate the caching trade-off based on Theorem 4.3, i.e., when the encoder does not know p and assumes it equals $\frac{1}{2}$. They key is that $(R_{\text{cache}}, \overline{R}_{\text{update}})^{\text{actual}}$ and $(R_{\text{cache}}, \overline{R}_{\text{update}})^{\text{unif}}$ diverge until $R_{\text{cache}} = R(1-\frac{\rho^2}{1-D_F}, D_F)$ and then converge. Note that $(R_{\text{cache}}, \overline{R}_{\text{update}})^{\text{actual}}$ and $(R_{\text{cache}}, \overline{R}_{\text{update}})^{\text{unif}}$ start and end in the same points, $(0, \frac{1}{2}\log\frac{1}{D_F})$ and $(\frac{1}{2}\log\frac{1-\rho^2}{D_F^2}, 0)$ respectively.

$(R_{\text{cache}}, \overline{R}_{\text{update}})^{\text{unif}}$ is a strictly convex curve on $R_{\text{cache}} \in [0, C_W(X_1,X_2)]$, after which it is a straight line connecting $(\frac{1}{2}\log\frac{1-|\rho|}{1+|\rho|}, \frac{1}{2}\log\frac{1-|\rho|}{D_F})$ to the end $(\frac{1}{2}\log\frac{1-\rho^2}{D_F^2}, 0)$. The fact that $p \neq \frac{1}{2}$ does not affect performance; p has been made irrelevant because there is complete symmetry in the sources, the end distortion constraints and the caching strategy, i.e. $D_1 = D_2$.

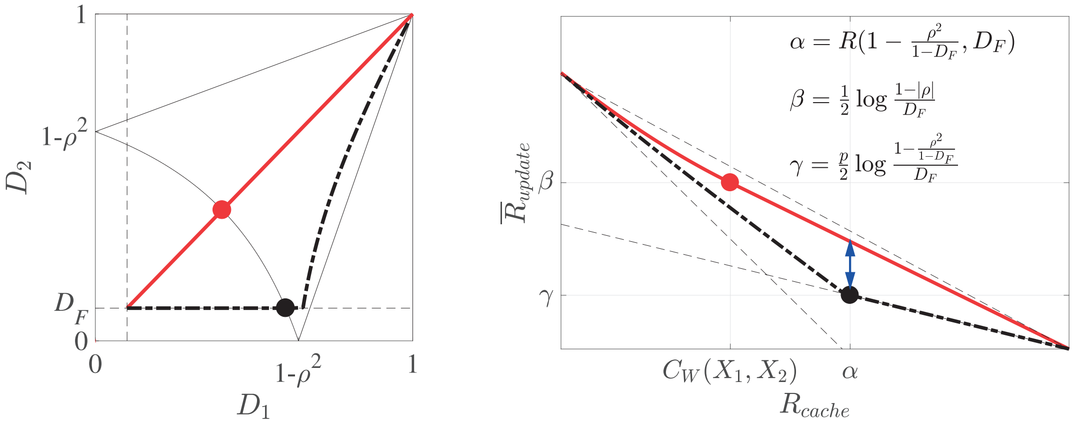
Then, $(R_{\text{cache}}, \overline{R}_{\text{update}})^{\text{actual}}$ is a strictly convex curve until Theorem 4.4 dictates to cache up to distortions lying inside $\mathcal{D}_1$. This happens at $R_{\text{cache}} = R(1-\frac{\rho^2}{1-D_F}, D_F) \ge C_W(X_1,X_2)$. From that point on, the trade-off is a straight line connecting $(\frac{1}{2}\log\frac{1-\rho^2}{(1-\frac{\rho^2}{1-D_F})D_F}, \frac{p}{2}\log\frac{1-\frac{\rho^2}{1-D_F}}{D_F})$ to the same end point as the uniform strategy.

Because $(R_{\text{cache}}, \overline{R}_{\text{update}})^{\text{actual}}$ moves into a straight line at an R_{cache} strictly larger and an $\overline{R}_{\text{update}}$ strictly smaller than where $(R_{\text{cache}}, \overline{R}_{\text{update}})^{\text{unif}}$ makes this change, combined with both curves being strictly convex before reaching that point, must mean that the curves diverge until $R_{\text{cache}} = R(1-\frac{\rho^2}{1-D_F}, D_F)$. The loss at this point equals

$$\begin{aligned}\Delta_{R_u} &= \frac{1}{2}\log\frac{\sqrt{(1-\frac{\rho^2}{1-D_F})D_F}}{D_F} - \frac{p}{2}\log\frac{1-\frac{\rho^2}{1-D_F}}{D_F} \\ &= \frac{1}{2}\log\left(\frac{1-\frac{\rho^2}{1-D_F}}{D_F}\right)^{\frac{1}{2}-p}.\end{aligned}$$

□

The geometry of this loss is depicted in Figure 4.13.



(a) Caching strategies enter $\mathcal{D}_1$ at these dots.

(b) The $(R_{\text{cache}}, \overline{R}_{\text{update}})$ trade-offs diverge until *both* strategies cache up to distortions inside $\mathcal{D}_1$; at that point loss is maximal.

Figure 4.13 – The geometry of how loss due to lack of knowledge of user preference is bounded. Red solid line stands for a uniform caching strategy, whereas the black dotted line stands for what the encouder should do. $p = \frac{1}{4}$ and $\rho = \frac{3}{5}$.

5 Caching of Multiple Gaussians under Uniform Preference Probabilities

This chapter[1] extends the caching of Gaussians under *uniform* preference probabilities to databases of more than two sources. In terms of the problem there is little difference with Chapter 4: If a user shows no preference for any of the files in the database, the encoder should cache the information that is shared between them. The information that is unique to a particular file is best left for the update phase. Like before, the concepts of Wyner's common information, Watanabe's total correlation and for Gaussians the Hadamard inequality are the measures that define what exactly this common core needs to be.

We saw in Section 4.3.4 that the encoder can successfully cache the total correlation of any two Gaussians by applying reverse water-filling on the eigenvalues of their correlation matrix. This *solution* does not generalize to larger databases; it is specific to having only two sources. We argue in this chapter that the common information of an arbitrary number of Gaussians is not (per sé) related to the eigendecomposition of the correlation matrix or a nice algebraic operation on it.

Also now we split the caching of Gaussians without user preference in a high and low cache rate regime, where again the turning point is when the cache rate exceeds the common information. This time, however, the key differentiator is complexity. In the high cache rate regime, figuring out what to cache is a convex problem, whereas for small R_{cache} it is non-convex. We show that a water filling procedure on the eigendecomposition of the correlation matrix is only an inner bound to achievable $(R_{\text{cache}}, \overline{R}_{\text{update}})-$pairs.

The chapter ends with the conjecture that this bound is tight if and only if the correlation matrix is circulant. Such a conjecture would include the results of Section 4.3, since any 2 × 2 correlation matrix is necessarily circulant. In that regard, the main argument of this chapter is thus that the clean caching procedure for a database of two sources is not the rule, but rather the exception.

[1] The material of this chapter appeared in [26].

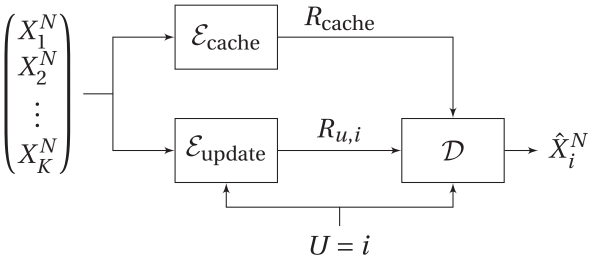

$\begin{pmatrix} X_1^N \\ X_2^N \\ \vdots \\ X_K^N \end{pmatrix}$
$\mathcal{E}_{\text{cache}}$
R_{cache}
$\mathcal{E}_{\text{update}}$
$R_{u,i}$
$\mathcal{D}$
$\hat{X}_i^N$
$U = i$

Figure 5.1 – The Caching Network for a database of K sources is the same as in Figure 4.1.

5.1 Problem Statement

The setup is equivalent to Chapter 4 (to be exact, Section 4.3), but is extended to a larger database. This time, the source is consists of K files like $\mathbf{X}^N \in \mathbb{R}^{K\times N}$. The user is only interested in one file X_i^N, but the encoder will send a cache message before the user gets the chance to announce this request.

A message m is an integer in the set $I_M = \{1,2,\cdots,M\}$. We have m_c for the 'cache' and $m_{u,1}, m_{u,2}, \cdots, m_{u,K}$ for the 'update'. The size of each of their message sets is denoted by the capital letter equivalent, i.e., M_c, $M_{u,i}$, etc. In any case m_c is transmitted, while for the update the decoder will only receive $m_{u,1}$ if it requests X_1^N (and similar for all other X_i^N). A code consists of encoder mappings

$$f: \quad \mathbb{R}^{K\times N} \to I_{M_c} \times I_{M_{u,1}} \times I_{M_{u,2}} \times \cdots \times I_{M_{u,K}}$$

and decoders

$$g_i: \quad I_{M_c} \times I_{M_{u,i}} \to \mathbb{R}^{1\times N} \qquad \text{for } i = 1,2,\cdots,K$$

of which the latter is meant to reconstruct $\hat{X}_i^N = g_i(m_c, m_{u,i})$ where i stands for the specific file requested.

A caching rate-distortion tuple $(R_{\text{cache}}, R_{u,1}, R_{u,2}, \cdots, R_{u,K}, D_F)$ is said to be achievable if for arbitrary $\epsilon > 0$ there exist such encoders and decoders that for both $i = 1,2$ we have

$$\begin{aligned} M_c &\leq 2^{N(R_{\text{cache}}+\epsilon)} \\ M_{u,i} &\leq 2^{N(R_{u,i}+\epsilon)} && \text{for } i = 1,2,\cdots,K \\ \frac{1}{N}\sum_{n=1}^{N} d(X_i(n), \hat{X}_i(n)) &\leq D_F + \epsilon && \text{for } i = 1,2,\cdots,K \end{aligned}$$

where $d_X(\cdot,\cdot)$ is some single letter distortion measure. For notational simplicity, each file is subjected to the same final distortion constraint.

Furthermore, we focus on the scenario in which the user chooses each file equally likely. We model this choice by a random variable $U \in \{1,\cdots,K\}$ which has the uniform distribution.

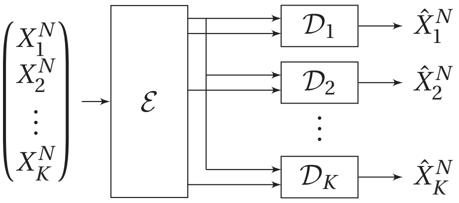


Figure 5.2 – The extended Gray–Wyner Network for K sources, equally many decoders and individual links, and one common link.

Therefore the *average* update rate still required after caching equals

$$\overline{R}_{\text{update}} \triangleq \frac{1}{K}\sum_{i=1}^{K} R_{u,i}. \tag{5.1}$$

A shorthand notation will be to say that $(R_{\text{cache}}, \overline{R}_{\text{update}}, D_F)$ is achievable to indicate that (at least one tuple) $(R_{\text{cache}}, R_{u,1}, R_{u,2}, \cdots, R_{u,K}, D_F)$ is, of which the average update rate equals $\overline{R}_{\text{update}}$.

5.1.1 The Extended Gray–Wyner Network

The analogy with the Gray–Wyner network also extends to higher dimensions, even though originally it was only defined for two sources [7]. The network was extended to K sources by Xu, Liu and Chen, first for discrete sources [29] and then for lossy source coding as well [30]. This extension consists of K decoders who are connected to the encoder by one common link and K individual links, like shown in Figure 5.2. In Xu *et al.*'s notation, R_0 is the rate spent on the common link and $R_1, R_2, \cdots, R_K$ relate to the individual links from the one encoder to each decoder.

Using the notation of the conditional rate-distortion function as defined in (3.4), the region of achievable communication was described as follows:

Lemma 5.1 (Xu, Liu and Chen [30])**.** *Communication on the extended Gray–Wyner network is achievable for rate-distortion tuples* $(R_0, R_1, \cdots, R_K, D_1, D_2, \cdots, D_K)$ *that satisfy*

$$\begin{aligned} &R_0 \geq I(\mathbf{X};\mathbf{V}) \\ &R_i \geq R_{X_i|\mathbf{V}}(D_i) \quad \textit{for } i = 1, 2, \cdots, K, \end{aligned}$$

for some conditional distribution $p(\mathbf{v}|\mathbf{x})$.

The extended Gray–Wyner network dictates the achievable rates for the caching problem. Namely, despite the user decoding only one X_i^N, any caching code must have update messages that work for all elements of $\mathbf{X}$ since any one could be requested. The equivalence is that the common link equals the cache and the individual links the update messages for each possible request:

$$\begin{aligned} R_0 &\leftrightarrow R_{\text{cache}} \\ R_i &\leftrightarrow R_{u,i} \qquad i = 1, 2, \cdots, K \end{aligned}$$

This gives us the following:

Theorem 5.1. *Given a joint density $p(\mathbf{x}, \mathbf{v})$, all caching rate-distortion tuples $(R_{cache}, \overline{R}_{update}, D_F)$ satisfying the following inequalities*

$$R_{cache} \geq I(\mathbf{X}; \mathbf{V}) \tag{5.2}$$

$$\overline{R}_{update} \geq \frac{1}{K} \sum_{i=1}^{K} R_{X_i|\mathbf{V}}(D_F) \tag{5.3}$$

are achievable. The closure of such achievable tuples over joint densities $p(\mathbf{x}, \mathbf{v})$ is denoted $\mathcal{R}_{caching}$.

The boundary of this achievable region lies inside the gray triangle of Figure 5.3. This shape is obtained by first and foremost the crossing of two outer bounds:

1. The condition $R_{\text{cache}} + \sum_{i=1}^{K} R_{u,i} \geq R(D_F, \cdots, D_F)$ and the false assumption this can be met with equality everywhere.
2. The straight line $(R_{\text{cache}}, \overline{R}_{\text{update}}) : (0, \frac{1}{K} \sum_{i=1}^{K} R_{X_i}(D_F)) \rightarrow (\frac{1}{K} \sum_{i=1}^{K} R_{X_i}(D_F), 0)$ that stems from a genie telling the cache encoder in advance which file will be requested.

These outer bounds are then paired with a time-sharing inner bound connecting the extremal ends of the trade-off to ultimately obtain the triangular shape.

5.1.2 The Gaussian Case

Now consider specifically a database $\mathbf{X}^N$ whose files are sequences of IID samples drawn from a Gaussian distribution with mean $\mathbf{0}$ and covariance

$$\Sigma_{\mathbf{X}} = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \cdots \\ \rho_{12} & 1 & \rho_{23} & \\ \rho_{13} & \rho_{23} & 1 & \\ \vdots & & & \ddots \end{bmatrix}. \tag{5.4}$$

In trying to understand the boundary of $\mathcal{R}_{\text{caching}}$ we restrict the discussion to Gaussian codebooks, which brings us to the following function:

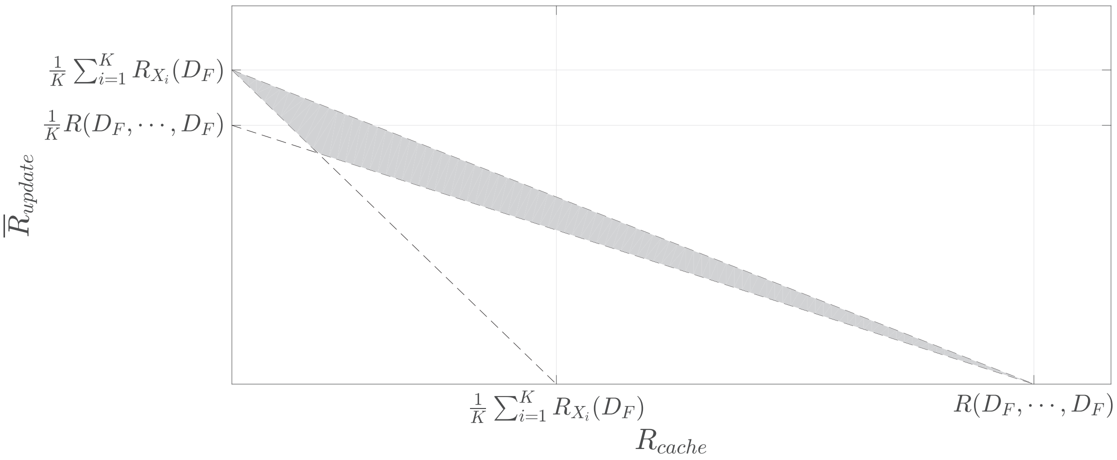


Figure 5.3 – The boundary of achievable $(R_{\text{cache}}, \overline{R}_{\text{update}})$−pairs lies inside this triangle. Example generated from a Gaussian $\mathbf{X}$ of $K = 3$ and $\Sigma_{\mathbf{X}}$ as defined in (5.4) with $\rho_{12} = 1/3$ and $\rho_{23} = \rho_{13} = 2/3$.

Corollary 5.1. *The Gaussian boundary of $\mathcal{R}_{caching}$ can be described by what we will call the Gaussian caching rate-distortion function:*

$$R_{cache}(d, D_F) = \min_{\mathbf{D}} \frac{1}{2}\log\frac{|\Sigma_{\mathbf{X}}|}{|\mathbf{D}|} \quad s.t. \quad \begin{cases} \mathbf{0} \preceq \mathbf{D} \preceq \Sigma_{\mathbf{X}} \\ \prod_{i=1}^{K} D_{i,i} \leq d \\ D_{i,i} \geq D_F, \quad \forall i = 1,2,\cdots,K \end{cases} \tag{5.5}$$

for a normalized parameter $d \in [D_F^K, 1]$ that relates back to $\overline{R}_{update}$ by picking $d = D_F^K 2^{2K\overline{R}_{update}}$.

Proof. The derivation is analogous to Corollary 4.1. The goal is again to express the (Gaussian) boundary of $\mathcal{R}_{\text{caching}}$ by minimizing R_{cache} while constraining $\overline{R}_{\text{update}}$.

Choosing a $\mathbf{V}$ that is jointly Gaussian with $\mathbf{X}$ in (5.2) yields:

$$I(\mathbf{X};\mathbf{V}) = \frac{1}{2}\log\frac{|\Sigma_{\mathbf{X}}|}{|\mathbf{D}|}, \tag{5.6}$$

where as always

$$\mathbf{D} = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}|\mathbf{V}])(\mathbf{X} - \mathbb{E}[\mathbf{X}|\mathbf{V}])^T] = \Sigma_{\mathbf{X}|\mathbf{V}} \preceq \Sigma_{\mathbf{X}}. \tag{5.7}$$

Furthermore, $p(\mathbf{x}|\mathbf{v})$ being Gaussian results in $R_{u,i}$ to be

$$R_{X_i|\mathbf{V}}(D_F) = \frac{1}{2}\log^{+}\frac{D_{i,i}}{D_F}. \tag{5.8}$$

Caching any X_i beyond D_F serves no purpose: it increases R_{cache} while not improving $R_{u,i}$, since it cannot be negative. Therefore, demanding $D_{i,i} \geq D_F$ ensures no cache rate is wasted on going beyond what is requested. This will thus ensure $R_{u,i} \geq 0$ and hence we arrive at

$$\overline{R}_{\text{update}} = \frac{1}{K}\sum_{i=1}^{K}\frac{1}{2}\log\frac{D_{i,i}}{D_F} \tag{5.9}$$

$$= \frac{1}{2K}\log\frac{\prod_{i=1}^{K} D_{i,i}}{D_F^K}. \tag{5.10}$$

Combining everything together: Minimizing R_{cache} over all jointly Gaussian distributions $p(\mathbf{v},\mathbf{x})$ equals minimizing (5.6) over distortion matrices $\mathbf{D} \preceq \Sigma_{\mathbf{X}}$. Constraining $\overline{R}_{\text{update}}$ means constraining $\prod_{i=1}^{K} D_{i,i}$ for which we introduce a parameter called d to constrain

$$\prod_{i=1}^{K} D_{i,i} = D_F^K 2^{2K\overline{R}_{\text{update}}} \leq d. \tag{5.11}$$

By picking $d \in [D_F^K, 1]$, one spans all possible values for $\overline{R}_{\text{update}}$, i.e. $\overline{R}_{\text{update}} \in [0, \frac{1}{2}\log\frac{1}{D_F}]$. This gives us the function as defined in the Corollary. We note that even though we constrain $\prod_{i=1}^{K} D_{i,i} \leq d$, the minimizer to (5.5) will result in equality by construction. □

5.1.3 Gaussian Caching is the Minimization of Total Conditional Correlation

The Hadamard inequality,

$$\prod_{i=1}^{K} D_{i,i} \geq |\mathbf{D}|, \tag{5.12}$$

still plays an important role in solving $R_{\text{cache}}(d, D_F)$ through its relationship with Watanabe's notion of total (conditional) correlation [9]. We introduced this notion before in Definition 4.3, where for two sources it reduced to mutual information. In higher dimensions, let us repeat that the definition is as follows (here denoted for a *conditional* distribution):

$$TC(\mathbf{X}|\mathbf{V}) \triangleq \sum_{i=1}^{K} h(X_i|\mathbf{V}) - h(\mathbf{X}|\mathbf{V}), \tag{5.13}$$

which for Gaussian distributions $p(\mathbf{x}|\mathbf{v})$ equals (using (5.7)):

$$TC(\mathbf{X}|\mathbf{V}) = \frac{1}{2}\log\frac{\prod_{i=1}^{K} D_{i,i}}{|\mathbf{D}|}. \tag{5.14}$$

Because $R_{\text{cache}}(d, D_F)$ maximizes the denominator in (5.14) while upper bounding the numerator by d, we have that an optimal caching strategy effectively minimizes the total conditional correlation. A distortion matrix $\mathbf{D}$ that closes the Hadamard inequality is associated to a Gaussian conditional distribution of which $TC(\mathbf{X}|\mathbf{V}) = 0$. This in turn implies that $p(\mathbf{x}|\mathbf{v}) = \prod_{i=1}^{K} p(x_i|\mathbf{v})$ or, in other words, that $\mathbf{V}$ makes the elements of $\mathbf{X}$ conditionally independent. Whether or not $TC(\mathbf{X}|\mathbf{V}) = 0$ is achievable will structure our discussion on $R_{\text{cache}}(d, D_F)$.

5.2 High Cache Rate Region & Common Information

The caching rate-distortion function is lower bounded as

$$R_{\text{cache}}(d, D_F) \geq \frac{1}{2}\log\frac{|\Sigma_{\mathbf{X}}|}{d}, \tag{5.15}$$

which follows from the higher-dimensional Hadamard inequality $d \geq \prod_{i=1}^{K} D_{i,i} \geq |\mathbf{D}|$. As stated in the previous subsection, this lower bound is associated to conditional independence. In this section we characterize in which regime of $(R_{\text{cache}}, \overline{R}_{\text{update}})-$pairs this bound is achievable.

To answer that question, we must better understand what is required to attain conditional independence. Wyner's common information was originally defined for two sources (4.21) and helped us characterize a minimum on R_{cache}. For higher dimensions, we scale this notion to multiple variables using the definition of total conditional correlation (5.13):

Definition 5.1. *We define the multivariate extension of Wyner's common information as*

$$C_W(\mathbf{X}) \triangleq \min_{\substack{\mathbf{V} \\ :TC(\mathbf{X}|\mathbf{V})=0}} I(\mathbf{X};\mathbf{V}). \tag{5.16}$$

For discrete random variables a similar extension appeared in [29].

Now, if $\mathbf{X}$ is Gaussian then the $\mathbf{V}$ that achieves $C_W(\mathbf{X})$ is necessarily Gaussian as well:

Theorem 5.2. *If $\mathbf{X}$ is Gaussian distributed then the common information equals:*

$$C_W(\mathbf{X}) = \min_{\mathbf{D}} \frac{1}{2}\log\frac{|\Sigma_{\mathbf{X}}|}{|\mathbf{D}|} \quad s.t. \quad \begin{cases} \mathbf{0} \preceq \mathbf{D} \preceq \Sigma_{\mathbf{X}}, \\ \mathbf{D} \textit{ is diagonal.} \end{cases} \tag{5.17}$$

Proof. The proof uses standard arguments. Consider any $\mathbf{V}$ that is jointly distributed with $\mathbf{X}$:

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

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 $\mathbb{E}[\mathbf{X}|\mathbf{V}]$, see [19, 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 does not guarantee (conditional) independence in general. For Gaussians, however, zero correlation and independence do have an if and only if relationship. Since any distortion matrix $\mathbf{D} \preceq \Sigma_{\mathbf{X}}$ is attainable by Gaussian distributions we can achieve equality throughout all steps. □

It is important to note that the constraint of $\mathbf{D}$ having to be diagonal is linear. Hence, the Gaussian common information can always be found efficiently via numerical methods, like the ones described by Boyd and Vandenberghe [22, 31]. Namely, the objective function is strictly convex and the search domain is convex as well. This stands in sharp contrast to the definition of $R_{\text{cache}}(d, D_F)$ in which the constraint $\prod_{i=1}^{K} D_{i,i} \leq d$ is non-linear, which renders that optimization non-convex.
We encourage the reader to solve $C_W(\mathbf{X})$ for oneself using the CVX package in MATLAB [32].

To see how the common information helps one to achieve equality in (5.15), define again the following:

Definition 5.2. *Let $\mathbf{D}_{C_W}$ be the matrix that optimizes* (5.17).

There exist correlation matrices $\Sigma_{\mathbf{X}}$ for which $\mathbf{D}_{C_W}$ is very asymmetric, i.e., some diagonal entries are very small in comparison to others. In this regard, one must not forget the other constraint inside $R_{\text{cache}}(d, D_F)$: the encoder should not cache any X_i up to a distortion $D_{i,i} < D_F$. A $\mathbf{D}_{C_W}$ of which one or some diagonal entries are below D_F can therefore never be an optimal caching strategy w.r.t. $R_{\text{cache}}(d, D_F)$. To overcome this, we include this practical coding constraint inside the Gaussian common information:

Definition 5.3. *We define the Gaussian* constrained *common information to be the following:*

$$C_W(\mathbf{X}, D_F) = \inf_{\mathbf{D}} \frac{1}{2} \log \frac{|\Sigma_{\mathbf{X}}|}{|\mathbf{D}|} \quad s.t. \quad \begin{cases} \mathbf{0} \preceq \mathbf{D} \preceq \Sigma_{\mathbf{X}}, \\ \mathbf{D} \text{ is diagonal}, \\ D_{i,i} \geq D_F, \quad \forall i = 1, 2, \cdots, K. \end{cases} \tag{5.18}$$

In contrast to $C_W(\mathbf{X})$, this minimum may not exist. It exists if and only if $D_F \cdot \mathbf{I} \preceq \Sigma_{\mathbf{X}}$, which is a consequence of combining all three constraints.

If this constrained common information exists, then we have

$$C_W(\mathbf{X}, D_F) \geq C_W(\mathbf{X}), \tag{5.19}$$

because the search domain of $C_W(\mathbf{X}, D_F)$ is a subset of that of $C_W(\mathbf{X})$. We note that the constrained common information is also a convex problem, which can be solved by the same numerical methods one would use for (5.17).

This constrained common information is closely related to the discussion on *lossy* common information by Viswanatha, Akyol and Rose [11]. They define for *two* random variables the notion $C_W(X_1, X_2, D_1, D_2)$ as the minimum common rate on the Gray–Wyner network needed such that communication of X_1, X_2 is achievable at a sum-rate that does not exceed the joint rate-distortion function $R(D_1, D_2)$. This notion was also introduced precisely to cover scenarios where the ‘lossless’ common information would result in distortions that are more strict than desired by the lossy coding application. Especially insightful is that Viswanatha *et al.* explicitly characterize the bivariate Gaussian case of their measure.

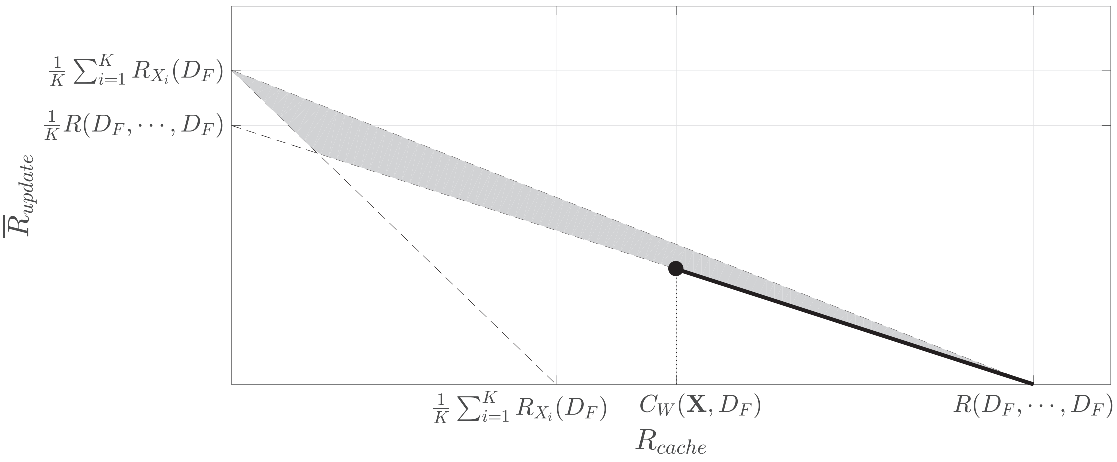


Figure 5.4 – If $R_{\text{cache}} \geq C_W(\mathbf{X}, D_F)$ and D_F is not too large, then Gaussian caching codes can achieve the outer bound. These $(R_{\text{cache}}, \overline{R}_{\text{update}})$-pairs are indicated by the thick black line.

Using this new constrained $C_W(\mathbf{X}, D_F)$, the high cache rate regime as referred to by the title of this subsection relates to the following theorem:

Theorem 5.3. *If $C_W(\mathbf{X}, D_F)$ exists, then for all $d \leq |\Sigma_\mathbf{X}|2^{-2C_W(\mathbf{X},D_F)}$ we have*

$$R_{cache}(d, D_F) = \frac{1}{2}\log\frac{|\Sigma_\mathbf{X}|}{d}. \tag{5.20}$$

If furthermore $D_F \leq \min(\text{diag}(\mathbf{D}_{C_W}))$, then the above holds for all $d \leq |\Sigma_\mathbf{X}|2^{-2C_W(\mathbf{X})}$.

Proof. As stated in its definition, $C_W(\mathbf{X}, D_F)$ exists if and only if $D_F \cdot \mathbf{I} \preceq \Sigma_\mathbf{X}$. This also implies that $R(D_F, \cdots, D_F) = R_{\text{cache}}(D_F^K, D_F) = \frac{1}{2}\log\frac{|\Sigma_\mathbf{X}|}{D_F^K}$.

Then, since $C_W(\mathbf{X}, D_F)$ is a minimum, a diagonal $\mathbf{D}$ and hence equality on (5.15) is not achievable for $R_{\text{cache}} < C_W(\mathbf{X}, D_F)$. On the contrary, for all $R_{\text{cache}} \geq C_W(\mathbf{X}, D_F)$ equality is achievable and this requires d to be smaller than the condition mentioned in the theorem. Denote by $\mathbf{D}'$ the diagonal matrix that solves $C_W(\mathbf{X}, D_F)$, then we have:

$$C_W(\mathbf{X}, D_F) = \frac{1}{2}\log\frac{|\Sigma_\mathbf{X}|}{|\mathbf{D}'|} \xrightarrow[\text{(in)equality}]{\text{Hadamard}} \prod_{i=1}^{K} D'_{i,i} = |\Sigma_\mathbf{X}|2^{-2C_W(\mathbf{X},D_F)}.$$

Hence, for $d' = |\Sigma_\mathbf{X}|2^{-2C_W(\mathbf{X},D_F)}$ we have that $R_{\text{cache}}(d', D_F) = \frac{1}{2}\log\frac{|\Sigma_\mathbf{X}|}{d'}$, because it is the lower bound (5.15) and it is achieved by diagonal distortion matrices. For all d smaller, the bound is also met with equality. Namely, there are infinitely many diagonal matrices $\mathbf{D}$ in the range of $D_F \cdot \mathbf{I} \preceq \mathbf{D} \preceq \mathbf{D}'$ that are

1. achievable, because $\mathbf{D} \preceq \mathbf{D}'$ and $\mathbf{D}' \preceq \Sigma_\mathbf{X}$ implies that $\mathbf{D} \preceq \Sigma_\mathbf{X}$,
2. and can span all $\prod_{i=1}^{K} D'_{i,i} \in [D_F^K, |\Sigma_\mathbf{X}|2^{-2C_W(\mathbf{X},D_F)}]$ and so all $R_{\text{cache}} \in [C_W(\mathbf{X}, D_F), \frac{1}{2}\log\frac{|\Sigma_\mathbf{X}|}{D_F^K}]$.

If $D_F \leq \min(\text{diag}(\mathbf{D}_{C_W}))$, then $C_W(\mathbf{X}) = C_W(\mathbf{X}, D_F)$ and $R_{\text{cache}} \geq C_W(\mathbf{X})$ is sufficient to achieve the lower bound (5.15). □

Corollary 5.2. *If $d \leq |\Sigma_{\mathbf{X}}| 2^{-2C_W(\mathbf{X}, D_F)}$, then the Gaussian caching rate-distortion function $R_{cache}(d, D_F)$ also characterizes the boundary of achievable $(R_{cache}, \overline{R}_{update})-$pairs in general.*

The use of Gaussian codebooks was an assumption in the desire to characterize the boundary of $\mathcal{R}_{\text{caching}}$. Whenever $R_{\text{cache}} \geq C_W(\mathbf{X}, D_F)$ their use is, however, sufficient for optimality. The proof of the corollary is a direct extension of the two-dimensional case of Corollary 4.2 and is related to the separability of the Gaussian rate-distortion function over all links of the Gray–Wyner network. Figure 5.4 shows an example of Theorem 5.3 and Corollary 5.2 for a case where $D_F \leq \min(\text{diag}(\mathbf{D}_{C_W}))$.

Theorems in Pictures

In higher dimensions, Wyner's common information has the same geometrical meaning as the two-dimensional case of Chapter 4 on Page 47: it corresponds to the *ellipsoid* of maximal *volume* that fits inside $\mathcal{E}_{\Sigma_\mathbf{X}}$ *and* that is straight, i.e., whose semi-principal axes align with the bases of the system.
One intriguing property of Gaussian common information is that of dimensionality. The condition that $\mathbf{0} \preceq \mathbf{D}_{C_W} \preceq \Sigma_\mathbf{X}$, implies that we can construct $\mathbf{D}_{C_W}$ as (a Schur-complement):

$$\mathbf{D}_{C_W} = \Sigma_\mathbf{X} - \bar{\Sigma},$$

where $\bar{\Sigma}$ must be positive semidefinite. Also noticing that $C_W(\mathbf{X})$ is associated to a minimum, we know that $\mathbf{D}_{C_W} \preceq \Sigma_\mathbf{X}$ cannot be strict (otherwise one can always find a better $\mathbf{D}$). The non-strictness eliminates the possibility of $\bar{\Sigma}$ being *full rank*. This in turn indicates that the $\mathbf{V}$ that optimizes (5.16) has a dimensionality between 1 and $K-1$, or that it can be compressed to a form of such limited dimensionality. This all relates again to Section 2.4.

Below are two example covariance matrices $\Sigma_\mathbf{1}$ and $\Sigma_\mathbf{2}$ of which the common information is associated to a matrix $\bar{\Sigma}$ of rank 1 and 2, respectively. The plots depict $\mathcal{E}_{\Sigma_\mathbf{X}}$ as a transparant ellipsoid circumfering $\mathcal{E}_{\mathbf{D}_{\mathbf{C_W}}}$ in color. The three smaller plots show the two-dimensional views from all sides of the bigger figure.
One intriguing open question is how the structure of a covariance matrix drives the dimensionality of the variable that attains the common information.

$$\Sigma_\mathbf{1} = \begin{bmatrix} 1 & {}^2/_3 & {}^1/_3 \\ {}^2/_3 & 1 & {}^1/_3 \\ {}^1/_3 & {}^1/_3 & 1 \end{bmatrix} \qquad \Sigma_\mathbf{2} = \begin{bmatrix} 1 & {}^2/_3 & {}^2/_3 \\ {}^2/_3 & 1 & {}^1/_3 \\ {}^2/_3 & {}^1/_3 & 1 \end{bmatrix}$$

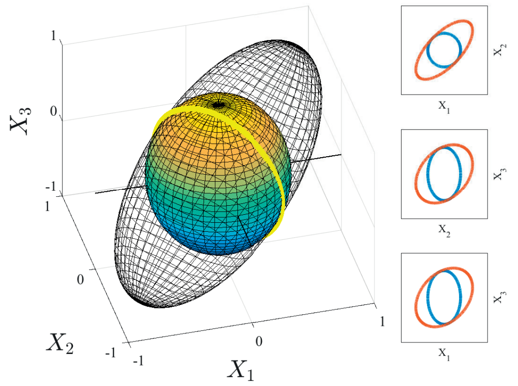


Figure 5.5 – Since rank($\bar{\Sigma}_1$) = 1, $\mathcal{E}_{\Sigma_1}$ and $\mathcal{E}_{\mathbf{D}_{\mathbf{C_W1}}}$ touch along two dimensions, i.e., a ring/ellipse along the surface of $\mathcal{E}_{\Sigma_1}$.

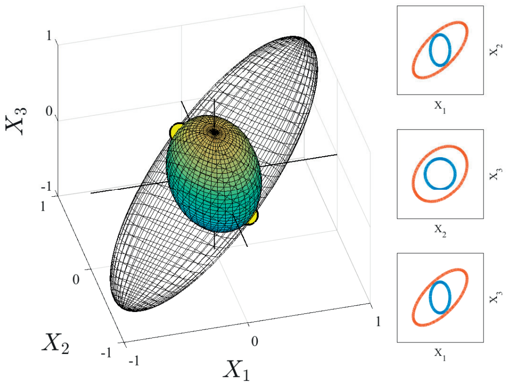


Figure 5.6 – Since rank($\bar{\Sigma}_2$) = 2, $\mathcal{E}_{\Sigma_1}$ and $\mathcal{E}_{\mathbf{D}_{\mathbf{C_W2}}}$ touch at two symmetric points only (a one-dimensional 'ellipse').

5.3 Low Cache Rate Region & Total Conditional Correlation

The difficulty of the Gaussian caching problem (5.5) is that even though the minimization of $-\log|\mathbf{D}|$ is strictly convex, the product-constraint on the diagonal entries of $\mathbf{D}$ renders the search domain non-convex. The peculiar consequence of the previous subsection is that when R_{cache} exceeds the common information, there is a turning point: There exists a $\mathbf{D} \preceq \Sigma_{\mathbf{X}}$ that closes the Hadamard inequality and the search for that particular matrix is a *convex* problem.

We call all rates $R_{\text{cache}} < C_W(\mathbf{X}, D_F)$ the 'low cache rate regime'. First of all, let us quickly summarize the defining features of this region: Whatever $\mathbf{V}$ that is jointly Gaussian with $\mathbf{X}$ we might consider, if $I(\mathbf{X};\mathbf{V}) < C_W(\mathbf{X}, D_F)$ then conditional independence between the elements of $\mathbf{X}$ is not attainable, which means that

$$TC(\mathbf{X}|\mathbf{V}) = \frac{1}{2}\log\frac{\prod_{i=1}^{K} D_{i,i}}{|\mathbf{D}|} > 0; \tag{5.21}$$

after caching there will always be some dependency left that the encoder no longer benefits from in the update phase. Consequently, the lower bound on $R_{\text{cache}}(d, D_F)$ of (5.15) *cannot* be met with equality.

For $R_{\text{cache}} < C_W(\mathbf{X})$, the non-convexity and hardness of the problem appears to persist. How to solve $R_{\text{cache}}(d, D_F)$ in this regime remains an open problem. In this section we argue that the solution we derived for $K = 2$ does not extend to higher dimensions. Instead we show it is merely a bound.

5.3.1 Counterexample to Eigenvalue Operations being Generally Optimal

When we say that the bivariate tactic of Chapter 4 does not extend to $K > 2$ we mean the following: operations on the eigenvalues of the correlation matrix is *not* the way to minimize total conditional correlation $TC(\mathbf{X}|\mathbf{V})$. Specifically, in Section 4.3 we derived that the encoder can do caching of two Gaussians by first transforming their covariance to a correlation matrix, followed by applying reverse water-filling on its eigenvalues. This no longer works.

Consider this counterexample: Assume D_F is small enough such that $\mathbf{D}_{C_W}$ is the optimal caching distortion matrix for some d for which $R_{\text{cache}}(d, D_F) = C_W(\mathbf{X})$. Then for $K = 3$ consider the following correlation matrix $\Sigma_{\mathbf{X}}$ and the $\mathbf{D}_{C_W}$ corresponding to its common information:

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

If the bivariates result did extend to higher dimensions, the implication would have been that

$\mathbf{D}_{C_W}$ is the result of an operation on the dominant eigenvalues of $\Sigma_{\mathbf{X}}$ only. Contrarily,

$$\mathbf{D}_{C_W} = \Sigma_{\mathbf{X}} - \frac{3}{2} \cdot \begin{bmatrix} {}^2\!/_3 & {}^2\!/_3 & {}^1\!/_3 \end{bmatrix} \begin{bmatrix} {}^2\!/_3 \\ {}^2\!/_3 \\ {}^1\!/_3 \end{bmatrix}.$$

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. For example, the dominant eigenvector of this $\Sigma_{\mathbf{X}}$ is

$$\mathbf{v}_{\text{dominant}} = \frac{1}{\sqrt{3+\sqrt{3}}} \begin{bmatrix} \frac{1}{2}(1+\sqrt{3}) \\ \frac{1}{2}(1+\sqrt{3}) \\ 1 \end{bmatrix},$$

which stands close but is not equal to the vector associated to the common information. This single point disproves that total conditional correlation is generally minimized by corrections that commute with the eigenspace of $\Sigma_{\mathbf{X}}$.

5.3.2 Reverse Water-Filling is an Inner Bound

In fact, a reverse water-filling procedure on the eigenvalues as what worked for $K = 2$ is only an inner bound to the caching problem for $K > 2$. It stems from the inequality of geometric and arithmetic means. To see why, define the following function similar to $R_{\text{cache}}(d, D_F)$, but in which we replace the product constraint by the trace:

$$R_{\text{trace}}(\gamma, D_F) \triangleq \min_{\mathbf{D}} \frac{1}{2} \log \frac{|\Sigma_{\mathbf{X}}|}{|\mathbf{D}|} \quad \text{s.t.} \quad \begin{cases} \mathbf{0} \preceq \mathbf{D} \preceq \Sigma_{\mathbf{X}} \\ \operatorname{tr}(\mathbf{D}) \leq \gamma \\ D_{i,i} \geq D_F, \quad \forall i = 1, 2, \cdots, K \end{cases} \tag{5.22}$$

Note that $\mathbf{D}$ being a squared error distortion matrix implies that $\operatorname{tr}(\mathbf{D})$ is the sum squared error:

$$\operatorname{tr}(\mathbf{D}) = \sum_{i=1}^{K} \mathbb{E}[(X_i - \mathbb{E}[X_i|\mathbf{V}])^2]. \tag{5.23}$$

Lemma 5.2. *The caching rate-distortion function is upper bounded as:*

$$R_{cache}(d, D_F) \leq R_{\text{trace}}(K d^{1/K}, D_F). \tag{5.24}$$

Proof. The trace of a matrix and the product of its diagonal entries are connected through the inequality of geometric and arithmetic means:

$$\left(\prod_{i=1}^{K} D_{i,i} \right)^{1/K} \leq \frac{1}{K} \operatorname{tr}(\mathbf{D}). \tag{5.25}$$

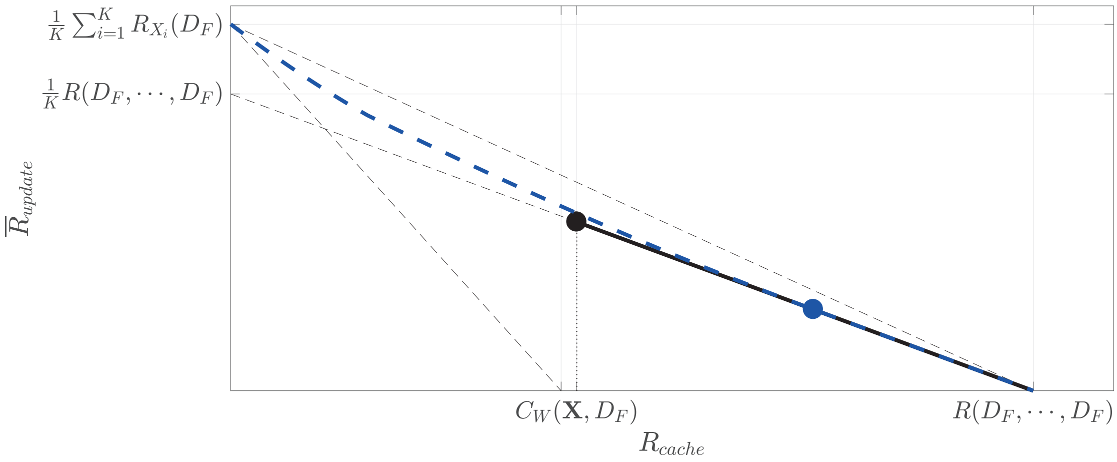


Figure 5.7 – Comparison of known achievable $(R_{\text{cache}}, \overline{R}_{\text{update}})$−pairs for all $R_{\text{cache}} \geq C_W(\mathbf{X}, D_F)$ (the tick black line) and the reverse water-filling inner bound of Lemma 5.2 (the dashed blue line). Example drawn for an $\mathbf{X}$ of which $K = 3$ and $\rho_{12} = \rho_{13} = \frac{1}{3}$, $\rho_{23} = \frac{5}{6}$.

Hence, if we plug in $Kd^{1/K}$ then $R_{\text{trace}}(Kd^{1/K}, D_F)$ must be solved by a matrix $\mathbf{D}$ of which

$$\prod_{i=1}^{K} D_{i,i} \leq \left(\frac{1}{K}\operatorname{tr}(\mathbf{D})\right)^{K} \leq d. \tag{5.26}$$

The chain of inequalities imply that the domain of feasible $\mathbf{D}$ for $R_{\text{trace}}(Kd^{1/K}, D_F)$ is a subset of the search domain of $R_{\text{cache}}(d, D_F)$. Since both functions minimize the same objective function, we must therefore have that $R_{\text{cache}}(d, D_F) \leq R_{\text{trace}}(Kd^{1/K}, D_F)$. □

The crux of Lemma 5.2 is the following: if D_F is 0 (or more exactly, is so small that it poses no active constraint in (5.22)), then $R_{\text{trace}}(d, D_F)$ is the ‘classic’ rate-distortion function of a Gaussian vector source subject to a sum squared error criterion ((3.7) and (3.14)). This function is minimized by a reverse water-filling procedure on the eigenvalues of the covariance matrix. Now if $\Sigma_{\mathbf{X}}$ is normalized to a correlation matrix, such water-filled distortion matrices were also optimal for caching for $K = 2$. The Lemma shows that for $K > 2$, though, they merely offer an inner bound. Experiments show that in general one can do caching more efficiently by picking a distortion matrix whose eigenbasis does *not* commute with the eigenspace of $\Sigma_{\mathbf{X}}$. Figure 5.7 shows an example of this trace-based inner bound compared to $(R_{\text{cache}}, \overline{R}_{\text{update}})$−pairs we know are achievable in the high cache rate regime through Theorem 5.3.

5.4 The Circulant Exception Conjecture

We conjecture that the inner bound of Lemma 5.2 is not tight except for one special case: when $\Sigma_{\mathbf{X}}$ is circulant. It would not have been unreasonable to expect that the eigendecomposition of $\Sigma_{\mathbf{X}}$ would capture the total conditional correlation $TC(\mathbf{X}|\mathbf{V})$, as it was the case for $K = 2$. Lemma 5.2 suggested, however, that one might have to look beyond the eigendecomposition. In this section, we conjecture that this is indeed the case.

Specifically, we argue that total conditional correlation is only minimized by taking a $\mathbf{D}$ whose eigenspace commutes with that of $\Sigma_{\mathbf{X}}$ if the eigenvectors are completely symmetric w.r.t. to the basis vectors of the system. This is only the case for circulant $\Sigma_{\mathbf{X}}$. In such a setting, each X_i is equivalent to any other and hence all X_i contribute equally to the total correlation.

If circulant $\Sigma_{\mathbf{X}}$ would indeed be a special case, then it would explain why a water-filling procedure on eigenvalues is optimal for the caching of bivariate Gaussians. After all, every 2×2 correlation matrix is circulant by construction. In other words, the ease of caching two Gaussians and its tightness on Lemma 5.2 would not be the rule, but rather the exception.

To start, we take a second look at the common information of K Gaussian sources. Xu *et al.* [30] were the first to extend common information to K sources. They found an analytic expression for (5.16) for 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, this result is part of a much wider class of sources, i.e., Gaussians whose correlation matrix is circulant:

Theorem 5.4. *For any $K-$dimensional Gaussian $\mathbf{X}$ whose covariance $\Sigma_{\mathbf{X}}$ is circulant, we have*

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

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

Proof. First, the solution to (5.17) is a unique distortion profile $\mathbf{D}$. In [33], among others, the strict convexity of the Gaussian channel was discussed. Our formulation, though, more closely resembles the geometric problem: $\min -\log|\mathbf{D}|$ under linear constraints corresponds to finding a maximum volume ellipsoid inside a *convex* body (as also discussed heavily in Chapter 2). It is called the (Löwner–)John ellipsoid and was shown in [34] to be unique.

Secondly, uniqueness implies that a circulant $\Sigma_{\mathbf{X}}$ will result in a circulant $\mathbf{D}$ as the minimizer of (5.17). Namely, 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 $D_{j,j}$, $\mathbf{D} \preceq \Sigma_{\mathbf{X}}$ would still hold and $|\mathbf{D}|$ would be unaffected.

Therefore, any $\mathbf{D}$ whose diagonal entries are not equal cannot be unique and thus cannot be the solution to (5.17). Concluding that all non-zero entries must be equal, $\mathbf{D} = \lambda_{\min}(\Sigma_{\mathbf{X}})\mathbf{I}$ has the largest determinant among all scaled identity matrices subject to $\mathbf{D} \preceq \Sigma_{\mathbf{X}}$. □

Theorems in Pictures

The plot below shows an example of Theorem 5.4.
The common information of a circulant covariance matrix is associated to $\mathbf{D}_{C_W} = \lambda_{\min}(\Sigma_{\mathbf{X}})\mathbf{I}$. This means that $\mathcal{E}_{\mathbf{D}_{\mathbf{C_W}}}$ is a sphere. From a geometric perspective it is not hard to see why this must be the case. Earlier, we established that $\mathcal{E}_{\mathbf{D}_{\mathbf{C_W}}}$ is the ellipsoid of the largest volume whose semi-principal axes are straight and that fits inside $\mathcal{E}_{\Sigma_{\mathbf{X}}}$. The eigenvectors of a circulant $\Sigma_{\mathbf{X}}$, however, are perfectly symmetric with respect to the basis vectors; most notably, the dominant eigenvector is $\mathbf{1}$ per definition, if K is even then the eigenvector of $\lambda_{\min}$ alternates $+1$ and -1, and all other eigenvalues come in multiplicity of two or multiples thereof. Hence, the largest straight ellipsoid inside $\mathcal{E}_{\Sigma_{\mathbf{X}}}$ must be a sphere simply because there is no 'room' to increase volume and be asymmetric in any particular direction.

For $K = 3$, the only circulant correlation matrix has equal correlation for all pairwise relations. Consider

$$\Sigma_{\mathbf{X}} = \begin{bmatrix} 1 & {}^2\!/_3 & {}^2\!/_3 \\ {}^2\!/_3 & 1 & {}^2\!/_3 \\ {}^2\!/_3 & {}^2\!/_3 & 1 \end{bmatrix}.$$

The following figure plots the matrix associated to common information of $\Sigma_{\mathbf{X}}$. Note in particular how each side view is the same.

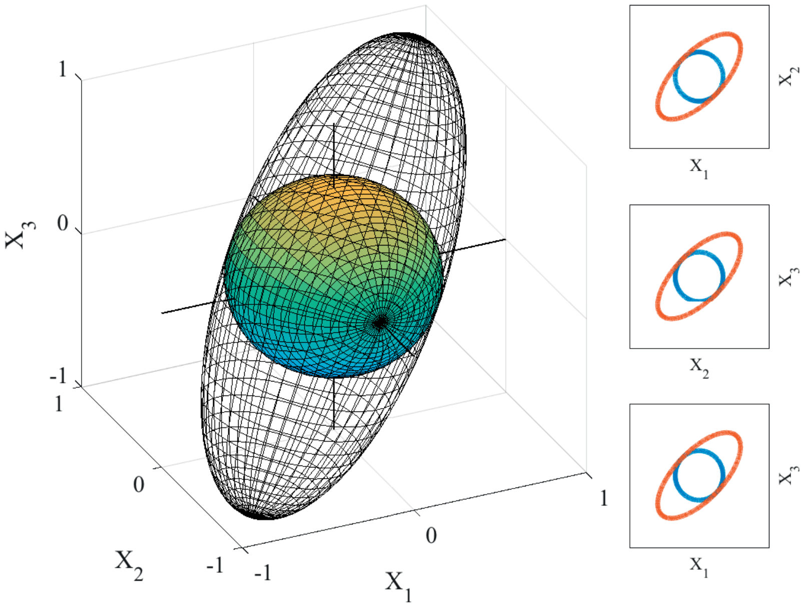


Figure 5.8 – The common information of a circulant covariance matrix is associated to the largest *sphere* that fits inside $\mathcal{E}_{\Sigma_{\mathbf{X}}}$.

As mentioned, we conjecture that total conditional correlation is minimized by a reverse waterfilling procedure on the eigenvalues of $\Sigma_{\mathbf{X}}$ if and only if $\Sigma_{\mathbf{X}}$ is circulant. In our caching coding setting, we say the bound of Lemma 5.2 is tight if this is the case, which is the same statement up to the inclusion of end distortion constraints D_F. We now develop this conjecture by first writing a lemma on the 'if'-part and then a conjecture on the 'only if' part of the statement.

Lemma 5.3. *If $R_{cache}(d, D_F)$ has a unique minimizer in the regime of $d > |\Sigma_{\mathbf{X}}|2^{-2C_W(\mathbf{X})}$, then*

$$R_{cache}(d, D_F) = R_{\text{trace}}(Kd^{1/K}, D_F)$$

if $\Sigma_{\mathbf{X}}$ is circulant.

Proof. The regime of $R_{\text{cache}} \geq C_W(\mathbf{X})$ is not the difficult part. First a technicality: observe the use of Wyner's common information rather than the constrained version $C_W(\mathbf{X}, D_F)$. There is no need for the latter, because $\mathbf{D}_{C_W} = \lambda_{\min}(\Sigma_{\mathbf{X}})\mathbf{I}$ has symmetric distortions and we defined $R_{\text{cache}}(d, D_F)$ for symmetric end distortions as well. Then, observe that

$$\frac{1}{2}\log\frac{|\Sigma_{\mathbf{X}}|}{d} \leq R_{\text{cache}}(d, D_F) \leq R_{\text{trace}}(Kd^{1/K}, D_F) \tag{5.27}$$

is met with equality in all steps for $d \leq |\Sigma_{\mathbf{X}}|2^{-2C_W(\mathbf{X})}$.

For $d > |\Sigma_{\mathbf{X}}|2^{-2C_W(\mathbf{X})}$ we invoke the same argument as for Theorem 5.4: if the $\mathbf{D}$ that solves $R_{\text{cache}}(d, D_F)$ is unique and $\Sigma_{\mathbf{X}}$ is circulant then $\mathbf{D}$ is necessarily circulant as well; otherwise one can swap two pairs of rows and columns and that new matrix would be equally feasible. This proves the lemma, because a circulant matrix achieves equality of the arithmetic and geometric means, as $D_{1,1} = D_{2,2} = \cdots = D_{K,K}$. This was the core argument of comparing $R_{\text{cache}}(d, D_F)$ and $R_{\text{trace}}(Kd^{1/K})$ through (5.25). □

At this point, it is unclear whether or not the minimizer of $R_{\text{cache}}(d, D_F)$ is unique. The actual conjecture consists of the assumption that uniqueness is indeed guaranteed and that the equivalence of $R_{\text{cache}}(d, D_F)$ and $R_{\text{trace}}(Kd^{1/K}, D_F)$ is a property solely held by circulant correlation matrices. Thusfar, we have not been able to prove this conjecture theoretically, nor disprove it by exhaustive search over possible distortion matrices that could solve $R_{\text{cache}}(d, D_F)$:

Conjecture 5.1. *We have*

$$R_{cache}(d, D_F) = R_{\text{trace}}(Kd^{1/K}, D_F),$$

if and only if the correlation matrix $\Sigma_{\mathbf{X}}$ is circulant.

The 'if'-part is contained in Lemma 5.3 and the assumption that the minimizer of $R_{\text{cache}}(d, D_F)$ is indeed unique. This assumption is supported by the assertion that $\mathbf{D} \preceq \Sigma_{\mathbf{X}}$ cannot be strict for $d > |\Sigma_{\mathbf{X}}|2^{-2C_W(\mathbf{X})}$ (otherwise there exists a better distortion matrix $\mathbf{D}$). However, the combination of the semidefinite ordering with the non-convex constraint $\prod_{i=1}^{K} D_{i,i} \leq d$ makes uniqueness difficult to prove.

For the 'only if', the argument is as follows: If D_F is not an active constraint then $R_{\text{trace}}(\gamma, D_F)$ coincides with the Gaussian rate-distortion function subject to a sum squared error, which is solved by reverse water-filling on the eigenvalues of $\Sigma_{\mathbf{X}}$. The aforementioned is known; what we now conjecture is that the construction of $R_{\text{cache}}(d, D_F)$ pushes the encoder to certainly *not* work on the eigendecomposition of $\Sigma_{\mathbf{X}}$, unless it has no other option, which we believe is what happens when the correlation matrix is circulant.

Namely, $R_{\text{cache}}(d, D_F)$ tries to find a $\mathbf{D}$ with $\max|\mathbf{D}|$, whose eigenvectors are as close to the identity matrix as possible (since this would render $\mathbf{D}$ diagonal) under the constraint $\mathbf{D} \preceq \Sigma_{\mathbf{X}}$. Consider any candidate $\mathbf{D}$. Multiplying $\mathbf{D}$ by an orthonormal rotation matrix leaves the determinant (and thus the objective value) intact, while $\prod_{i=1}^{K} D_{i,i}$ may improve (hence providing room on the constraint). The encoder will always try to rotate a distortion profile away from the eigenbasis of $\Sigma_{\mathbf{X}}$ in an attempt to diagonalize it. If the dominant eigenvectors of $\Sigma_{\mathbf{X}}$ lean towards any of the basis vectors, then this gives room on the constraint $\mathbf{D} \preceq \Sigma_{\mathbf{X}}$ to do so.

All $\Sigma_{\mathbf{X}}$ have dominant eigenvectors that lean in some direction, except for circulant matrices: $\lambda_0(\Sigma_{\mathbf{X}})$ is associated to the eigenvector $\mathbf{1}$, $\lambda_{K-1}(\Sigma_{\mathbf{X}})$ to $[+1, -1, +1, -1 \cdots]^T$ (if K even) and all other λ come in pairs of two and have no uniquely defined eigenvectors. Consequently, the encoder sees no room on the constraint $\mathbf{D} \preceq \Sigma_{\mathbf{X}}$ to rotate the eigenbasis of $\mathbf{D}$ in any direction (which could potentially improve the product-constraint) without compromising on $|\mathbf{D}|$ (which hurts the objective value). One can therefore observe in simulations that the $\mathbf{D}$ that minimizes $R_{\text{cache}}(d, D_F)$ 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 identity as R_{cache} increases.
The next section of 'Theorems in Pictures' will visualize this difficult argument.

Connections Supporting the Conjecture
The conjecture does not contradict any of the results derived before and connects in the following way:

1. As stated, it explains the result for the caching of two Gaussians, because every 2 × 2 correlation matrix is necessarily circulant.

2. As for $K \geq 2$, the conjecture connects the low cache-rate regime to the start of the high cache-rate regime at Theorem 5.4. Namely, a reverse water filling procedure on the eigenvalues of $\Sigma_{\mathbf{X}}$ will only result in a diagonal distortion matrix once $\mathbf{D} = \lambda_{\min}(\Sigma_{\mathbf{X}}) \cdot \mathbf{I}$ is reached.

3. In Section 5.3.1 we argued by counterexample that, in general, the notion of total (conditional) correlation is not minimized by operations that commute with the eigenspace of $\Sigma_{\mathbf{X}}$. The conjecture would strengthen this anecdotal evidence by a clear segregation of when it is and when it is not.

Theorems in Pictures

In this box, we illustrate the 'only if' part of Conjecture 5.1. We resort back to examples from $K = 2$ as we know the optimal caching strategy for correlation matrices (which are circulant per construction) and non-circulant covariance matrices (by normalizing variance and then doing caching, as described in Section 4.3.4). Plotted are $\mathcal{E}_{\Sigma_\mathbf{X}}$ and the ellipse of one particular optimal caching distortion profile for some $R_{\text{cache}} \leq C_W(\mathbf{X})$. The left shows a circulant correlation matrix, the right a non-circulant covariance. Observe that in the right Figure, the dominant eigenvector leans towards one of the two basis vectors.

Recall that in optimizing $R_{\text{cache}}(d, D_F)$, the encoder tries to cache as much of the correlation as possible (as measured by evaluating $TC(\mathbf{X}|\mathbf{V})$ (5.14)). This corresponds to an achievable distortion profile that closes the Hadamard inequality on diag($\mathbf{D}$) as much as possible. In geometry, a diagonal matrix corresponds to a straight ellipse. The encoder thus tries to find an ellipse that maximizes volume (= large determinant = low R_{cache}) while being 'as straight as possible'. If the eigenvectors of $\Sigma_\mathbf{X}$ are leaning towards any basis vector(s), the encoder can use this as wiggle room to find large ellipses that are 'more straight' than $\mathcal{E}_{\Sigma_\mathbf{X}}$ without violating $\mathbf{D} \preceq \Sigma_\mathbf{X}$. Circulant (correlation) matrices offer no such wiggle room, because their eigenvectors are perfectly symmetric w.r.t. all basis vectors. Therefore, the encoder simply has no other option than to pick a $\mathbf{D}$ whose eigenspace commutes with that of $\Sigma_\mathbf{X}$. That is the crux of Conjecture 5.1.

Figure 4.8 in Chapter 4 also shows how the eigenbasis of the optimal $\mathbf{D}$ rotates as R_{cache} increases. It rotates in the direction of the basis vector to which the dominant eigenvector leans, i.e., where the most room is offered by the convex hull (the '$\preceq \Sigma_\mathbf{X}$'-constraint).

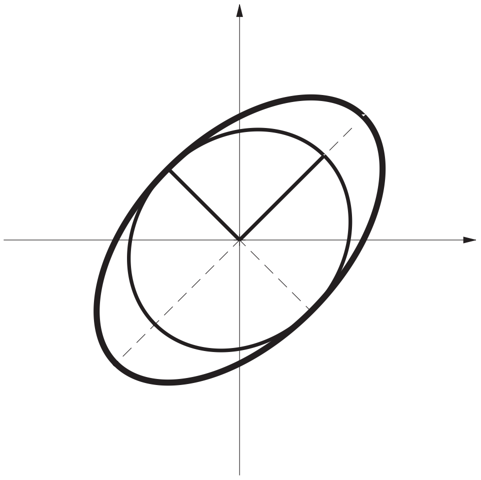


Figure 5.9 – Circulant $\Sigma_\mathbf{X}$, there is no 'wiggle room' to straighten $\mathcal{E}_\mathbf{D}$ without compromising volume.

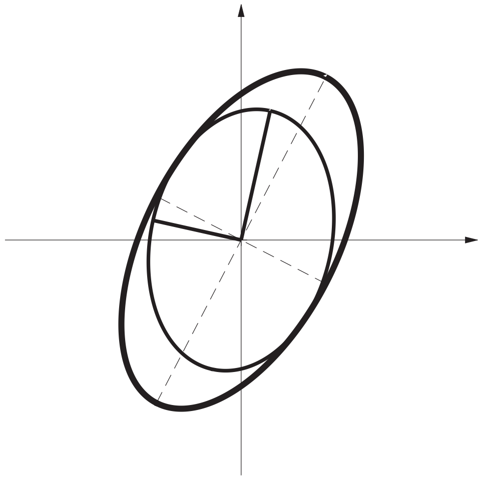


Figure 5.10 – The dominant eigenvector leans towards $[0,1]^T$. This 'wiggle room' allows $\mathcal{E}_\mathbf{D}$ to straighten through rotation.

6 Total Correlation of Gaussian Vectors

In caching under uniform user preference, Watanabe's notion of total conditional correlation measures how well a caching strategy captures the shared information between the files in a database. Whereas in Chapter 5 we increased the size of the database from two to K sources, now instead we increase the size of the files[1]. We study the common information and total correlation that exists in a library of *vectors*, $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \cdots, \mathbf{X}^{(K)}$ (though for the greatest part we take $K = 2$).

These thoughts are inspired by the work of Satpathy and Cuff on secure source coding [36]. They found *en passant* a closed-form expression for Wyner's common information of two Gaussian vectors. They used a transformation that turns two length d vectors $(\mathbf{X}, \mathbf{Y})$ into d independent $(\tilde{X}_i, \tilde{Y}_i)$ pairs . The common information is then the sum of that of all the pairs:

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

Does this mechanism also apply to capturing *parts* of the correlation through the measure of total conditional correlation, rather than capturing *all* through common information? The answer is yes. Restricting again to Gaussian auxiliaries, total conditional correlation of two vectors falls apart into a convex and non-convex part. The latter is the minimization of correlation of each $(\tilde{X}_i, \tilde{Y}_i)-$pair. The former is a convex distribution problem of which pairs need to be tackled first.

The most efficient way to minimize total conditional correlation of two vectors is by minimizing the correlation of each pair separately in a *reverse water-filling* fashion. Unlike traditional Gaussian water filling, this procedure is not with respect to eigenvalues, but to which pair has the most common information. The emphasis of this Chapter is on the lossless statistical concepts, though at the end we also fit this building block into our caching coding problem on the Gray–Wyner network and show a fit if the end distortion constraints are not too large.

[1] The material of this chapter appeared in [35]

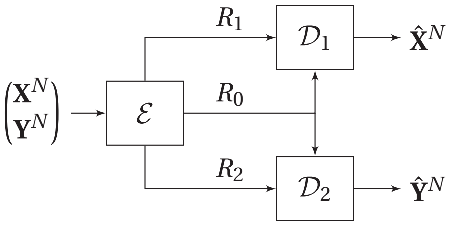


Figure 6.1 – The Gray–Wyner network for 2 sources uses the notation $(\mathbf{X},\mathbf{Y})$.

6.1 Problem Statement

The topic of study of this chapter is twofold, with each side touching upon different dimensionality:

1. A simple generalization of Wyner's common information for K Gaussian vectors.
2. The minimization of Watanabe's total conditional correlation for 2 Gaussian vectors.

The second caries the most weight and hence most of the discussion will be told from the bivariate perspective. At the very end we shall also apply that minimization of total correlation as a building block inside the caching setting, sticking as well to dimensionality 2.

6.1.1 Notation

The motivation of the discussion is again anchored in the Gray–Wyner network of Figure 6.1 and its extension to K sources of Figure 6.2. In general, the information source of interest consists of K Gaussian vectors $(\mathbf{X}^{(1)},\mathbf{X}^{(2)},\cdots,\mathbf{X}^{(K)})$, each of length d. Their joint covariance is the following:

$$\Sigma = \begin{bmatrix} \Sigma_{\mathbf{X}^{(1)}} & \Sigma_{\mathbf{X}^{(1)}\mathbf{X}^{(2)}} & \cdots \\ \Sigma^T_{\mathbf{X}^{(1)}\mathbf{X}^{(2)}} & \Sigma_{\mathbf{X}^{(2)}} & \\ \vdots & & \ddots \end{bmatrix}.$$

The indexing makes notation undesirably complex. So whenever we deal with only two Gaussian vectors, we prefer to use letters, i.e., $(\mathbf{X},\mathbf{Y})$, and define the covariance as:

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

No subscript indicates the $2d \times 2d$ matrix corresponding to the joint distribution $p(\mathbf{x},\mathbf{y})$, whereas a subscript is to refer to a corner of that matrix. The same applies to the $Kd \times Kd$ matrix and its block structure. The same style of matrix indexing will apply in general, for example to distortions.

The total correlation, as used before, follows for these Gaussian vectors the following expres-

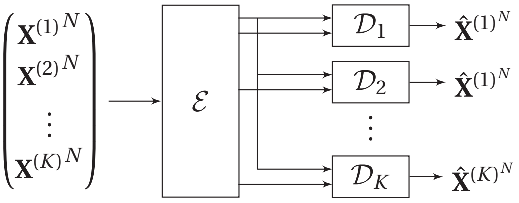


Figure 6.2 – A $K-$extended Gray–Wyner network uses the notation $(\mathbf{X}^{(1)},\mathbf{X}^{(2)},\cdots,\mathbf{X}^{(K)})$.

sion:

$$TC(\mathbf{X}^{(1)},\mathbf{X}^{(2)},\cdots,\mathbf{X}^{(K)}) \triangleq \sum_{i=1}^{K} h(\mathbf{X}^{(i)}) - h(\mathbf{X}^{(1)},\mathbf{X}^{(2)},\cdots,\mathbf{X}^{(K)}) \tag{6.1}$$

$$= \frac{1}{2}\log\frac{\prod_{i=1}^{K}|\Sigma_{\mathbf{X}^{(i)}}|}{|\Sigma|}, \tag{6.2}$$

or more compactly if we are only considering two:

$$\begin{aligned} TC(\mathbf{X},\mathbf{Y}) &\triangleq I(\mathbf{X};\mathbf{Y}) \\ &= \frac{1}{2}\log\frac{|\Sigma_{\mathbf{X}}|\cdot|\Sigma_{\mathbf{Y}}|}{|\Sigma|}. \end{aligned} \tag{6.3}$$

In words: total correlation is thus driven by the ratio of the product of determinants of each block on the diagonal of Σ and the determinant of Σ itself. As for total *conditional* correlation,

$$TC(\mathbf{X}^{(1)},\mathbf{X}^{(2)},\cdots,\mathbf{X}^{(K)}|\mathbf{V}) = \sum_{i=1}^{K} h(\mathbf{X}^{(i)}|\mathbf{V}) - h(\mathbf{X}^{(1)},\mathbf{X}^{(2)},\cdots,\mathbf{X}^{(K)}|\mathbf{V}). \tag{6.4}$$

It decomposes into the same Gaussian expression based on the covariance matrix of the conditional distribution $p(\mathbf{x},\mathbf{y}|\mathbf{v})$. Studying this last entity is the main topic of this chapter.

6.1.2 The Gray–Wyner Network in Vector Notation

Even though the emphasis of this chapter is on the statistical concepts of common information and total correlation, let us briefly address the application/coding side as well to get everything in vector notation. For the formal definitions of codes on the Gray–Wyner network we refer the reader back to Section 4.1. Furthermore, for brevity and with the rest of the chapter in mind, we stick to the $K = 2$ setting.

Consider Figure 6.1. A rate-distortion tuple $(R_0, R_1, R_2, D_x, D_y)$ is achievable if for some $p(\mathbf{v}|\mathbf{x},\mathbf{y})$ the following conditions hold [7, 12, 30]:

$$\begin{aligned} R_0 &\geq I(\mathbf{X},\mathbf{Y};\mathbf{V}) \\ R_1 &\geq R_{\mathbf{X}|\mathbf{V}}(D_x) \\ R_2 &\geq R_{\mathbf{Y}|\mathbf{V}}(D_y). \end{aligned}$$

The union of all such tuples over all densities $p(\mathbf{v}|\mathbf{x},\mathbf{y})$ characterizes the entire achievable rate region.

Working with Gaussian vector sources, let us choose for the distortion metric a trace-constraint, i.e.,

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

the same for $\mathbf{Y}$. With these constraints, the rate-distortion function under *individual* distortion criteria is the following expression:

$$R_{\mathbf{X},\mathbf{Y}}(D_x,D_y) = \min_{\mathbf{D}} \frac{1}{2}\log\frac{|\Sigma|}{|\mathbf{D}|} \quad \text{s.t.} \begin{cases} \mathbf{0} \preceq \mathbf{D} \preceq \Sigma, \\ \operatorname{tr}(\mathbf{D}_{\mathbf{X}}) \le D_x, \\ \operatorname{tr}(\mathbf{D}_{\mathbf{Y}}) \le D_y. \end{cases} \tag{6.5}$$

The above is a direct consequence of [19], as explained in Section 3.1. Observe that these trace-constraints on the corners of $\mathbf{D}$ respect semidefinite ordering, i.e., $\mathbf{D}_1 \preceq \mathbf{D}_2 \Rightarrow d(\mathbf{D}_1) \le d(\mathbf{D}_2)$; therefore Gaussian codebooks are optimal to achieve optimality in the rate-distortion sense by Theorem 3.1.

To justify the motivation of studying total conditional correlation, observe that this measure relates closely to the loss incurred in distributing rate between R_0 and the two individual branches R_1+R_2. For example, if one picks a $\mathbf{V}$ for the common branch that is jointly Gaussian with $(\mathbf{X},\mathbf{Y})$ then $R_{\mathbf{X}|\mathbf{V}}(D_x)$ and $R_{\mathbf{Y}|\mathbf{V}}(D_y)$ are necessarily solved by Gaussian distributions as well. Let $\mathbf{K}$ be the covariance of the conditional distribution $p(\mathbf{x},\mathbf{y}|\mathbf{v})$ and let $\mathbf{D}_{\mathbf{X}},\mathbf{D}_{\mathbf{Y}}$ be the distortion matrices resulting from the conditional rate-distortion functions on the individual branches. Then we have,

$$R_0 = \frac{1}{2}\log\frac{|\Sigma|}{|\mathbf{K}|} \tag{6.6}$$

$$R_1 + R_2 = \frac{1}{2}\log\frac{|\mathbf{K}_{\mathbf{X}}|\cdot|\mathbf{K}_{\mathbf{Y}}|}{|\mathbf{D}_{\mathbf{X}}|\cdot|\mathbf{D}_{\mathbf{Y}}|}. \tag{6.7}$$

Observe that

$$R_0 + R_1 + R_2 = \frac{1}{2}\log\frac{|\Sigma_{\mathbf{X}}|}{|\mathbf{D}_{\mathbf{X}}|\cdot|\mathbf{D}_{\mathbf{Y}}|} + \frac{1}{2}\log\frac{|\mathbf{K}_{\mathbf{X}}|\cdot|\mathbf{K}_{\mathbf{Y}}|}{|\mathbf{K}|} \tag{6.8}$$

$$= \frac{1}{2}\log\frac{|\Sigma_{\mathbf{X}}|}{|\mathbf{D}_{\mathbf{X}}|\cdot|\mathbf{D}_{\mathbf{Y}}|} + TC(\mathbf{X},\mathbf{Y}|\mathbf{V}). \tag{6.9}$$

$T(\mathbf{X},\mathbf{Y}|V)$ characterizes the loss due to the encoder's inability to leverage on the individual branches the correlation between $\mathbf{X}$ and $\mathbf{Y}$ that still exists after conditioning on $\mathbf{V}$ (the common message). We observed the same in the caching problem in which we compared the caching rate to the average of updating either file. The rest of this chapter focuses more on the measure $T(\mathbf{X},\mathbf{Y}|\mathbf{V})$ itself.

6.1.3 Vectors to Pairs decomposition

The key operation of this chapter is a set of operations to turn two vectors $(\mathbf{X}, \mathbf{Y})$ into a set of d independent unit-variance pairs $(\tilde{X}_i, \tilde{Y}_i)$, which Satpathy and Cuff used to characterize the common information [36]. They used this method as a *transform*, but we -for convenience later on- will rather *decompose* $(\mathbf{X}, \mathbf{Y})$. To start, pull out the variance:

$$\mathbf{X} = \Sigma_{\mathbf{X}}^{1/2} \bar{\mathbf{X}}, \tag{6.10}$$

$$\mathbf{Y} = \Sigma_{\mathbf{Y}}^{1/2} \bar{\mathbf{Y}}. \tag{6.11}$$

The components of $\bar{\mathbf{X}}, \bar{\mathbf{Y}}$ are necessarily unit-variance independent Gaussians, but their cross-correlation does not disappear by this step. Namely, their covariance takes on this shape:

$$\bar{\Sigma} = \begin{bmatrix} \mathbf{I} & \Sigma_{\mathbf{X}}^{-1/2} \Sigma_{\mathbf{XY}} \Sigma_{\mathbf{Y}}^{-1/2} \\ \Sigma_{\mathbf{X}}^{-1/2} \Sigma_{\mathbf{XY}} \Sigma_{\mathbf{Y}}^{-1/2} & \mathbf{I} \end{bmatrix}. \tag{6.12}$$

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

$$\Sigma_{\bar{\mathbf{X}}\bar{\mathbf{Y}}} = \Sigma_{\mathbf{X}}^{-1/2} \Sigma_{\mathbf{XY}} \Sigma_{\mathbf{Y}}^{-1/2} = \mathbf{B}_{\mathbf{X}} \Lambda \mathbf{B}_{\mathbf{Y}}, \tag{6.13}$$

where now Λ is diagonal and positive, and $\mathbf{B}_{\mathbf{X}}, \mathbf{B}_{\mathbf{Y}}$ are orthonormal matrices. This gives us

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

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

$$\tilde{\Sigma} = \begin{bmatrix} \mathbf{I} & \Lambda \\ \Lambda & \mathbf{I} \end{bmatrix}, \tag{6.14}$$

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

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

$$\begin{aligned} h(\mathbf{X}, \mathbf{Y}) &= \frac{1}{2} \log(2\pi e)^{2d} |\Sigma| \\ &= \frac{1}{2} \log(2\pi e)^{2d} |\Sigma_{\mathbf{X}}||\Sigma_{\mathbf{Y}}||\tilde{\Sigma}| \\ &= \sum_{i=1}^{d} h(\tilde{X}_i, \tilde{Y}_i) + \frac{1}{2} \log |\Sigma_{\mathbf{X}}||\Sigma_{\mathbf{Y}}|. \end{aligned} \tag{6.15}$$

6.2 Common Information of a Database of Vectors

First, we would like to add one comment to the work of Satpathy and Cuff on the common information for Gaussian vectors [36]. For two scalars, we know that the closed-form solution equals [28, 30]

$$C_W(X,Y) = \min_{X-V-Y} I(X,Y;V) = \frac{1}{2}\log\frac{1+|\rho|}{1-|\rho|}. \tag{6.16}$$

Furthermore, we showed in Section 5.2 that the common information of $K > 2$ jointly Gaussian random variables is necessarily Gaussian as well, but may not have a 'nice' analytic expression. It is however a convex problem, which at least can be solved efficiently by linear programming [26]. For Gaussian *vectors*, this scales up entirely in the same manner.

For *two* Gaussian vectors, Satpathy and Cuff derived the following:

Lemma 6.1 (Satpathy and Cuff [36])**.** *For jointly Gaussian* $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^d$*, Wyner's common information is given by*

$$C_W(\mathbf{X},\mathbf{Y}) = \min_{\mathbf{X}-\mathbf{V}-\mathbf{Y}} I(\mathbf{X},\mathbf{Y};\mathbf{V}) = \frac{1}{2}\sum_{i=1}^{d}\log\frac{1+|\rho_i|}{1-|\rho_i|}, \tag{6.17}$$

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

In short, the vector common information equals the sum of the common information of all $(\tilde{X}_i, \tilde{Y}_i)$-pairs obtained by the vector-to-pairs decomposition described in Section 6.1.3. Unfortunately, such a transformation cannot apply when there are more than two vectors to make independent; it breaks, because the singular value decomposition of (6.13) can diagonalize a single cross-correlation block in the corner of $\bar{\Sigma}$, but not multiple blocks simultaneously in the case of $K > 2$.

For higher dimensions, define:

$$\begin{aligned} C_W(\mathbf{X}^{(1)},\cdots,\mathbf{X}^{(K)}) \triangleq \min_{\mathbf{V}} \quad & I(\mathbf{X}^{(1)},\cdots,\mathbf{X}^{(K)};\mathbf{V}) \\ \text{s.t.} \quad & TC(\mathbf{X}^{(1)},\cdots,\mathbf{X}^{(K)}|\mathbf{V}) = 0. \end{aligned} \tag{6.18}$$

Note that $TC(\mathbf{X}^{(1)},\cdots,\mathbf{X}^{(K)}|\mathbf{V}) = 0$ if and only if $p(\mathbf{x}^{(1)},\cdots,\mathbf{x}^{(K)}|\mathbf{v}) = \prod_{i=1}^{K} p(\mathbf{x}^{(i)}|\mathbf{v})$.

Theorem 6.1. *For jointly Gaussian* $\mathbf{X}^{(1)},\cdots,\mathbf{X}^{(K)} \in \mathbb{R}^d$ *with* $Kd \times Kd$ *joint covariance* Σ*, Wyner's common information is given by*

$$C_W(\mathbf{X}^{(1)},\cdots,\mathbf{X}^{(M)}) = \min_{\mathbf{D}} \frac{1}{2}\log\frac{|\Sigma|}{|\mathbf{D}|} \quad \textit{such that} \begin{cases} 0 \preceq \mathbf{D} \preceq \Sigma, \\ \mathbf{D} \textit{ is block-diagonal.} \end{cases} \tag{6.19}$$

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

conditionally independent. Let furthermore $\mathbf{D}^{(i)} = \mathbb{E}[(\mathbf{X}^{(i)} - \mathbb{E}[\mathbf{X}^{(i)}|\mathbf{V}])(\mathbf{X}^{(i)} - \mathbb{E}[\mathbf{X}^{(i)}|\mathbf{V}])^T]$ Then,

$$\begin{aligned}
I(\mathbf{X}^{(1)}, \cdots, \mathbf{X}^{(K)}; \mathbf{V}) &= h(\mathbf{X}^{(1)}, \cdots, \mathbf{X}^{(K)}) - h(\mathbf{X}^{(1)}, \cdots, \mathbf{X}^{(K)}|\mathbf{V}) \\
&= h(\mathbf{X}^{(1)}, \cdots, \mathbf{X}^{(K)}) - \sum_{i=1}^{K} h(\mathbf{X}^{(i)}|\mathbf{V}) \\
&= \sum_{i=1}^{K} \Big(h(\mathbf{X}^{(i)}) - h(\mathbf{X}^{(i)}|\mathbf{V}) \Big) - TC(\mathbf{X}^{(1)}, \cdots, \mathbf{X}^{(K)}) \\
&= \sum_{i=1}^{K} I(\mathbf{X}^{(i)}; \mathbf{V}) - TC(\mathbf{X}^{(1)}, \cdots, \mathbf{X}^{(K)}) \\
&\geq \sum_{i=1}^{K} I(\mathbf{X}^{(i)}; \mathbb{E}[\mathbf{X}^{(i)}|\mathbf{V}]) - TC(\mathbf{X}^{(1)}, \cdots, \mathbf{X}^{(K)}) \qquad (6.20) \\
&\geq \sum_{i=1}^{K} R_{\mathbf{X}^{(i)}}(\mathbf{D}^{(i)}) - TC(\mathbf{X}^{(1)}, \cdots, \mathbf{X}^{(K)}) \qquad (6.21) \\
&= \frac{1}{2} \log \frac{\prod_{i=1}^{K} |\Sigma_{\mathbf{X}^{(i)}}|}{\prod_{i=1}^{K} |\mathbf{D}^{(i)}|} + \frac{1}{2} \log \frac{|\Sigma|}{\prod_{i=1}^{K} |\Sigma_{\mathbf{X}^{(i)}}|} \\
&= \frac{1}{2} \log \frac{|\Sigma|}{\prod_{i=1}^{K} |\mathbf{D}^{(i)}|}.
\end{aligned}$$

(6.20) is the data processing inequality.
(6.21) is the Gaussian rate-distortion function with respect to a matrix distortion constraint. Furthermore, let the covariance of the conditional distribution $p(\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(K)}|\mathbf{v}) = \prod_{i=1}^{K} p(\mathbf{x}^{(i)}|\mathbf{v})$ be denoted by the $Kd \times Kd$ matrix $\mathbf{D}$. Then by the law of total covariance we must have $\mathbf{D} \preceq \Sigma$. We have equality in all steps by picking $\mathbf{V}$ to be jointly Gaussian with $(\mathbf{X}^{(1)}, \cdots, \mathbf{X}^{(M)})$ and $\mathbf{D}$ to be block-diagonal. Hence one has to maximize $\prod_{i=1}^{K} |\mathbf{D}^{(i)}|$ subject to $0 \preceq \mathbf{D} \preceq \Sigma$, and it will be attainable by a Gaussian distribution. □

Another way of looking at it would be as follows: Conditional independence requires zero correlation. So one has to minimize mutual information subject to a shape-constraint on the covariance, which necessarily leads to Gaussian distributions. Now in a general case having zero correlation is necessary, but not sufficient for independence; for Gaussians, on the other hand, it is.

The theorem in itself is not so much a revelation as is the insight that the problem of common information is again a strictly convex MaxDet problem, constrained by only linear constraints [31]. Consequently, as we argued for scalar Gaussians, the common information and the distortion matrix that attains it can be found efficiently by linear programming. Using again the popular CVX package for MATLAB [32], one can easily verify for oneself that for $K = 2$ the optimization of Theorem 6.1 leads to the analytically found result by Satpathy and Cuff. For $K > 2$ we know no analytic expression for the optimal distortion matrix $\mathbf{D}$, though numerically the problem remains equally tractable.

6.3 Total Conditional Correlation of Two Vectors

Whereas common information characterizes a 'cost' of making random variables conditionally independent, total conditional correlation is instead a *measure* of how dependent random variables still are after conditioning. In this chapter we generalize this concept in a similar way: we wish to minimize a cost function such that the total correlation after conditioning does not exceed a certain level. This is similar to how in Chapters 4–5 we minimized cache rates while constraining average update rate (4.7)–(5.5), but here we strip the core mechanism bare of the application context.

The hypothesis is that by using the same cost function as $C_W(\mathbf{X},\mathbf{Y})$, the vectors-to-pair decomposition (Section 6.1.3) is also the right tool to capture as much of the total correlation as possible. This claim turns out to be true (given one restricts one's attention to jointly Gaussian auxiliaries). To that end, let us rewrite the notion of total conditional correlation into a cost minimization:

$$T_{\mathbf{X},\mathbf{Y}}(\gamma) \triangleq \min_{\mathbf{V}} I(\mathbf{X},\mathbf{Y};\mathbf{V}) \quad \text{s.t.} \quad h(\mathbf{X}|\mathbf{V}) + h(\mathbf{Y}|\mathbf{V}) \leq \gamma. \tag{6.22}$$

The choice for $I(\mathbf{X},\mathbf{Y};\mathbf{V})$ as the cost stems -of course- from the Gray–Wyner network and the similarity to Wyner's common information. $T_{\mathbf{X},\mathbf{Y}}(\gamma)$ is convex and non-increasing in γ, however, the constraint-function is concave in $\mathbf{V}$. So far, it is still unclear whether $\mathbf{X},\mathbf{Y}$ being Gaussian implies it suffices to also take $\mathbf{V}$ Gaussian. We therefore take Gaussianity as an assumption.

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

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

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

$$T_{\mathbf{X},\mathbf{Y}}(\gamma) = \min_{\mathbf{K}} \frac{1}{2}\log\frac{|\Sigma|}{|\mathbf{K}|} \quad \text{s.t.} \quad \begin{cases} 0 \preceq \mathbf{K} \preceq \Sigma, \\ |\mathbf{K}_\mathbf{X}||\mathbf{K}_\mathbf{Y}| \leq \gamma, \end{cases} \tag{6.23}$$

for which now the parameter γ is nicely bounded to $\gamma \in [0, |\Sigma_\mathbf{X}| \cdot |\Sigma_\mathbf{Y}|]$.

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

$$h(\mathbf{X},\mathbf{Y}|\mathbf{V}) \leq h(\mathbf{X}|\mathbf{V}) + h(\mathbf{Y}|\mathbf{V}), \tag{6.24}$$

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

$$|\mathbf{K}| \le |\mathbf{K_X}||\mathbf{K_Y}|. \tag{6.25}$$

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

Lemma 6.2. *For jointly Gaussian* $\mathbf{X},\mathbf{Y} \in \mathbb{R}^d$,

$$T_{\mathbf{X},\mathbf{Y}}\left(\gamma'\right) = C_W(\mathbf{X},\mathbf{Y}), \tag{6.26}$$

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

Proof. Conditional independence is equivalent to the condition $h(\mathbf{X}|\mathbf{V}) + h(\mathbf{Y}|\mathbf{V}) = h(\mathbf{X},\mathbf{Y}|\mathbf{V})$. For Gaussian distributions, this equality means that the covariance matrix associated to $p(\mathbf{x},\mathbf{y}|\mathbf{v})$ satisfies $|\mathbf{K}| = |\mathbf{K_X}||\mathbf{K_Y}|$. For minimizing total conditional correlation, this equality is the best one can achieve, as can be seen in (6.23). Filling equality in (6.25) into (6.23) gives:

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

The **V** that achieves common information corresponds to a matrix **K** that is diagonal after the transformation of Section 6.1.3, and is of the form $\tilde{\mathbf{K}}_\mathbf{X} = \tilde{\mathbf{K}}_\mathbf{Y} = \mathrm{diag}(\{1-|\rho_i|\}_{i=1}^d)$. So the **V** that achieves $C_W(\mathbf{X},\mathbf{Y})$ results in[2]:

$$\begin{aligned} h(\mathbf{X}|\mathbf{V}) + h(\mathbf{Y}|\mathbf{V}) &= \frac{1}{2}\log(2\pi e)^{2d}|\mathbf{K_X}||\mathbf{K_Y}| \\ &= \frac{1}{2}\log(2\pi e)^{2d}|\Sigma_\mathbf{X}||\Sigma_\mathbf{Y}|\prod_{i=1}^d(1-|\rho_i|)^2. \end{aligned} \tag{6.27}$$

Hence, choosing γ equal to this value of $|\mathbf{K_X}||\mathbf{K_Y}|$ gives $T_{\mathbf{X},\mathbf{Y}}(\gamma) = C_W(\mathbf{X},\mathbf{Y})$. □

For $\gamma < |\Sigma_\mathbf{X}||\Sigma_\mathbf{Y}|\prod_{i=1}^d(1-|\rho_i|)^2$ there is still equality in both (6.25) and (6.24), which implies that the choice of picking **V** jointly Gaussian with **X**, **Y** is not just an assumption anymore, it is also optimal in the general formulation of (6.22). Note, however, that for such small γ it holds that that $T_{\mathbf{X},\mathbf{Y}}(\gamma) > C_W(\mathbf{X},\mathbf{Y})$. Another implication is that this regime of small γ and conditional independence is not as challenging as large γ, where the total correlation remains strictly positive after conditioning on **V**.

We now focus on $\gamma > |\Sigma_\mathbf{X}||\Sigma_\mathbf{Y}|\prod_{i=1}^d(1-|\rho_i|)^2$, which brings us to the main result: $T_{\mathbf{X},\mathbf{Y}}(\gamma)$ is minimized by a reverse water-filling procedure on the common information of each $(\tilde{X}_i,\tilde{Y}_i)-$pair found by the decomposition of Section 6.1.3:

[2] Recall for these steps that entropy is affected by scaling if one applies the decomposition, like we showed before in (6.15).

Theorem 6.2. *For jointly Gaussian* $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^d$*, the total conditional correlation is minimized by a reverse water-filling procedure until* $\mathbf{X}, \mathbf{Y}$ *become conditionally independent, i.e.*

$$T_{\mathbf{X},\mathbf{Y}}(\gamma) = \begin{cases} \frac{1}{2}\log\frac{|\Sigma|}{\gamma} & \gamma \le |\Sigma_{\mathbf{X}}||\Sigma_{\mathbf{Y}}|\prod_{i=1}^d (1-|\rho_i|)^2 \\ \sum_{i=1}^d R_i & \textit{otherwise} \end{cases} \tag{6.28}$$

where

$$R_i = \max(C_W(\tilde{X}_i, \tilde{Y}_i) - \theta, 0), \tag{6.29}$$

and θ *is a positive constant chosen such that*

$$|\Sigma_{\mathbf{X}}||\Sigma_{\mathbf{Y}}|\prod_{i=1}^d \left(\frac{1}{2}\left(2^{-2R_i}(1+|\rho_i|) + (1-|\rho_i|)\right)\right)^2 = \gamma, \tag{6.30}$$

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

Proof. For γ small such that $T_{\mathbf{X},\mathbf{Y}}(\gamma) \ge C_W(\mathbf{X},\mathbf{Y})$, conditional independence and equality in (6.25) (and, as a matter of fact, also (6.24)) is attainable, see Lemma 6.2. Hence the following chain is met with equality:

$$\frac{1}{2}\log\frac{|\Sigma|}{|\mathbf{K}|} \ge \frac{1}{2}\log\frac{|\Sigma|}{|\mathbf{K}_{\mathbf{X}}||\mathbf{K}_{\mathbf{Y}}|} \ge \frac{1}{2}\log\frac{|\Sigma|}{\gamma}. \tag{6.31}$$

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

$$\max_{\mathbf{0}\preceq\mathbf{K}\preceq\Sigma} |\mathbf{K}| \quad \text{s.t.} \quad |\mathbf{K}_{\mathbf{X}}||\mathbf{K}_{\mathbf{Y}}| \le \gamma. \tag{6.32}$$

Without loss of generality, one can do a change of variable by applying the vectors-to-pairs decomposition of Section 6.1.3 to the source covariance Σ and the *same* transformation to the variable $\mathbf{K}$. Since this decomposition scales out the variance, which are *not* orthonormal matrices, the objective and constraint are affected:

$$\max_{\mathbf{0}\preceq\tilde{\mathbf{K}}\preceq\tilde{\Sigma}} |\Sigma_{\mathbf{X}}||\Sigma_{\mathbf{Y}}||\tilde{\mathbf{K}}| \quad \text{s.t.} \quad |\Sigma_{\mathbf{X}}||\Sigma_{\mathbf{Y}}||\tilde{\mathbf{K}}_{\mathbf{X}}||\tilde{\mathbf{K}}_{\mathbf{Y}}| \le \gamma.$$

Both are, however, affected *equally* and we can restrict our attention to finding a suitable $\tilde{\mathbf{K}} \preceq \tilde{\Sigma}$ such that $|\tilde{\mathbf{K}}_{\mathbf{X}}||\tilde{\mathbf{K}}_{\mathbf{Y}}| \le \frac{\gamma}{|\Sigma_{\mathbf{X}}||\Sigma_{\mathbf{Y}}|}$.

Note the special structure of $\tilde{\Sigma}$ (6.14): there only exists correlation between $(\tilde{X}_i, \tilde{Y}_i)$-pairs. It is essentially a permuted block-diagonal matrix of 2 × 2 matrices that relate these pairs:

$$\tilde{\Sigma}^{(i)} = \begin{bmatrix} 1 & \rho_i \\ \rho_i & 1 \end{bmatrix}, \tag{6.33}$$

where $\{\rho_i\}$ are the singular values of $\Sigma_{\mathbf{X}}^{-1/2}\Sigma_{\mathbf{XY}}\Sigma_{\mathbf{Y}}^{-1/2}$ found in the top-right and bottom-left corners of $\tilde{\Sigma}$ (6.14). In this notation $\tilde{\mathbf{K}}^{(i)}$ and $\tilde{\Sigma}^{(i)}$ are the principal submatrices of $\tilde{\mathbf{K}}, \tilde{\Sigma}$ by keeping the i'th and $(i+d)$'th row and column.

Our hypothesis is that without loss of optimality $\tilde{\mathbf{K}}$ has the same eigenbasis that generates a block-matrix with diagonal matrices in all its four corners. Consider the following relaxation (in which for brevity we remove the constants from the objective function):

$$\max_{\tilde{\mathbf{K}}} |\tilde{\mathbf{K}}| \text{ s.t. } \begin{cases} \mathbf{0} \preceq \tilde{\mathbf{K}} \preceq \tilde{\Sigma} \\ |\tilde{\mathbf{K}}_{\mathbf{X}}||\tilde{\mathbf{K}}_{\mathbf{Y}}| \leq \frac{\gamma}{|\Sigma_{\mathbf{X}}||\Sigma_{\mathbf{Y}}|} \end{cases} \leq \max_{\tilde{\mathbf{K}}} \prod_{i=1}^{d} |\tilde{\mathbf{K}}^{(i)}| \text{ s.t. } \begin{cases} \mathbf{0} \preceq \tilde{\mathbf{K}}^{(i)} \preceq \tilde{\Sigma}^{(i)}, \quad i = 1, \cdots, d \\ \mathbf{0} \preceq \tilde{\mathbf{K}}_{\mathbf{X}} \preceq \tilde{\Sigma}_{\mathbf{X}} = \mathbf{I} \\ \mathbf{0} \preceq \tilde{\mathbf{K}}_{\mathbf{Y}} \preceq \tilde{\Sigma}_{\mathbf{Y}} = \mathbf{I} \\ |\tilde{\mathbf{K}}_{\mathbf{X}}||\tilde{\mathbf{K}}_{\mathbf{Y}}| \leq \frac{\gamma}{|\Sigma_{\mathbf{X}}||\Sigma_{\mathbf{Y}}|} \end{cases} \tag{6.34}$$

we simultaneously upper-bounded the objective function by the block-equivalent of the Hadamard inequality, as well as relaxed the semidefinite ordering constraint to only consider a subset of principal submatrices.

Now, a candidate $\tilde{\mathbf{K}}$ does not need to feature correlation between $(\tilde{X}_i, \tilde{Y}_j)_{i \neq j}$-pair, because it neither affects the objective, nor the constraint. Furthermore, observe that without loss of optimality one can assume $\tilde{\mathbf{K}}_{\mathbf{X}}, \tilde{\mathbf{K}}_{\mathbf{Y}}$ to be diagonal, because it is *the least restrictive*. Namely, the last constraint is unaffected by any basis rotation, while $\tilde{\mathbf{K}}_{\mathbf{X}} \preceq \tilde{\Sigma}_{\mathbf{X}}$ (and the one for $\mathbf{Y}$) would drop out as it would be implicitly included in $\tilde{\mathbf{K}}^{(i)} \preceq \tilde{\Sigma}^{(i)}$.

Hence, there is no added value in a $\tilde{\mathbf{K}}$ with a different eigenbasis than $\tilde{\Sigma}$ and one arrives at the modular approach of looking for a 2 × 2 distortion matrix $\tilde{\mathbf{K}}^{(i)}$ for each $(\tilde{X}_i, \tilde{Y}_i)$-pair. For a pair of Gaussians, we derived in the uniform caching problem of Section 4.3.3 that $\tilde{\mathbf{K}}^{(i)}$ should be a rank-one correction along the dominant eigenvector of the correlation matrix of $(\tilde{X}_i, \tilde{Y}_i)$. That leaves the question of which $(\tilde{X}_i, \tilde{Y}_i)-$pairs have the biggest impact on minimizing the *total* conditional correlation of the vectors $\mathbf{X}$ and $\mathbf{Y}$. This distribution problem turns out to be convex, but it requires the proper variable to expose so. To that end, let

$$R_i \triangleq I(\tilde{X}_i, \tilde{Y}_i; V_i) = \frac{1}{2} \log \frac{1-\rho_i^2}{|\tilde{\mathbf{K}}^{(i)}|}. \tag{6.35}$$

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

$$|\tilde{\mathbf{K}}_{\mathbf{X}}||\tilde{\mathbf{K}}_{\mathbf{Y}}| = \prod_{i=1}^{d} (\tilde{K}_{X_i}^{(i)} \tilde{K}_{Y_i}^{(i)}) = \prod_{i=1}^{d} \left(\frac{1}{2} \left(2^{-2R_i}(1+|\rho_i|) + (1-|\rho_i|) \right) \right)^2. \tag{6.36}$$

Applying logarithms (in base 2), also the constraint-function is convex and we use Lagrangian

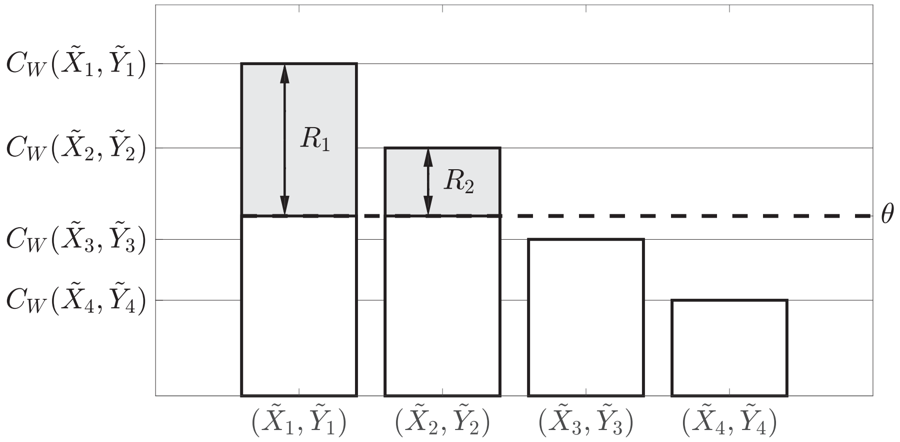


Figure 6.3 – Example of the reverse water filling procedure of Theorem 6.2. The bars represent the common information of each $(\tilde{X}_i, \tilde{Y}_i)$-pair and the shaded area equals R_i. In this example, $d = 4$, $\rho_i \in \{0.9, 0.8, 0.6, 0.4\}$ (f.l.t.r.) and θ is such that $R_3 = R_4 = 0$.

multipliers to construct the following expressions of which we set the derivative to zero:

$$J = \sum_{i=1}^{d} R_i + \lambda \sum_{i=1}^{d} \log\left(\frac{1}{4}\left(2^{-2R_i}(1+|\rho_i|) + (1-|\rho_i|)\right)^2\right), \tag{6.37}$$

followed by

$$\frac{\partial J}{\partial R_i} = 1 - 4\lambda \frac{(1+|\rho_i|)2^{-2R_i}}{2^{-2R_i}(1+|\rho_i|) + (1-|\rho_i|)} = 0. \tag{6.38}$$

Rewriting the above expression then leads to

$$\begin{aligned} R_i &= \frac{1}{2}\log\frac{1+|\rho_i|}{1-|\rho_i|} - \frac{1}{2}\log\left(\frac{1}{4\lambda - 1}\right) \\ &= C_W(\tilde{X}_i, \tilde{Y}_i) - \theta. \end{aligned} \tag{6.39}$$

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

The rate–distortion function of a Gaussian vector $\mathbf{X}$ subject to a trace distortion constraint, i.e. $\operatorname{tr}\left(\mathbb{E}[(\mathbf{X}-\hat{\mathbf{X}})(\mathbf{X}-\hat{\mathbf{X}})^T]\right) \le D$, is a classic result that also admits a reverse water-filling procedure (3.14). We attend the reader to a subtle difference: the Gaussian vector rate-distortion function applies reverse water filling to the *eigenvalues* of the covariance matrix Σ, whereas the minimization of total conditional correlation uses $R_i = I(\tilde{X}_i, \tilde{Y}_i; V_i)$ as the variable. Consequently, one will not observe similar thresholding behavior by plotting the evolution of the eigenvalues. The right way to plot the water-filling of total conditional correlation is by plotting the common information of each $(\tilde{X}_i, \tilde{Y}_i)$-pair as a bar graph. An example is shown in Figure 6.3.

6.4 The Caching of Two Gaussian Databases

This section is to serve as an example of how the essentially lossless definition of $T_{\mathbf{X},\mathbf{Y}}(\gamma)$ applies to our lossy caching problem on the Gray–Wyner network. Let us extend the same model of Section 4.3 in which a user chooses **X** or **Y** equally likely and base it on the vector-notated version of the Gray–Wyner network as paraphrased in Section 6.1.2. In doing so, one finds:

$$\begin{aligned} R_{\text{cache}} &\geq I(\mathbf{X},\mathbf{Y};\mathbf{V}), \qquad (6.40)\\ \overline{R}_{\text{update}} &\geq \frac{1}{2}\left(R_{\mathbf{X}|\mathbf{V}}(D_x)+R_{\mathbf{Y}|\mathbf{V}}(D_y)\right). \end{aligned}$$

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

$$\begin{aligned} R_{\text{cache}} &\geq \frac{1}{2}\log\frac{|\Sigma|}{|\mathbf{K}|} \qquad (6.41)\\ \overline{R}_{\text{update}} &\geq \frac{1}{4}\log\frac{|\mathbf{K}_{\mathbf{X}}||\mathbf{K}_{\mathbf{Y}}|}{|\mathbf{D}_{\mathbf{X}}||\mathbf{D}_{\mathbf{Y}}|} \qquad (6.42)\\ D_x &\geq \operatorname{tr}(\mathbf{D}_{\mathbf{X}})\\ D_y &\geq \operatorname{tr}(\mathbf{D}_{\mathbf{Y}}), \end{aligned}$$

over positive semidefinite matrices $\mathbf{D}_{\mathbf{X}}, \mathbf{D}_{\mathbf{Y}} \in \mathbb{R}^{d\times d}$ and $\mathbf{K} \in \mathbb{R}^{2d\times 2d}$ satisfying $\mathbf{D}_{\mathbf{X}} \preceq \mathbf{K}_{\mathbf{X}}, \mathbf{D}_{\mathbf{Y}} \preceq \mathbf{K}_{\mathbf{Y}}$, and $\mathbf{K} \preceq \Sigma$.

Observe that the interaction between R_{cache} and $\overline{R}_{\text{update}}$ features the trade-off we studied in $T_{\mathbf{X},\mathbf{Y}}(\gamma) : |\mathbf{K}| \leftrightarrow |\mathbf{K}_{\mathbf{X}}||\mathbf{K}_{\mathbf{Y}}|$. However, the end distortion D_x, D_y also influence what choice of **K** provides the most efficient $(R_{\text{cache}}, \overline{R}_{\text{update}})$ trade-off. The lossless concept of $T_{\mathbf{X},\mathbf{Y}}(\gamma)$ is therefore not *directly* applicable to this coding problem on the Gray–Wyner network, but it is under the following conditions:

Corollary 6.1. *Let $\mathbf{K}^{C_W}$ be the distortion matrix that attains the common information $C_W(\mathbf{X},\mathbf{Y})$. Then, the Gaussian trade-off between R_{cache}* (6.41) *and $\overline{R}_{update}$* (6.42) *can be controlled by a parameter γ such that*

$$\begin{aligned} R_{cache} &\geq T_{\mathbf{X},\mathbf{Y}}(\gamma), \qquad (6.43)\\ \overline{R}_{update} &\geq \frac{1}{4}\log\frac{\gamma}{|\mathbf{D}_{\mathbf{X}}||\mathbf{D}_{\mathbf{Y}}|}, \end{aligned}$$

for the regime of end distortion constraints satisfying:

$$\begin{aligned} D_x &\leq d\cdot\lambda_{min}(\mathbf{K}_{\mathbf{X}}^{C_W}), \qquad (6.44)\\ D_y &\leq d\cdot\lambda_{min}(\mathbf{K}_{\mathbf{Y}}^{C_W}). \end{aligned}$$

Proof. The rate-distortion theorem for Gaussian multivariates under a trace-constraint dictates the update phase is most efficiently coded via a reverse water filling procedure on the

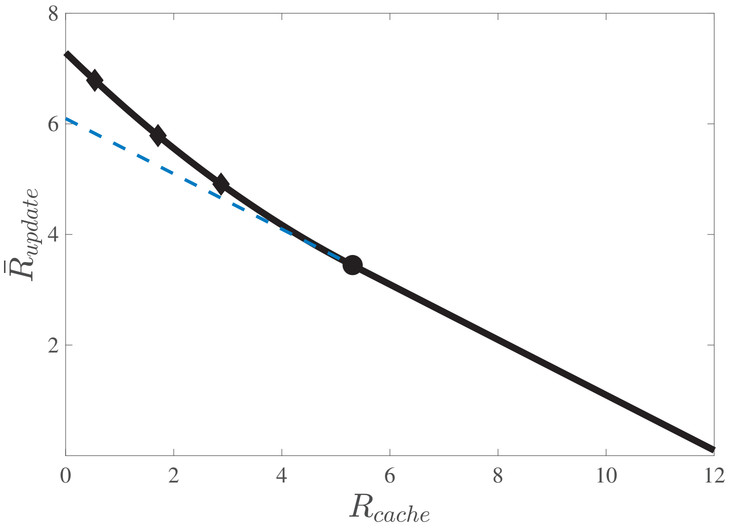


Figure 6.4 – Example of the caching trade-off with $d = 4$ and $\rho_i \in \{0.9, 0.8, 0.6, 0.4\}$. The diamonds correspond f.l.t.r. to the points were respectively R_1, R_2 and R_3 become positive, following the waterfilling of Theorem 6.2. The circle corresponds to $R_{\text{cache}} = C_W(\mathbf{X}, \mathbf{Y})$.

eigenvalues of $\mathbf{K_X}$ and $\mathbf{K_Y}$ (3.14). Hence, for large D_x, D_y the optimal choice of a $\mathbf{K}$ not only depends on the determinants, but also the *specific spectra* of the submatrices in its top-left and bottom-right corner, $\mathbf{K_X}$ and $\mathbf{K_Y}$. If $D_x \le d \cdot \lambda_{\min}(\mathbf{K_X})$, the distortion matrix $\mathbf{D_X}$ that minimizes the update rate does not depend on the spectrum of $\mathbf{K_X}$, but equals $\mathbf{D_X} = (\frac{D_x}{d}) \cdot \mathbf{I} \preceq \mathbf{K_X}$.

If indeed $D_x \le d \cdot \lambda_{\min}(\mathbf{K}_{\mathbf{X}}^{C_W})$ then the choice of $\mathbf{K}$ and $\mathbf{D_X}, \mathbf{D_Y}$ that minimize R_{cache} and $\overline{R}_{\text{update}}$ decouple. Namely, in the regime $R_{\text{cache}} \in [0,\ C_W(\mathbf{X}, \mathbf{Y})]$ the trade-off $|\mathbf{K}| \leftrightarrow |\mathbf{K_X}||\mathbf{K_Y}|$ solved by $T_{\mathbf{X},\mathbf{Y}}(\gamma)$ produces a $\mathbf{K}'$ that satisfies $\mathbf{K}' \succeq \mathbf{K}^{C_W}$, a consequence of Theorem 6.2. Therefore, the optimal choice of $\mathbf{D_X}$ remains $\mathbf{D_X} = (\frac{D_x}{d}) \cdot \mathbf{I}$. The same for $\mathbf{Y}$. In the regime $R_{\text{cache}} \ge C_W(\mathbf{X}, \mathbf{Y})$, $\mathbf{X}$ and $\mathbf{Y}$ can become conditionally independent and the trade-off between cache and update rate comes without rate loss. □

An example of this trade-off is plotted in Figure 6.4 for the same example as used in Figure 6.3. The diamonds mark the points where the water-filling procedure hits a new threshold and starts including another $(\tilde{X}_i, \tilde{Y}_i)$–pair into the coding process. Once $R_{\text{cache}} \ge C_W(\mathbf{X}, \mathbf{Y})$ the trade-off between R_{cache} and $\overline{R}_{\text{update}}$ coincides with the straight line connecting the points of $\left(R_{\text{cache}}, \overline{R}_{\text{update}}\right) = \left(R_{\mathbf{X},\mathbf{Y}}(D_x, D_y), 0\right)$ and $\left(0, \frac{1}{2} R_{\mathbf{X},\mathbf{Y}}(D_x, D_y)\right)$.

7 Practical Caching of Discrete Sources with Convolutional Codes

As a final chapter, we briefly touch upon the ancient question one gets when facing relatives: “So as of when will we actually use such algorithms?”. Random coding limits and coding theorems for continuous information sources in particular may appear to those relatives as distant from the real world. The concepts of all these theoretic ideas of caching, however, may not have to be so. This chapter serves as a proof-of-concept: caching can be practical.

As quite an abrupt change, we shift the discussion to the world of discrete information. The discrete Gray–Wyner-based caching model was the topic of, as referenced before, Wang, Lim and Gastpar [3]. The concepts of Wyner’s common information and total conditional correlation are also the drivers of this lossless problem. Among general coding results, the authors also derived precise constructions on how to cache a doubly symmetric binary source. This we seek to actually build as an algorithm with practical block lengths and run times.

We argue in this chapter that the caching of two files generated by a doubly symmetric binary source is a natural fit for convolution codes. Their common information can be captured in a single bit per every pair of bits (from the two files), carrying the majority symbol followed by compression. We explain in this chapter how this leads to a symmetric model on which linear codes perform well in taking care of this compression part. In addition, the tracking of the majority symbol lends itself quite naturally for trellis decoding in particular, motivating our choice for convolution codes.

The barrier ahead, however, is that linear codes provide a practical tool for compression, but one would first need to know *what* to compress. This is the biggest challenge of scaling practical caching to databases of more than two files, like one would probably encounter in real applications. We close this chapter by a discussion on a general class of *circularly symmetric binary sources*, whose common information is also a single bit regardless of how many sources are included. Specifically we study a trio of files and implement its caching with convolutional codes as well.

7.1 Problem Statement and Recap of Theory

We consider the same caching model as Figure 4.1 and 5.1 subject to uniform request probabilities for all files. Two changes are made:

1. All alphabets are discrete.
2. The user requires a *lossless* copy of the samples produced by X_k^N.

This first section is a quick recap of the work by Wang, Lim and Gastpar [3].

7.1.1 Notation and General Model

Let $\mathbf{X}$ be a length-K discrete memoryless source taking values on alphabet $\mathcal{X}$, which produces a length-N sequence of samples. Similarly to our earlier discussed models, the encoder encodes a cache message $I_{M_c} \in [1, \cdots, 2^{NR_{\text{cache}}}]$ which the decoder receives in any case. Later, the user announces a request for X_k^N, after which the encoder sends an update message $I_{M_{u,k}} \in [1, \cdots, 2^{NR_{u,k}}]$ tailored towards that choice. Whereras [3] considers several models for the user's request and the reconstruction requirement we restrict out attention to the user requesting each X_k^N equally likely, in its entirety and up to perfect lossless precision. Consequently, after the cache phase the encoder still spends the following:

$$\overline{R}_{\text{update}} \triangleq \frac{1}{K} \sum_{k=1}^{K} R_{u,k}. \tag{7.1}$$

Theorem 7.1 (Wang, Lim and Gastpar [3])**.** *Caching is achievable for* $(R_{cache}, \overline{R}_{update})$ *satisfying*

$$R_{cache} \geq I(\mathbf{X};\mathbf{V}) \tag{7.2}$$

$$\overline{R}_{update} \geq \frac{1}{K} \sum_{k=1}^{K} H(X_k|\mathbf{V}), \tag{7.3}$$

for a conditional pmf $p(\mathbf{v}|\mathbf{x})$ *where* $|\mathcal{V}| \leq |\mathcal{X}| + 1$.

Note that again the notion of total conditional correlation $TC(\mathbf{X}|\mathbf{V}) = \sum_{k=1}^{K} H(X_k|\mathbf{V}) - H(\mathbf{X}|\mathbf{V})$ characterizes the move from cache to update phase, even more explicitly than in the Gaussian case. This leads to an equivalent characterization of achievable $(R_{\text{cache}}, \overline{R}_{\text{update}})$:

Corollary 7.1 (Wang, Lim and Gastpar [3])**.** *The boundary of the achievable rate region can be described by a parameter* $r \in [0, \ H(\mathbf{X})]$*:*

$$R_{cache} = r \tag{7.4}$$

$$\overline{R}_{update} = \frac{1}{K}\left(H(\mathbf{X}) - r + \psi(r)\right), \tag{7.5}$$

where

$$\psi(r) = \min_{p(\mathbf{v}|\mathbf{x})} TC(\mathbf{X}|\mathbf{V}) \qquad s.t. \ I(\mathbf{X};\mathbf{V}) = r. \tag{7.6}$$

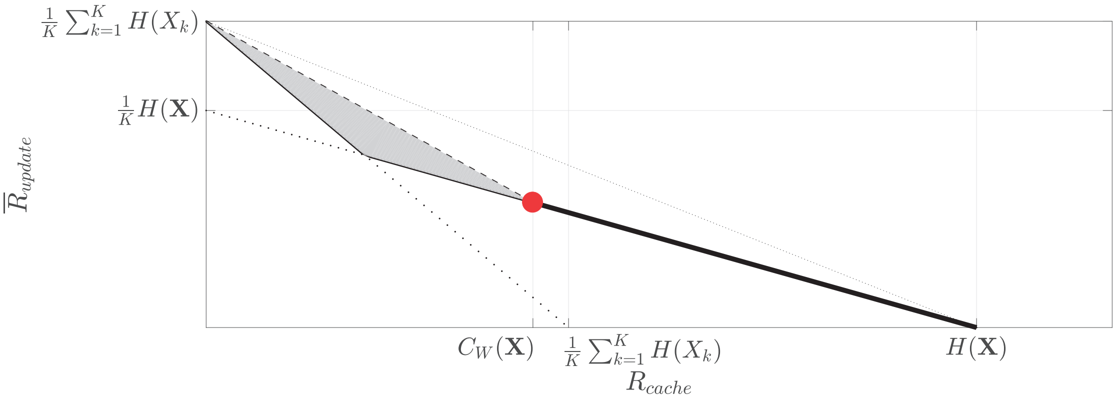


Figure 7.1 – The boundary of achievable $(R_{\text{cache}}, \overline{R}_{\text{update}})$−pairs is the thick black line right of $C_W(\mathbf{X})$, and an unknown convex curve inside the gray triangle related to solving $\psi(r)$ (7.6).

In this notation $\psi(r)$ minimizes total conditional correlation by fixing R_{cache} and minimizing $\overline{R}_{\text{update}}$, as opposed to the other way around like we did in Chapters 4 to 6.

When $\psi(r) \geq 0$ is met with equality then the system moves from the cache into the update phase without any loss of rate. This is attainable when $R_{\text{cache}} \geq C_W(\mathbf{X})$ where $C_W(\mathbf{X})$ is the (extended) Wyner's common information [3, 29, 30]:

$$C_W(\mathbf{X}) \triangleq \min_{p(\mathbf{v}|\mathbf{x})} I(\mathbf{X};\mathbf{V}) \qquad \text{s.t.} \;\; TC(\mathbf{X}|\mathbf{V}) = 0. \tag{7.7}$$

Therefore, for $R_{\text{cache}} \geq C_W(\mathbf{X})$ one has $\psi(r) = 0$ and the $(R_{\text{cache}}, \overline{R}_{\text{update}})$ trade-off becomes a straight line, regardless of K or the distribution of $\mathbf{X}$:

$$\begin{cases} R_{\text{cache}} & \geq r \\ \overline{R}_{\text{update}} & \geq \frac{1}{K}(H(\mathbf{X}) - r). \end{cases} \tag{7.8}$$

All together this constitutes Figure 7.1. For details on the plotted outer bounds we refer to [3].

7.1.2 A DSBS-pair of Information Sources

Consider specifically the case of two doubly symmetric binary memoryless source sequences.The remark of 'doubly symmetric' refers to (X_1, X_2) following the distribution for some $q \in [0, \frac{1}{2}]$:

$$p(x_1, x_2) = \begin{bmatrix} \frac{1-q}{2} & \frac{q}{2} \\ \frac{q}{2} & \frac{1-q}{2} \end{bmatrix}. \tag{7.9}$$

The common information for the DSBS source was already derived by Wyner in his original

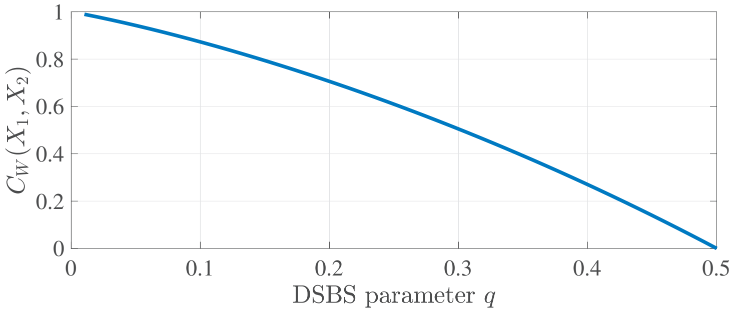


Figure 7.2 – $C_W(X_1,X_2)$ of a doubly symmetric binary source as a function of q (7.9).

paper [8]:

$$C_W(X_1,X_2) = \min_{\substack{p(v|x_1,x_2)\\:X_1-V-X_2}} = 1 + h_2(q) - 2h_2\left(\frac{1}{2}(1-\sqrt{1-2q})\right), \tag{7.10}$$

where $h_2(\cdot)$ is the binary entropy:

$$h_2(p) \triangleq p\log\frac{1}{p} + (1-p)\log\frac{1}{1-p}. \tag{7.11}$$

Figure 7.2 characterizes $C_W(X_1,X_2)$ as a function of q.

For $R_{\text{cache}} < C_W(X_1,X_2)$, Wang, Lim and Gastpar conjecture the following choice of V is optimal [3]:

$$V = \begin{cases} X_1 \oplus U & \text{if } X_1 \oplus X_2 = 0, \\ W & \text{if } X_1 \oplus X_2 \neq 0. \end{cases} \tag{7.12}$$

U, W are binary and independent of (X_1,X_2). $V \sim \text{Bern}(\frac{1}{2})$ and $p_U(1) \in [\frac{1}{2} - \frac{\sqrt{1-2q}}{2(1-q)}, \frac{1}{2}]$ can be controlled to cover different values of R_{cache}. Setting specifically $p_U(1) = \frac{1}{2} - \frac{\sqrt{1-2q}}{2(1-q)}$ results in this construction attaining Wyner's common information. Formally, this setup for V is a conjecture, but numerical search over all distribution for V confirms this is indeed optimal. Such a search is feasible as it must hold that $|\mathcal{V}| \leq |\mathcal{X}| + 1$ by Theorem 7.1.

Note that $H(X_i|V) = H(X_i \oplus V)$, which simplifies any practical implementation: whereas the auxiliary random variable V is used for the cache phase, the update encoder can simply compute and compress $X_k^N \oplus V^N$, where $k \in \{1,2\}$ marks the desired file.

7.2 Convolutional Codes: A Natural Fit

7.2.1 Motivation for Convolutional Codes

The construction of V (7.12) is such that random coding is not necessary; linear codes should do well on this problem. Namely, the IID and symmetric natures makes joint typicality tests simple: for the cache, select a codeword that agrees as much as possible with the majority symbol of X_1 and X_2 on all positions and ignore those where $X_1 \neq X_2$. In addition, linear codes are sufficient to generate the codebook, since $p_V(0) = p_V(1) = \frac{1}{2}$.

Moreover, the construction of V (7.12) closely resembles to what in the literature on linear codes for compression is known as *binary erasure quantization*: a symbol is either zero, one or it is erased. In our caching setting X_1 and X_2 are either both zero or one, or they are different in which case V flips a coin. Martinian and Yedidia designed for this application Low Density Generator Matrix (LDGM) codes, the dual of LDPC [37]. Later also binary symmetric sources were considered [38] and more elaborate LDGM compound codes [39, 40].

Nonetheless, a prevalent barrier LDGM suffers from in terms of practicality is that belief propagation, which contributed to the success of LDPC, does not translate well to source coding. The issue is that to-be-encoded sequences are spread uniformly and not necessarily close to codewords, like what would be realistic in channel coding. Message parsing relies on a reasonable initial state in order to converge. An excellent review on the implementation challenges of LDGM can be read in the PhD thesis of Korada [41, Section 1.2.2].

Convolutional codes, on the other hand, fit caching like a glove. First, *convergence* of trellis decoding/encoding does not depend on the distance of to-be-encoded sequences from possible codewords. Second, the Viterbi algorithm can be applied to source coding as it would for channel coding without any changes [42, 43]. Namely, encoding a sequence can be done by 'pretending' the source samples came from a channel output. Source coding is in that sense even simpler than channel coding, because the input to a Viterbi algorithm can be discrete levels instead of a soft input like -for example- AWGN would generate.

For caching in particular, encoding V can happen by mimicking a BPSK channel: if $X_1 = X_2$ map their symbol to a fixed positive or negative value, if $X_1 \neq X_2$ input the Viterbi algorithm with 0. The rate of the convolutional code used then dictates the level of compression. Despite these advantages in implementation, though, one must remark that the *computational* complexity of the Viterbi algorithm scales exponentially in trellis length.

7.2.2 The Empirical Caching Model

The system we build follows the structure of Figure 7.3 and uses the CML package (in C) for MATLAB to do large scale simulations of the Viterbi algorithm [44].

Cache Encoder

For the cache encoder side, the two source sequences X_1^N and X_2^N pass through three stages to empirically extract and compress the common information:

$$\begin{aligned} f_{\text{majority}}: &\quad \{0,1\}^N \times \{0,1\}^N \to \{0,?,1\}^N, &(7.13)\\ f_{\text{channel}}: &\quad \{0,?,1\}^N \to \{-1,0,+1\}^N, &(7.14)\\ f_{\text{Viterbi}}: &\quad \{-1,0,+1\}^N \to \{0,1\}^N. &(7.15)\end{aligned}$$

The first stage extracts the majority bit similar to the theoretical V (7.12):

$$f_{\text{majority}}(X_1, X_2) = \begin{cases} X_1 & \text{if } X_1 \oplus X_2 = 0, \\ ? & \text{if } X_1 \oplus X_2 \neq 0. \end{cases} \tag{7.16}$$

The second prepares the sequence to an input suitable for the Viterbi algorithm by mapping the majority-symbol sequence to 'log-likelihood ratios'. In channel coding, this LLR of a symbol gives the decoder an indication of how strongly that symbol is believed to be a zero or one. Our compression setting, though, requires only hard/discrete beliefs as opposed to soft information (as if it were BPSK before passing through a noisy channel if you will): Namely, $p_{X_1,X_2}(0,0) = p_{X_1,X_2}(1,1)$, in which case V should match that symbol up to a desired level of compression (7.12). The symmetry of the probability of this happening indicates that in these cases the input to the Viterbi encoder should be symmetric for the $0-$ or $1-$majority. If $X_1 \neq X_2$, a coin flip suffices. This leads to the following mapping:

$$f_{\text{channel}}(\cdot): \begin{cases} 0 \to -1, \\ ? \to 0, \\ 1 \to +1. \end{cases} \tag{7.17}$$

The values ±1 are arbitrary; the symmetry is what counts. Mapping a ? to a 0 indicates to the Viterbi-compressor that the symbol is a 'don't care', which by construction and sufficiently large block length will lead to the desired $\sim \text{Bern}(\frac{1}{2})$ distribution for these $X_1 \neq X_2$ conflicts.

In the last stage, this ternary sequence is pulled through the Viterbi algorithm to compress it to a binary codeword. We denote this cache-codeword by V_C^N. This is in principle a length-N sequence, but can be communicated more compactly by the index of the codeword in the codebook.

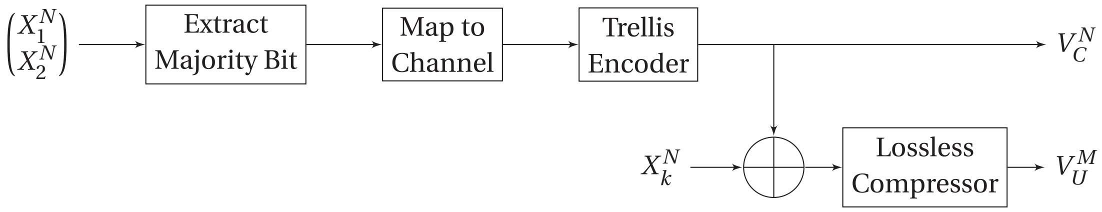


Figure 7.3 – The schematic of cache and update encoders in the case the user requests the samples of X_k^N for a $k \in \{1,2\}$.

An example of these steps is the following chain:

$$\left.\begin{array}{l} X_1^N = 01011011\cdots \\ X_2^N = 01110011\cdots \end{array}\right\} \xrightarrow{f_{\text{majority}}} 01?1?011\cdots \xrightarrow{f_{\text{channel}}} \begin{array}{cccccccc} & +1 & & +1 & & & +1 & +1 \\ & & 0 & & 0 & & & \\ -1 & & & & & -1 & & \end{array} \cdots \xrightarrow{f_{\text{Viterbi}}} V_C^N \tag{7.18}$$

Update Encoder

The update phase consists of two steps:

1. Identify the discrepancy between the cache V_C^N and the desired file X_k^N by computing

$$\Delta_k^N = V_C^N \oplus X_k^N, \tag{7.19}$$

   where $k \in \{1,2\}$.

2. Compress Δ_k^N using any lossless universal compression algorithm[1].

The above produces an update codeword V_U^M, where $M \leq N$.

Decoder: the User End

The user is provided with the cache codeword V_C^N and the one for the update V_U^M. Subsequently, the user first decompresses V_U^M back to the length-N sequence Δ_k^N by the inverse steps of the universal compression algorithm used. Afterwards, the user retrieves the k'th file losslessly by computing

$$X_k^N = V_C^N \oplus \Delta_k^N. \tag{7.20}$$

[1] Note the subtlety here: whereas the caching phase is a lossy encoding, the update phase is lossless.

7.3 Experimental Results

Simulations are provided in Figure 7.4–7.6. Plotted in black is the ensemble of (7.8) and the cache/update rates resulting from a construction based on (7.12). The two sections are separated by the red dot indicating $R_{\text{cache}} = C_W(X_1, X_2)$.

The blue triangles correspond to the empirical caching model described in the previous section. R_{cache} is the rate of the convolutional codes used. $\overline{R}_{\text{update}}$ is computed as the average conditional entropy of Δ_1^N and Δ_2^N, based on their empirical distribution:

$$\bar{p}_{\Delta_k}(1) = \frac{1}{N}\sum_{i=1}^{N} \Delta_k^N(i) \qquad k = \{1,2\}. \tag{7.21}$$

This results in:

$$R_U = \frac{1}{2}\left(h_2(\bar{p}_{\Delta_1}(1)) + h_2(\bar{p}_{\Delta_2}(1))\right). \tag{7.22}$$

Block length is $N = 30,000$. Convolutional codes used have rates equal to (f.l.t.r.) $\frac{1}{5}, \frac{1}{4}, \frac{1}{3}$ and $\frac{1}{2}$.

In these Figures, we plot different instances for q (7.9). Note in particular $q = 0.3$. Under this condition we have $C_W(X_1, X_2) \approx 0.5048$; the convolution code of rate $\frac{1}{2}$ comes exceptionally close to achieving Wyner's common information. In general, the experiments do not fully attain the theoretic performance (for this N), but approach it closely.

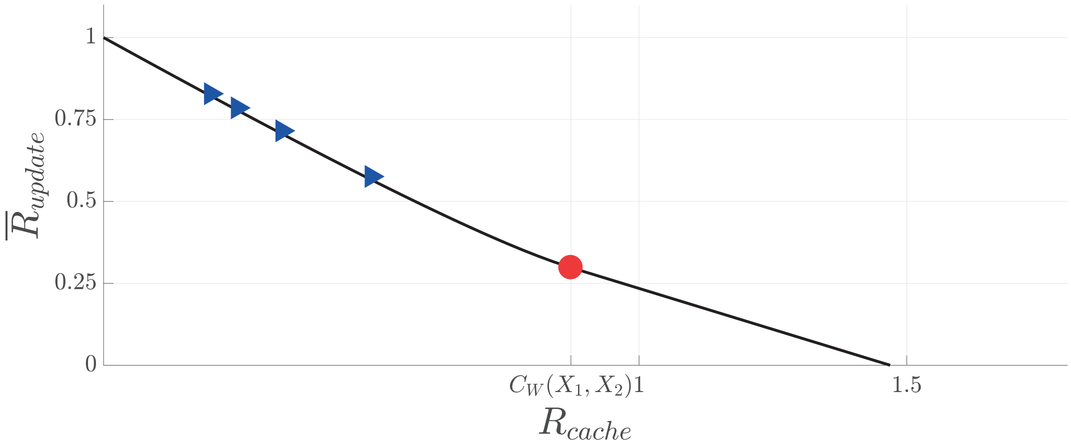


Figure 7.4 – Caching a DSBS-duo with $q = 0.1$.

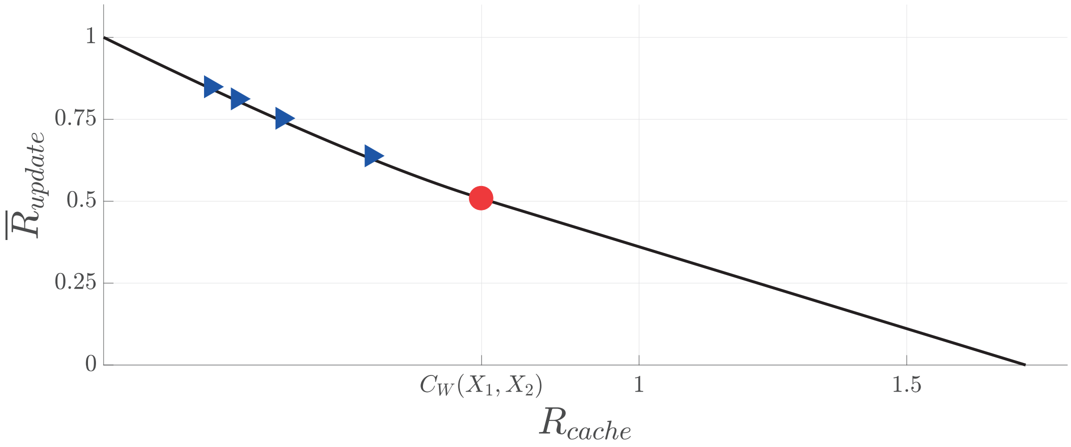


Figure 7.5 – Caching a DSBS-duo with $q = 0.2$.

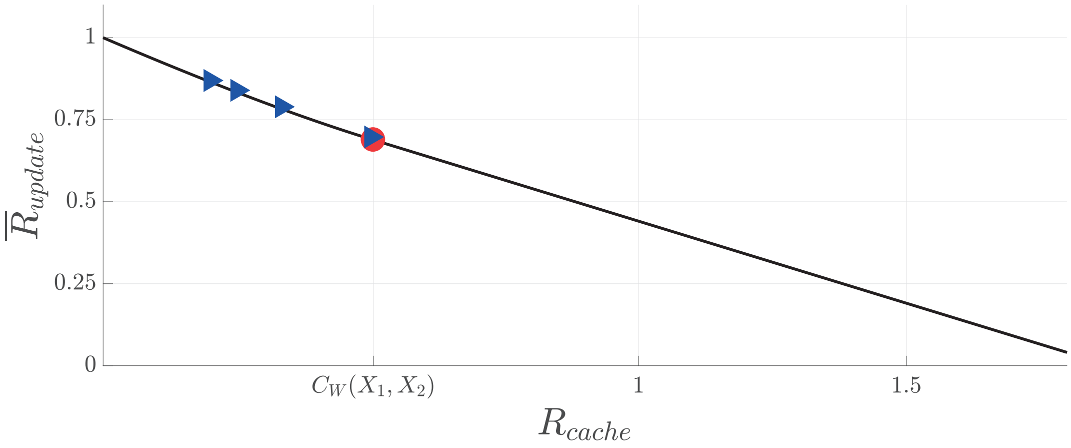


Figure 7.6 – Caching a DSBS-duo with $q = 0.3$.

7.4 Barriers Ahead: Total Correlation beyond Two Files

Even though the convolutional caching of two binary symmetric files has proven to be feasible, this does not yet pave the way for larger, more complex caching systems involving $K > 2$ files. The caching phase of Figure 7.3 consists of two distinct parts: empirically extracting a perfect common sequence of two files, followed by compression. Linear codes like convolutional codes can easily be implemented in bigger settings to take care of the second task, but what about the first?

The model of Figure 7.3 bases itself on a theoretical understanding of the notion of total conditional correlation and common information through V (7.12): for a DSBS, both of these measures are the majority symbol of X_1 and X_2, corrupted by more or less Bernoulli noise. For three files or more is the common information still captured in a single binary random variable, or should V have a higher cardinality? If the correlation structure of several random variables is asymmetric are the symbols of each file then equally important in determining the common information or do some contribute more?

The theory of total conditional correlation and common information beyond $K > 2$ is scarce, which indicates the difficulty of the matter. The challenge of practical caching algorithms therefore does not lie in *how* to cache, but rather *what* to cache. Currently it is not just unclear what would be optimal, but even simply what would be smart.

7.4.1 (Three) Circularly Symmetric Binary Sources

A special case might be random variables whose common information is known to be one bit, as first discussed by Liu, Xu and Chen [29]. Their approach was one of reverse engineering: create a common binary random variable V and construct several X_i by processing V through independent but identical bit-flip channels; they call these X_i *circularly symmetric* binary sources (CSBS).

Consider $K = 3$, a CSBS-triplet is defined by the following pmf for some $q \in [0, \frac{1}{2}]$:

$$p(x_1, x_2, x_3) = \begin{cases} \frac{1}{2} - \frac{3}{4}q & \text{if } x_1 = x_2 = x_3, \\ \frac{1}{4}q & \text{otherwise.} \end{cases} \tag{7.23}$$

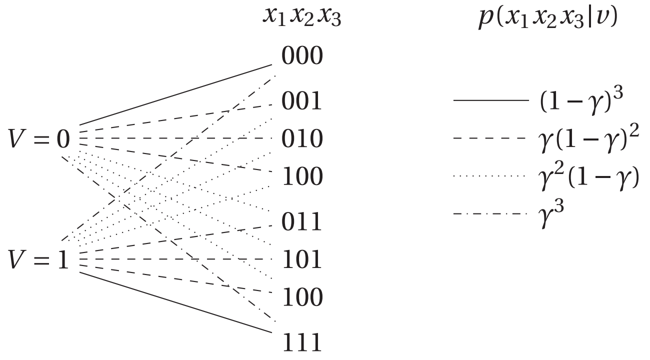
Liu, Xu and Chen proved that the Wyner's common information of such a triplet equals:

$$C_W(X_1, X_2, X_3) = \underbrace{1 + q + h(q) + (1-q)h\left(\frac{q}{2(1-q)}\right)}_{=H(\mathbf{X})} - 3h(\gamma), \tag{7.24}$$

where

$$\gamma = \frac{1}{2}\left(1 - \sqrt{1-2q}\right).$$

Reasoning in reverse: Let V be a $\sim \mathrm{Bern}(\frac{1}{2})$ random variable. Then each X_i is the result of an independent but identical binary symmetric channel with V as an input and γ as the crossover probability[2]. This leads to the following construction of a $(X_1, X_2, X_3)-$triplet:



The main proof of Liu, Xu and Chen is that there exists no other common random variable V that independently constitutes all X_i with a lower value of $I(\mathbf{X}; V)$. In other words, this V not only attains conditional independence for $\mathbf{X}$, it is also indeed the minimizer for $C_W(X_1, X_2, X_3)$.

7.4.2 Caching Inner Bound for a CSBS-triplet

We take the knowledge of Liu, Xu and Chen on the common information of a CSBS-triplet and extrapolate it to a proposed construction to capture total conditional correlation. Then, we put it to the test in our convolutional caching experiment.

First, observe that the V that achieves $C_W(X_1, X_2, X_3)$ can be rewritten in a form similar to the DSBS construction (7.12):

$$V = \begin{cases} X_1 \oplus U_1 & \text{if } X_1 = X_2 = X_3, \\ \text{(majority symbol)} \oplus U_2 & \text{if otherwise.} \end{cases} \tag{7.25}$$

And

$$U_1 \sim \mathrm{Bern}\left(\frac{\gamma^3}{1-\frac{3}{2}q}\right), \tag{7.26}$$

$$U_2 \sim \mathrm{Bern}\left(\frac{2(1-\gamma)\gamma^2}{q}\right). \tag{7.27}$$

Since common information is only a single bit, we conjecture a similar construction as for the DSBS case: for $R_{\mathrm{cache}} \leq C_W(X_1, X_2, X_3)$ it is sufficient to cache a further compressed version of this V and one should not have to consider variables of larger cardinality.

The optimal way of compression would be to tune $p_{U_1}(1)$ and $p_{U_2}(1)$ separately and evaluate

[2] Note that any duo subset of this triplet constitutes a DSBS as considered earlier in this chapter.

the impact on R_{cache} (7.2) and $\overline{R}_{\text{update}}$ (7.3). Instead, we start with a simple construction:

$$W = V \oplus U_3, \tag{7.28}$$

where $U_3 \sim \text{Bern}(\alpha)$ and we control this $\alpha \in [0\ \frac{1}{2}]$. Then, thanks to the total symmetry the rate equations simplify:

$$R_{\text{cache}} = I(\mathbf{X}; W), \tag{7.29}$$

$$\overline{R}_{\text{update}} = H(X_1|W). \tag{7.30}$$

The resulting rate-pairs, as a function of α span the black curves in Figures 7.7–7.9 to the left of the big red dot corresponding to $C_W(X_1, X_2, X_3)$. One can verify visually (or mathematically if preferred) that as a function of α this construction gives a smooth convex curve covering the range of

$$R_{\text{cache}} \in [0\ \ C_W(X_1, X_2, X_3)]$$

and

$$\overline{R}_{\text{update}} \in [H(X_1)\ \ \frac{1}{3}(H(\mathbf{X}) - C_W(X_1, X_2, X_3))].$$

7.4.3 Empirical Caching Setup for a CSBS-triplet

For the *practical* side we apply the same principle as the DSBS: provide the Viterbi encoder with a ‘perfect’ common sequence and let the convolutional code handle the compression. For a CSBS-triplet this procedure is more complex, since we either have that all symbols of X_1, X_2, X_3 are equal or there is a majority of only two being equal. The common information through V (7.25) compressed these states with different U_1, U_2 of which $p_{U_1}(1) \le p_{U_2}(1)$. In words: If all X_i take on the same value, it is more likely that V takes on that same value than when only two X_i are equal.

Therefore, when compressing and caching three CSBS-generated files one must input the encoder with this level of nuance. The Viterbi algorithm as programmed in the CML package for MATLAB requires log-likelihood-ratios as an input [44]. In the $K = 2-$case we only required symmetry, now these need to be properly computed:

$$LLR(x_1, x_2, x_3) = \log \frac{p(x_1x_2x_3|v=1)}{p(x_1x_2x_3|v=0)} = \begin{cases} 3\log\frac{1-\gamma}{\gamma} & \text{if } x_1x_2x_3 = 111 \\ \log\frac{1-\gamma}{\gamma} & \text{if } x_1x_2x_3 \in \{110, 101, 011\} \\ \log\frac{\gamma}{1-\gamma} & \text{if } x_1x_2x_3 \in \{001, 010, 100\} \\ 3\log\frac{\gamma}{1-\gamma} & \text{if } x_1x_2x_3 = 000 \end{cases} \tag{7.31}$$

Calling

$$\lambda \triangleq \log\frac{1-\gamma}{\gamma}, \tag{7.32}$$

we do the following symbol-by-symbol mapping from three CSBS-generated files X_1^N, X_2^N, X_3^N to suitable input the Viterbi algorithm:

$$f_{\text{symbols-to-Viterbi}}(x_1, x_2, x_3) = \begin{cases} 3\lambda & \text{if } x_1x_2x_3 = 111 \\ \lambda & \text{if } x_1x_2x_3 \in \{110, 101, 011\} \\ -\lambda & \text{if } x_1x_2x_3 \in \{001, 010, 100\} \\ -3\lambda & \text{if } x_1x_2x_3 = 000 \end{cases} \tag{7.33}$$

7.4.4 Experimental Results for a CSBS-triplet

The theory and experiment combined constitute Figures 7.7–7.9. The big red dot stands for the turning point of common information (7.24). The black curve is the theory (or the random coding characterization if you will): the section right of the red dot stands for the straight line (7.8), to the left the proposed 'blunt' compression of the common information V (7.25) through (7.28). The dotted straight line connects the points $(R_{\text{cache}}, \overline{R}_{\text{update}}) = (0, \frac{1}{3}\sum_{i=1}^{3} H(X_i))$ to $(C_W(X_1, X_2, X_3), \frac{1}{3}(H(\mathbf{X}) - C_W(X_1, X_2, X_3)))$, acting as a benchmark to beat. The blue triangles correspond to the tests with convolutional codes. Plotted are the rate of those codes versus the average entropy of Δ_k^N as in (7.22). Also here $N = 30,000$.

Observe a subtle difference with the DSBS experiments of Figure 7.4–7.6: in those earlier tests the convolutional code approached the theoretical construction. In this CSBS-triplet case the empirical caching *beats* the inner bound created through (7.28) by a small margin. This can be best viewed in the plot for $q = 0.1$. In other words, the convolutional code finds a better way to compress the common information than the 'blunt' version of (7.28), while still using only one bit to cache three (X_1 through X_3). This proves that total conditional correlation must be captured by a random variable that is more nuanced. By construction, for small R_{cache} the Viterbi algorithm gives a higher priority to positions where all X_k agree on the same symbol than positions where only two of them agree.

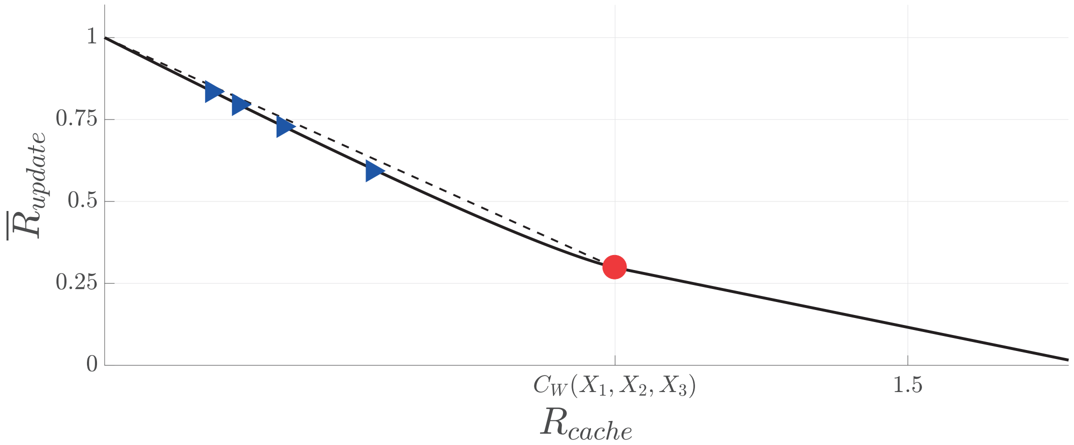


Figure 7.7 – Caching a CSBS-triplet with $q = 0.1$.

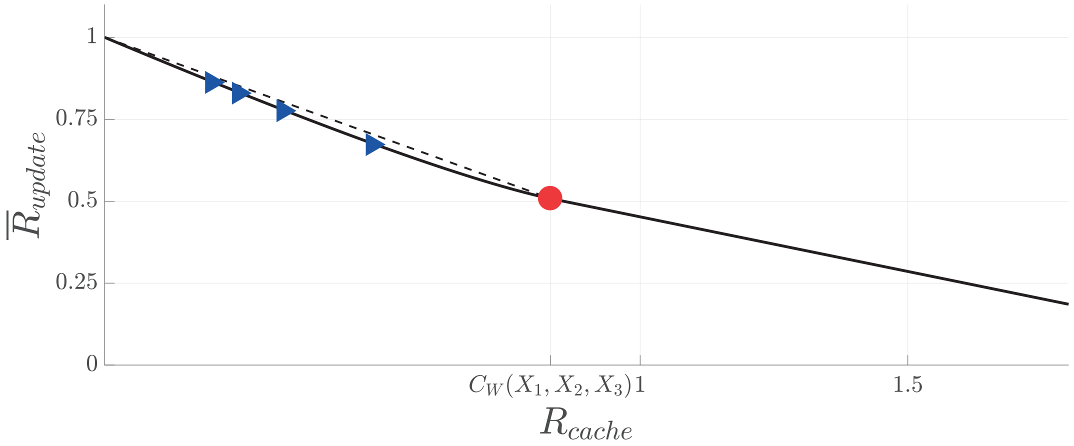


Figure 7.8 – Caching a CSBS-triplet with $q = 0.2$.

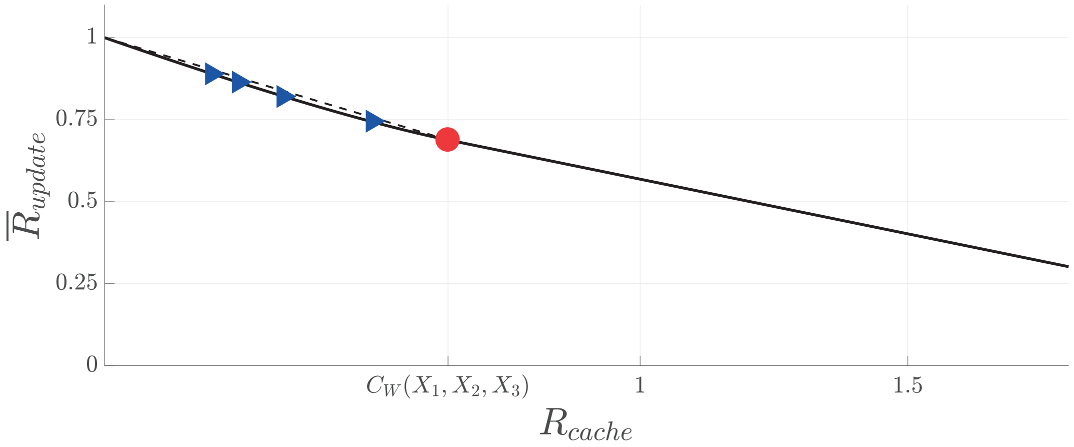


Figure 7.9 – Caching a CSBS-triplet with $q = 0.3$.

7.5 Closing Thoughts: Analogies between the Gaussian and Binary Case

On page 71 we posed an open question: how does correlation structure in $\Sigma_{\mathbf{X}}$ drive the dimensionality of the random variable $\mathbf{V}$ that attains Wyner's common information? All that is known is that for K dependent Gaussians, the dimensionality of that V is between 1 and $K-1$. Through experiments we found plenty of examples of $\Sigma_{\mathbf{X}}$ whose common information is associated to any dimensionality in that range.

In the discrete world the question is rather: how does the joint distribution $p(\mathbf{x})$ drive the *cardinality* $|\mathcal{V}|$ of the $\mathbf{V}$ that attains $C_W(\mathbf{X})$? Also the discrete question is bounded, as $|\mathcal{V}| \leq |\mathcal{X}| + 1$, mentioned in, e.g., Theorem 7.1. This chapter studied in particular doubly symmetric binary sources and their extension to K dimensions called circularly symmetric. A property inherent to their construction is that their common information is associated to a single binary random variable $V \sim \text{Bern}(\frac{1}{2})$.

The CSBS source is one of which each X_i is indistinguishable from any other, which might falsely remind one of the Gaussians with a circulant correlation matrix we studied heavily in Chapter 5. Though tempting, the actual analogy is with respect to a particular Gaussian distribution inside the class of circulants, those with all-equal correlation:

$$\Sigma_{\mathbf{X}} = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & & \vdots \\ \vdots & & \ddots & \rho \\ \rho & \cdots & \rho & 1 \end{bmatrix}.$$

We know $C_W(\mathbf{X})$ for circulant covariances in closed form by Theorem 5.4; it is associated to a distortion matrix $\mathbf{D}_{C_W} = \lambda_{\min}(\Sigma_{\mathbf{X}})\mathbf{I}$. The covariance above is special in the sense that there is only one unique dominant eigenvalue and $K-1$ repeated other ones. The common information of this class of Gaussians is associated to a V that captures the contribution along this one top eigenvector. It has dimensionality 1 regardless of K, just like the common information of a CSBS is a single binary variable regardless of K.

7.6 Appendix: Generator Matrices Used

The following generator matrices were the basis of the convolutional codes used in this chapter:

$$G_{\frac{1}{2}} = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 1 \end{bmatrix}$$

$$G_{\frac{1}{3}} = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 \end{bmatrix}$$

$$G_{\frac{1}{4}} = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 1 \end{bmatrix}$$

$$G_{\frac{1}{5}} = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 1 \end{bmatrix}$$

8 Conclusions

Caching Gaussian sources is hard whenever an encoder wishes to only cache a little and easy when it can cache a lot. This fierce but slightly vague statement is carried through all perspectives we have taken on in this thesis. When we considered the interplay of correlation and user preference in our complete bivariate model, we derived this separation in terms of difficulty/complexity in Theorem 4.4: for small R_{cache} optimal caching distortions lie on a peculiar line in the $\mathcal{D}-$plane without a clear closed-form expression (Figure 4.11a). When we looked at uniform user requests we observed similar behavior: caching *all* the shared information through Wyner's common information is a convex optimization problem, whereas caching only *some* through total conditional correlation is non-convex. Luckily, for bivariates this non-convex problem turned out to be still manageable (Lemma 4.3).

The uniform caching model for bivariate Gaussians is intuitively pleasing in particular: optimal caching strategies need to capture as much of the total correlation as possible and we derived that the encoder can do this by caching the contribution along the dominant eigenvector of the correlation matrix first. This result is reminiscent of the intuition behind the Gaussian multivariate rate-distortion function subject to a trace-constraint: do an eigendecomposition on the covariance to identify which components contribute the most to variance and code those first. This analogy begged a grander question: would there be a similar decomposition of the *correlation* matrix to tell one which components of $\mathbf{X}$ contribute most to $TC(\mathbf{X})$?

Chapter 5 tells us that for multivariates of arbitrary length the *eigen*decomposition of the correlation matrix is *not* the right tool to capture total correlation. In higher dimensions it is merely a bound to potentially better decompositions, as we saw by bounding our caching problem by the trace-based rate-distortion function in Lemma 5.2. We conjectured that only for circulant matrices, whose correlation structure is completely symmetric, the eigenbasis offers the right decomposition.

For the moment, the separation in difficulty between capturing Wyner's common information and Watanabe's total correlation persists and keeps us from fully understanding the latter. An underlining of this was shown in Chapter 6. There we transformed two Gaussian vectors $\mathbf{X}$ and

$\mathbf{Y}$ into independent sets of $(X_i, Y_i)-$ pairs. The difficulty of capturing total correlation then split into a convex problem of which pairs to tackle first, followed by the same non-convex bivariate problem of minimizing the correlation between each X_i and Y_i. It illustrated that the size of the random variables does not make the problem harder, but rather that the hardness is fundamentally and only in the $X - Y$ interaction.

As a closing remark we point out two possible future directions:

- **The algorithmic perspective**: can total conditional correlation be efficiently captured in a suboptimal fashion?

  One question we barely touched upon is whether or not the non-convex optimization problems involving total (conditional) correlation like $R_{\text{cache}}(d, D_F)$ can be relaxed in a manner that is both meaningful and efficient. In his master's thesis, Rohan Pote proposed a change of variable that does not avoid the non-convexity of the feasible set, but reshuffles it in such a way that optimizing $R_{\text{cache}}(d, D_F)$ generally avoids these parts of the set [45]. Experiments in his report have proven to be successful in *tricking* interior point algorithms into believing the problem is actually convex.

  In addition, even though the geometry of capturing total correlation is not convex, it does have structure that can potentially be leveraged. Namely, consider the following rewriting:

$$\max_{\mathbf{D}} |\mathbf{D}| \text{ s.t. } \begin{cases} \mathbf{0} \preceq \mathbf{D} \preceq \Sigma_{\mathbf{X}} \\ \prod_{i=1}^{K} D_{i,i} \leq d \end{cases} = \max_{\substack{D_1,\cdots,D_K \\ :\prod_{i=1}^{K} D_i \leq d}} \max_{\substack{\mathbf{D} \\ :\text{diag}(\mathbf{D}) \leq [D_1,\cdots,D_K]}} |\mathbf{D}| \text{ s.t. } \mathbf{0} \preceq \mathbf{D} \preceq \Sigma_{\mathbf{X}}. \tag{8.1}$$

  The outer optimization walks over hyper-dimensional surfaces $\prod_{i=1}^{K} D_i = d$, while the inner optimization maximizes a matrix inside the convex intersection of the constraints $\text{diag}(\mathbf{D}) \leq [D_1, \cdots, D_K]$ and $\mathbf{D} \preceq \Sigma_{\mathbf{X}}$. There is a sense of monotonic increase in the size of this intersection as one walks over the surface $\prod_{i=1}^{K} D_i = d$ towards an optimum. This structure shows there is more to this barrier of non-convexity than meets the eye; it can potentially inspire the writing of clever approximation algorithms.

- **Further analogies between the Gaussian and binary case**: in Section 7.5 we argued that the CSBS sources of Chapter 7 are analogous to Gaussian multivariates with all-equal correlation. This class of Gaussians is a special case in the grander set of those with circulant correlation matrices of which we understand the common information in analytic closed-form (Theorem 5.4). Perhaps the understanding of how symmetry in correlation drives Gaussian common information can help one to study the discrete equivalent for some natural extension of CSBS sources.

Bibliography

[1] M. A. Maddah-Ali and U. Niesen, “Fundamental limits of caching,” *IEEE Transactions on Information Theory*, vol. 60, pp. 2856–2867, May 2014.

[2] A. Hartung, “Netflix - the turnaround story of 2012!,” *Forbes*, January 2013. [Online].

[3] C. Y. Wang, S. H. Lim, and M. Gastpar, “Information-theoretic caching: Sequential coding for computing,” *IEEE Transactions on Information Theory*, vol. 62, pp. 6393–6406, Nov 2016.

[4] R. Timo, S. S. Bidokhti, M. A. Wigger, and B. C. Geiger, “A rate-distortion approach to caching,” *CoRR*, vol. abs/1610.07304, 2016.

[5] P. Hassanzadeh, E. Erkip, J. Llorca, and A. Tulino, “Distortion-memory tradeoffs in cache-aided wireless video delivery,” in *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 1150–1157, Sept 2015.

[6] Q. Yang and D. Gündüz, “Centralized coded caching for heterogeneous lossy requests,” in *2016 IEEE International Symposium on Information Theory (ISIT)*, pp. 405–409, July 2016.

[7] R. Gray and A. Wyner, “Source coding for a simple network,” *Bell System Technical Journal, The*, vol. 53, pp. 1681–1721, Nov 1974.

[8] A. Wyner, “The common information of two dependent random variables,” *IEEE Transactions on Information Theory*, vol. 21, pp. 163–179, Mar 1975.

[9] S. Watanabe, “Information theoretical analysis of multivariate correlation,” *IBM Journal of Research and Development*, vol. 4, pp. 66–82, Jan 1960.

[10] P. Gács and J. Körner, “Common information is far less than mutual information,” *Problems of Control and Information Theory*, vol. 2, no. 2, pp. 149–162, 1973.

[11] K. Viswanatha, E. Akyol, and K. Rose, “The lossy common information of correlated sources,” *IEEE Transactions on Information Theory*, vol. 60, pp. 3238–3253, June 2014.

[12] G. Xu, W. Liu, and B. Chen, “A lossy source coding interpretation of wyner’s common information,” *IEEE Transactions on Information Theory*, vol. 62, pp. 754–768, Feb 2016.

[13] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics).* Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[14] 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, pp. 1769–1781, April 2010.

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

[16] T. Ando and D. Petz, "Gaussian markov triplets approached by block matrices," *Acta Scientiarum Mathematicarum*, vol. 75, no. 1-2, pp. 329–345, 2009.

[17] M. Friendly, G. Monette, and J. Fox, "Elliptical insights: Understanding statistical methods through elliptical geometry," *Statistical Science*, vol. 28, no. 1, pp. 1–39, 2013.

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

[19] J. Xiao and Q. Luo, "Compression of correlated gaussian sources under individual distortion criteria," in *43rd Allerton Conference on Communication, Control, and Computing*, pp. 438–447, 2005.

[20] A. Kolmogorov, "On the shannon theory of information transmission in the case of continuous signals," *Information Theory, IRE Transactions on*, vol. 2, pp. 102–108, December 1956.

[21] H. Wang and P. Viswanath, "Vector gaussian multiple description with individual and central receivers," *Information Theory, IEEE Transactions on*, vol. 53, pp. 2133–2153, June 2007.

[22] S. Boyd and L. Vandenberghe, *Convex Optimization.* New York, NY, USA: Cambridge University Press, 2004.

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

[24] J. Nayak and E. Tuncel, "Successive coding of correlated sources," *IEEE Transactions on Information Theory*, vol. 55, pp. 4286–4298, Sept 2009.

[25] G. J. Op 't Veld and M. C. Gastpar, "Caching (a pair of) gaussians," in *36th WIC Symposium on Information Theory in the Benelux*, pp. 4–11, May 2015.

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

[27] G. J. Op 't Veld and M. C. Gastpar, "Caching of bivariate gaussians with non-uniform preference probabilities," in *2017 Symposium on Information Theory and Signal Processing in the Benelux*, pp. 176–183, May 2017.

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

[29] W. Liu, G. Xu, and B. Chen, "The common information of N dependent random variables," in *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 836–843, Sept 2010.

[30] 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.

[31] L. Vandenberghe, S. Boyd, and S. P. Wu, "Determinant maximization with linear matrix inequality constraints," *SIAM Journal on Matrix Analysis and Applications*, vol. 19, no. 2, pp. 499–533, 1998.

[32] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1." http://cvxr.com/cvx, Mar. 2014.

[33] Y. Kim and S. Kim, "On the convexity of log det (I + K x^{-1})," *CoRR*, vol. abs/cs/0611043, 2006.

[34] 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.

[35] G. J. Op 't Veld and M. C. Gastpar, "Total correlation of gaussian vector sources on the gray-wyner network," in *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 385–392, Sept 2016.

[36] S. Satpathy and P. Cuff, "Gaussian secure source coding and wyner's common information," in *2015 IEEE International Symposium on Information Theory (ISIT)*, pp. 116–120, June 2015.

[37] E. Martinian and J. Yedidia, "Iterative quantization using codes on graphs," in *Allerton Conference on Communication, Control, and Computing*, Oct. 2003.

[38] M. J. Wainwright, E. Maneva, and E. Martinian, "Lossy source compression using low-density generator matrix codes: Analysis and algorithms," *IEEE Transactions on Information Theory*, vol. 56, pp. 1351–1368, March 2010.

[39] E. Martinian and M. Wainwright, "Low density codes achieve the rate-distortion bound," in *Data Compression Conference (DCC'06)*, pp. 153–162, March 2006.

[40] E. Martinian and M. J. Wainwright, "Analysis of LDGM and compound codes for lossy compression and binning," *CoRR*, vol. abs/cs/0602046, 2006.

[41] S. B. Korada, *"Polar codes for channel and source coding"*. PhD thesis, EPFL (Lausanne), 2009.

[42] A. Viterbi, “Error bounds for convolutional codes and an asymptotically optimum decoding algorithm,” *IEEE Transactions on Information Theory*, vol. 13, pp. 260–269, April 1967.

[43] A. Viterbi and J. Omura, “Trellis encoding of memoryless discrete-time sources with a fidelity criterion,” *IEEE Transactions on Information Theory*, vol. 20, pp. 325–332, May 1974.

[44] Iterative Solutions, “CML, the coded modulation library.” http://www.iterativesolutions.com/, 2005.

[45] R. Pote, “K-gaussians coded caching problem: Inner bound and its optimality,” Master’s thesis, École Polytechnique Fédéral de Lausanne, Lausanne, Switzerland, 2017.

[46] G. J. Op ’t Veld and M. C. Gastpar, “Successive refinement of gaussian projections,” in *35th WIC Symposium on Information Theory in the Benelux*, pp. 42–49, May 2014.

[47] G. J. Op ’t Veld, “On a new compressed sensing paradigm in the modulated wideband converter,” in *Eurocon 2013*, pp. 2140–2145, July 2013.

Curriculum Vitae

Guillaume Jean (Giel) Op ’t Veld

School of Computer and Communication Sciences
École Polytechnique Fédéral de Lausanne (EPFL)
CH-1015, Lausanne, Switzerland
Email: giel.optveld@epfl.ch / gieloptveld@gmail.com

Education

2013-2017	**École Polytechnique Fédéral de Lausanne**, Switzerland PhD candidate, Communication Sciences
2010-2012	**Eindhoven University of Technology**, the Netherlands MSc, Electrical Engineering graduated with honors
2007-2010	**Eindhoven University of Technology**, the Netherlands BSc, Electrical Engineering

Research Experience

Summer 2011	**Data Storage Institute**, A*Star, Singapore internship under the supervision of prof. Cai Kui ‘Numerical Analysis on Floating Codes for Flash and other Non-Volatile Memories’

Publications

[27] G. J. Op ’t Veld and M. C. Gastpar, “Caching of bivariate gaussians with non-uniform preference probabilities,” in *2017 Symposium on Information Theory and Signal Processing in the Benelux*, pp. 176–183, May 2017

[35] G. J. Op ’t Veld and M. C. Gastpar, “Total correlation of gaussian vector sources on

the gray-wyner network," in *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 385–392, Sept 2016

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

[25] G. J. Op 't Veld and M. C. Gastpar, "Caching (a pair of) gaussians," in *36th WIC Symposium on Information Theory in the Benelux*, pp. 4–11, May 2015

[46] G. J. Op 't Veld and M. C. Gastpar, "Successive refinement of gaussian projections," in *35th WIC Symposium on Information Theory in the Benelux*, pp. 42–49, May 2014

[47] G. J. Op 't Veld, "On a new compressed sensing paradigm in the modulated wideband converter," in *Eurocon 2013*, pp. 2140–2145, July 2013