# UNIVERSITÉ PARIS-EST

**École Doctorale MSTIC**

MATHÉMATIQUES ET SCIENCES ET TECHNOLOGIES
DE L'INFORMATION ET DE LA COMMUNICATION


**DISSERTATION**

In Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

Presented on 22 November 2019 by:


## Yiğit UĞUR


---

# An Information-Theoretic Approach to Distributed Learning. Distributed Source Coding Under Logarithmic Loss

---

**Jury :**

| | | | |
|---|---|---|---|
| *Advisor :* | Prof. Abdellatif Zaidi | - | Université Paris-Est, France |
| *Thesis Director :* | Prof. Abderrezak Rachedi | - | Université Paris-Est, France |
| *Reviewers :* | Prof. Giuseppe Caire | - | Technical University of Berlin, Germany |
| | Prof. Gerald Matz | - | Vienna University of Technology, Austria |
| | Dr. Aline Roumy | - | Inria, France |
| *Examiners :* | Prof. David Gesbert | - | Eurecom, France |
| | Prof. Michel Kieffer | - | Université Paris-Sud, France |

# Acknowledgments

First, I would like to express my gratitude to my advisor Abdellatif Zaidi for his guidance and support. It was a pleasure to benefit and learn from his knowledge and vision through my studies.

I want to thank my colleague Iñaki Estella Aguerri. I enjoyed very much collaborating with him. He was very helpful, and tried to share his experience whenever I need.

My Ph.D. was in the context of a *CIFRE* contract. I appreciate my company Huawei Technologies France for supporting me during my education. It was a privilege to be a part of the Mathematical and Algorithmic Sciences Lab, Paris Research Center, and to work with scientists coming from different parts of the world. It was a unique experience to be within a very competitive international working environment.

During my Ph.D. studies, Paris gave me a pleasant surprise, the sincerest coincidence of meeting with Özge. I would like to thank her for always supporting me and sharing the Parisian life with me.

Last, and most important, my deepest thanks are to my family: my parents Mustafa and Kıymet, and my brother Kağan. They have been always there to support me whenever I need. I could not have accomplished any of this without them. Their infinite love and support is what make it all happen.

# Abstract

One substantial question, that is often argumentative in learning theory, is how to choose a 'good' loss function that measures the fidelity of the reconstruction to the original. Logarithmic loss is a natural distortion measure in the settings in which the reconstructions are allowed to be 'soft', rather than 'hard' or deterministic. In other words, rather than just assigning a deterministic value to each sample of the source, the decoder also gives an assessment of the degree of confidence or reliability on each estimate, in the form of weights or probabilities. This measure has appreciable mathematical properties which establish some important connections with lossy universal compression. Logarithmic loss is widely used as a penalty criterion in various contexts, including clustering and classification, pattern recognition, learning and prediction, and image processing. Considering the high amount of research which is done recently in these fields, the logarithmic loss becomes a very important metric and will be the main focus as a distortion metric in this thesis.

In this thesis, we investigate a distributed setup, so-called the Chief Executive Officer (CEO) problem under logarithmic loss distortion measure. Specifically, $K \geq 2$ agents observe independently corrupted noisy versions of a remote source, and communicate independently with a decoder or CEO over rate-constrained noise-free links. The CEO also has its own noisy observation of the source and wants to reconstruct the remote source to within some prescribed distortion level where the incurred distortion is measured under the logarithmic loss penalty criterion.

One of the main contributions of the thesis is the explicit characterization of the rate-distortion region of the vector Gaussian CEO problem, in which the source, observations and side information are jointly Gaussian. For the proof of this result, we first extend Courtade-Weissman's result on the rate-distortion region of the discrete memoryless (DM) $K$-encoder CEO problem to the case in which the CEO has access to a correlated side information stream which is such that the agents' observations are independent conditionally given the side information and remote source. Next, we obtain an outer bound on the region of the vector Gaussian CEO problem by evaluating the outer bound of the DM model by means of a technique that relies on the de Bruijn identity and the properties of Fisher information. The approach is similar to Ekrem-Ulukus outer bounding technique for the vector Gaussian CEO problem under quadratic distortion measure, for which it was there found generally non-tight; but it is shown here to yield a complete characterization of the region for the case of logarithmic loss measure. Also, we show that Gaussian test channels with time-sharing exhaust the Berger-Tung inner bound, which is optimal. Furthermore, application of our results allows us to find the complete solutions of three related problems: the quadratic vector Gaussian CEO problem with *determinant* constraint, the vector Gaussian distributed hypothesis testing against conditional independence problem and the vector Gaussian distributed Information Bottleneck problem.

With the known relevance of the logarithmic loss fidelity measure in the context of learning and prediction, developing algorithms to compute the regions provided in this thesis may find usefulness in a variety of applications where learning is performed distributively. Motivated from this fact, we develop two type algorithms: i) Blahut-Arimoto (BA) type iterative numerical algorithms for both discrete and Gaussian models in which the joint distribution of the sources are known; and ii) a variational inference type algorithm in which the encoding mappings are parameterized by neural networks and the variational bound approximated by Monte Carlo sampling and optimized with stochastic gradient descent for the case in which there is only a set of training data is available. Finally, as an application, we develop an unsupervised generative clustering framework that uses the variational Information Bottleneck (VIB) method and models the latent space as a mixture of Gaussians. This generalizes the VIB which models the latent space as an isotropic Gaussian which is generally not expressive enough for the purpose of unsupervised clustering. We illustrate the efficiency of our algorithms through some numerical examples.

***Keywords:*** *Multiterminal source coding, CEO problem, rate-distortion region, logarithmic loss, quadratic loss, hypothesis testing, Information Bottleneck, Blahut-Arimoto algorithm, distributed learning, classification, unsupervised clustering.*

# Contents

# List of Figures

# List of Algorithms

# List of Tables

# Notation

Throughout the thesis, we use the following notation. Upper case letters are used to denote random variables, e.g., $X$; lower case letters are used to denote realizations of random variables, e.g., $x$; and calligraphic letters denote sets, e.g., $\mathcal{X}$. The cardinality of a set $\mathcal{X}$ is denoted by $|\mathcal{X}|$. The closure of a set $\mathcal{A}$ is denoted by $\overline{A}$. The probability distribution of the random variable $X$ taking the realizations $x$ over the set $\mathcal{X}$ is denoted by $P_X(x) = \Pr[X = x]$; and, sometimes, for short, as $p(x)$. We use $\mathcal{P}(\mathcal{X})$ to denote the set of discrete probability distributions on $\mathcal{X}$. The length-$n$ sequence $(X_1, \ldots, X_n)$ is denoted as $X^n$; and, for integers $j$ and $k$ such that $1 \leq k \leq j \leq n$, the sub-sequence $(X_k, X_{k+1}, \ldots, X_j)$ is denoted as $X_k^j$. We denote the set of natural numbers by $\mathbb{N}$, and the set of positive real numbers by $\mathbb{R}_+$. For an integer $K \geq 1$, we denote the set of natural numbers smaller or equal $K$ as $\mathcal{K} = \{k \in \mathbb{N} \,:\, 1 \leq k \leq K\}$. For a set of natural numbers $\mathcal{S} \subseteq \mathcal{K}$, the complementary set of $\mathcal{S}$ is denoted by $\mathcal{S}^c$, i.e., $\mathcal{S}^c = \{k \in \mathbb{N} \,:\, k \in \mathcal{K} \setminus \mathcal{S}\}$. Sometimes, for convenience we use $\bar{\mathcal{S}}$ defined as $\bar{\mathcal{S}} = \{0\} \cup \mathcal{S}^c$. For a set of natural numbers $\mathcal{S} \subseteq \mathcal{K}$; the notation $X_{\mathcal{S}}$ designates the set of random variables $\{X_k\}$ with indices in the set $\mathcal{S}$, i.e., $X_{\mathcal{S}} = \{X_k\}_{k \in \mathcal{S}}$. Boldface upper case letters denote vectors or matrices, e.g., $\mathbf{X}$, where context should make the distinction clear. The notation $\mathbf{X}^\dagger$ stands for the conjugate transpose of $\mathbf{X}$ for complex-valued $\mathbf{X}$, and the transpose of $\mathbf{X}$ for real-valued $\mathbf{X}$. We denote the covariance of a zero mean, complex-valued, vector $\mathbf{X}$ by $\mathbf{\Sigma}_{\mathbf{x}} = \mathbb{E}[\mathbf{X}\mathbf{X}^\dagger]$. Similarly, we denote the cross-correlation of two zero-mean vectors $\mathbf{X}$ and $\mathbf{Y}$ as $\mathbf{\Sigma}_{\mathbf{x},\mathbf{y}} = \mathbb{E}[\mathbf{X}\mathbf{Y}^\dagger]$, and the conditional correlation matrix of $\mathbf{X}$ given $\mathbf{Y}$ as $\mathbf{\Sigma}_{\mathbf{x}|\mathbf{y}} = \mathbb{E}\left[(\mathbf{X} - \mathbb{E}[\mathbf{X}|\mathbf{Y}])(\mathbf{X} - \mathbb{E}[\mathbf{X}|\mathbf{Y}])^\dagger\right]$, i.e., $\mathbf{\Sigma}_{\mathbf{x}|\mathbf{y}} = \mathbf{\Sigma}_{\mathbf{x}} - \mathbf{\Sigma}_{\mathbf{x},\mathbf{y}}\mathbf{\Sigma}_{\mathbf{y}}^{-1}\mathbf{\Sigma}_{\mathbf{y},\mathbf{x}}$. For matrices $\mathbf{A}$ and $\mathbf{B}$, the notation $\mathrm{diag}(\mathbf{A}, \mathbf{B})$ denotes the block diagonal matrix whose diagonal elements are the matrices $\mathbf{A}$ and $\mathbf{B}$ and its off-diagonal elements are the all zero matrices. Also, for a set of integers $\mathcal{J} \subset \mathbb{N}$ and a family of matrices $\{\mathbf{A}_i\}_{i \in \mathcal{J}}$ of the same size, the notation $\mathbf{A}_{\mathcal{J}}$ is used to denote the (super) matrix obtained by concatenating vertically the matrices $\{\mathbf{A}_i\}_{i\in\mathcal{J}}$, where the indices are sorted in the ascending order, e.g, $\mathbf{A}_{\{0,2\}} = [\mathbf{A}_0^\dagger, \mathbf{A}_2^\dagger]^\dagger$. We use $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to denote a real multivariate Gaussian random variable with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and $\mathcal{CN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to denote a circularly symmetric complex multivariate Gaussian random variable with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

# Acronyms

| | |
|---|---|
| ACC | Clustering Accuracy |
| AE | Autoencoder |
| BA | Blahut-Arimoto |
| BSC | Binary Symmetric Channel |
| CEO | Chief Executive Officer |
| C-RAN | Cloud Radio Acces Netowrk |
| DEC | Deep Embedded Clustering |
| DM | Discrete Memoryless |
| DNN | Deep Neural Network |
| ELBO | Evidence Lower Bound |
| EM | Expectation Maximization |
| GMM | Gaussian Mixture Model |
| IB | Information Bottleneck |
| IDEC | Improved Deep Embedded Clustering |
| KKT | Karush-Kuhn-Tucker |
| KL | Kullback-Leibler |
| LHS | Left Hand Side |
| MDL | Minimum Description Length |

| | |
|---|---|
| MIMO | Multiple-Input Multiple-Output |
| MMSE | Minimum Mean Square Error |
| NN | Neural Network |
| PCA | Principal Component Analysis |
| PMF | Probability Mass Function |
| RHS | Right Hand Side |
| SGD | Stochastic Gradient Descent |
| SUM | Successive Upper-bound Minimization |
| VaDE | Variational Deep Embedding |
| VAE | Variational Autoencoder |
| VIB | Variational Information Bottleneck |
| VIB-GMM | Variational Information Bottleneck with Gaussian Mixture Model |
| WZ | Wyner-Ziv |

# Chapter 1

# Introduction and Main Contributions

The Chief Executive Officer (CEO) problem – also called as the *indirect multiterminal source coding problem* – was first studied by Berger *et al.* in [2]. Consider the vector Gaussian CEO problem shown in Figure 1.1. In this model, there is an arbitrary number $K \geq 2$ of encoders (so-called agents) each having a noisy observation of a vector Gaussian source $\mathbf{X}$. The goal of the agents is to describe the source to a central unit (so-called CEO), which wants to reconstruct this source to within a prescribed distortion level. The incurred distortion is measured according to some loss measure $d : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}$, where $\hat{\mathcal{X}}$ designates the reconstruction alphabet. For quadratic distortion measure, i.e.,

$$d(x, \hat{x}) = |x - \hat{x}|^2$$

the rate-distortion region of the vector Gaussian CEO problem is still unknown in general, except in few special cases the most important of which is perhaps the case of scalar sources, i.e., scalar Gaussian CEO problem, for which a complete solution, in terms of characterization of the optimal rate-distortion region, was found independently by Oohama in [3] and by Prabhakaran *et al.* in [4]. Key to establishing this result is a judicious application of the entropy power inequality. The extension of this argument to the case of vector Gaussian sources, however, is not straightforward as the entropy power inequality is known to be non-tight in this setting. The reader may refer also to [5, 6] where non-tight outer bounds on the rate-distortion region of the vector Gaussian CEO problem under quadratic distortion measure are obtained by establishing some extremal inequalities that

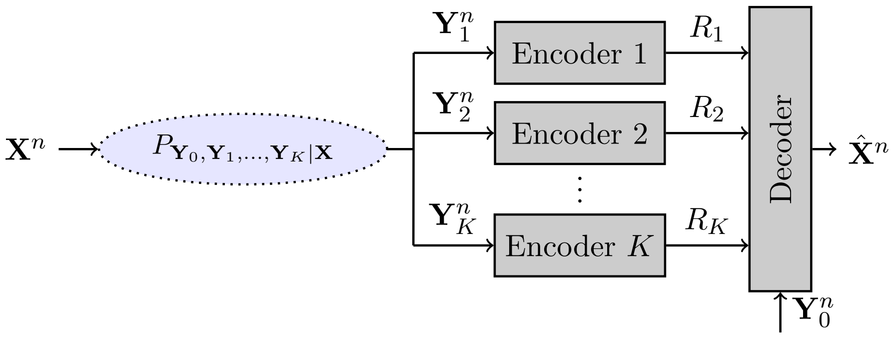



Figure 1.1: Chief Executive Officer (CEO) source coding problem with side information.

are similar to Liu-Viswanath [7], and to [8] where a strengthened extremal inequality yields a complete characterization of the region of the vector Gaussian CEO problem in the special case of trace distortion constraint.

In this thesis, our focus will be mainly on the memoryless CEO problem with side information at the decoder of Figure 1.1 in the case in which the distortion is measured using the logarithmic loss criterion, i.e.,

$$d^{(n)}(x^n, \hat{x}^n) = \frac{1}{n}\sum_{i=1}^{n} d(x_i, \hat{x}_i) \ ,$$

with the letter-wise distortion given by

$$d(x, \hat{x}) = \log\left(\frac{1}{\hat{x}(x)}\right) \ ,$$

where $\hat{x}(\cdot)$ designates a probability distribution on $\mathcal{X}$ and $\hat{x}(x)$ is the value of this distribution evaluated for the outcome $x \in \mathcal{X}$. The logarithmic loss distortion measure plays a central role in settings in which reconstructions are allowed to be ‘soft’, rather than ‘hard’ or deterministic. That is, rather than just assigning a deterministic value to each sample of the source, the decoder also gives an assessment of the degree of confidence or reliability on each estimate, in the form of weights or probabilities. This measure was introduced in the context of rate-distortion theory by Courtade *et al.* [9, 10] (see Chapter 2.1 for a detailed discussion on the logarithmic loss).

## 1.1 Main Contributions

One of the main contributions of this thesis is a complete characterization of the rate-distortion region of the vector Gaussian CEO problem of Figure 1.1 under logarithmic loss distortion measure. In the special case in which there is no side information at the decoder, the result can be seen as the counterpart, to the vector Gaussian case, of that by Courtade and Weissman [10, Theorem 10] who established the rate-distortion region of the CEO problem under logarithmic loss in the discrete memoryless (DM) case. For the proof of this result, we derive a matching outer bound by means of a technique that relies of the de Bruijn identity, a connection between differential entropy and Fisher information, along with the properties of minimum mean square error (MMSE) and Fisher information. By opposition to the case of quadratic distortion measure, for which the application of this technique was shown in [11] to result in an outer bound that is generally non-tight, we show that this approach is successful in the case of logarithmic distortion measure and yields a complete characterization of the region. On this aspect, it is noteworthy that, in the specific case of scalar Gaussian sources, an alternate converse proof may be obtained by extending that of the scalar Gaussian many-help-one source coding problem by Oohama [3] and Prabhakaran *et al.* [4] by accounting for side information and replacing the original mean square error distortion constraint with conditional entropy. However, such approach does not seem to lead to a conclusive result in the vector case as the entropy power inequality is known to be generally non-tight in this setting [12, 13]. The proof of the achievability part simply follows by evaluating a straightforward extension to the continuous alphabet case of the solution of the DM model using Gaussian test channels and *no* time-sharing. Because this does *not* necessarily imply that Gaussian test channels also exhaust the Berger-Tung inner bound, we investigate the question and we show that they *do* if time-sharing is allowed.

Besides, we show that application of our results allows us to find complete solutions to three related problems:

1) The first is a quadratic vector Gaussian CEO problem with reconstruction constraint on the *determinant* of the error covariance matrix that we introduce here, and for which we also characterize the optimal rate-distortion region. Key to establishing this result, we show that the rate-distortion region of vector Gaussian CEO problem under logarithmic loss which is found in this paper translates into an outer bound on the rate region of the quadratic vector Gaussian CEO problem with *determinant* constraint. The reader may refer to, e.g., [14] and [15] for examples of usage of such a determinant constraint in the context of equalization and others.

2) The second is the $K$-encoder hypothesis testing against conditional independence problem that was introduced and studied by Rahman and Wagner in [16]. In this problem, $K$ sources $(\mathbf{Y}_1, \dots, \mathbf{Y}_K)$ are compressed distributively and sent to a detector that observes the pair $(\mathbf{X}, \mathbf{Y}_0)$ and seeks to make a decision on whether $(\mathbf{Y}_1, \dots, \mathbf{Y}_K)$ is independent of $\mathbf{X}$ conditionally given $\mathbf{Y}_0$ or not. The aim is to characterize all achievable encoding rates and exponents of the Type II error probability when the Type I error probability is to be kept below a prescribed (small) value. For both DM and vector Gaussian models, we find a full characterization of the optimal rates-exponent region when $(\mathbf{X}, \mathbf{Y}_0)$ induces conditional independence between the variables $(\mathbf{Y}_1, \dots, \mathbf{Y}_K)$ under the null hypothesis. In both settings, our converse proofs show that the Quantize-Bin-Test scheme of [16, Theorem 1], which is similar to the Berger-Tung distributed source coding, is optimal. In the special case of one encoder, the assumed Markov chain under the null hypothesis is non-restrictive; and, so, we find a complete solution of the vector Gaussian hypothesis testing against conditional independence problem, a problem that was previously solved in [16, Theorem 7] in the case of scalar-valued source and testing against independence (note that [16, Theorem 7] also provides the solution of the scalar Gaussian many-help-one hypothesis testing against independence problem).

3) The third is an extension of Tishby's single-encoder Information Bottleneck (IB) method [17] to the case of multiple encoders. Information theoretically, this problem is known to be essentially a remote source coding problem with logarithmic loss distortion measure [18]; and, so, we use our result for the vector Gaussian CEO problem under logarithmic loss to infer a full characterization of the optimal trade-off between *complexity* (or rate) and *accuracy* (or information) for the distributed vector Gaussian IB problem.

On the algorithmic side, we make the following contributions.

1) For both DM and Gaussian settings in which the joint distribution of the sources is known, we develop Blahut-Arimoto (BA) [19, 20] type iterative algorithms that allow to compute (approximations of) the rate regions that are established in this thesis; and prove their convergence to stationary points. We do so through a variational formulation that allows to determine the set of self-consistent equations

that are satisfied by the stationary solutions. In the Gaussian case, we show that the algorithm reduces to an appropriate updating rule of the parameters of noisy linear projections. This generalizes the Gaussian Information Bottleneck projections [21] to the distributed setup. We note that the computation of the rate-distortion regions of multiterminal and CEO source coding problems is important *per-se* as it involves non-trivial optimization problems over distributions of auxiliary random variables. Also, since the logarithmic loss function is instrumental in connecting problems of multiterminal rate-distortion theory with those of distributed learning and estimation, the algorithms that are developed in this paper also find usefulness in emerging applications in those areas. For example, our algorithm for the DM CEO problem under logarithm loss measure can be seen as a generalization of Tishby's IB method [17] to the distributed learning setting. Similarly, our algorithm for the vector Gaussian CEO problem under logarithm loss measure can be seen as a generalization of that of [21, 22] to the distributed learning setting. For other extension of the BA algorithm in the context of multiterminal data transmission and compression, the reader may refer to related works on point-to-point [23, 24] and broadcast and multiple access multiterminal settings [25, 26].

2) For the cases in which the joint distribution of the sources is not known (instead only a set of training data is available), we develop a variational inference type algorithm, so-called D-VIB. In doing so: i) we develop a variational bound on the optimal information-rate function that can be seen as a generalization of IB method, the evidence lower bound (ELBO) and the $\beta$-VAE criteria [27, 28] to the distributed setting, ii) the encoders and the decoder are parameterized by deep neural networks (DNN), and iii) the bound approximated by Monte Carlo sampling and optimized with stochastic gradient descent. This algorithm makes usage of Kingma *et al.*'s reparameterization trick [29] and can be seen as a generalization of the variational Information Bottleneck (VIB) algorithm in [30] to the distributed case.

Finally, we study an application to the unsupervised learning, which is a generative clustering framework that combines variational Information Bottleneck and the Gaussian Mixture Model (GMM). Specifically, we use the variational Information Bottleneck method and model the latent space as a mixture of Gaussians. Our approach falls into the class in which clustering is performed over the latent space representations rather than the data itself. We derive a bound on the cost function of our model that generalizes the ELBO; and provide a variational inference type algorithm that allows to compute it. Our algorithm, so-called Variational Information Bottleneck with Gaussian Mixture Model (VIB-GMM), generalizes the variational deep embedding (VaDE) algorithm of [31] which is based on variational autoencoders (VAE) and performs clustering by maximizing the ELBO, and can be seen as a specific case of our algorithm obtained by setting $s = 1$. Besides, the VIB-GMM also generalizes the VIB of [30] which models the latent space as an isotropic Gaussian which is generally not expressive enough for the purpose of unsupervised clustering. Furthermore, we study the effect of tuning the hyperparameter $s$, and propose an annealing-like algorithm [32], in which the parameter $s$ is increased gradually with iterations. Our algorithm is applied to various datasets, and we observed a better performance in term of the clustering accuracy (ACC) compared to the state of the art algorithms, e.g., VaDE [31], DEC [33].

## 1.2 Outline

The chapters of the thesis and the content in each of them are summarized in what follows.

#### Chapter 2

The aim of this chapter is to explain some preliminaries for the point-to-point case before presenting our contributions in the distributed setups. First, we explain the logarithmic loss distortion measure, which plays an important role on the theory of learning. Then, the remote source coding problem [34] is presented, which is eventually the Information Bottleneck problem with the choice of logarithmic loss as a distortion measure. Later, we explain the Tishby's Information Bottleneck problem for the discrete memoryless [17] and Gaussian cases [21], also present the Blahut-Arimoto type algorithms [19, 20] to compute the IB curves. Besides, there is shown the connections of the IB with some well-known information-theoretical source coding problems, e.g., common reconstruction [35], information combining [36–38], the Wyner-Ahlswede-Körner problem [39, 40], the efficiency of investment information [41], and the privacy funnel problem [42]. Finally, we present the learning via IB section, which includes a brief explanation of representation learning [43], finite-sample bound on the generalization gap, as well as, the variational bound method which leads the IB to a learning algorithm, so-called the variational IB (VIB) [30] with the usage of neural reparameterization and Kingma *et al.*'s reparameterization trick [29].

**Chapter 3**

In this chapter, we study the discrete memoryless CEO problem with side information under logarithmic loss. First, we provide a formal description of the DM CEO model that is studied in this chapter, as well as some definitions that are related to it. Then, the Courtade-Weissman's result [10, Theorem 10] on the rate-distortion region of the DM $K$-encoder CEO problem is extended to the case in which the CEO has access to a correlated side information stream which is such that the agents' observations are conditionally independent given the decoder's side information and the remote source. This will be instrumental in the next chapter to study the vector Gaussian CEO problem with side information under logarithmic loss. Besides, we study a two-encoder case in which the decoder is interested in estimation of encoder observations. For this setting, we find the rate-distortion region that extends the result of [10, Theorem 6] for the two-encoder multiterminal source coding problem with average logarithmic loss distortion constraints on $Y_1$ and $Y_2$ and no side information at the decoder to the setting in which the decoder has its own side information $Y_0$ that is arbitrarily correlated with $(Y_1, Y_2)$. Furthermore, we study the distributed pattern classification problem as an example of the DM two-encoder CEO setup and we find an upper bound on the probability of misclassification. Finally, we look another closely related problem called the distributed hypothesis testing against conditional independence, specifically the one studied by Rahman and Wagner in [16]. We characterize the rate-exponent region for this problem by providing a converse proof and show that it is achieved using the Quantize-Bin-Test scheme of [16].

**Chapter 4**

In this chapter, we study the vector Gaussian CEO problem with side information under logarithmic loss. First, we provide a formal description of the vector Gaussian CEO problem that is studied in this chapter. Then, we present one of the main results of the thesis, which is an explicit characterization of the rate-distortion region of the vector Gaussian CEO problem with side information under logarithmic loss. In doing so, we use a similar approach to Ekrem-Ulukus outer bounding technique [11] for the vector Gaussian CEO problem under quadratic distortion measure, for which it was there found generally non-tight; but it is shown here to yield a complete characterization of the region for the case of logarithmic loss measure. We also show that Gaussian test channels with time-sharing exhaust the Berger-Tung rate region which is optimal. In this chapter, we also use our results on the CEO problem under logarithmic loss to infer complete solutions of three related problems: the quadratic vector Gaussian CEO problem with a determinant constraint on the covariance matrix error, the vector Gaussian distributed hypothesis testing against conditional independence problem, and the vector Gaussian distributed Information Bottleneck problem.

**Chapter 5**

This chapter contains a description of two algorithms and architectures that were developed in [1] for the distributed learning scenario. We state them here for reasons of completeness. In particular, the chapter provides: i) Blahut-Arimoto type iterative algorithms that allow to compute numerically the rate-distortion or relevance-complexity regions of the DM and vector Gaussian CEO problems that are established in previous chapters for the case in which the joint distribution of the data is known perfectly or can be estimated with a high accuracy; and ii) a variational inference type algorithm in which the encoding mappings are parameterized by neural networks and the variational bound approximated by Monte Carlo sampling and optimized with stochastic gradient descent for the case in which there is only a set of training data is available. The second algorithm, so-called D-VIB [1], can be seen as a generalization of the variational Information Bottleneck (VIB) algorithm in [30] to the distributed case. The advantage of D-VIB over centralized VIB can be explained by the advantage of training the latent space embedding for each observation separately, which allows to adjust better the encoding and decoding parameters to the statistics of each observation, justifying the use of D-VIB for multi-view learning [44, 45] even if the data is available in a centralized manner.

**Chapter 6**

In this chapter, we study an unsupervised generative clustering framework that combines variational Information Bottleneck and the Gaussian Mixture Model for the point-to-point case (e.g., the CEO problem with one encoder). The variational inference type algorithm provided in the previous chapter assumes that there is access to the labels (or remote sources), and the latent space therein is modeled with an isotropic Gaussian. Here, we turn our attention to the case in which there is no access to the labels at all. Besides, we use a more expressive model for the latent space, e.g., Gaussian Mixture Model. Similar to the previous chapter, we derive a bound on the cost function of our model that generalizes the evidence lower bound (ELBO); and provide a variational inference type algorithm that allows to compute it. Furthermore, we show how tuning the trade-off parameter $s$ appropriately by gradually increasing its value with iterations (number of epochs) results in a better accuracy. Finally, our algorithm is applied to various datasets, including the MNIST [46], REUTERS [47] and STL-10 [48], and it is seen that our algorithm outperforms the state of the art algorithms, e.g., VaDE [31], DEC [33] in term of clustering accuracy.

**Chapter 7**

In this chapter, we propose and discuss some possible future research directions.

## Publications

The material of the thesis has been published in the following works.

- Yiğit Uğur, Iñaki Estella Aguerri and Abdellatif Zaidi, "Vector Gaussian CEO Problem Under Logarithmic Loss and Applications," accepted for publication in *IEEE Transactions on Information Theory*, January 2020.
- Yiğit Uğur, Iñaki Estella Aguerri and Abdellatif Zaidi, "Vector Gaussian CEO Problem Under Logarithmic Loss," in *Proceedings of IEEE Information Theory Workshop*, pages 515 – 519, November 2018.
- Yiğit Uğur, Iñaki Estella Aguerri and Abdellatif Zaidi, "A Generalization of Blahut-Arimoto Algorithm to Compute Rate-Distortion Regions of Multiterminal Source Coding Under Logarithmic Loss," in *Proceedings of IEEE Information Theory Workshop*, pages 349 – 353, November 2017.
- Yiğit Uğur, George Arvanitakis and Abdellatif Zaidi, "Variational Information Bottleneck for Unsupervised Clustering: Deep Gaussian Mixture Embedding," *Entropy*, vol. 22, no. 2, article number 213, February 2020.

# Chapter 2

# Logarithmic Loss Compression and Connections

## 2.1 Logarithmic Loss Distortion Measure

Shannon's rate-distortion theory gives the optimal trade-off between compression rate and fidelity. The rate is usually measured in terms of the bits per sample and the fidelity of the reconstruction to the original can be measured by using different distortion measures, e.g., mean-square error, mean-absolute error, quadratic error, etc., preferably chosen according to requirements of the setting where it is used. The main focus in this thesis will be on the logarithmic loss, which is a natural distortion measure in the settings in which the reconstructions are allowed to be 'soft', rather than 'hard' or deterministic. That is, rather than just assigning a deterministic value to each sample of the source, the decoder also gives an assessment of the degree of confidence or reliability on each estimate, in the form of weights or probabilities. This measure, which was introduced in the context of rate-distortion theory by Courtade *et al.* [9, 10] (see also [49, 50] for closely related works), has appreciable mathematical properties [51, 52], such as a deep connection to lossless coding for which fundamental limits are well developed (e.g., see [53] for recent results on universal lossy compression under logarithmic loss that are built on this connection). Also, it is widely used as a penalty criterion in various contexts, including clustering and classification [17], pattern recognition, learning and prediction [54], image processing [55], secrecy [56] and others.

Let random variable $X$ denote the source with finite alphabet $\mathcal{X} = \{x_1, \ldots, x_n\}$ to be compressed. Also, let $\mathcal{P}(\mathcal{X})$ denote the reconstruction alphabet, which is the set of probability measures on $\mathcal{X}$. The logarithmic loss distortion between $x \in \mathcal{X}$ and its reconstruction $\hat{x} \in \mathcal{P}(\mathcal{X})$, $l_{\log} : \mathcal{X} \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}_+$, is given by

$$l_{\log}(x, \hat{x}) = \log \frac{1}{\hat{x}(x)} \ , \tag{2.1}$$

where $\hat{x}(\cdot)$ designates a probability distribution on $\mathcal{X}$ and $\hat{x}(x)$ is the value of this distribution evaluated for the outcome $x \in \mathcal{X}$. We can interpret the logarithmic loss distortion measure as the remaining uncertainty about $x$ given $\hat{x}$. Logarithmic loss is also known as the *self-information loss* in literature.

Motivated by the increasing interest for problems of learning and prediction, a growing body of works study point-to-point and multiterminal source coding models under logarithmic loss. In [51], Jiao *et al.* provide a fundamental justification for inference using logarithmic loss, by showing that under some mild conditions (the loss function satisfying some data processing property and alphabet size larger than two) the reduction in optimal risk in the presence of side information is uniquely characterized by mutual information, and the corresponding loss function coincides with the logarithmic loss. Somewhat related, in [57] Painsky and Wornell show that for binary classification problems the logarithmic loss dominates "universally" any other convenient (i.e., smooth, proper and convex) loss function, in the sense that by minimizing the logarithmic loss one minimizes the regret that is associated with any such measures. More specifically, the divergence associated any smooth, proper and convex loss function is shown to be bounded from above by the Kullback-Leibler divergence, up to a multiplicative normalization constant. In [53], the authors study the problem of universal lossy compression under logarithmic loss, and derive bounds on the non-asymptotic fundamental limit of fixed-length universal coding with respect to a family of distributions that generalize the well-known minimax bounds for universal lossless source coding. In [58], the minimax approach is studied for a problem of remote prediction and is shown to correspond to a one-shot minimax noisy source coding problem. The setting of remote prediction of [58] provides an approximate one-shot operational interpretation of the Information Bottleneck method of [17], which is also sometimes interpreted as a remote source coding problem under logarithmic loss [18].

Logarithmic loss is also instrumental in problems of data compression under a mutual information constraint [59], and problems of relaying with relay nodes that are constrained not to know the users' codebooks (sometimes termed "oblivious" or nomadic processing) which is studied in the single user case first by Sanderovich *et al.* in [60] and then by Simeone *et al.* in [61], and in the multiple user multiple relay case by Aguerri *et al.* in [62] and [63]. Other applications in which the logarithmic loss function can be used include secrecy and privacy [56,64], hypothesis testing against independence [16,65–68] and others.

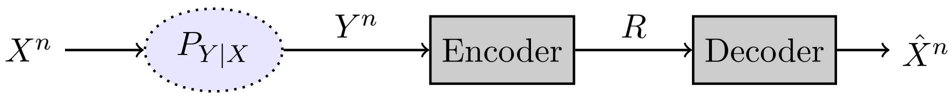



Figure 2.1: Remote, or indirect, source coding problem.

## 2.2 Remote Source Coding Problem

Consider the remote source coding problem [34] depicted in Figure 2.1. Let $X^n$ designates a memoryless remote source sequence, i.e., $X^n := \{X_i\}_{i=1}^n$, with alphabet $\mathcal{X}^n$. An encoder observes the sequence $Y^n$ with alphabet $\mathcal{Y}^n$ that is a noisy version of $X^n$ and obtained from $X^n$ passing through the channel $P_{Y|X}$. The encoder describes its observation using the following encoding mapping

$$\phi^{(n)} \ : \ \mathcal{Y}^n \to \{1, \ldots, M^{(n)}\} \ , \tag{2.2}$$

and sends to a decoder through an error-free link of the capacity $R$. The decoder produces $\hat{X}^n$ with alphabet $\hat{\mathcal{X}}^n$ which is the reconstruction of the remote source sequence through the following decoding mapping

$$\psi^{(n)} \ : \ \{1, \ldots, M^{(n)}\} \to \hat{\mathcal{X}}^n \ . \tag{2.3}$$

The decoder is interested in reconstructing the remote source $X^n$ to within an average distortion level $D$, i.e.,

$$\mathbb{E}_{P_{X,Y}} \left[ d^{(n)}(x^n, \hat{x}^n) \right] \leq D \ , \tag{2.4}$$

for some chosen fidelity criterion $d^{(n)}(x^n, \hat{x}^n)$ obtained from the per-letter distortion function $d(x_i, \hat{x}_i)$, as

$$d^{(n)}(x^n, \hat{x}^n) = \frac{1}{n} \sum_{i=1}^{n} d(x_i, \hat{x}_i) \ . \tag{2.5}$$

The rate-distortion function is defined as the minimum rate $R$ such that the average distortion between the remote source sequence and its reconstruction does not exceed $D$, as there exists a blocklength $n$, an encoding function (2.2) and a decoding function (2.3).

**Remote Source Coding Under Logarithmic Loss**

Here we consider the remote source coding problem in which the distortion measure is chosen as the logarithmic loss.

Let $\zeta(y) = Q(\cdot|y) \in \mathcal{P}(\mathcal{X})$ for every $y \in \mathcal{Y}$. It is easy to see that

$$\begin{aligned}
\mathbb{E}_{P_{X,Y}}\left[l_{\log}(X,Q)\right] &= \sum_x \sum_y P_{X,Y}(x,y) \log \frac{1}{Q(x|y)} \\
&= \sum_x \sum_y P_{X,Y}(x,y) \log \frac{1}{P_{X|Y}(x|y)} + \sum_x \sum_y P_{X,Y}(x,y) \log \frac{P_{X|Y}(x|y)}{Q(x|y)} \\
&= H(X|Y) + D_{\mathrm{KL}}(P_{Y|X} \| Q) \\
&\geq H(X|Y) \;, \qquad (2.6)
\end{aligned}$$

with equality if and only of $\zeta(Y) = P_{X|Y}(\cdot|y)$.

Now let the stochastic mapping $\phi^{(n)} : \mathcal{Y}^n \to \mathcal{U}^n$ be the encoder, i.e., $\|\phi^{(n)}\| \leq nR$ for some prescribed complexity value $R$. Then, $U^n = \phi^{(n)}(X^n)$. Also, let the stochastic mapping $\psi^{(n)} : \mathcal{U}^n \to \mathcal{X}^n$ be the decoder. Thus, the expected logarithmic loss can be written as

$$D \overset{(a)}{\geq} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{P_{X,Y}}\left[l_{\log}(Y, \psi(U))\right] \overset{(b)}{\geq} H(X|U) \;, \qquad (2.7)$$

where $(a)$ follows from (2.4) and (2.5), and $(b)$ follows due to (2.6).

Hence, the rate-distortion of the remote source coding problem under logarithmic loss is given by the union of all pairs $(R, D)$ that satisfy

$$\begin{aligned}
R &\geq I(U;Y) \\
D &\geq H(X|U) \;,
\end{aligned} \qquad (2.8)$$

where the union is over all auxiliary random variables $U$ that satisfy the Markov chain $U -\!\!\circ\!\!- Y -\!\!\circ\!\!- X$. Also, using the substitution $\Delta := H(X) - D$, the region can be written equivalently as the union of all pairs $(R, \Delta)$ that satisfy

$$\begin{aligned}
R &\geq I(U;Y) \\
\Delta &\leq I(U;X) \;.
\end{aligned} \qquad (2.9)$$

This gives a clear connection between the remote source coding problem under logarithmic and the Information Bottleneck problem, which will be explained in the next section.

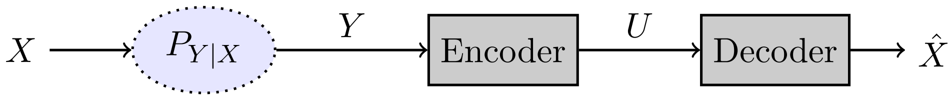



Figure 2.2: Information Bottleneck problem.

## 2.3 Information Bottleneck Problem

Tishby *et al.* in [17] present the Information Bottleneck (IB) framework, which can be considered as a remote source coding problem in which the distortion measure is logarithmic loss. By the choice of distortion metric as the logarithmic loss defined in (2.1), the connection of the rate-distortion problem with the IB is studied in [18, 52, 69]. Next, we explain the IB problem for the discrete memoryless and Gaussian cases.

### 2.3.1 Discrete Memoryless Case

The IB method depicted in Figure 2.2 formulates the problem of extracting the relevant information that a random variable $Y \in \mathcal{Y}$ captures about another one $X \in \mathcal{X}$ such that finding a representation $U$ that is maximally informative about $X$ (i.e., large mutual information $I(U;X)$), meanwhile minimally informative about $Y$ (i.e., small mutual information $I(U;Y)$). The term $I(U;X)$ is referred as *relevance* and $I(U;Y)$ is referred as *complexity*. Finding the representation $U$ that maximizes $I(U;X)$ while keeping $I(U;Y)$ smaller than a prescribed threshold can be formulated as the following optimization problem

$$\Delta(R) := \max_{P_{U|Y}\,:\, I(U;Y)\leq R} \quad I(U;X)\ . \tag{2.10}$$

Optimizing (2.10) is equivalent to solving the following Lagrangian problem

$$\mathcal{L}_s^{\mathrm{IB}} \ : \ \max_{P_{U|Y}} \quad I(U;X) - sI(U;Y)\ , \tag{2.11}$$

where $\mathcal{L}_s^{\mathrm{IB}}$ can be called as the IB objective, and $s$ designates the Lagrange multiplier.

For a known joint distribution $P_{X,Y}$ and a given trade-off parameter $s \geq 0$, the optimal mapping $P_{U|Y}$ can be found by solving the Lagrangian formulation (2.11). As shown in [17, Theorem 4], the optimal solution for the IB problem satisfies the self-consistent equations

$$p(u|y) = p(u)\frac{\exp[-D_{\mathrm{KL}}(P_{X|y}\|P_{X|u})]}{\sum_u p(u)\exp[-D_{\mathrm{KL}}(P_{X|y}\|P_{X|u})]} \tag{2.12a}$$

$$p(u) = \sum_{y} p(u|y)p(y) \tag{2.12b}$$

$$p(x|u) = \sum_{y} p(x|y)p(y|u) = \sum_{y} p(x,y)\frac{p(u|y)}{p(u)} \ . \tag{2.12c}$$

The self consistent equations in (2.12) can be iterated, similar to Blahut-Arimoto algorithm[1], for finding the optimal mapping $P_{U|Y}$ which maximizes the IB objective in (2.11). To do so, first $P_{U|Y}$ is initialized randomly, and then self-consistent equations (2.12) are iterated until convergence. This process is summarized hereafter as

$$P_{U|Y}^{(0)} \to P_{U}^{(1)} \to P_{X|U}^{(1)} \to P_{U|Y}^{(1)} \to \ldots \to P_{U}^{(t)} \to P_{X|U}^{(t)} \to P_{U|Y}^{(t)} \to \ldots \to P_{U|Y}^{\star} \ .$$

### 2.3.2 Gaussian Case

Chechik *et al.* in [21] study the *Gaussian Information Bottleneck* problem (see also [22, 70, 71]), in which the pair $(\mathbf{X}, \mathbf{Y})$ is jointly multivariate Gaussian variables of dimensions $n_x, n_y$. Let $\mathbf{\Sigma_x}, \mathbf{\Sigma_y}$ denote the covariance matrices of $\mathbf{X}, \mathbf{Y}$; and let $\mathbf{\Sigma_{x,y}}$ denote their cross-covariance matrix.

It is shown in [21,22,70] that if $\mathbf{X}$ and $\mathbf{Y}$ are jointly Gaussian, the optimal representation $\mathbf{U}$ is the linear transformation of $\mathbf{Y}$ and jointly Gaussian with $\mathbf{Y}$ [2]. Hence, we have

$$\mathbf{U} = \mathbf{A}\mathbf{Y} + \mathbf{Z} \ , \qquad \mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma_z}) \ . \tag{2.13}$$

Thus, $\mathbf{U} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma_u})$ with $\mathbf{\Sigma_u} = \mathbf{A}\mathbf{\Sigma_y}\mathbf{A}^\dagger + \mathbf{\Sigma_z}$.

The Gaussian IB curve defines the optimal trade-off between compression and preserved relevant information, and is known to have an analytical closed form solution. For a given trade-off parameter $s$, the parameters of the optimal projection of the Gaussian IB

[1] Blahut-Arimoto algorithm [19, 20] is originally developed for computation of the channel capacity and the rate-distortion function, and for these cases it is known to converge to the optimal solution. These iterative algorithms can be generalized to many other situations, e.g., including the IB problem. However, it only converges to stationary points in the context of IB.

[2] One of the main contribution of this thesis is the generalization of this result to the distributed case. The distributed Gaussian IB problem can be considered as the vector Gaussian CEO problem that we study in Chapter 4. In Theorem 4, we show that the optimal test channels are Gaussian when the sources are jointly multivariate Gaussian variables.

problem is found in [21, Theorem 3.1], and given by $\mathbf{\Sigma_z} = \mathbf{I}$ and

$$\mathbf{A} = \begin{cases} \left[\ \mathbf{0}^\dagger\ ;\ \mathbf{0}^\dagger\ ;\ \mathbf{0}^\dagger\ ;\ \ldots\ ;\ \mathbf{0}^\dagger\ \right] & 0 \leq s \leq \beta_1^{\mathrm{c}} \\ \left[\ \alpha_1 \mathbf{v}_1^\dagger\ ;\ \mathbf{0}^\dagger\ ;\ \mathbf{0}^\dagger\ ;\ \ldots\ ;\ \mathbf{0}^\dagger\ \right] & \beta_1^{\mathrm{c}} \leq s \leq \beta_2^{\mathrm{c}} \\ \left[\ \alpha_1 \mathbf{v}_1^\dagger\ ;\ \alpha_2 \mathbf{v}_2^\dagger\ ;\ \mathbf{0}^\dagger\ ;\ \ldots\ ;\ \mathbf{0}^\dagger\ \right] & \beta_2^{\mathrm{c}} \leq s \leq \beta_3^{\mathrm{c}} \\ \qquad\qquad \vdots & \qquad \vdots \end{cases}, \tag{2.14}$$

where $\{\mathbf{v}_1^\dagger, \ldots, \mathbf{v}_{n_y}^\dagger\}$ are the left eigenvectors of $\mathbf{\Sigma_{y|x}} \mathbf{\Sigma_y}^{-1}$ sorted by their corresponding ascending eigenvalues $\lambda_1, \ldots, \lambda_{n_y}$; $\beta_i^{\mathrm{c}} = \frac{1}{1-\lambda_i}$ are critical $s$ values; $\alpha_i$ are coefficients defined by $\alpha_i = \sqrt{\frac{s(1-\lambda_i)-1}{\lambda_i \mathbf{v}_i^\dagger \mathbf{\Sigma_y} \mathbf{v}_i}}$; $\mathbf{0}^\dagger$ is an $n_y$ dimensional row vectors of zeros; and semicolons separate rows in the matrix $\mathbf{A}$.

Alternatively, we can use a BA-type iterative algorithm to find the optimal relevance-complexity tuples. By doing so, we leverage on the optimality of Gaussian test channel, to restrict the optimization of $P_{\mathbf{U}|\mathbf{Y}}$ to Gaussian distributions, which are represented by parameters, namely its mean and covariance (e.g., $\mathbf{A}$ and $\mathbf{\Sigma_z}$). For a given trade-off parameter $s$, the optimal representation can be found by finding its representing parameters iterating over the following update rules

$$\mathbf{\Sigma}_{\mathbf{z}^{t+1}} = \left( \mathbf{\Sigma}^{-1}_{\mathbf{u}^t|x} - \frac{(s-1)}{s} \mathbf{\Sigma}^{-1}_{\mathbf{u}^t} \right)^{-1} \tag{2.15a}$$

$$\mathbf{A}^{t+1} = \mathbf{\Sigma}_{\mathbf{z}^{t+1}} \mathbf{\Sigma}^{-1}_{\mathbf{u}^t|x} \mathbf{A}^t \left( \mathbf{I} - \mathbf{\Sigma_{x|y}} \mathbf{\Sigma_y}^{-1} \right) . \tag{2.15b}$$

### 2.3.3 Connections

In this section, we review some interesting information theoretic connections that were reported originally in [72]. For instance, it is shown that the IB problem has strong connections with the problems of common reconstruction, information combining, the Wyner-Ahlswede-Körner problem and the privacy funnel problem.

#### Common Reconstruction

Here we consider the source coding problem with side information at the decoder, also called the Wyner-Ziv problem [73], under logarithmic loss distortion measure. Specifically, an encoder observes a memoryless source $Y$ and communicates with a decoder over a rate-constrained noise-free link. The decoder also observes a statistically correlated side information $X$. The encoder uses $R$ bits per sample to describe its observation $Y$ to the decoder. The decoder wants to reconstruct an estimate of $Y$ to within a prescribed fidelity level $D$. For the general distortion metric, the rate-distortion function of the Wyner-Ziv problem is given by

$$R^{\mathrm{WZ}}_{Y|X}(D) = \min_{P_{U|Y}\,:\,\mathbb{E}[d(Y,\psi(U,X))]\leq D} I(U;Y|X)\ , \tag{2.16}$$

where $\psi : \mathcal{U} \times \mathcal{X} \to \hat{\mathcal{Y}}$ is the decoding mapping.

The optimal coding coding scheme utilizes standard Wyner-Ziv compression at the encoder, and the decoding mapping $\psi$ is given by

$$\psi(U,X) = \Pr[Y = y|U,X]\ . \tag{2.17}$$

Then, note that with such a decoding mapping we have

$$\mathbb{E}[l_{\log}(Y,\psi(U,X))] = H(Y|U,X)\ . \tag{2.18}$$

Now we look at the source coding problem under the requirement such that the encoder is able to produce an exact copy of the compressed source constructed by the decoder. This requirement, termed as *common reconstruction* (CR), is introduced and studied by Steinberg in [35] for various source coding models, including Wyner-Ziv setup under a general distortion measure. For the Wyner-Ziv problem under logarithmic loss, such a common reconstruction constraint causes some rate loss because the reproduction rule (2.17) is not possible anymore. The Wyner-Ziv problem under logarithmic loss with common reconstruction constraint can be written as follows

$$R^{\mathrm{CR}}_{Y|X}(D) = \min_{P_{U|Y}\,:\,H(Y|U)\leq D} I(U;Y|X)\ , \tag{2.19}$$

for some auxiliary random variable $U$ for which the Markov chain $U -\!\!\circ\!\!- Y -\!\!\circ\!\!- X$ holds. Due to this Markov chain, we have $I(U;Y|X) = I(U;Y) - I(U;X)$. Besides, observe that the constrain $H(Y|U) \leq D$ is equivalent to $I(U;Y) \geq H(Y) - D$. Then, we can rewrite (2.19) as

$$R^{\mathrm{CR}}_{Y|X}(D) = \min_{P_{U|Y}\,:\,I(U;Y)\geq H(Y)-D} I(U;Y) - I(U;X)\ . \tag{2.20}$$

Under the constraint $I(U;Y) = H(Y) - D$, minimizing $I(U;Y|X)$ is equivalent to maximizing $I(U;X)$, which connects the problem of CR readily with the IB.

In the above, the side information $X$ is used for binning but not for the estimation at the decoder. If the encoder ignores whether $X$ is present at the decoder, the benefit of binning is reduced – see the Heegard-Berger model with CR [74, 75].

**Information Combining**

Here we consider the IB problem, in which one seeks to find a suitable representation $U$ that maximizes the relevance $I(U;X)$ for a given prescribed complexity level, e.g., $I(U;Y) = R$. For this setup, we have

$$\begin{aligned} I(Y;U,X) &= I(Y;U) + I(Y;X|U) \\ &= I(Y;U) + I(X;Y,U) - I(X;U) \\ &\stackrel{(a)}{=} I(Y;U) + I(X;Y) - I(X;U) \end{aligned} \tag{2.21}$$

where $(a)$ holds due the Markov chain $U -\!\!\circ\!\!- Y -\!\!\circ\!\!- X$. Hence, in the IB problem (2.11), for a given complexity level, e.g., $I(U;Y) = R$, maximizing the relevance $I(U;X)$ is equivalent of minimizing $I(Y;U,X)$. This is reminiscent of the problem of *information combining* [36–38], where $Y$ can be interpreted as a source transferred through two channels $P_{U|Y}$ and $P_{X|Y}$. The outputs of these two channels are conditionally independent given $Y$; and they should be processed in a manner such that, when combined, they capture as much as information about $Y$.

**Wyner-Ahlswede-Körner Problem**

In the Wyner-Ahlswede-Körner problem, two memoryless sources $X$ and $Y$ are compressed separately at rates $R_X$ and $R_Y$, respectively. A decoder gets the two compressed streams and aims at recovering $X$ in a lossless manner. This problem was solved independently by Wyner in [39] and Ahlswede and Körner in [40]. For a given $R_Y = R$, the minimum rate $R_X$ that is needed to recover $X$ losslessly is given as follows

$$R_X^\star(R) = \min_{P_{U|Y}\,:\,I(U;Y)\leq R} H(X|U)\ . \tag{2.22}$$

Hence, the connection of Wyner-Ahlswede-Körner problem (2.22) with the IB (2.10) can be written as

$$\Delta(R) = \max_{P_{U|Y}\,:\,I(U;Y)\leq R} I(U;X) = H(X) + R_X^\star(R)\ . \tag{2.23}$$

**Privacy Funnel Problem**

Consider the pair $(X,Y)$ where $X \in \mathcal{X}$ be the random variable representing the private (or sensitive) data that is not meant to be revealed at all, or else not beyond some level $\Delta$;

and $Y \in \mathcal{Y}$ be the random variable representing the non-private (or nonsensitive) data that is shared with another user (data analyst). Assume that $X$ and $Y$ are correlated, and this correlation is captured by the joint distribution $P_{X,Y}$. Due to this correlation, releasing data $Y$ is directly to the data analyst may cause that the analyst can draw some information about the private data $X$. Therefore, there is a trade-off between the amount of information that the user keeps private about $X$ and shares about $Y$. The aim is to find a mapping $\phi : \mathcal{Y} \to \mathcal{U}$ such that $U = \phi(Y)$ is maximally informative about $Y$, meanwhile minimally informative about $X$.

The analyst performs an adversarial inference attack on the private data $X$ from the disclosed data $U$. For a given arbitrary distortion metric $d : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}_+$ and the joint distribution $P_{X,Y}$, the average inference cost gain by the analyst after observing $U$ can be written as

$$\Delta C(d, P_{X,Y}) := \inf_{\hat{x} \in \hat{\mathcal{X}}} \mathbb{E}_{P_{X,Y}}[d(X, \hat{x})] - \inf_{\hat{X}(\phi(Y))} \mathbb{E}_{P_{X,Y}}[d(X, \hat{X})|U] \,. \tag{2.24}$$

The quantity $\Delta C$ was proposed as a general privacy metric in [76], since it measures the improvement in the quality of the inference of the private data $X$ due to the observation $U$. In [42] (see also [77]), it is shown that for any distortion metric $d$, the inference cost gain $\Delta C$ can be upper bounded as

$$\Delta C(d, P_{X,Y}) \leq 2\sqrt{2} L \sqrt{I(U;X)} \,, \tag{2.25}$$

where $L$ is a constant. This justifies the use of the logarithmic loss as a privacy metric since the threat under any bounded distortion metric can be upper bounded by an explicit constant factor of the mutual information between the private and disclosed data. With the choice of logarithmic loss, we have

$$I(U;X) = H(X) - \inf_{\hat{X}(U)} \mathbb{E}_{P_{X,Y}}[l_{\log}(X, \hat{X})] \,. \tag{2.26}$$

Under the logarithmic loss function, the design of the mapping $U = \phi(Y)$ should strike a right balance between the utility for inferring the non-private data $Y$ as measured by the mutual information $I(U;Y)$ and the privacy threat about the private data $X$ as measured by the mutual information $I(U;X)$. That is refereed as the *privacy funnel* method [42], and can be formulated as the following optimization

$$\min_{P_{U|Y} \,:\, I(U;Y) \geq R} \quad I(U;X) \,. \tag{2.27}$$

Notice that this is an opposite optimization to the Information Bottleneck (2.10).

## 2.4 Learning via Information Bottleneck

### 2.4.1 Representation Learning

The performance of learning algorithms highly depends on the characteristics and properties of the data (or features) on which the algorithms are applied. Due to this fact, feature engineering, i.e., preprocessing operations – that may include sanitization and transferring the data on another space – is very important to obtain good results from the learning algorithms. On the other hand, since these preprocessing operations are both task- and data-dependent, feature engineering is high labor-demanding and this is one of the main drawbacks of the learning algorithms. Despite the fact that it can be sometimes considered as helpful to use feature engineering in order to take advantage of human know-how and knowledge on the data itself, it is highly desirable to make learning algorithms less dependent on feature engineering to make progress towards true artificial intelligence.

Representation learning [43] is a sub-field of learning theory which aims at learning representations by extracting some useful information from the data, possibly without using any resources of feature engineering. Learning good representations aims at disentangling the underlying explanatory factors which are hidden in the observed data. It may also be useful to extract expressive low-dimensional representations from high-dimensional observed data. The theory behind the elegant IB method may provide a better understanding of the representation learning.

Consider a setting in which for a given data $\mathbf{Y}$ we want to find a representation $\mathbf{U}$, which is a function of $\mathbf{Y}$ (possibly non-deterministic) such that $\mathbf{U}$ preserves some desirable information regarding to a task $\mathbf{X}$ in view of the fact that the representation $\mathbf{U}$ is more convenient to work or expose relevant statistics.

Optimally, the representation should be as good as the original data for the task, however, should not contain the parts that are irrelevant to the task. This is equivalent finding a representation $\mathbf{U}$ satisfying the following criteria [78]:

($i$) $\mathbf{U}$ is a function of $\mathbf{Y}$, the Markov chain $\mathbf{X} -\!\!\circ\!\!- \mathbf{Y} -\!\!\circ\!\!- \mathbf{U}$ holds.

($ii$) $\mathbf{U}$ is *sufficient* for the task $\mathbf{X}$, that means $I(\mathbf{U};\mathbf{X}) = I(\mathbf{Y};\mathbf{X})$.

($iii$) $\mathbf{U}$ discards all variability in $\mathbf{Y}$ that is not relevant to task $\mathbf{X}$, i.e., *minimal* $I(\mathbf{U};\mathbf{Y})$.

Besides, ($ii$) is equivalent to $I(\mathbf{Y};\mathbf{X}|\mathbf{U}) = 0$ due to the Markov chain in ($i$). Then, the optimal representation $\mathbf{U}$ satisfying the conditions above can be found by solving the following optimization

$$\min_{P_{\mathbf{U}|\mathbf{Y}}\,:\,I(\mathbf{Y};\mathbf{X}|\mathbf{U})=0} \quad I(\mathbf{U};\mathbf{Y})\ . \tag{2.28}$$

However, (2.28) is very hard to solve due to the constrain $I(\mathbf{Y};\mathbf{X}|\mathbf{U}) = 0$. Tishby's IB method solves (2.28) by relaxing the constraint as $I(\mathbf{U};\mathbf{X}) \geq \Delta$, which stands for that the representation $\mathbf{U}$ contains relevant information regarding the task $\mathbf{X}$ larger than a threshold $\Delta$. Eventually, (2.28) boils down to minimizing the following Lagrangian

$$\begin{aligned} &\min_{P_{\mathbf{U}|\mathbf{Y}}} H(\mathbf{X}|\mathbf{U}) + sI(\mathbf{U};\mathbf{Y}) && \text{(2.29a)}\\ &\quad = \min_{P_{\mathbf{U}|\mathbf{Y}}} \mathbb{E}_{P_{\mathbf{X},\mathbf{Y}}}\Big[\mathbb{E}_{P_{\mathbf{U}|\mathbf{Y}}}[-\log P_{\mathbf{X}|\mathbf{U}}] + sD_{\mathrm{KL}}(P_{\mathbf{U}|\mathbf{Y}}\|P_{\mathbf{U}})\Big]\ . && \text{(2.29b)}\end{aligned}$$

In representation learning, *disentanglement of hidden factors* is also desirable in addition to *sufficiency* (*ii*) and *minimality* (*iii*) properties. The disentanglement can be measured with the *total correlation* (TC) [79, 80], defined as

$$\mathrm{TC}(\mathbf{U}) := D_{\mathrm{KL}}(P_{\mathbf{U}}\|\prod_j P_{U_j})\ , \tag{2.30}$$

where $U_j$ denotes the $j$-th component of $\mathbf{U}$, and $\mathrm{TC}(\mathbf{U}) = 0$ when the components of $\mathbf{U}$ are independent.

In order to obtain a more disentangled representation, we add (2.30) as a penalty in (2.29). Then, we have

$$\min_{P_{\mathbf{U}|\mathbf{Y}}} \mathbb{E}_{P_{\mathbf{X},\mathbf{Y}}}\Big[\mathbb{E}_{P_{\mathbf{U}|\mathbf{Y}}}[-\log P_{\mathbf{X}|\mathbf{U}}] + sD_{\mathrm{KL}}(P_{\mathbf{U}|\mathbf{Y}}\|P_{\mathbf{U}})\Big] + \beta D_{\mathrm{KL}}\Big(P_{\mathbf{U}}\|\prod_j P_{U_j}\Big)\ , \tag{2.31}$$

where $\beta$ is the Lagrangian for TC constraint (2.30). For the case in which $\beta = s$, it is easy to see that the minimization (2.31) is equivalent to

$$\min_{P_{\mathbf{U}|\mathbf{Y}}} \mathbb{E}_{P_{\mathbf{X},\mathbf{Y}}}\left[\mathbb{E}_{P_{\mathbf{U}|\mathbf{Y}}}[-\log P_{\mathbf{X}|\mathbf{U}}] + sD_{\mathrm{KL}}\Big(P_{\mathbf{U}|\mathbf{Y}}\|\prod_j P_{U_j}\Big)\right]\ . \tag{2.32}$$

In other saying, optimizing the original IB problem (2.29) with the assumption of independent representations, i.e., $P_{\mathbf{U}} = \prod_j P_{U_j}(u_j)$, is equivalent forcing representations to be more disentangled. Interestingly, we note that this assumption is already adopted for the simplicity in many machine learning applications.

### 2.4.2 Variational Bound

The optimization of the IB cost (2.11) is generally computationally challenging. In the case in which the true distribution of the source pair is known, there are two notable exceptions explained in Chapter 2.3.1 and 2.3.2: the source pair $(X, Y)$ is discrete memoryless [17] and the multivariate Gaussian [21, 22]. Nevertheless, these assumptions on the distribution of the source pair severely constrain the class of learnable models. In general, only a set of training samples $\{(x_i, y_i)\}_{i=1}^{n}$ is available, which makes the optimization of the original IB cost (2.11) intractable. To overcome this issue, Alemi *et al.* in [30] present a variational bound on the IB objective (2.11), which also enables a neural network reparameterization for the IB problem, which will be explained in Chapter 2.4.4.

For the variational distribution $Q_U$ on $\mathcal{U}$ (instead of unknown $P_U$), and a variational stochastic decoder $Q_{X|U}$ (instead of the unknown optimal decoder $P_{X|U}$), let define $\mathbf{Q} := \{Q_{X|U}, Q_U\}$. Besides, for convenience let $\mathbf{P} := \{P_{U|Y}\}$. We define the variational IB cost $\mathcal{L}_s^{\mathrm{VIB}}(\mathbf{P}, \mathbf{Q})$ as

$$\mathcal{L}_s^{\mathrm{VIB}}(\mathbf{P}, \mathbf{Q}) := \mathbb{E}_{P_{X,Y}}\Big[\mathbb{E}_{P_{U|Y}}[\log Q_{X|U}] - sD_{\mathrm{KL}}(P_{U|Y}\|Q_U)\Big] \,. \tag{2.33}$$

Besides, we note that maximizing $\mathcal{L}_s^{\mathrm{IB}}$ in (2.11) over $\mathbf{P}$ is equivalent to maximizing

$$\tilde{\mathcal{L}}_s^{\mathrm{IB}}(\mathbf{P}) := -H(X|U) - sI(U;Y) \,. \tag{2.34}$$

Next lemma states that $\mathcal{L}_s^{\mathrm{VIB}}(\mathbf{P}, \mathbf{Q})$ is a lower bound on $\tilde{\mathcal{L}}_s^{\mathrm{IB}}(\mathbf{P})$ for all distributions $\mathbf{Q}$.

**Lemma 1.**

$$\mathcal{L}_s^{\mathrm{VIB}}(\mathbf{P}, \mathbf{Q}) \le \tilde{\mathcal{L}}_s^{\mathrm{IB}}(\mathbf{P}) \,, \quad \textit{for all pmfs } \mathbf{Q} \,.$$

*In addition, there exists a unique* $\mathbf{Q}$ *that achieves the maximum* $\max_{\mathbf{Q}} \mathcal{L}_s^{\mathrm{VIB}}(\mathbf{P}, \mathbf{Q}) = \tilde{\mathcal{L}}_s^{\mathrm{IB}}(\mathbf{P})$*, and is given by*

$$Q^*_{X|U} = P_{X|U} \,, \quad Q^*_U = P_U \,. \qquad \blacksquare$$

Using Lemma 1, the optimization in (2.11) can be written in term of the variational IB cost as follows

$$\max_{\mathbf{P}} \mathcal{L}_s^{\mathrm{IB}}(\mathbf{P}) = \max_{\mathbf{P}} \max_{\mathbf{Q}} \mathcal{L}_s^{\mathrm{VIB}}(\mathbf{P}, \mathbf{Q}) \,. \tag{2.35}$$

### 2.4.3 Finite-Sample Bound on the Generalization Gap

The IB method requires that the joint distribution $P_{X,Y}$ is known, although this is not the case for most of the time. In fact, there is only access to a finite sample, e.g., $\{(x_i, y_i)\}_{i=1}^n$. The generalization gap is defined as the difference between the empirical risk (average risk over a finite training sample) and the population risk (average risk over the true joint distribution).

It has been shown in [81], and revisited in [82], that it is possible to generalize the IB as a learning objective for finite samples in the course of bounded representation complexity (e.g., the cardinality of $U$). In the following, $\hat{I}(\cdot\,;\,\cdot)$ denotes the empirical estimate of the mutual information based on finite sample distribution $\hat{P}_{X,Y}$ for a given sample size of $n$. In [81, Theorem 1], a finite-sample bound on the generalization gap is provided, and we state it below.

Let $U$ be a fixed probabilistic function of $Y$, determined by a fixed and known conditional probability $P_{U|Y}$. Also, let $\{(x_i, y_i)\}_{i=1}^n$ be samples of size $n$ drawn from the joint probability distribution $P_{X,Y}$. For given $\{(x_i, y_i)\}_{i=1}^n$ and any confidence parameter $\delta \in (0, 1)$, the following bounds hold with a probability of at least $1 - \delta$,

$$|I(U;Y) - \hat{I}(U;Y)| \leq \frac{(|\mathcal{U}| \log n + \log |\mathcal{U}|)\sqrt{\log \frac{4}{\delta}}}{\sqrt{2n}} + \frac{|\mathcal{U}| - 1}{n} \tag{2.36a}$$

$$|I(U;X) - \hat{I}(U;X)| \leq \frac{(3|\mathcal{U}| + 2) \log n \sqrt{\log \frac{4}{\delta}}}{\sqrt{2n}} + \frac{(|\mathcal{X}| + 1)(|\mathcal{U}| + 1) - 4}{n}\,. \tag{2.36b}$$

Observe that the generalization gaps decreases when the cardinality of representation $U$ get smaller. This means the optimal IB curve can be well estimated if the representation space has a simple model, e.g., $|\mathcal{U}|$ is small. On the other hand, the optimal IB curve is estimated badly for learning complex representations. It is also observed that the bounds does not depend on the cardinality of $Y$. Besides, as expected for larger sample size $n$ of the training data, the optimal IB curve is estimated better.

### 2.4.4 Neural Reparameterization

The aforementioned BA-type algorithms works for the cases in which the joint distribution of the data pair $P_{\mathbf{X},\mathbf{Y}}$ is known. However, this is a very tight constraint which is very unusual to meet, especially for real-life applications. Here we explain the neural reparameterization and evolve the IB method to a learning algorithm to be able to use it with real datasets.

Let $P_\theta(\mathbf{u}|\mathbf{y})$ denote the encoding mapping from the observation $\mathbf{Y}$ to the bottleneck representation $\mathbf{U}$, parameterized by a DNN $f_\theta$ with parameters $\theta$ (e.g., the weights and biases of the DNN). Similarly, let $Q_\phi(\mathbf{x}|\mathbf{u})$ denote the decoding mapping from the representation $\mathbf{U}$ to the reconstruction of the label $\mathbf{Y}$, parameterized by a DNN $g_\phi$ with parameters $\phi$. Furthermore, let $Q_\psi(\mathbf{u})$ denote the prior distribution of the latent space, which does not depend on a DNN. By using this neural reparameterization of the encoder $P_\theta(\mathbf{u}|\mathbf{y})$, decoder $Q_\phi(\mathbf{x}|\mathbf{u})$ and prior $Q_\psi(\mathbf{u})$, the optimization in (2.35) can be written as

$$\max_{\theta,\phi,\psi} \quad \mathbb{E}_{P_{\mathbf{X},\mathbf{Y}}}\left[\mathbb{E}_{P_\theta(\mathbf{U}|\mathbf{Y})}[\log Q_\phi(\mathbf{X}|\mathbf{U})] - sD_{\mathrm{KL}}(P_\theta(\mathbf{U}|\mathbf{Y})\|Q_\psi(\mathbf{U}))\right] \ . \tag{2.37}$$

Then, for a given dataset consists of $n$ samples, i.e., $\mathcal{D} := \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$, the optimization of (2.37) can be approximated in terms of an empirical cost as follows

$$\max_{\theta,\phi,\psi} \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_{s,i}^{\mathrm{emp}}(\theta,\phi,\psi) \ , \tag{2.38}$$

where $\mathcal{L}_{s,i}^{\mathrm{emp}}(\theta,\phi,\psi)$ is the empirical IB cost for the $i$-th sample of the training set $\mathcal{D}$, and given by

$$\mathcal{L}_{s,i}^{\mathrm{emp}}(\theta,\phi,\psi) = \mathbb{E}_{P_\theta(\mathbf{U}_i|\mathbf{Y}_i)}[\log Q_\phi(\mathbf{X}_i|\mathbf{U}_i)] - sD_{\mathrm{KL}}(P_\theta(\mathbf{U}_i|\mathbf{Y}_i)\|Q_\psi(\mathbf{U}_i)) \ . \tag{2.39}$$

Now, we investigate the possible choices of the parametric distributions. The encoder can be chosen as a multivariate Gaussian, i.e., $P_\theta(\mathbf{u}|\mathbf{y}) = \mathcal{N}(\mathbf{u};\boldsymbol{\mu}_\theta,\boldsymbol{\Sigma}_\theta)$. So, it can be modeled with a DNN $f_\theta$, which maps the observation $\mathbf{y}$ to the parameters of a multivariate Gaussian, namely the mean $\boldsymbol{\mu}_\theta$ and the covariance $\boldsymbol{\Sigma}_\theta$, i.e., $(\boldsymbol{\mu}_\theta,\boldsymbol{\Sigma}_\theta) = f_\theta(\mathbf{y})$. The decoder $Q_\phi(\mathbf{x}|\mathbf{u})$ can be a categorical distribution parameterized by a DNN $f_\phi$ with a softmax operation in the last layer, which outputs the probabilities of dimension $|\mathcal{X}|$, i.e., $\hat{\mathbf{x}} = g_\phi(\mathbf{u})$. The prior of the latent space $Q_\psi(\mathbf{u})$ can be chosen as a multivariate Gaussian (e.g., $\mathcal{N}(\mathbf{0},\mathbf{I})$) such that the KL divergence $D_{\mathrm{KL}}(P_\theta(\mathbf{U}|\mathbf{Y})\|Q_\psi(\mathbf{U}))$ has a closed form solution and is easy to compute.

With the aforementioned choices, the first term of the RHS of (2.39) can be computed using Monte Carlo sampling and the reparameterization trick [29] as

$$\mathbb{E}_{P_\theta(\mathbf{U}_i|\mathbf{Y}_i)}[\log Q_\phi(\mathbf{X}_i|\mathbf{U}_i)] = \frac{1}{m}\sum_{j=1}^{m}\log Q_\phi(\mathbf{x}_i|\mathbf{u}_{i,j}) \ , \quad \mathbf{u}_{i,j} = \boldsymbol{\mu}_{\theta,i} + \boldsymbol{\Sigma}_{\theta,i}^{\frac{1}{2}}\cdot\boldsymbol{\epsilon}_j \ , \quad \boldsymbol{\epsilon}_j \sim \mathcal{N}(\mathbf{0},\mathbf{I}) \ ,$$

where $m$ is the number of samples for the Monte Carlo sampling step. The second term of the RHS of (2.39) – the KL divergence between two multivariate Gaussian distributions –

has a closed form. For convenience, in the specific case in which the covariance matrix is diagonal, i.e., $\mathbf{\Sigma}_{\theta,i} := \mathrm{diag}(\{\sigma^2_{\theta,i,k}\}_{k=1}^{n_u})$, with $n_u$ denoting the latent space dimension, the RHS of (2.39) can be computed as follows

$$\frac{1}{2}\sum_{k=1}^{n_u}\left[\mu_{\theta,i,k} - \log\sigma^2_{\theta,i,k} - 1 + \sigma^2_{\theta,i,k}\right] \ . \tag{2.40}$$

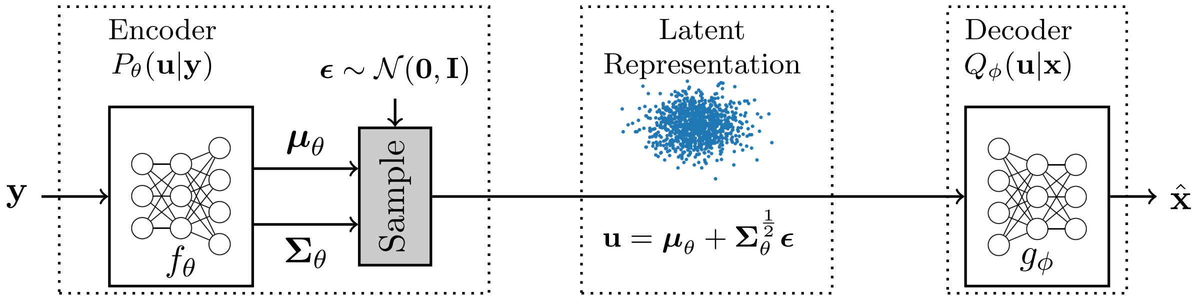



Figure 2.3: Representation learning.

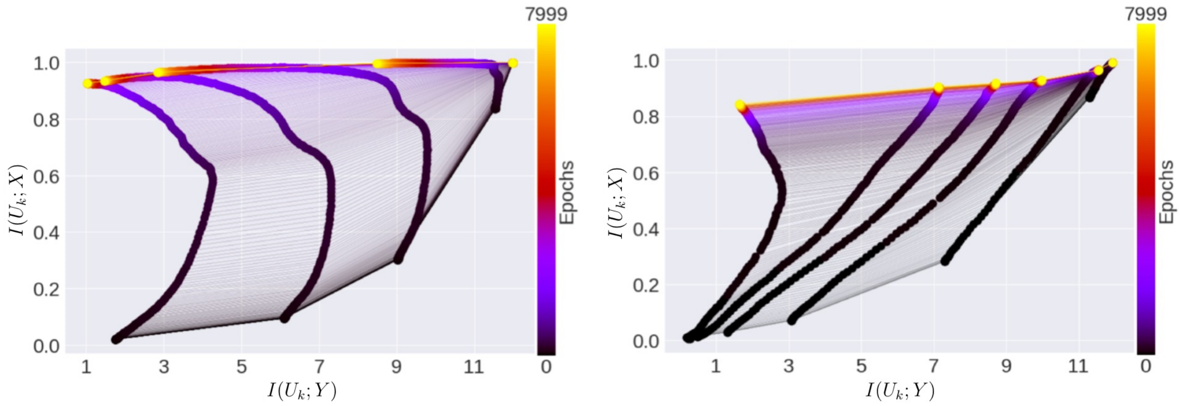
Altogether, we have the following cost to be trained over DNN parameters $\theta, \phi$ using stochastic gradient descent methods (e.g., SGD or ADAM [83]),

$$\max_{\theta,\phi} \quad \frac{1}{m}\sum_{j=1}^{m}\log Q_\phi(\mathbf{x}_i|\mathbf{u}_{i,j}) - \frac{s}{2}\sum_{k=1}^{n_u}\left[\mu_{\theta,i,k} - \log\sigma^2_{\theta,i,k} - 1 + \sigma^2_{\theta,i,k}\right] \ . \tag{2.41}$$

Note that, without loss of generality, the prior is fixed to $Q_\psi(\mathbf{u}) = \mathcal{N}(\mathbf{0},\mathbf{I})$, hence the optimization is not over the prior parameter $\psi$. So the VIB learning algorithm optimizes the DNN parameters for a given training dataset $\mathcal{D}$ and a parameter $s$. After the convergence of the parameters to $\theta^\star, \phi^\star$, the representation $\mathbf{U}$ can be inferred by sampling from the encoder $P_{\theta^\star}(\mathbf{U}|\mathbf{Y})$ and then the soft estimate of the target variable $\mathbf{X}$ can be calculated using the decoder $Q_{\phi^\star}(\mathbf{X}|\mathbf{U})$ for a new data $\mathbf{Y}$. An example of learning architecture which can be trained to minimize cost (2.41) using neural networks is shown in Figure 2.3.

### 2.4.5 Opening the Black Box

Learning algorithms using DNNs is getting more and more popular due to its remarkable success in many practical problems. However, it is not well studied how algorithms using DNNs improves the state of the art, and there is no rigorous understanding about what it is going inside of DNNs. Due to the lack of this understanding, the DNN is usually treated as a black box and integrated into various algorithms as a block in which it is not known exactly what it is going on. Schwartz-Ziv and Tishby in [84] (also Tishby and Zaslavsky in a preliminary work [82]) suggested to use an information-theoretical approach to 'open the black box', where the IB principle is used to explain theory of deep learning. In [84], it is proposed to analyze the *information plane* – where $I(U;X)$ versus $I(U;Y)$ is plotted – due to useful insights about the trade-off between prediction and compression.




(a) Tanh activation function. (b) ReLU activation function.

Figure 2.4: The evolution of the layers with the training epochs in the information plane. In the x-axis, the mutual information between each layer and the input, i.e., $I(U_k;Y)$, is plotted. In the y-axis, the mutual information between each layer and the label, i.e., $I(U_k;X)$, is plotted. The colors indicate training time in epochs. The curve on the far corresponds the mutual information with the output layer; and the curve on the far right corresponds the mutual information with the input layer. Figures are taken from [85].

Now consider a NN with $K$ layers and let $U_k$ be a random variable denoting the representation, which is the output of $k$-th hidden layer. Then, the Markov chain $X -\!\!\circ\!\!- Y -\!\!\circ\!\!- U_1 -\!\!\circ\!\!- \cdots -\!\!\circ\!\!- U_K -\!\!\circ\!\!- \hat{X}$ holds. In particular, a fully connected NN with 5 hidden layers with dimensions $12 - 10 - 7 - 5 - 4 - 3 - 2$ is trained using SGD to make a binary classification from a 12-dimensional input. All except the last layers are activated with the hyperbolic tangent function (tanh); and sigmoid function is used for the last (i.e., output) layer. In order to calculate the mutual information of layers with respect to input and output variables, neuron's tanh output activations are binned into 30 equal intervals between -1 and 1. Then, these discretized values in each layer is used to calculate the joint distributions $P_{U_i,Y}$ and $P_{U_i,X}$ over the $2^{12}$ equally likely input patterns and true output labels. Using these discrete joint distributions, the mutual informations $I(U_k;Y)$

and $I(U_k;X)$ are calculated, and depicted in Figure 2.4a. In Figure 2.4a, a transition is observed between an initial *fitting* phase and a subsequent *compression* phase. In the fitting phase, the relevance between representations in each layer and label (e.g., the mutual information $I(U_k;X)$) increases. The fitting phase is shorter, needs less epochs. During the compression phase, the mutual information between representations and the input, i.e., $I(U_k;Y)$, decreases.

In a recent work [85], Saxe *et al.* reports that these fitting and compression phases mentioned in [84] are not observed for all activation functions. To show that, the same experiment is repeated, however the tanh activations are interchanged with ReLU. The mutual information between each layer with the input $Y$ and the label $X$ over epochs is plotted in Figure 2.4b. It is observed that except the curve on the far left in Figure 2.4b which corresponds the output layer with sigmoid activation, the mutual information with the input monotonically increases in all ReLU layers, hence the compression phase is not visible here.

## 2.5 An Example Application: Text clustering

In this section, we present a deterministic annealing-like algorithm [32, Chapter 3.2], and also an application of it to the text clustering. The annealing-like IB is an algorithm which works by tuning the parameter $s$. First, we recall the IB objective

$$\mathcal{L}_s^{\mathrm{IB}} \ : \ \min_{P_{U|Y}} \quad I(U;Y) - sI(U;X) \ . \tag{2.42}$$

When $s \to 0$, the representation $U$ is designed with the most compact form, i.e., $|\mathcal{U}| = 1$, which corresponds the maximum compression. By gradually increasing the parameter $s$, the emphasization on the relevance term $I(U;X)$ increases, and at a critical value of $s$, the optimization focuses on not only the compression but also the relevance term. To fulfill the demand on the relevance term, this results that the cardinality of $U$ bifurcates. This is referred as a *phase transition* of the system. The further increases in the value of $s$ will cause other phase transitions, hence additional splits of $U$ until it reaches the desired level, e.g., $|\mathcal{U}| = |\mathcal{X}|$.

The main difficulty is how to identify these critical phase transition values of $s$. In [32], the following procedure offered for detecting phase transition values: At each step, the previous solution – which is found for the previous value of $s$ – is taken as an initialization; and each value of $U$ is duplicated. Let $u_1$ and $u_2$ be such duplicated values of $u$. Then,

$$\begin{aligned} p(u_1|y) &= p(u|y)\left(\frac{1}{2} + \alpha\,\hat{\epsilon}(u,y)\right) \\ p(u_2|y) &= p(u|y)\left(\frac{1}{2} - \alpha\,\hat{\epsilon}(u,y)\right), \end{aligned} \tag{2.43}$$

where $\hat{\epsilon}(u,x)$ is a random noise term uniformly selected in the range $[-1/2, 1/2]$ and $\alpha$ is a small scalar. Thus, the $p(u_1|y)$ and $p(u_2|y)$ is slightly perturbed values of $p(u|y)$. If these perturbed version of distributions are different enough, i.e., $D_{\mathrm{JS}}^{(\frac{1}{2},\frac{1}{2})}(P_{X|U_1}\|P_{X|U_2}) \geq \tau$, where $\tau$ is a threshold value and $D_{\mathrm{JS}}$ is the Jensen - Shannon divergence given by

$$D_{\mathrm{JS}}^{(\pi_1,\pi_2)}(P_X, Q_X) = \pi_1 D_{\mathrm{KL}}(P_X\|\tilde{P}_X) + \pi_2 D_{\mathrm{KL}}(Q_X\|\tilde{P}_X), \text{ where } \tilde{P}_X = \pi_1 P_X + \pi_2 Q_X\,, \tag{2.44}$$

the corresponding value of $s$ is a phase transition value and $u$ is splitted into $u_1$ and $u_2$. Otherwise, both perturbed values collapse to the same solution. Finally, the value of $s$ is increased and the whole procedure is repeated. This algorithm is called deterministic annealing IB and stated in Algorithm 1. We note that tuning $s$ parameter is very critical, such that the step size in update of $s$ should be chosen carefully, otherwise cluster splits (phase transitions) might be skipped.

**Algorithm 1** Deterministic annealing-like IB algorithm

1: **input:** pmf $P_{X,Y}$, parameters $\alpha, \tau, \epsilon_s$.
2: **output:** Optimal $P^{\star}_{U|Y}$. (soft partitions $U$ of $Y$ into $M$ clusters)
3: **initialization** Set $s \to 0$ and $|\mathcal{U}| = 1$, $p(u|y) = 1$, $\forall y \in \mathcal{Y}$.
4: **repeat**
5: Update $s$, $s = (1+\epsilon_s)s_{\mathrm{old}}$.
6: Duplicate clusters according to (2.43).
7: Apply IB algorithm by using iteration rules (2.12).
8: Check for splits. If $D_{\mathrm{JS}}^{(\frac{1}{2},\frac{1}{2})}(P_{X|U_1}\|P_{X|U_2}) \geq \tau$, then $\mathcal{U} \leftarrow \{\mathcal{U} \setminus \{u\}\} \cup \{u_1, u_2\}$.
9: **until** $|\mathcal{U}| \geq M$.

Now, we apply the annealing-like algorithm to the 20 newsgroups dataset for word clustering according to their topics. For convenience, we use a tiny version of 20 newsgroups dataset, in which the most informative 100 words selected which come from 4 different topics listed in Table 2.1. By using the the number of occurrences of words in topics, the joint probability $P_{X,Y}$ is calculated. With the choice of parameters $\alpha = 0.005, \epsilon_s = 0.001$

and $\tau = 1/s$, the annealing IB algorithm is run and Figure 2.5 shows the corresponding IB curve, as well as, the phase transitions. Besides, the resulting complexity-relevance pairs are plotted with the application of $K$-means algorithm for different number of clusters. The obtained clusters are given in Table 2.2.

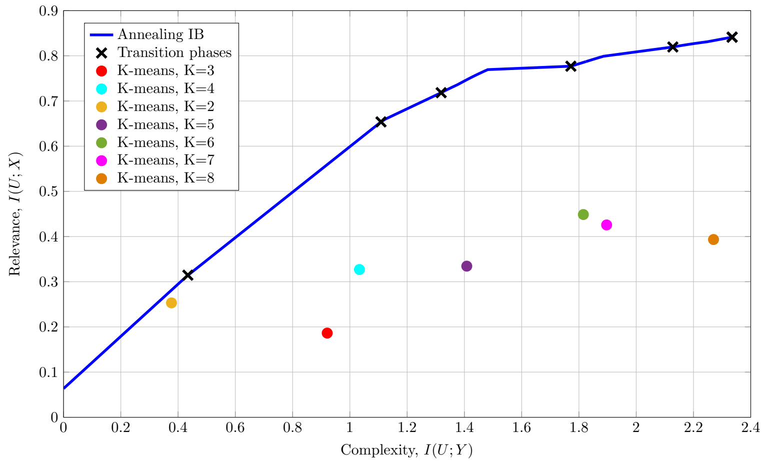



Figure 2.5: Annealing IB algorithm for text clustering.

| Topics | Sub-Topics |
|---|---|
| Group 1 (comp) | comp.graphics, comp.os.ms-windows.misc, comp.sys.ibm.pc.hardware, comp.sys.mac.hardware, comp.windows.x |
| Group 2 (rec) | rec.autos, rec.motorcycles, rec.sport.baseball, rec.sport.hockey |
| Group 3 (sci) | sci.crypt, sci.electronics, sci.med, sci.spacesci.space |
| Group 4 (talk) | talk.politics.misc, talk.politics.guns, talk.politics.mideast, talk.religion.misc |

Table 2.1: The topics of 100 words in the the subgroup of 20 newsgroup dataset.

| | Words |
|---|---|
| Cluster 1 | card, computer, data, disk, display, dos, drive, driver, email, files, format, ftp, graphics, help, image, mac, memory, number, pc, phone, problem, program, scsi, server, software, system, version, video, windows |
| Cluster 2 | baseball, bmw, car, engine, fans, games, hit, hockey, honda, league, nhl, players, puck, season, team, win, won |
| Cluster 3 | cancer, disease, doctor, insurance, launch, lunar, mars, medicine, mission, moon, msg, nasa, orbit, patients, research, satellite, science, shuttle, solar, space, studies, technology, vitamin |
| Cluster 4 | aids, bible, case, children, christian, course, dealer, earth, evidence, fact, food, god, government, gun, health, human, israel, jesus, jews, law, oil, power, president, question, religion, rights, state, university, war, water, world |

Table 2.2: Clusters obtained through the application of the annealing IB algorithm on the subgroup of 20 newsgroup dataset.

## 2.6 Design of Optimal Quantizers

The IB method has been used in many fields, and in this section we present an application in communications, which is an optimal quantizer design based on the IB method [86, 87]. The main idea is adapted from the deterministic IB, which was first proposed in [32] for text clustering (which is presented in the previous section). Here, the IB method compresses an observation $Y$ to a quantized variable $U$ while preserving the relevant information with a random variable $X$. We consider the case in which the variable $U$ is quantized with $q \in \mathbb{N}$ bits, i.e., $|\mathcal{U}| = 2^q$. The aim is to find the deterministic quantizer mapping $P_{U|Y}$ which maps the discrete observation $Y$ to a quantized variable $U$ which maximizes the relevance $I(U;X)$ under a cardinality constraint $|\mathcal{U}|$. This is equivalent to finding the optimal clustering of $Y$ which maximizes the mutual information $I(U;X)$.

So we initialize randomly by grouping $Y$ into $|\mathcal{U}|$ clusters. The algorithm takes one of the elements into a new cluster – so-called the singleton cluster. Due to this change, the probabilities $P_{X|U}$ and $P_U$ are changed, and the new values are calculated using the IB updates rules (2.12). Then, the deterministic IB is applied to decide on which one of the original $|\mathcal{U}|$ clusters that the singleton cluster will be merged. The possible $|\mathcal{U}|$ choices corresponds to merger costs given by

$$C(\mathcal{Y}_{\text{sing}}, \mathcal{Y}_k) = \psi D_{\text{JS}}^{(\pi_1,\pi_2)}(P_{X|y} \| P_{X|t}) \ , \qquad k = 1, \ldots, |\mathcal{U}| \ , \tag{2.45}$$

where $D_{\text{JS}}^{(\pi_1,\pi_2)}$ is the Jensen - Shannon divergence given in (2.44) and

$$\psi = \Pr(Y = y) + \Pr(U = u) \tag{2.46a}$$

$$\pi_1 = \Pr(Y = y)/\psi \tag{2.46b}$$

$$\pi_2 = \Pr(U = u)/\psi \ . \tag{2.46c}$$

The singleton cluster merges with the one which has a smaller merger cost.
The algorithm is a greedy algorithm, which repeats the draw and merge steps for all $Y$ until the obtained clusters are the same. Since the IB method does not converge to the global optimum, it should be run several times and the clustering (quantization) should be done with the best outcome, i.e., the mapping which maximize the IB cost (2.11).

Now we consider an example of finding the optimum channel quantizers for the binary input additive white Gaussian noise (AWGN) channel [86, Section III], in which a code bit $x \in \{0,1\}$ from a binary LDPC codeword is transmitted over a binary symmetric AWGN channel with binary shift keying (BPSK) modulation, i.e., $s(x) = -2x+1$. Symbol $s(x)$ is transmitted over the channel, and the continuous channel output $y$ is observed. The prior distribution of the code bits is assumed to be Bernoulli-(1/2), i.e., $p(x=0) = p(x=1) = 1/2$. Then the joint distribution $p(x,y)$ is given by

$$p(x,y) = \frac{1}{2\sqrt{2\pi\sigma_n^2}} \exp\left(-\frac{|y-s(x)|^2}{2\sigma_n^2}\right) , \tag{2.47}$$

where $\sigma_n^2$ is the channel noise variance. We note that the deterministic method offered for the optimum channel quantizers is valid for only the discrete variables, so $Y$ needs to be discretized with a fine resolution. The channel output is discretized into uniformly spaced representation values. Figure 2.6 illustrates an example in which the channel output interval $[-M, M]$ is discretized into 20 values, i.e., $|\mathcal{Y}| = 20$, and these values are represented by using unsigned integers.

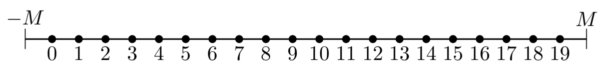



Figure 2.6: Discretization of the channel output.

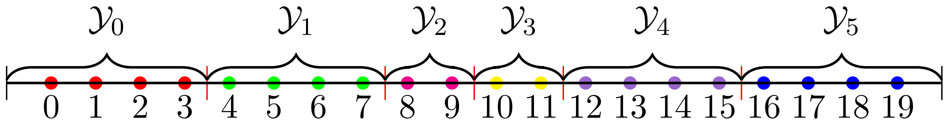



Figure 2.7: Visualization of clusters $\{\mathcal{Y}_k\}_{k=1}^{|\mathcal{U}|}$ separated by boundaries $|$, that are to be optimized.

The idea is to build a quantizer which uses a deterministic mapping $P_{U|Y}$ which maps from the discrete output $Y$ to the quantized value $U$, such that the quantized values are as much as informative about X (i.e., large mutual information $I(U:X)$) under the resolution constraint of the quantizer, i.e., $|\mathcal{U}|$. Finding the mapping $P_{U|Y}$ which maximizes $I(U;X)$ corresponds to finding the optimum boundaries separating the clusters $\mathcal{Y}_k$, as illustrated in Figure 2.7. For example, after the random initialization of clusters, at the first step, the rightmost element of $\mathcal{Y}_0$ is taken into the singleton cluster, and the merger costs are calculated for putting it back into $\mathcal{Y}_0$ and putting it to its neighbor cluster $\mathcal{Y}_1$. The cluster which makes the merger cost smaller is chosen. At each iteration, an element on the border is taken into the singleton cluster, which will be merged into the one with a smaller cost among the original and neighbor clusters. These steps are repeated until the resulting cluster does not change anymore. This algorithm is detailed in [86, Algorithm 1].

In digital communication systems, a continuous channel output is fed into an analog-to-digital converter to obtain a discrete valued sample – depicted in Figure 2.8. In theory, it is assumed that the quantizer has a very high resolution so the effect of quantization is generally ignored. However, this is not the case in real life. A few bits are desired in the implementations, hence the quantizer becomes a bottleneck in the communication system.

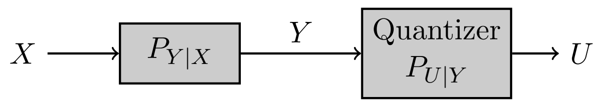



Figure 2.8: Memoryless channel with subsequent quantizer.

The state of the art low-density parity-check (LDPC) decoders execute the node operations by processing the quasi-continuous LLRs, which makes belief propagation decoding challenging. The IB method is proposed in [86] to overcome this complexity issues. The main idea is to pass compressed but highly informative integer-valued messages along the edges of a Tanner graph. To do so, Lewandowsky and Bauch use the IB method [86], and construct discrete message passing decoders for LDPC codes; and they showed that these decoders outperform state of the art decoders.

We close this section by mentioning the implementation issues of DNNs which are used for many artificial intelligence (AI) algorithms. The superior success of DNNs comes at the cost of high complexity (computational- and memory-wise). Although the devices, e.g., smartphones, get more and more powerful compared to a few year ago with the significant improvement of the chipsets, the implementation of DNNs is still a challenging task. The proposed approach seems particularly promising for the implementation of DNN algorithms on chipsets.

# Chapter 3

# Discrete Memoryless CEO Problem with Side Information

In this chapter, we study the $K$-encoder DM CEO problem with side information shown in Figure 3.1. Consider a $(K+2)$-dimensional memoryless source $(X, Y_0, Y_1, \ldots, Y_K)$ with finite alphabet $\mathcal{X} \times \mathcal{Y}_0 \times \mathcal{Y}_1 \times \ldots \times \mathcal{Y}_K$ and joint probability mass function (pmf) $P_{X,Y_0,Y_1,\ldots,Y_K}(x, y_0, y_1, \ldots, y_K)$. It is assumed that for all $\mathcal{S} \subseteq \mathcal{K} := \{1, \ldots, K\}$,

$$Y_{\mathcal{S}} -\!\!\!\ominus\!\!\!- (X, Y_0) -\!\!\!\ominus\!\!\!- Y_{\mathcal{S}^c} \;, \tag{3.1}$$

forms a Markov chain in that order. Also, let $\{(X_i, Y_{0,i}, Y_{1,i}, \ldots, Y_{K,i})\}_{i=1}^n$ be a sequence of $n$ independent copies of $(X, Y_0, Y_1, \ldots, Y_K)$, i.e., $(X^n, Y_0^n, Y_1^n, \ldots, Y_K^n) \sim \prod_{i=1}^n P_{X,Y_0,Y_1,\ldots,Y_K}$ $(x_i, y_{0,i}, y_{1,i}, \ldots, y_{K,i})$. In the model studied in this chapter, Encoder (or agent) $k$, $k \in \mathcal{K}$, observes the memoryless source $Y_k^n$ and uses $R_k$ bits per sample to describe it to the decoder. The decoder observes a statistically dependent memoryless side information stream, in the form of the sequence $Y_0^n$, and wants to reconstruct the remote source $X^n$ to within a prescribed fidelity level. Similar to [10], in this thesis we take the reproduction alphabet $\hat{\mathcal{X}}$ to be equal to the set of probability distributions over the source alphabet $\mathcal{X}$. Thus, for a vector $\hat{X}^n \in \hat{\mathcal{X}}^n$, the notation $\hat{X}_j(x)$ means the $j^{th}$-coordinate of $\hat{X}^n$, $1 \leq j \leq n$, which is a probability distribution on $\mathcal{X}$, evaluated for the outcome $x \in \mathcal{X}$. In other words, the decoder generates 'soft' estimates of the remote source's sequences. We consider the logarithmic loss distortion measure defined as in (2.5), where the letter-wise distortion measure is given by (2.1).

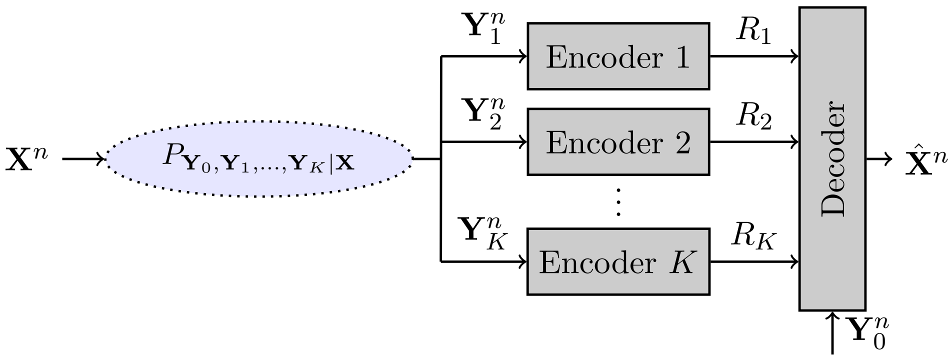



Figure 3.1: CEO source coding problem with side information.

**Definition 1.** *A rate-distortion code (of blocklength $n$) for the model of Figure 3.1 consists of $K$ encoding functions*

$$\phi_k^{(n)} : \mathcal{Y}_k^n \to \{1, \ldots, M_k^{(n)}\} , \quad \text{for } k = 1, \ldots, K ,$$

*and a decoding function*

$$\psi^{(n)} : \{1, \ldots, M_1^{(n)}\} \times \ldots \times \{1, \ldots, M_K^{(n)}\} \times \mathcal{Y}_0^n \to \hat{\mathcal{X}}^n . \qquad \blacksquare$$

**Definition 2.** *A rate-distortion tuple $(R_1, \ldots, R_K, D)$ is achievable for the DM CEO source coding problem with side information if there exist a blocklength $n$, encoding functions $\{\phi_k^{(n)}\}_{k=1}^K$ and a decoding function $\psi^{(n)}$ such that*

$$\begin{aligned} R_k &\geq \frac{1}{n} \log M_k^{(n)} , \quad \text{for } k = 1, \ldots, K , \\ D &\geq \mathbb{E}\big[d^{(n)}\big(X^n, \psi^{(n)}(\phi_1^{(n)}(Y_1^n), \ldots, \phi_K^{(n)}(Y_K^n), Y_0^n)\big)\big] . \end{aligned}$$

*The rate-distortion region $\mathcal{RD}^\star_{\mathrm{CEO}}$ of the model of Figure 3.1 is defined as the closure of all non-negative rate-distortion tuples $(R_1, \ldots, R_K, D)$ that are achievable.* $\blacksquare$

## 3.1 Rate-Distortion Region

The following theorem gives a single-letter characterization of the rate-distortion region $\mathcal{RD}^\star_{\mathrm{CEO}}$ of the DM CEO problem with side information under logarithmic loss measure.

**Definition 3.** *For given tuple of auxiliary random variables $(U_1, \ldots, U_K, Q)$ with distribution $P_{U_\mathcal{K},Q}(u_\mathcal{K}, q)$ such that $P_{X,Y_0,Y_\mathcal{K},U_\mathcal{K},Q}(x, y_0, y_\mathcal{K}, u_\mathcal{K}, q)$ factorizes as*

$$P_{X,Y_0}(x, y_0) \prod_{k=1}^K P_{Y_k|X,Y_0}(y_k|x, y_0)\, P_Q(q) \prod_{k=1}^K P_{U_k|Y_k,Q}(u_k|y_k, q) , \tag{3.2}$$

*define* $\mathcal{RD}_{\text{CEO}}(U_1,\dots,U_K,Q)$ *as the set of all non-negative rate-distortion tuples* $(R_1,\dots,R_K,D)$ *that satisfy, for all subsets* $\mathcal{S}\subseteq\mathcal{K}$*,*

$$\sum_{k\in\mathcal{S}} R_k + D \geq \sum_{k\in\mathcal{S}} I(Y_k;U_k|X,Y_0,Q) + H(X|U_{\mathcal{S}^c},Y_0,Q)\;. \qquad \blacksquare$$

**Theorem 1.** *The rate-distortion region for the DM CEO problem under logarithmic loss is given by*

$$\mathcal{RD}^{\star}_{\text{CEO}} = \bigcup \mathcal{RD}_{\text{CEO}}(U_1,\dots,U_K,Q)\;,$$

*where the union is taken over all tuples* $(U_1,\dots,U_K,Q)$ *with distributions that satisfy* (3.2)*.*

*Proof.* The proof of Theorem 1 is given in Appendix A. □

**Remark 1.** *To exhaust the region of Theorem 1, it is enough to restrict* $\{U_k\}_{k=1}^K$ *and* $Q$ *to satisfy* $|\mathcal{U}_k|\leq|\mathcal{Y}_k|$ *for* $k\in\mathcal{K}$ *and* $|\mathcal{Q}|\leq K+2$ *(see [10, Appendix A]).* ■

**Remark 2.** *Theorem 1 extends the result of [10, Theorem 10] to the case in which the decoder has, or observes, its own side information stream* $Y_0^n$ *and the agents' observations are conditionally independent given the remote source* $X^n$ *and* $Y_0^n$*, i.e.,* $Y_{\mathcal{S}}^n -\!\!\circ\!\!- (X^n,Y_0^n) -\!\!\circ\!\!- Y_{\mathcal{S}^c}^n$ *holds for all subsets* $\mathcal{S}\subseteq\mathcal{K}$*. The rate-distortion region of this problem can be obtained readily by applying [10, Theorem 10], which provides the rate-distortion region of the model without side information at decoder, to the modified setting in which the remote source is* $\tilde{\mathbf{X}}=(\mathbf{X},\mathbf{Y}_0)$*, another agent (agent* $K+1$*) observes* $\mathbf{Y}_{K+1}=\mathbf{Y}_0$ *and communicates at large rate* $R_{K+1}=\infty$ *with the CEO, which wishes to estimates* $\tilde{\mathbf{X}}$ *to within average logarithmic distortion* $D$ *and has no own side information stream*[1]*.* ■

## 3.2 Estimation of Encoder Observations

In this section, we focus on the two-encoder case, i.e., $K=2$. Suppose the decoder wants to estimate the encoder observations $(Y_1,Y_2)$, i.e., $X=(Y_1,Y_2)$. Note that in this case the side information $Y_0$ can be chosen *arbitrarily* correlated to $(Y_1,Y_2)$ and is *not* restricted to satisfy any Markov structure, since the Markov chain $Y_1 -\!\!\circ\!\!- (X,Y_0) -\!\!\circ\!\!- Y_2$ is satisfied for all choices of $Y_0$ that are arbitrarily correlated with $(Y_1,Y_2)$.

[1]Note that for the modified CEO setting the agents' observations are conditionally independent given the remote source $\tilde{\mathbf{X}}$.

If a distortion of $D$ bits is tolerated on the joint estimation of the pair $(Y_1, Y_2)$, then the achievable rate-distortion region can be obtained easily from Theorem 1, as a slight variation of the Slepian-Wolf region, namely the set of non-negative rate-distortion triples $(R_1, R_2, D)$ such that

$$R_1 \geq H(Y_1|Y_0, Y_2) - D \tag{3.3a}$$

$$R_2 \geq H(Y_2|Y_0, Y_1) - D \tag{3.3b}$$

$$R_1 + R_2 \geq H(Y_1, Y_2|Y_0) - D\ . \tag{3.3c}$$

The following theorem gives a characterization of the set of rate-distortion quadruples $(R_1, R_2, D_1, D_2)$ that are achievable in the more general case in which a distortion $D_1$ is tolerated on the estimation of the source component $Y_1$ and a distortion $D_2$ is tolerated on the estimation of the source component $Y_2$, i.e., the rate-distortion region of the two-encoder DM multiterminal source coding problem with arbitrarily correlated side information at the decoder.

**Theorem 2.** *If $X = (Y_1, Y_2)$, the component $Y_1$ is to be reconstructed to within average logarithmic loss distortion $D_1$ and the component $Y_2$ is to be reconstructed to within average logarithmic loss distortion $D_2$, the rate-distortion region $\mathcal{RD}^{\star}_{\mathrm{MT}}$ of the associated two-encoder DM multiterminal source coding problem with correlated side information at the decoder under logarithmic loss is given by the set of all non-negative rate-distortion quadruples $(R_1, R_2, D_1, D_2)$ that satisfy*

$$\begin{aligned}
R_1 &\geq I(U_1; Y_1|U_2, Y_0, Q) \\
R_2 &\geq I(U_2; Y_2|U_1, Y_0, Q) \\
R_1 + R_2 &\geq I(U_1, U_2; Y_1, Y_2|Y_0, Q) \\
D_1 &\geq H(Y_1|U_1, U_2, Y_0, Q) \\
D_2 &\geq H(Y_2|U_1, U_2, Y_0, Q)\ ,
\end{aligned}$$

*for some joint measure of the form $P_{Y_0,Y_1,Y_2}(y_0, y_1, y_2)P_Q(q)P_{U_1|Y_1,Q}(u_1|y_1, q)P_{U_2|Y_2,Q}(u_2|y_2, q)$.*

*Proof.* The proof of Theorem 2 is given in Appendix B. □

**Remark 3.** *The auxiliary random variables of Theorem 2 are such that $U_1 -\!\!\!\ominus\!\!\!- (Y_1, Q) -\!\!\!\ominus\!\!\!- (Y_0, Y_2, U_2)$ and $U_2 -\!\!\!\ominus\!\!\!- (Y_2, Q) -\!\!\!\ominus\!\!\!- (Y_0, Y_1, U_1)$ form Markov chains.* ■

**Remark 4.** *The result of Theorem 2 extends that of [10, Theorem 6] for the two-encoder source coding problem with average logarithmic loss distortion constraints on $Y_1$ and $Y_2$ and no side information at the decoder to the setting in which the decoder has its own side information $Y_0$ that is arbitrarily correlated with $(Y_1, Y_2)$. It is noteworthy that while the Berger-Tung inner bound is known to be non-tight for more than two encoders, as it is not optimal for the lossless modulo-sum problem of Korner and Marton [88], Theorem 2 shows that it is tight for the case of three encoders if the observation of the third encoder is encoded at large (infinite) rate.* ■

In the case in which the sources $Y_1$ and $Y_2$ are conditionally independent given $Y_0$, i.e., $Y_1 -\!\!\circ\!\!- Y_0 -\!\!\circ\!\!- Y_2$ forms a Markov chain, it can be shown easily that the result of Theorem 2 reduces to the set of rates and distortions that satisfy

$$R_1 \geq I(U_1; Y_1) - I(U_1; Y_0) \tag{3.4}$$

$$R_2 \geq I(U_2; Y_2) - I(U_2; Y_0) \tag{3.5}$$

$$D_1 \geq H(Y_1|U_1, Y_0) \tag{3.6}$$

$$D_2 \geq H(Y_2|U_2, Y_0) \;, \tag{3.7}$$

for some measure of the form $P_{Y_0,Y_1,Y_2}(y_0, y_1, y_2)P_{U_1|Y_1}(u_1|y_1)P_{U_2|Y_2}(u_2|y_2)$.
This result can also be obtained by applying [89, Theorem 6] with the reproduction functions therein chosen as

$$f_k(U_k, Y_0) := \Pr[Y_k = y_k|U_k, Y_0] \;, \quad \text{for } k = 1, 2 \;. \tag{3.8}$$

Then, note that with this choice we have

$$\mathbb{E}[d(Y_k, f_k(U_k, Y_0)] = H(Y_k|U_k, Y_0) \;, \quad \text{for } k = 1, 2 \;. \tag{3.9}$$

## 3.3 An Example: Distributed Pattern Classification

Consider the problem of distributed pattern classification shown in Figure 3.2. In this example, the decoder is a predictor whose role is to guess the unknown class $X \in \mathcal{X}$ of a measurable pair $(Y_1, Y_2) \in \mathcal{Y}_1 \times \mathcal{Y}_2$ on the basis of inputs from two learners as well as its own observation about the target class, in the form of some correlated $Y_0 \in \mathcal{Y}_0$. It is assumed that $Y_1 -\!\!\circ\!\!- (X, Y_0) -\!\!\circ\!\!- Y_2$. The first learner produces its input based only

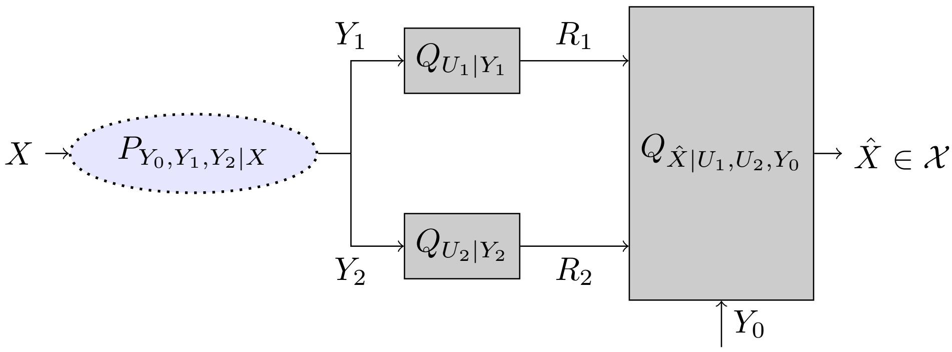



Figure 3.2: An example of distributed pattern classification.

on $Y_1 \in \mathcal{Y}_1$; and the second learner produces its input based only on $Y_2 \in \mathcal{Y}_2$. For the sake of a smaller *generalization gap*[2], the inputs of the learners are restricted to have description lengths that are no more than $R_1$ and $R_2$ bits per sample, respectively. Let $Q_{U_1|Y_1} : \mathcal{Y}_1 \longrightarrow \mathcal{P}(\mathcal{U}_1)$ and $Q_{U_2|Y_2} : \mathcal{Y}_2 \longrightarrow \mathcal{P}(\mathcal{U}_2)$ be two (stochastic) such learners. Also, let $Q_{\hat{X}|U_1,U_2,Y_0} : \mathcal{U}_1 \times \mathcal{U}_2 \times \mathcal{Y}_0 \longrightarrow \mathcal{P}(\mathcal{X})$ be a soft-decoder or predictor that maps the pair of representations $(U_1, U_2)$ and $Y_0$ to a probability distribution on the label space $\mathcal{X}$. The pair of learners and predictor induce a classifier

$$\begin{aligned} Q_{\hat{X}|Y_0,Y_1,Y_2}(x|y_0,y_1,y_2) &= \sum_{u_1\in\mathcal{U}_1} Q_{U_1|Y_1}(u_1|y_1) \sum_{u_2\in\mathcal{U}_2} Q_{U_2|Y_2}(u_2|y_2) Q_{\hat{X}|U_1,U_2,Y_0}(x|u_1,u_2,y_0) \\ &= \mathbb{E}_{Q_{U_1|Y_1}} \mathbb{E}_{Q_{U_2|Y_2}} [Q_{\hat{X}|U_1,U_2,Y_0}(x|U_1,U_2,y_0)] \,, \end{aligned} \tag{3.10}$$

whose probability of classification error is defined as

$$P_{\mathcal{E}}(Q_{\hat{X}|Y_0,Y_1,Y_2}) = 1 - \mathbb{E}_{P_{X,Y_0,Y_1,Y_2}}[Q_{\hat{X}|Y_0,Y_1,Y_2}(X|Y_0,Y_1,Y_2)] \,. \tag{3.11}$$

Let $\mathcal{RD}^\star_{\mathrm{CEO}}$ be the rate-distortion region of the associated two-encoder DM CEO problem with side information as given by Theorem 1. The following proposition shows that there exists a classifier $Q^\star_{\hat{X}|Y_0,Y_1,Y_2}$ for which the probability of misclassification can be upper bounded in terms of the minimal average logarithmic loss distortion that is achievable for the rate pair $(R_1, R_2)$ in $\mathcal{RD}^\star_{\mathrm{CEO}}$.

[2]The generalization gap, defined as the difference between the empirical risk (average risk over a finite training sample) and the population risk (average risk over the true joint distribution), can be upper bounded using the mutual information between the learner's inputs and outputs, see, e.g., [90,91] and the recent [92], which provides a fundamental justification of the use of the *minimum description length* (MDL) constraint on the learners mappings as a regularizer term.

**Proposition 1.** *For the problem of distributed pattern classification of Figure 3.2, there exists a classifier $Q^{\star}_{\hat{X}|Y_0,Y_1,Y_2}$ for which the probability of classification error satisfies*

$$P_{\mathcal{E}}(Q^{\star}_{\hat{X}|Y_0,Y_1,Y_2}) \leq 1 - \exp\Big( - \inf\{D \,:\, (R_1, R_2, D) \in \mathcal{RD}^{\star}_{\mathrm{CEO}}\}\Big) \,,$$

*where $\mathcal{RD}^{\star}_{\mathrm{CEO}}$ is the rate-distortion region of the associated two-encoder DM CEO problem with side information as given by Theorem 1.*

*Proof.* Let a triple mappings $(Q_{U_1|Y_1}, Q_{U_2|Y_2}, Q_{\hat{X}|U_1,U_2,Y_0})$ be given. It is easy to see that the probability of classification error of the classifier $Q_{\hat{X}|Y_0,Y_1,Y_2}$ as defined by (3.11) satisfies

$$P_{\mathcal{E}}(Q_{\hat{X}|Y_0,Y_1,Y_2}) \leq \mathbb{E}_{P_{X,Y_0,Y_1,Y_2}}[-\log Q_{\hat{X}|Y_0,Y_1,Y_2}(X|Y_0,Y_1,Y_2)] \,. \tag{3.12}$$

Applying Jensen's inequality on the right hand side (RHS) of (3.12), using the concavity of the logarithm function, and combining with the fact that the exponential function increases monotonically, the probability of classification error can be further bounded as

$$P_{\mathcal{E}}(Q_{\hat{X}|Y_0,Y_1,Y_2}) \leq 1 - \exp\Big( - \mathbb{E}_{P_{X,Y_0,Y_1,Y_2}}[-\log Q_{\hat{X}|Y_0,Y_1,Y_2}(X|Y_0,Y_1,Y_2)]\Big) \,. \tag{3.13}$$

Using (3.10) and continuing from (3.13), we get

$$\begin{aligned} P_{\mathcal{E}}(Q_{\hat{X}|Y_0,Y_1,Y_2}) &\leq 1 - \exp\Big( - \mathbb{E}_{P_{X,Y_0,Y_1,Y_2}}[-\log \mathbb{E}_{Q_{U_1|Y_1}}\mathbb{E}_{Q_{U_2|Y_2}}[Q_{\hat{X}|U_1,U_2,Y_0}(X|U_1,U_2,Y_0)]]\Big) \\ &\leq 1 - \exp\Big( - \mathbb{E}_{P_{X,Y_0,Y_1,Y_2}}\mathbb{E}_{Q_{U_1|Y_1}}\mathbb{E}_{Q_{U_2|Y_2}}[-\log[Q_{\hat{X}|U_1,U_2,Y_0}(X|U_1,U_2,Y_0)]]\Big) \,, \end{aligned} \tag{3.14}$$

where the last inequality follows by applying Jensen's inequality and using the concavity of the logarithm function.

Noticing that the term in the exponential function in the RHS of (3.14),

$$\mathcal{D}(Q_{U_1|Y_1}, Q_{U_1|Y_1}, Q_{\hat{X}|U_1,U_2,Y_0}) := \mathbb{E}_{P_{XY_0Y_1Y_2}}\mathbb{E}_{Q_{U_1|Y_1}}\mathbb{E}_{Q_{U_2|Y_2}}[-\log Q_{\hat{X}|U_1,U_2,Y_0}(X|U_1,U_2,Y_0)] \,,$$

is the average logarithmic loss, or cross-entropy risk, of the triple $(Q_{U_1|Y_1}, Q_{U_2|Y_2}, Q_{\hat{X}|U_1,U_2,Y_0})$; the inequality (3.14) implies that minimizing the average logarithmic loss distortion leads to classifier with smaller (bound on) its classification error. Using Theorem 1, the minimum average logarithmic loss, minimized over all mappings $Q_{U_1|Y_1} \,:\, \mathcal{Y}_1 \longrightarrow \mathcal{P}(\mathcal{U}_1)$ and $Q_{U_2|Y_2} \,:\, \mathcal{Y}_2 \longrightarrow \mathcal{P}(\mathcal{U}_2)$ that have description lengths no more than $R_1$ and $R_2$ bits per-sample, respectively, as well as all choices of $Q_{\hat{X}|U_1,U_2,Y_0} \,:\, \mathcal{U}_1 \times \mathcal{U}_2 \times \mathcal{Y}_0 \longrightarrow \mathcal{P}(\mathcal{X})$, is

$$D^{\star}(R_1, R_2) = \inf\{D \,:\, (R_1, R_2, D) \in \mathcal{RD}^{\star}_{\mathrm{CEO}}\} \,. \tag{3.15}$$

Thus, the direct part of Theorem 1 guarantees the existence of a classifier $Q^{\star}_{\hat{X}|Y_0,Y_1,Y_2}$ whose probability of error satisfies the bound given in Proposition 1. $\square$

To make the above example more concrete, consider the following scenario where $Y_0$ plays the role of information about the sub-class of the label class $X \in \{0,1,2,3\}$. More specifically, let $S$ be a random variable that is uniformly distributed over $\{1,2\}$. Also, let $X_1$ and $X_2$ be two random variables that are independent between them and from $S$, distributed uniformly over $\{1,3\}$ and $\{0,2\}$ respectively. The state $S$ acts as a random switch that connects $X_1$ or $X_2$ to $X$, i.e.,

$$X = X_S \,. \tag{3.16}$$

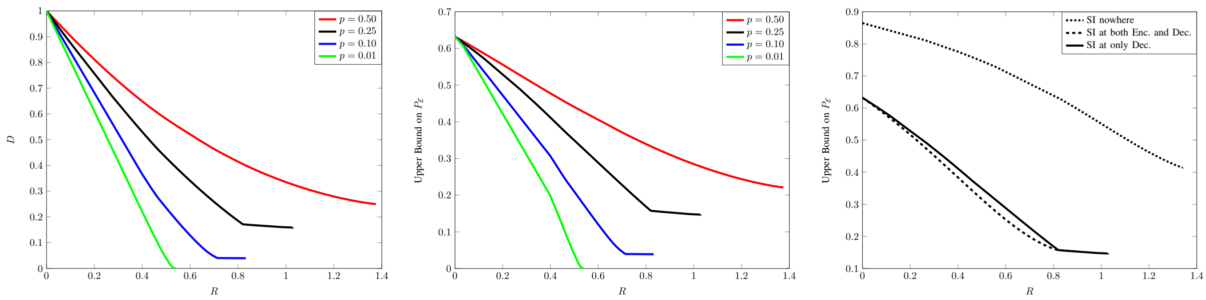
That is, if $S = 1$ then $X = X_1$, and if $S = 2$ then $X = X_2$. Thus, the value of $S$ indicates whether $X$ is odd- or even-valued (i.e., the sub-class of $X$). Also, let

$$Y_0 = S \tag{3.17a}$$

$$Y_1 = X_S \oplus Z_1 \tag{3.17b}$$

$$Y_2 = X_S \oplus Z_2 \,, \tag{3.17c}$$

where $Z_1$ and $Z_2$ are Bernoulli-($p$) random variables, $p \in (0,1)$, that are independent between them, and from $(S, X_1, X_2)$, and the addition is modulo 4. For simplification, we let $R_1 = R_2 = R$. We numerically approximate the set of $(R, D)$ pairs such that $(R, R, D)$ is in the rate-distortion region $\mathcal{RD}^{\star}_{\mathrm{CEO}}$ corresponding to the CEO network of this example. The algorithm that we use for the computation will be described in detail in Chapter 5.1.1. The lower convex envelope of these $(R, D)$ pairs is plotted in Figure 3.3a for $p \in \{0.01, 0.1, 0.25, 0.5\}$. Continuing our example, we also compute the upper bound on the probability of classification error according to Proposition 1. The result is given in Figure 3.3b. Observe that if $Y_1$ and $Y_2$ are high-quality estimates of $X$ (e.g., $p = 0.01$), then a small increase in the *complexity* $R$ results in a large relative improvement of the (bound on) the probability of classification error. On the other hand, if $Y_1$ and $Y_2$ are low-quality estimates of $X$ (e.g., $p = 0.25$) then we require a large increase of $R$ in order to obtain an appreciable reduction in the error probability. Recalling that larger $R$ implies lesser generalization capability [90–92], these numerical results are consistent with the fact that classifiers should strike a good balance between accuracy and their ability to generalize well to unseen data. Figure 3.3c quantifies the value of side information $S$ given to both learners and predictor, none of them, or only the predictor, for $p = 0.25$.




(a) (b) (c)

Figure 3.3: Illustration of the bound on the probability of classification error of Proposition 1 for the example described by (3.16) and (3.17).

(a) Distortion-rate function of the network of Figure 3.2 computed for $p \in \{0.01, 0.1, 0.25, 0.5\}$.

(b) Upper bound on the probability of classification error computed according to Proposition 1.

(c) Effect of side information (SI) $Y_0$ when given to both learners and the predictor, only the predictor or none of them.

## 3.4 Hypothesis Testing Against Conditional Independence

Consider the multiterminal detection system shown in Figure 3.4, where a memoryless vector source $(X, Y_0, Y_1, \ldots, Y_K)$, $K \geq 2$, has a joint distribution that depends on two hypotheses, a null hypothesis $H_0$ and an alternate hypothesis $H_1$. A detector that observes directly the pair $(X, Y_0)$ but only receives summary information of the observations $(Y_1, \ldots, Y_K)$, seeks to determine which of the two hypotheses is true. Specifically, Encoder $k$, $k = 1, \ldots, K$, which observes an i.i.d. string $Y_k^n$, sends a message $M_k$ to the detector a finite rate of $R_k$ bits per observation over a noise-free channel; and the detector makes its decision between the two hypotheses on the basis of the received messages $(M_1, \ldots, M_K)$ as well as the available pair $(X^n, Y_0^n)$. In doing so, the detector can make two types of error: Type I error (guessing $H_1$ while $H_0$ is true) and Type II error (guessing $H_0$ while $H_1$ is true). The Type II error probability decreases exponentially fast with the size $n$ of the i.i.d. strings, say with an exponent $E$; and, classically, one is interested is characterizing the set of achievable rate-exponent tuples $(R_1, \ldots, R_K, E)$ in the regime in which the

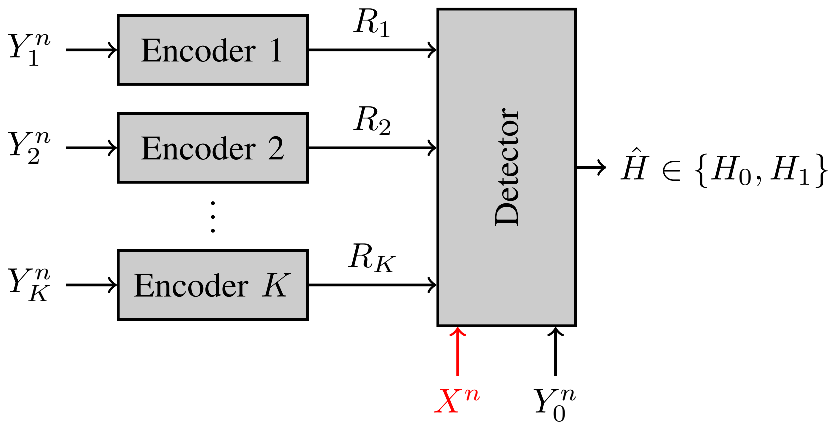



Figure 3.4: Distributed hypothesis testing against conditional independence.

probability of the Type I error is kept below a prescribed small value $\epsilon$. This problem, which was first introduced by Berger [93], and then studied further in [65, 66, 94], arises naturally in many applications (for recent developments on this topic, the reader may refer to [16, 67, 68, 95–99] and references therein).

In this section, we are interested in a class of the hypothesis testing problem studied in [16][3] obtained by restricting the joint distribution of the variables to satisfy the Markov chain

$$Y_{\mathcal{S}} -\!\!\circ\!\!- (X, Y_0) -\!\!\circ\!\!- Y_{\mathcal{S}^c}, \quad \text{for all } \mathcal{S} \subseteq \mathcal{K} := \{1, \ldots, K\}\ , \tag{3.18}$$

under the null hypothesis $H_0$; and $X$ and $(Y_1, \ldots, Y_K)$ are independent conditionally given $Y_0$ under the alternate hypothesis $H_1$, i.e.,

$$H_0 : P_{X,Y_0,Y_1\ldots,Y_K} = P_{X,Y_0} \prod_{i=1}^{K} P_{Y_k|X,Y_0} \tag{3.19a}$$

$$H_1 : Q_{X,Y_0,Y_1\ldots,Y_K} = P_{Y_0} P_{X|Y_0} P_{Y_1,\ldots,Y_K|Y_0}\ . \tag{3.19b}$$

Let $\{(X_i, Y_{0,i}, Y_{1,i}, \ldots, Y_{K,i})\}_{i=1}^n$ be an i.i.d. sequence of random vectors with the distribution at a single stage being the same as the generic vector $(X, Y_0, Y_1, \ldots, Y_K)$. As shown in Figure 3.4, Encoder $k \in \mathcal{K}$ observes $Y_k^n$ and then sends a message to the detector using an encoding function

$$\breve{\phi}_k^{(n)}\ :\ \mathcal{Y}_k^n \to \{1, \ldots, M_k^{(n)}\}\ . \tag{3.20}$$

[3] In fact, the model of [12] also involves a random variable $Y_{K+1}$, which is chosen here to be deterministic as it is not relevant for the analysis and discussion that will follow in this thesis (see Remark 5).

The pair $(X^n, Y_0^n)$ is available at the detector which uses it together with the messages from the encoders to make a decision between the two hypotheses based on a decision rule

$$\breve{\psi}^{(n)} \, : \{1, \ldots, M_1^{(n)}\} \times \ldots \times \{1, \ldots, M_K^{(n)}\} \times \mathcal{X}^n \times \mathcal{Y}_0^n \rightarrow \{H_0, H_1\} \, . \tag{3.21}$$

The mapping (3.21) is such that $\breve{\psi}^{(n)}(m_1, \ldots, m_K, x^n, y_0^n) = H_0$ if $(m_1, \ldots, m_K, x^n, y_0^n) \in \mathcal{A}_n$ and $H_1$ otherwise, with

$$\mathcal{A}_n \subseteq \prod_{k=1}^{n} \{1, \ldots, M_k^{(n)}\} \times \mathcal{X}^n \times \mathcal{Y}_0^n \, ,$$

designating the acceptance region for $H_0$. The encoders $\{\breve{\phi}_k^{(n)}\}_{k=1}^K$ and the detector $\breve{\psi}^{(n)}$ are such that the Type I error probability does not exceed a prescribed level $\epsilon \in [0, 1]$, i.e.,

$$P_{\breve{\phi}_1^{(n)}(Y_1^n), \ldots, \breve{\phi}_K^{(n)}(Y_K^n), X^n, Y_0^n}(\mathcal{A}_n^c) \leq \epsilon \, , \tag{3.22}$$

and the Type II error probability does not exceed $\beta$, i.e.,

$$Q_{\breve{\phi}_1^{(n)}(Y_1^n), \ldots, \breve{\phi}_K^{(n)}(Y_K^n), X^n, Y_0^n}(\mathcal{A}_n) \leq \beta \, . \tag{3.23}$$

**Definition 4.** *A rate-exponent tuple $(R_1, \ldots, R_K, E)$ is achievable for a fixed $\epsilon \in [0, 1]$ and any positive $\delta$ if there exist a sufficiently large blocklength $n$, encoders $\{\breve{\phi}_k^{(n)}\}_{k=1}^K$ and a detector $\breve{\psi}^{(n)}$ such that*

$$\frac{1}{n} \log M_k^{(n)} \leq R_k + \delta \, , \quad \textit{for } k = 1, \ldots, K \, , \tag{3.24a}$$

$$-\frac{1}{n} \log \beta \geq E - \delta \, . \tag{3.24b}$$

*The rate-exponent region $\mathcal{R}_{\mathrm{HT}}$ is defined as*

$$\mathcal{R}_{\mathrm{HT}} := \bigcap_{\epsilon > 0} \mathcal{R}_{\mathrm{HT}, \epsilon} \, , \tag{3.25}$$

*where $\mathcal{R}_{\mathrm{HT}, \epsilon}$ is the set of all achievable rate-exponent vectors for a fixed $\epsilon \in (0, 1]$.* ■

We start with an entropy characterization of the rate-exponent $\mathcal{R}_{\mathrm{HT}}$ as defined by (3.25). Let

$$\mathcal{R}^\star = \bigcup_n \bigcup_{\{\breve{\phi}_k^{(n)}\}_{k \in \mathcal{K}}} \mathcal{R}^\star\Big(n, \{\breve{\phi}_k^{(n)}\}_{k \in \mathcal{K}}\Big) \, , \tag{3.26}$$

where

$$\mathcal{R}^\star\Big(n, \{\breve{\phi}_k^{(n)}\}_{k\in\mathcal{K}}\Big) = \Bigg\{(R_1,\dots,R_K,E) \ \text{s.t.}$$
$$R_k \geq \frac{1}{n}\log|\breve{\phi}_k^{(n)}(Y_k^n)| \ , \quad \text{for } k=1,\dots,K \ ,$$
$$E \leq \frac{1}{n} I(\{\breve{\phi}_k^{(n)}(Y_k^n)\}_{k\in\mathcal{K}}; X^n|Y_0^n) \quad \Bigg\} \ .$$

We have the following proposition, whose proof is essentially similar to that of [65, Theorem 5] and, hence, is omitted.

**Proposition 2.** $\mathcal{R}_{HT} = \overline{\mathcal{R}^\star}$ . ■

Now, recall the CEO source coding problem under logarithmic loss of Figure 3.1 and its rate-distortion region $\mathcal{RD}^\star_{\text{CEO}}$ as given by Theorem 1 in the case in which the Markov chain (3.1) holds. The following proposition states that $\mathcal{R}_{\text{HT}}$ and $\mathcal{RD}^\star_{\text{CEO}}$ can be inferred from each other.

**Proposition 3.** *$(R_1,\dots,R_K,E) \in \mathcal{R}_{HT}$ if and only if $(R_1,\dots,R_K,H(X|Y_0)-E) \in \mathcal{RD}^\star_{\text{CEO}}$.*

*Proof.* The proof of Proposition 3 appears in Appendix C. □

The result of the next theorem follows easily by using Theorem 1 and Proposition 3.

**Theorem 3.** *[100, Theorem 1] For the distributed hypothesis testing against conditional independence problem of Figure 3.4, the rate-exponent region is given by the union of all non-negative tuples $(R_1,\dots,R_K,E)$ that satisfy, for all subsets $\mathcal{S} \subseteq \mathcal{K}$,*

$$E \leq I(U_{\mathcal{S}^c}; X|Y_0,Q) + \sum_{k\in\mathcal{S}} \big(R_k - I(Y_k;U_k|X,Y_0,Q)\big) \ ,$$

*for some auxiliary random variables $(U_1,\dots,U_K,Q)$ with distribution $P_{U_\mathcal{K},Q}(u_\mathcal{K},q)$ such that $P_{X,Y_0,Y_\mathcal{K},U_\mathcal{K},Q}(x,y_0,y_\mathcal{K},u_\mathcal{K},q)$ factorizes as*

$$P_{X,Y_0}(x,y_0)\prod_{k=1}^K P_{Y_k|X,Y_0}(y_k|x,y_0)\, P_Q(q)\prod_{k=1}^K P_{U_k|Y_k,Q}(u_k|y_k,q) \ .$$ ■

**Remark 5.** *In [16], Rahman and Wagner study the hypothesis testing problem of Figure 3.4 in the case in which* $X$ *is replaced by a two-source* $(Y_{K+1}, X)$ *such that, like in our setup (which corresponds to* $Y_{K+1}$ *deterministic),* $Y_0$ *induces conditional independence between* $(Y_1, \ldots, Y_K, Y_{K+1})$ *and* $X$ *under the alternate hypothesis* $H_1$*. Under the null hypothesis* $H_0$*, however, the model studied by Rahman and Wagner in [16] assumes a more general distribution than ours in which* $(Y_1, \ldots, Y_K, Y_{K+1})$ *are arbitrarily correlated among them and with the pair* $(X, Y_0)$*. More precisely, the joint distributions of* $(X, Y_1, \ldots, Y_K, Y_{K+1})$ *under the null and alternate hypotheses as considered in [16] are*

$$H_0 : \tilde{P}_{X,Y_0,Y_1\ldots,Y_K,Y_{K+1}} = P_{Y_0} P_{X,Y_1,\ldots,Y_K,Y_{K+1}|Y_0} \tag{3.28a}$$

$$H_1 : \tilde{Q}_{X,Y_0,Y_1\ldots,Y_K,Y_{K+1}} = P_{Y_0} P_{X|Y_0} P_{Y_1,\ldots,Y_K,Y_{K+1}|Y_0} \, . \tag{3.28b}$$

*For this model, they provide inner and outer bounds on the rate-exponent region which do not match in general (see [16, Theorem 1] for the inner bound and [16, Theorem 2] for the outer bound). The inner bound of [16, Theorem 1] is based on a scheme, named Quantize-Bin-Test scheme therein, that is similar to the Berger-Tung distributed source coding scheme [101, 102]; and whose achievable rate-exponent region can be shown through submodularity arguments to be equivalent to the region stated in Theorem 3 (with* $Y_{K+1}$ *set to be deterministic). The result of Theorem 3 then shows that if the joint distribution of the variables under the null hypothesis is restricted to satisfy* (3.19a)*, i.e., the encoders' observations* $\{Y_k\}_{k\in\mathcal{K}}$ *are independent conditionally given* $(X, Y_0)$*, then the Quantize-Bin-Test scheme of [16, Theorem 1] is optimal. We note that, prior to this work, for general distributions under the null hypothesis (i.e., without the Markov chain* (3.1) *under this hypothesis) the optimality of the Quantize-Bin-Test scheme of [16] for the problem of testing against conditional independence was known only for the special case of a single encoder, i.e.,* $K = 1$*, (see [16, Theorem 3]), a result which can also be recovered from Theorem 3.* ■

# Chapter 4

# Vector Gaussian CEO Problem with Side Information

In this chapter, we study the $K$-encoder vector Gaussian CEO problem with side information shown in Figure 4.1. The remote vector source $\mathbf{X}$ is complex-valued, has $n_x$-dimensions, and is assumed to be Gaussian with zero mean and covariance matrix $\mathbf{\Sigma_x} \succeq \mathbf{0}$. $\mathbf{X}^n = (\mathbf{X}_1, \dots, \mathbf{X}_n)$ denotes a collection of $n$ independent copies of $\mathbf{X}$. The agents' observations are Gaussian noisy versions of the remote vector source, with the observation at agent $k \in \mathcal{K}$ given by

$$\mathbf{Y}_{k,i} = \mathbf{H}_k \mathbf{X}_i + \mathbf{N}_{k,i} \ , \quad \text{for } \ i = 1, \dots, n \ , \tag{4.1}$$

where $\mathbf{H}_k \in \mathbb{C}^{n_k \times n_x}$ represents the channel matrix connecting the remote vector source to the $k$-th agent; and $\mathbf{N}_{k,i} \in \mathbb{C}^{n_k}$ is the noise vector at this agent, assumed to be i.i.d. Gaussian with zero-mean and independent from $\mathbf{X}_i$. The decoder has its own noisy observation of the remote vector source, in the form of a correlated jointly Gaussian side information stream $\mathbf{Y}_0^n$, with

$$\mathbf{Y}_{0,i} = \mathbf{H}_0 \mathbf{X}_i + \mathbf{N}_{0,i} \ , \quad \text{for } \ i = 1, \dots, n \ , \tag{4.2}$$

where, similar to the above, $\mathbf{H}_0 \in \mathbb{C}^{n_0 \times n_x}$ is the channel matrix connecting the remote vector source to the CEO; and $\mathbf{N}_{0,i} \in \mathbb{C}^{n_0}$ is the noise vector at the CEO, assumed to be Gaussian with zero-mean and covariance matrix $\mathbf{\Sigma}_0 \succeq \mathbf{0}$ and independent from $\mathbf{X}_i$. In this chapter, it is assumed that the agents' observations are independent conditionally given

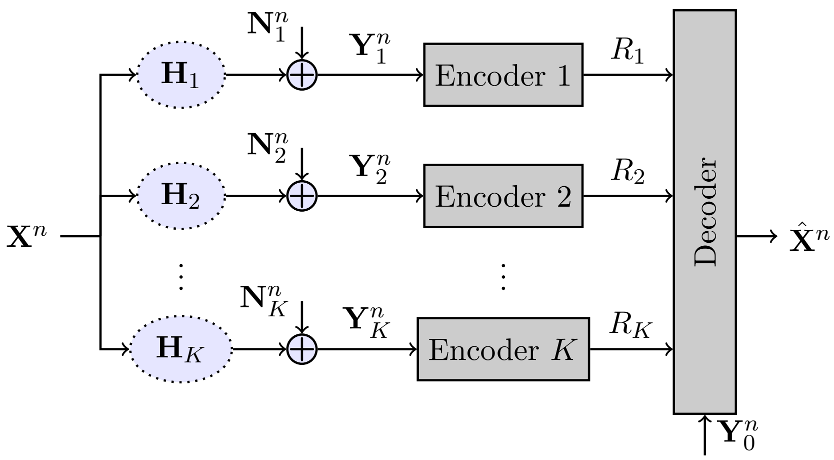



Figure 4.1: Vector Gaussian CEO problem with side information.

the remote vector source $\mathbf{X}^n$ and the side information $\mathbf{Y}_0^n$, i.e., for all $\mathcal{S} \subseteq \mathcal{K}$,

$$\mathbf{Y}_{\mathcal{S}}^n -\!\!\circ\!\!- (\mathbf{X}^n, \mathbf{Y}_0^n) -\!\!\circ\!\!- \mathbf{Y}_{\mathcal{S}^c}^n \ . \tag{4.3}$$

Using (4.1) and (4.2), it is easy to see that the assumption (4.3) is equivalent to that the noises at the agents are independent conditionally given $\mathbf{N}_0$. For notational simplicity, $\mathbf{\Sigma}_k$ denotes the conditional covariance matrix of the noise $\mathbf{N}_k$ at the $k$-th agent given $\mathbf{N}_0$, i.e., $\mathbf{\Sigma}_k := \mathbf{\Sigma}_{\mathbf{n}_k|\mathbf{n}_0}$. Recalling that for a set $\mathcal{S} \subseteq \mathcal{K}$, the notation $\mathbf{N}_{\mathcal{S}}$ designates the collection of noise vectors with indices in the set $\mathcal{S}$, in what follows we denote the covariance matrix of $\mathbf{N}_{\mathcal{S}}$ as $\mathbf{\Sigma}_{\mathbf{n}_{\mathcal{S}}}$.

## 4.1 Rate-Distortion Region

We first state the following proposition which essentially extends the result of Theorem 1 to the case of sources with continuous alphabets.

**Definition 5.** *For given tuple of auxiliary random variables $(U_1, \ldots, U_K, Q)$ with distribution $P_{U_{\mathcal{K}},Q}(u_{\mathcal{K}}, q)$ such that $P_{\mathbf{X},\mathbf{Y}_0,\mathbf{Y}_{\mathcal{K}},U_{\mathcal{K}},Q}(\mathbf{x}, \mathbf{y}_0, \mathbf{y}_{\mathcal{K}}, u_{\mathcal{K}}, q)$ factorizes as*

$$P_{\mathbf{X},\mathbf{Y}_0}(\mathbf{x}, \mathbf{y}_0) \prod_{k=1}^{K} P_{\mathbf{Y}_k|\mathbf{X},\mathbf{Y}_0}(\mathbf{y}_k|\mathbf{x}, \mathbf{y}_0)\ P_Q(q) \prod_{k=1}^{K} P_{U_k|\mathbf{Y}_k,Q}(u_k|\mathbf{y}_k, q)\ , \tag{4.4}$$

*define $\widetilde{\mathcal{RD}}_{\mathrm{CEO}}^{\mathrm{I}}(U_1, \ldots, U_K, Q)$ as the set of all non-negative rate-distortion tuples $(R_1, \ldots, R_K, D)$ that satisfy, for all subsets $\mathcal{S} \subseteq \mathcal{K}$,*

$$D + \sum_{k \in \mathcal{S}} R_k \geq \sum_{k \in \mathcal{S}} I(\mathbf{Y}_k; U_k|\mathbf{X}, \mathbf{Y}_0, Q) + h(\mathbf{X}|U_{\mathcal{S}^c}, \mathbf{Y}_0, Q)\ . \tag{4.5}$$

*Also, let* $\widetilde{\mathcal{RD}}_{\mathrm{CEO}}^{\mathrm{I}} := \bigcup \widetilde{\mathcal{RD}}_{\mathrm{CEO}}^{\mathrm{I}}(U_1, \dots, U_K, Q)$ *where the union is taken over all tuples* $(U_1, \dots, U_K, Q)$ *with distributions that satisfy* (4.4). ■

**Definition 6.** *For given tuple of auxiliary random variables* $(V_1, \dots, V_K, Q')$ *with distribution* $P_{V_{\mathcal{K}},Q'}(v_{\mathcal{K}}, q')$ *such that* $P_{\mathbf{X},\mathbf{Y}_0,\mathbf{Y}_{\mathcal{K}},V_{\mathcal{K}},Q'}(\mathbf{x}, \mathbf{y}_0, \mathbf{y}_{\mathcal{K}}, v_{\mathcal{K}}, q')$ *factorizes as*

$$P_{\mathbf{X},\mathbf{Y}_0}(\mathbf{x}, \mathbf{y}_0) \prod_{k=1}^{K} P_{\mathbf{Y}_k|\mathbf{X},\mathbf{Y}_0}(\mathbf{y}_k|\mathbf{x}, \mathbf{y}_0)\, P_{Q'}(q') \prod_{k=1}^{K} P_{V_k|\mathbf{Y}_k,Q'}(v_k|\mathbf{y}_k, q')\ , \tag{4.6}$$

*define* $\widetilde{\mathcal{RD}}_{\mathrm{CEO}}^{\mathrm{II}}(V_1, \dots, V_K, Q')$ *as the set of all non-negative rate-distortion tuples* $(R_1, \dots, R_K, D)$ *that satisfy, for all subsets* $\mathcal{S} \subseteq \mathcal{K}$,

$$\begin{aligned} \sum_{k\in\mathcal{S}} R_k &\geq I(\mathbf{Y}_{\mathcal{S}}; V_{\mathcal{S}}|V_{\mathcal{S}^c}, \mathbf{Y}_0, Q') \\ D &\geq h(\mathbf{X}|V_1, \dots, V_K, \mathbf{Y}_0, Q')\ . \end{aligned}$$

*Also, let* $\widetilde{\mathcal{RD}}_{\mathrm{CEO}}^{\mathrm{II}} := \bigcup \widetilde{\mathcal{RD}}_{\mathrm{CEO}}^{\mathrm{II}}(V_1, \dots, V_K, Q')$ *where the union is taken over all tuples* $(V_1, \dots, V_K, Q')$ *with distributions that satisfy* (4.6). ■

**Proposition 4.** *The rate-distortion region for the vector Gaussian CEO problem under logarithmic loss is given by*

$$\mathcal{RD}_{\text{VG-CEO}}^{\star} = \widetilde{\mathcal{RD}}_{\mathrm{CEO}}^{\mathrm{I}} = \widetilde{\mathcal{RD}}_{\mathrm{CEO}}^{\mathrm{II}}\ .$$

*Proof.* The proof of Proposition 4 is given in Appendix D. □

For convenience, we now introduce the following notation which will be instrumental in what follows. Let, for every set $\mathcal{S} \subseteq \mathcal{K}$, the set $\bar{\mathcal{S}} := \{0\} \cup \mathcal{S}^c$. Also, for $\mathcal{S} \subseteq \mathcal{K}$ and given matrices $\{\mathbf{\Omega}_k\}_{k=1}^K$ such that $\mathbf{0} \preceq \mathbf{\Omega}_k \preceq \mathbf{\Sigma}_k^{-1}$, let $\mathbf{\Lambda}_{\bar{\mathcal{S}}}$ designate the block-diagonal matrix given by

$$\mathbf{\Lambda}_{\bar{\mathcal{S}}} := \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathrm{diag}(\{\mathbf{\Sigma}_k - \mathbf{\Sigma}_k\mathbf{\Omega}_k\mathbf{\Sigma}_k\}_{k\in\mathcal{S}^c}) \end{bmatrix}\ , \tag{4.7}$$

where $\mathbf{0}$ in the principal diagonal elements is the $n_0 \times n_0$-all zero matrix.

The following theorem gives an explicit characterization of the rate-distortion region of the vector Gaussian CEO problem with side information under logarithmic loss measure that we study in this chapter.

**Theorem 4.** *The rate-distortion region $\mathcal{RD}^\star_{\text{VG-CEO}}$ of the vector Gaussian CEO problem under logarithmic loss is given by the set of all non-negative rate-distortion tuples $(R_1,\dots,R_K,D)$ that satisfy, for all subsets $\mathcal{S}\subseteq\mathcal{K}$,*

$$D+\sum_{k\in\mathcal{S}}R_k \geq \sum_{k\in\mathcal{S}}\log\frac{1}{|\mathbf{I}-\mathbf{\Omega}_k\mathbf{\Sigma}_k|}+\log\left|(\pi e)\left(\mathbf{\Sigma}_{\mathbf{x}}^{-1}+\mathbf{H}_{\bar{\mathcal{S}}}^\dagger\mathbf{\Sigma}_{\mathbf{n}_{\bar{\mathcal{S}}}}^{-1}(\mathbf{I}-\mathbf{\Lambda}_{\bar{\mathcal{S}}}\mathbf{\Sigma}_{\mathbf{n}_{\bar{\mathcal{S}}}}^{-1})\mathbf{H}_{\bar{\mathcal{S}}}\right)^{-1}\right|,$$

*for matrices $\{\mathbf{\Omega}_k\}_{k=1}^K$ such that $\mathbf{0}\preceq\mathbf{\Omega}_k\preceq\mathbf{\Sigma}_k^{-1}$, where $\bar{\mathcal{S}}=\{0\}\cup\mathcal{S}^c$ and $\mathbf{\Lambda}_{\bar{\mathcal{S}}}$ is as defined by (4.7).*

*Proof.* The proof of the direct part of Theorem 4 follows simply by evaluating the region $\widetilde{\mathcal{RD}}^{\mathrm{I}}_{\text{CEO}}$ as described by the inequalities (4.5) using Gaussian test channels and no time-sharing. Specifically, we set $Q=\emptyset$ and $p(u_k|\mathbf{y}_k,q)=\mathcal{CN}(\mathbf{y}_k,\mathbf{\Sigma}_k^{1/2}(\mathbf{\Omega}_k-\mathbf{I})\mathbf{\Sigma}_k^{1/2})$, $k\in\mathcal{K}$. The proof of the converse appears in Appendix E. □

In the case in which the noises at the agents are independent among them and from the noise $\mathbf{N}_0$ at the CEO, the result of Theorem 4 takes a simpler form which is stated in the following corollary.

**Corollary 1.** *Consider the vector Gaussian CEO problem described by (4.1) and (4.2) with the noises $(\mathbf{N}_1,\dots,\mathbf{N}_K)$ being independent among them and with $\mathbf{N}_0$. Under logarithmic loss, the rate-distortion region this model is given by the set of all non-negative tuples $(R_1,\dots,R_K,D)$ that satisfy, for all subsets $\mathcal{S}\subseteq\mathcal{K}$,*

$$D+\sum_{k\in\mathcal{S}}R_k \geq \sum_{k\in\mathcal{S}}\log\frac{1}{|\mathbf{I}-\mathbf{\Omega}_k\mathbf{\Sigma}_k|}+\log\left|(\pi e)\big(\mathbf{\Sigma}_{\mathbf{x}}^{-1}+\mathbf{H}_0^\dagger\mathbf{\Sigma}_0^{-1}\mathbf{H}_0+\sum_{k\in\mathcal{S}^c}\mathbf{H}_k^\dagger\mathbf{\Omega}_k\mathbf{H}_k\big)^{-1}\right|,$$

*for some matrices $\{\mathbf{\Omega}_k\}_{k=1}^K$ such that $\mathbf{0}\preceq\mathbf{\Omega}_k\preceq\mathbf{\Sigma}_k^{-1}$.* ■

**Remark 6.** *The direct part of Theorem 4 shows that Gaussian test channels and no-time sharing exhaust the region. For the converse proof of Theorem 4, we derive an outer bound on the region $\widetilde{\mathcal{RD}}^{\mathrm{I}}_{\text{CEO}}$. In doing so, we use the de Bruijn identity, a connection between differential entropy and Fisher information, along with the properties of MMSE and Fisher information. By opposition to the case of quadratic distortion measure for which the application of this technique was shown in [11] to result in an outer bound that is generally non-tight, Theorem 4 shows that the approach is successful in the case of logarithmic loss distortion measure as it yields a complete characterization of the region. On this aspect, note that in the specific case of scalar Gaussian sources, an alternate* *converse proof may be obtained by extending that of the scalar Gaussian many-help-one source coding problem by Oohama [3] and Prabhakaran et al. [4] through accounting for additional side information at CEO and replacing the original mean square error distortion constraint with conditional entropy. However, such approach does not seem conclusive in the vector case, as the entropy power inequality is known to be generally non-tight in this setting [12, 13].* ■

**Remark 7.** *The result of Theorem 4 generalizes that of [59] which considers the case of only one agent, i.e., the remote vector Gaussian Wyner-Ziv model under logarithmic loss, to the case of an arbitrarily number of agents. The converse proof of [59], which relies on the technique of orthogonal transform to reduce the vector setting to one of parallel scalar Gaussian settings, seems insufficient to diagonalize all the noise covariance matrices simultaneously in the case of more than one agent. The result of Theorem 4 is also connected to recent developments on characterizing the capacity of multiple-input multiple-output (MIMO) relay channels in which the relay nodes are connected to the receiver through error-free finite-capacity links (i.e., the so-called cloud radio access networks). In particular, the reader may refer to [103, Theorem 4] where important progress is done, and [62] where compress-and-forward with joint decompression-decoding is shown to be optimal under the constraint of oblivious relay processing.* ■

## 4.2 Gaussian Test Channels with Time-Sharing Exhaust the Berger-Tung Region

Proposition 4 shows that the union of all rate-distortion tuples that satisfy (4.5) for all subsets $\mathcal{S} \subseteq \mathcal{K}$ coincides with the Berger-Tung inner bound in which time-sharing is used. The direct part of Theorem 4 is obtained by evaluating (4.5) using Gaussian test channels and no time-sharing, i.e., $Q = \emptyset$, not the Berger-Tung inner bound. The reader may wonder: i) whether Gaussian test channels also exhaust the Berger-Tung inner bound for the vector Gaussian CEO problem that we study here, and ii) whether time-sharing is needed with the Berger-Tung scheme. In this section, we answer both questions in the affirmative. In particular, we show that the Berger-Tung coding scheme with Gaussian test channels and time-sharing achieves distortion levels that are not larger than any other coding scheme. That is, Gaussian test channels *with* time-sharing exhaust the region

$\widetilde{\mathcal{RD}}^{\mathrm{II}}_{\mathrm{CEO}}$ as defined in Definition 6.

**Proposition 5.** *The rate-distortion region for the vector Gaussian CEO problem under logarithmic loss is given by*

$$\mathcal{RD}^{\star}_{\text{VG-CEO}} = \bigcup \widetilde{\mathcal{RD}}^{\mathrm{II}}_{\mathrm{CEO}}(V_1^{\mathrm{G}}, \ldots, V_K^{\mathrm{G}}, Q') \ ,$$

*where* $\widetilde{\mathcal{RD}}^{\mathrm{II}}_{\mathrm{CEO}}(\cdot)$ *is as given in Definition 6 and the superscript* G *is used to denote that the union is taken over Gaussian distributed* $V_k^{\mathrm{G}} \sim p(v_k|\mathbf{y}_k, q')$ *conditionally on* $(\mathbf{Y}_k, Q')$.

*Proof.* For the proof of Proposition 5, it is sufficient to show that, for fixed Gaussian conditional distributions $\{p(u_k|\mathbf{y}_k)\}_{k=1}^K$, the extreme points of the polytopes defined by (4.5) are *dominated* by points that are in $\widetilde{\mathcal{RD}}^{\mathrm{II}}_{\mathrm{CEO}}$ and which are achievable using Gaussian conditional distributions $\{p(v_k|\mathbf{y}_k, q')\}_{k=1}^K$. Hereafter, we give a brief outline of proof for the case $K = 2$. The reasoning for $K \geq 2$ is similar and is provided in Appendix F. Consider the inequalities (4.5) with $Q = \emptyset$ and $(U_1, U_2) := (U_1^{\mathrm{G}}, U_2^{\mathrm{G}})$ chosen to be Gaussian (see Theorem 4). Consider now the extreme points of the polytopes defined by the obtained inequalities:

$$\begin{aligned}
P_1 &= (0, 0, I(\mathbf{Y}_1; U_1^{\mathrm{G}}|\mathbf{X}, \mathbf{Y}_0) + I(\mathbf{Y}_2; U_2^{\mathrm{G}}|\mathbf{X}, \mathbf{Y}_0) + h(\mathbf{X}|\mathbf{Y}_0))\\
P_2 &= (I(\mathbf{Y}_1; U_1^{\mathrm{G}}|\mathbf{Y}_0), 0, I(U_2^{\mathrm{G}}; \mathbf{Y}_2|\mathbf{X}, \mathbf{Y}_0) + h(\mathbf{X}|U_1^{\mathrm{G}}, \mathbf{Y}_0))\\
P_3 &= (0, I(\mathbf{Y}_2; U_2^{\mathrm{G}}|\mathbf{Y}_0), I(U_1^{\mathrm{G}}; \mathbf{Y}_1|\mathbf{X}, \mathbf{Y}_0) + h(\mathbf{X}|U_2^{\mathrm{G}}, \mathbf{Y}_0))\\
P_4 &= (I(\mathbf{Y}_1; U_1^{\mathrm{G}}|\mathbf{Y}_0), I(\mathbf{Y}_2; U_2^{\mathrm{G}}|U_1^{\mathrm{G}}, \mathbf{Y}_0), h(\mathbf{X}|U_1^{\mathrm{G}}, U_2^{\mathrm{G}}, \mathbf{Y}_0))\\
P_5 &= (I(\mathbf{Y}_1; U_1^{\mathrm{G}}|U_2^{\mathrm{G}}, \mathbf{Y}_0), I(\mathbf{Y}_2; U_2^{\mathrm{G}}|\mathbf{Y}_0), h(\mathbf{X}|U_1^{\mathrm{G}}, U_2^{\mathrm{G}}, \mathbf{Y}_0)) \ ,
\end{aligned}$$

where the point $P_j$ is a a triple $(R_1^{(j)}, R_2^{(j)}, D^{(j)})$. It is easy to see that each of these points is *dominated* by a point in $\widetilde{\mathcal{RD}}^{\mathrm{II}}_{\mathrm{CEO}}$, i.e., there exists $(R_1, R_2, D) \in \widetilde{\mathcal{RD}}^{\mathrm{II}}_{\mathrm{CEO}}$ for which $R_1 \leq R_1^{(j)}$, $R_2 \leq R_2^{(j)}$ and $D \leq D^{(j)}$. To see this, first note that $P_4$ and $P_5$ are both in $\widetilde{\mathcal{RD}}^{\mathrm{II}}_{\mathrm{CEO}}$. Next, observe that the point $(0, 0, h(\mathbf{X}|\mathbf{Y}_0))$ is in $\widetilde{\mathcal{RD}}^{\mathrm{II}}_{\mathrm{CEO}}$, which is clearly achievable by letting $(V_1, V_2, Q') = (\emptyset, \emptyset, \emptyset)$, dominates $P_1$. Also, by using letting $(V_1, V_2, Q') = (U_1^{\mathrm{G}}, \emptyset, \emptyset)$, we have that the point $(I(\mathbf{Y}_1; U_1|\mathbf{Y}_0), 0, h(\mathbf{X}|U_1, \mathbf{Y}_0))$ is in $\widetilde{\mathcal{RD}}^{\mathrm{II}}_{\mathrm{CEO}}$, and dominates the point $P_2$. A similar argument shows that $P_3$ is dominated by a point in $\widetilde{\mathcal{RD}}^{\mathrm{II}}_{\mathrm{CEO}}$. The proof is terminated by observing that, for all above corner points, $V_k$ is set either equal $U_k^{\mathrm{G}}$ (which is Gaussian distributed conditionally on $\mathbf{Y}_k$) or a constant. $\square$

**Remark 8.** *Proposition 5 shows that for the vector Gaussian CEO problem with side information under a logarithmic loss constraint, vector Gaussian quantization codebooks with time-sharing are optimal. In the case of quadratic distortion constraint, however, a characterization of the rate-distortion region is still to be found in general, and it is not known yet whether vector Gaussian quantization codebooks (with or without time-sharing) are optimal, except in few special cases such as that of scalar Gaussian sources or the case of only one agent, i.e., the remote vector Gaussian Wyner-Ziv problem whose rate-distortion region is found in [59]. In [59], Tian and Chen also found the rate-distortion region of the remote vector Gaussian Wyner-Ziv problem under logarithmic loss, which they showed achievable using Gaussian quantization codebooks that are different from those (also Gaussian) that are optimal in the case of quadratic distortion. As we already mentioned, our result of Theorem 4 generalizes that of [59] to the case of an arbitrary number of agents.* ■

**Remark 9.** *One may wonder whether giving the decoder side information* $\mathbf{Y}_0$ *to the encoders is beneficial. Similar to the well known result in Wyner-Ziv source coding of scalar Gaussian sources, our result of Theorem 4 shows that encoder side information does not help.* ■

## 4.3 Quadratic Vector Gaussian CEO Problem with Determinant Constraint

We now turn to the case in which the distortion is measured under quadratic loss. In this case, the mean square error matrix is defined by

$$\mathbf{D}^{(n)} := \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[(\mathbf{X}_i - \hat{\mathbf{X}}_i)(\mathbf{X}_i - \hat{\mathbf{X}}_i)^{\dagger}] \ . \tag{4.8}$$

Under a (general) error constraint of the form

$$\mathbf{D}^{(n)} \preceq \mathbf{D} \ , \tag{4.9}$$

where $\mathbf{D}$ designates here a prescribed positive definite error matrix, a complete solution is still to be found in general. In what follows, we replace the constraint (4.9) with one on the *determinant* of the error matrix $\mathbf{D}^{(n)}$, i.e.,

$$|\mathbf{D}^{(n)}| \leq D \ , \tag{4.10}$$

($D$ is a scalar here). We note that since the error matrix $\mathbf{D}^{(n)}$ is minimized by choosing the decoding as

$$\hat{\mathbf{X}}_i = \mathbb{E}[\mathbf{X}_i | \breve{\phi}_1^{(n)}(\mathbf{Y}_1^n), \ldots, \breve{\phi}_K^{(n)}(\mathbf{Y}_K^n), \mathbf{Y}_0^n] \;, \tag{4.11}$$

where $\{\breve{\phi}_k^{(n)}\}_{k=1}^K$ denote the encoding functions, without loss of generality we can write (4.8) as

$$\mathbf{D}^{(n)} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{mmse}(\mathbf{X}_i | \breve{\phi}_1^{(n)}(\mathbf{Y}_1^n), \ldots, \breve{\phi}_K^{(n)}(\mathbf{Y}_K^n), \mathbf{Y}_0^n) \;. \tag{4.12}$$

**Definition 7.** *A rate-distortion tuple $(R_1, \ldots, R_K, D)$ is achievable for the quadratic vector Gaussian CEO problem with determinant constraint if there exist a blocklength $n$, $K$ encoding functions $\{\breve{\phi}_k^{(n)}\}_{k=1}^K$ such that*

$$\begin{aligned} R_k &\geq \frac{1}{n} \log M_k^{(n)}, \quad \text{for } k = 1, \ldots, K, \\ D &\geq \left| \frac{1}{n} \sum_{i=1}^{n} \mathrm{mmse}(\mathbf{X}_i | \breve{\phi}_1^{(n)}(\mathbf{Y}_1^n), \ldots, \breve{\phi}_K^{(n)}(\mathbf{Y}_K^n), \mathbf{Y}_0^n) \right| \;. \end{aligned}$$

*The rate-distortion region $\mathcal{RD}_{\text{VG-CEO}}^{\det}$ is defined as the closure of all non-negative tuples $(R_1, \ldots, R_K, D)$ that are achievable.* ■

The following theorem characterizes the rate-distortion region of the quadratic vector Gaussian CEO problem with determinant constraint.

**Theorem 5.** *The rate-distortion region $\mathcal{RD}_{\text{VG-CEO}}^{\det}$ of the quadratic vector Gaussian CEO problem with determinant constraint is given by the set of all non-negative tuples $(R_1, \ldots, R_K, D)$ that satisfy, for all subsets $\mathcal{S} \subseteq \mathcal{K}$,*

$$\log \frac{1}{D} \leq \sum_{k \in \mathcal{S}} R_k + \log |\mathbf{I} - \mathbf{\Omega}_k \mathbf{\Sigma}_k| + \log \left| \mathbf{\Sigma}_{\mathbf{x}}^{-1} + \mathbf{H}_{\bar{\mathcal{S}}}^{\dagger} \mathbf{\Sigma}_{\mathbf{n}_{\bar{\mathcal{S}}}}^{-1} (\mathbf{I} - \mathbf{\Lambda}_{\bar{\mathcal{S}}} \mathbf{\Sigma}_{\mathbf{n}_{\bar{\mathcal{S}}}}^{-1}) \mathbf{H}_{\bar{\mathcal{S}}} \right| \;,$$

*for matrices $\{\mathbf{\Omega}_k\}_{k=1}^K$ such that $\mathbf{0} \preceq \mathbf{\Omega}_k \preceq \mathbf{\Sigma}_k^{-1}$, where $\bar{\mathcal{S}} = \{0\} \cup \mathcal{S}^c$ and $\mathbf{\Lambda}_{\bar{\mathcal{S}}}$ is as defined by (4.7).*

*Proof.* The proof of Theorem 5 is given in Appendix G. □

**Remark 10.** *It is believed that the approach of this section, which connects the quadratic vector Gaussian CEO problem to that under logarithmic loss, can also be exploited to possibly infer other new results on the quadratic vector Gaussian CEO problem. Alternatively, it can also be used to derive new converses on the quadratic vector Gaussian CEO problem. For example, in the case of scalar sources, Theorem 5, and Lemma 15, readily provide* *an alternate converse proof to those of [3, 4] for this model. Similar connections were made in [104, 105] where it was observed that the results of [10] can be used to recover known results on the scalar Gaussian CEO problem (such as the sum rate-distortion region of [106]) and the scalar Gaussian two-encoder distributed source coding problem. We also point out that similar information constraints have been applied to log-determinant reproduction constraints previously in [107].* ■

| Two-Encoder Rate Region | $K$-Encoder Rate Region |
|---|---|
| Cooperative bound [trivial] | Oohama '98 [108], Prabhakaran *et al.* '04 [4]<br>scalar |
| Wagner *et al.* '08 [106]<br>scalar, *sum-rate* | Tavildar *et al.* '10 [109]<br>scalar, *tree-structure* constraint |
| Rahman and Wagner '15 [110]<br>vector | Ekrem and Ulukus '14 [11]<br>vector, *outer bound* |
| | **Ugur *et al.* '19**<br>**vector, *determinant* constraint** |

Table 4.1: Advances in the resolution of the rate region of the quadratic Gaussian CEO problem.

We close this section by presenting Table 4.1, where advances in the resolution of the rate region of the quadratic Gaussian CEO problem is summarized.

## 4.4 Hypothesis Testing Against Conditional Independence

In this section we study the continuous case of the hypothesis testing problem presented in Chapter 3.4. Here, $(\mathbf{X}, \mathbf{Y}_0, \mathbf{Y}_1, \ldots, \mathbf{Y}_K)$ is a zero-mean Gaussian random vector such that

$$\mathbf{Y}_0 = \mathbf{H}_0\mathbf{X} + \mathbf{N}_0 \,, \tag{4.13}$$

where $\mathbf{H}_0 \in \mathbb{C}^{n_0 \times n_x}$, $\mathbf{X} \in \mathbb{C}^{n_x}$ and $\mathbf{N}_0 \in \mathbb{C}^{n_0}$ are independent Gaussian vectors with zero-mean and covariance matrices $\mathbf{\Sigma}_{\mathbf{x}} \succeq \mathbf{0}$ and $\mathbf{\Sigma}_0 \succeq \mathbf{0}$, respectively. The vectors $(\mathbf{Y}_1, \ldots, \mathbf{Y}_K)$ and $\mathbf{X}$ are correlated under the null hypothesis $H_0$ and are independent under the alternate hypothesis $H_1$, with

$$H_0 \;: \mathbf{Y}_k = \mathbf{H}_k\mathbf{X} + \mathbf{N}_k, \quad \text{for all } k \in \mathcal{K} \tag{4.14a}$$

$$H_1 \;: (\mathbf{Y}_1, \ldots, \mathbf{Y}_K) \text{ independent from } \mathbf{X} \text{ conditionally given } \mathbf{Y}_0 \,. \tag{4.14b}$$

The noise vectors $(\mathbf{N}_1, \ldots, \mathbf{N}_K)$ are jointly Gaussian with zero mean and covariance matrix $\mathbf{\Sigma}_{\mathbf{n}_\mathcal{K}} \succeq \mathbf{0}$. They are assumed to be independent from $\mathbf{X}$ but correlated among them and with $\mathbf{N}_0$, with for every $\mathcal{S} \subseteq \mathcal{K}$,

$$\mathbf{N}_\mathcal{S} -\!\!\!\ominus\!\!\!- \mathbf{N}_0 -\!\!\!\ominus\!\!\!- \mathbf{N}_{\mathcal{S}^c} \ . \tag{4.15}$$

Let $\mathbf{\Sigma}_k$ denote the conditional covariance matrix of noise $\mathbf{N}_k$ given $\mathbf{N}_0$ , $k \in \mathcal{K}$. Also, let $\mathcal{R}_{\text{VG-HT}}$ denote the rate-exponent region of this vector Gaussian hypothesis testing against conditional independence problem. The following theorem gives an explicit characterization of $\mathcal{R}_{\text{VG-HT}}$. The proof uses Proposition 3 and Theorem 4 in a manner that is essentially similar to that in the proof of Theorem 5; and, hence, it is omitted for brevity.

**Theorem 6.** *[100, Theorem 2] The rate-exponent region $\mathcal{R}_{\text{VG-HT}}$ of the vector Gaussian hypothesis testing against conditional independence problem is given by the set of all non-negative tuples $(R_1, \ldots, R_K, E)$ that satisfy, for all subsets $\mathcal{S} \subseteq \mathcal{K}$,*

$$\begin{aligned} E \leq & \sum_{k \in \mathcal{S}} \left[ R_k + \log |\mathbf{I} - \mathbf{\Omega}_k \mathbf{\Sigma}_k| \,\right] + \log \left| \mathbf{I} + \mathbf{\Sigma}_\mathbf{x} \mathbf{H}_{\bar{\mathcal{S}}}^\dagger \mathbf{\Sigma}_{\mathbf{n}_{\bar{\mathcal{S}}}}^{-1} (\mathbf{I} - \mathbf{\Lambda}_{\bar{\mathcal{S}}} \mathbf{\Sigma}_{\mathbf{n}_{\bar{\mathcal{S}}}}^{-1}) \mathbf{H}_{\bar{\mathcal{S}}} \right| \\ & - \log \left| \mathbf{I} + \mathbf{\Sigma}_\mathbf{x} \mathbf{H}_0^\dagger \mathbf{\Sigma}_0^{-1} \mathbf{H}_0 \right| \ , \end{aligned}$$

*for matrices $\{\mathbf{\Omega}_k\}_{k=1}^K$ such that $\mathbf{0} \preceq \mathbf{\Omega}_k \preceq \mathbf{\Sigma}_k^{-1}$, where $\bar{\mathcal{S}} = \{0\} \cup \mathcal{S}^c$ and $\mathbf{\Lambda}_{\bar{\mathcal{S}}}$ is given by (4.7).* ■

**Remark 11.** *An alternate proof of Theorem 6, which is direct, can be obtained by evaluating the region of Proposition 3 for the model (4.14), and is provided in [100, Section V-B]. Specifically, in the proof of the direct part we set $Q = \emptyset$ and $p(u_k|\mathbf{y}_k) = \mathcal{CN}(\mathbf{y}_k, \mathbf{\Sigma}_k^{1/2}(\mathbf{\Omega}_k - \mathbf{I})\mathbf{\Sigma}_k^{1/2})$ for $k \in \mathcal{K}$. The proof of the converse part follows by using Proposition 3 and proceeding along the lines of the converse part of Theorem 4 in Appendix E.* ■

In what follows, we elaborate on two special cases of Theorem 6, i) the one-encoder vector Gaussian testing against conditional independence problem (i.e., $K = 1$) and ii) the $K$-encoder scalar Gaussian testing against independence problem.

#### One-encoder vector Gaussian testing against conditional independence problem

Let us first consider the case $K = 1$. In this case, the Markov chain (4.15) which is to be satisfied under the null hypothesis is non-restrictive; and Theorem 6 then provides a complete solution of the (general) one-encoder vector Gaussian testing against conditional independence problem. More precisely, in this case the optimal trade-off between rate and Type II error exponent is given by the set of pairs $(R_1, E)$ that satisfy

$$\begin{aligned}
&E \leq R_1 + \log|\mathbf{I} - \mathbf{\Omega}_1\mathbf{\Sigma}_1| \\
&E \leq \log\left|\mathbf{I} + \mathbf{\Sigma}_{\mathbf{x}}\mathbf{H}_{\{0,1\}}^{\dagger}\mathbf{\Sigma}_{\mathbf{n}_{\{0,1\}}}^{-1}(\mathbf{I} - \mathbf{\Lambda}_{\{0,1\}}\mathbf{\Sigma}_{\mathbf{n}_{\{0,1\}}}^{-1})\mathbf{H}_{\{0,1\}}\right| - \log\left|\mathbf{I} + \mathbf{\Sigma}_{\mathbf{x}}\mathbf{H}_0^{\dagger}\mathbf{\Sigma}_0^{-1}\mathbf{H}_0\right| ,
\end{aligned} \tag{4.16}$$

for some $n_1 \times n_1$ matrix $\mathbf{\Omega}_1$ such that $\mathbf{0} \preceq \mathbf{\Omega}_1 \preceq \mathbf{\Sigma}_1^{-1}$, where $\mathbf{H}_{\{0,1\}} = [\mathbf{H}_0^{\dagger}, \mathbf{H}_1^{\dagger}]^{\dagger}$, $\mathbf{\Sigma}_{\mathbf{n}_{\{0,1\}}}$ is the covariance matrix of noise $(\mathbf{N}_0, \mathbf{N}_1)$ and

$$\mathbf{\Lambda}_{\{0,1\}} := \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_1 - \mathbf{\Sigma}_1\mathbf{\Omega}_1\mathbf{\Sigma}_1 \end{bmatrix} , \tag{4.17}$$

with the $\mathbf{0}$ in its principal diagonal denoting the $n_0 \times n_0$-all zero matrix. In particular, for the setting of testing against independence, i.e., $\mathbf{Y}_0 = \emptyset$ and the decoder's task reduced to guessing whether $\mathbf{Y}_1$ and $\mathbf{X}$ are independent or not, the optimal trade-off expressed by (4.16) reduces to the set of $(R_1, E)$ pairs that satisfy, for some $n_1 \times n_1$ matrix $\mathbf{\Omega}_1$ such that $\mathbf{0} \preceq \mathbf{\Omega}_1 \preceq \mathbf{\Sigma}_1^{-1}$,

$$E \leq \min\left\{R_1 + \log|\mathbf{I} - \mathbf{\Omega}_1\mathbf{\Sigma}_1| ,\ \log\left|\mathbf{I} + \mathbf{\Sigma}_{\mathbf{x}}\mathbf{H}_1^{\dagger}\mathbf{\Omega}_1\mathbf{H}_1\right|\right\} . \tag{4.18}$$

Observe that (4.16) is the counter-part, to the vector Gaussian setting, of the result of [16, Theorem 3] which provides a single-letter formula for the Type II error exponent for the one-encoder DM testing against conditional independence problem. Similarly, (4.18) is the solution of the vector Gaussian version of the one-encoder DM testing against independence problem which is studied, and solved, by Ahlswede and Csiszar in [65, Theorem 2]. Also, we mention that, perhaps non-intuitive, in the one-encoder vector Gaussian testing against independence problem swapping the roles of $\mathbf{Y}_1$ and $\mathbf{X}$ (i.e., giving $\mathbf{X}$ to the encoder and the noisy (under the null hypothesis) $\mathbf{Y}_1$ to the decoder) does not result in an increase of the Type II error exponent which is then identical to (4.18). Note that this is in sharp contrast with the related[1] setting of standard lossy source reproduction, i.e., the decoder aiming to reproduce the source observed at the encoder to within some average squared error distortion level using the sent compression message and its own side information,

[1]The connection, which is sometimes misleading, consists in viewing the decoder in the hypothesis testing against independence problem considered here as one that computes a binary-valued function of $(\mathbf{X}, \mathbf{Y}_1)$.

for which it is easy to see that, for given $R_1$ bits per sample, smaller distortion levels are allowed by having the encoder observe $\mathbf{X}$ and the decoder observe $\mathbf{Y}_1$, instead of the encoder observing the noisy $\mathbf{Y}_1 = \mathbf{H}_1\mathbf{X} + \mathbf{N}_1$ and the decoder observing $\mathbf{X}$.

#### $K$-encoder scalar Gaussian testing against independence problem

Consider now the special case of the setup of Theorem 6 in which $K \geq 2$, $Y_0 = \emptyset$, and the sources and noises are all scalar complex-valued, i.e., $n_x = 1$ and $n_k = 1$ for all $k \in \mathcal{K}$. The vector $(Y_1, \ldots, Y_K)$ and $X$ are correlated under the null hypothesis $H_0$ and independent under the alternate hypothesis $H_1$, with

$$H_0 \; : Y_k = X + N_k, \quad \text{for all } \; k \in \mathcal{K} \tag{4.19a}$$

$$H_1 \; : (Y_1, \ldots, Y_K) \; \text{independent from} \; X \; . \tag{4.19b}$$

The noises $N_1, \ldots, N_K$ are zero-mean jointly Gaussian, mutually independent and independent from $X$. Also, we assume that the variances $\sigma_k^2$ of noise $N_k$, $k \in \mathcal{K}$, and $\sigma_x^2$ of $X$ are all positive. In this case, it can be easily shown that Theorem 6 reduces to

$$\mathcal{R}_{\text{SG-HT}} = \Bigg\{ (R_1, \ldots, R_K, E) \; : \; \exists \, (\gamma_1, \ldots, \gamma_K) \in \mathbb{R}_+^K \quad \text{s.t.} \\ \gamma_k \leq \frac{1}{\sigma_k^2}, \; \forall \, k \in \mathcal{K} \\ \sum_{k \in \mathcal{S}} R_k \geq E - \log \Big( (1 + \sigma_x^2 \sum_{k \in \mathcal{S}^c} \gamma_k) \prod_{k \in \mathcal{S}} [1 - \gamma_k \sigma_k^2] \Big), \; \forall \, \mathcal{S} \subseteq \mathcal{K} \quad \Bigg\} . \tag{4.20}$$

The region $\mathcal{R}_{\text{SG-HT}}$ as given by (4.20) can be used to, e.g., characterize the centralized rate region, i.e., the set of rate vectors $(R_1, \ldots, R_K)$ that achieve the centralized Type II error exponent

$$I(Y_1, \ldots, Y_K; X) = \sum_{k=1}^{K} \log \frac{\sigma_x^2}{\sigma_k^2} \; . \tag{4.21}$$

We close this section by mentioning that, implicit in Theorem 6, the Quantize-Bin-Test scheme of [16, Theorem 1] with Gaussian test channels and time-sharing is optimal for the vector Gaussian $K$-encoder hypothesis testing against conditional independence problem (4.14). Furthermore, we note that Rahman and Wagner also characterized the optimal rate-exponent region of a different[2] Gaussian hypothesis testing against independence problem, called the Gaussian many-help-one hypothesis testing against independence problem therein, in the case of scalar valued sources [16, Theorem 7]. Specialized to the case $K = 1$, the result of Theorem 6 recovers that of [16, Theorem 7] in the case of no helpers; and extends it to vector-valued sources and testing against conditional independence in that case.

## 4.5 Distributed Vector Gaussian Information Bottleneck

Consider now the vector Gaussian CEO problem with side information, and let the logarithmic loss distortion constraint be replaced by the mutual information constraint

$$I\Big(\mathbf{X}^n; \psi^{(n)}\big(\phi_1^{(n)}(Y_1^n), \ldots, \phi_K^{(n)}(Y_K^n), Y_0^n\big)\Big) \geq n\Delta \ . \tag{4.22}$$

In this case, the region of optimal tuples $(R_1, \ldots, R_K, \Delta)$ generalizes the *Gaussian Information Bottleneck Function* of [21,22] as given by (4.24) to the setting in which the decoder observes correlated side information $\mathbf{Y}_0$ and the inference is done in a distributed manner by $K$ learners. This region can be obtained readily from Theorem 4 by substituting therein $\Delta := h(\mathbf{X}) - D$. The following corollary states the result, which was first established in [1, 111].

**Corollary 2.** *[111, Theorem 2] For the problem of distributed Gaussian Information Bottleneck with side information at the predictor, the complexity-relevance region is given by the union of all non-negative tuples $(R_1, \ldots, R_K, \Delta)$ that satisfy, for every $\mathcal{S} \subseteq \mathcal{K}$,*

$$\Delta \leq \sum_{k \in \mathcal{S}} \big[R_k + \log|\mathbf{I} - \mathbf{\Omega}_k \mathbf{\Sigma}_k|\big] + \log\big|\mathbf{I} + \mathbf{\Sigma}_{\mathbf{x}} \mathbf{H}_{\bar{\mathcal{S}}}^{\dagger} \mathbf{\Sigma}_{\mathbf{n}_{\bar{\mathcal{S}}}}^{-1}\big(\mathbf{I} - \mathbf{\Lambda}_{\bar{\mathcal{S}}} \mathbf{\Sigma}_{\mathbf{n}_{\bar{\mathcal{S}}}}^{-1}\big)\mathbf{H}_{\bar{\mathcal{S}}}\big| \ ,$$

*for matrices $\{\mathbf{\Omega}_k\}_{k=1}^K$ such that $\mathbf{0} \preceq \mathbf{\Omega}_k \preceq \mathbf{\Sigma}_k^{-1}$, where $\bar{\mathcal{S}} = \{0\} \cup \mathcal{S}^c$ and $\mathbf{\Lambda}_{\bar{\mathcal{S}}}$ is given by (4.7).* ■

In particular, if $K = 1$ and $\mathbf{Y}_0 = \emptyset$, with the substitutions $\mathbf{Y} := \mathbf{Y}_1$, $R := R_1$, $\mathbf{H} := \mathbf{H}_1$, $\mathbf{\Sigma} := \mathbf{\Sigma}_1$, and $\mathbf{\Omega}_1 := \mathbf{\Omega}$, the rate-distortion region of Theorem 4 reduces to the set of

[2] This problem is related to the Gaussian many-help-one problem [3, 4, 106]. Here, different from the setup of Figure 3.4, the source $X$ is observed directly by a *main encoder* who communicates with a detector that observes $Y$ in the aim of making a decision on whether $X$ and $Y$ are independent or not. Also, there are helpers that observe independent noisy versions of $X$ and communicate with the detector in the aim of facilitating that test.

rate-distortion pairs $(R, D)$ that satisfy

$$D \geq \log \left|(\pi e)\left(\mathbf{\Sigma}_{\mathbf{x}}^{-1} + \mathbf{H}^{\dagger}\mathbf{\Omega}\mathbf{H}\right)^{-1}\right| \tag{4.23a}$$

$$R + D \geq \log \frac{1}{|\mathbf{I} - \mathbf{\Omega}\mathbf{\Sigma}|} + \log \left|(\pi e)\mathbf{\Sigma}_{\mathbf{x}}\right| , \tag{4.23b}$$

for some matrix $\mathbf{\Omega}$ such that $\mathbf{0} \preceq \mathbf{\Omega} \preceq \mathbf{\Sigma}^{-1}$. Alternatively, by making the substitution $\Delta := h(\mathbf{X}) - D$, the trade-off expressed by (4.23) can be written equivalently as

$$\Delta \leq \log \left|\mathbf{I} + \mathbf{\Sigma}_{\mathbf{x}}\mathbf{H}^{\dagger}\mathbf{\Omega}\mathbf{H}\right| \tag{4.24a}$$

$$\Delta \leq R + \log \left|\mathbf{I} - \mathbf{\Omega}\mathbf{\Sigma}\right| , \tag{4.24b}$$

for some matrix $\mathbf{\Omega}$ such that $\mathbf{0} \preceq \mathbf{\Omega} \preceq \mathbf{\Sigma}^{-1}$.

Expression (4.24) is known as the *Gaussian Information Bottleneck Function* [21, 22], which is the solution of the Information Bottleneck method of [17] in the case of jointly Gaussian variables. More precisely, using the terminology of [17], the inequalities (4.24) describe the optimal trade-off between the complexity (or rate) $R$ and the relevance (or accuracy) $\Delta$. The concept of Information Bottleneck was found useful in various learning applications, such as for data clustering [112], feature selection [113] and others.

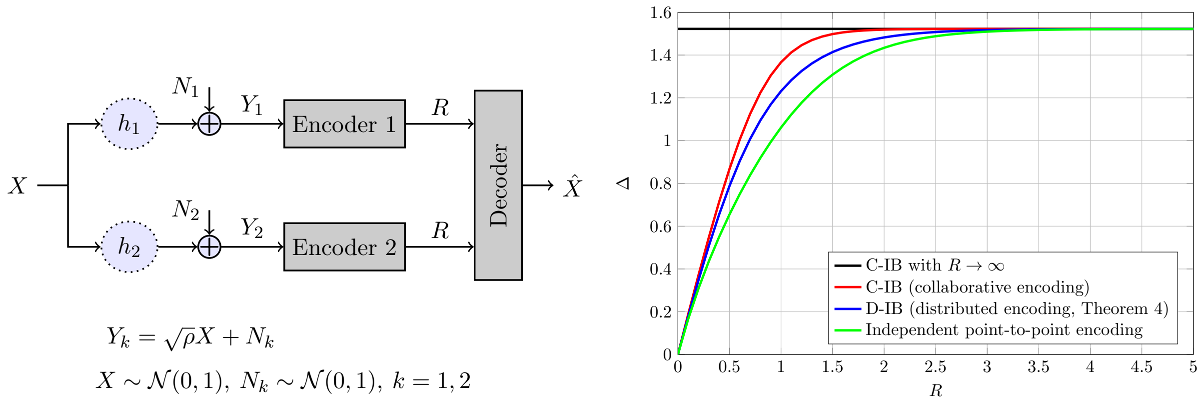
Furthermore, if in (4.1) and (4.2) the noises are independent among them and from $\mathbf{N}_0$, the relevance-complexity region of Corollary 2 reduces to the union of all non-negative tuples $(R_1, \dots, R_K, \Delta)$ that satisfy, for every $\mathcal{S} \subseteq \mathcal{K}$,

$$\Delta \leq \sum_{k \in \mathcal{S}} \left[R_k + \log |\mathbf{I} - \mathbf{\Omega}_k\mathbf{\Sigma}_k|\right] + \log \left|\mathbf{I} + \mathbf{\Sigma}_{\mathbf{x}}\left(\mathbf{H}_0^{\dagger}\mathbf{\Sigma}_0^{-1}\mathbf{H}_0 + \sum_{k \in \mathcal{S}^c} \mathbf{H}_k^{\dagger}\mathbf{\Omega}_k\mathbf{H}_k\right)\right| , \tag{4.25}$$

for some matrices $\{\mathbf{\Omega}_k\}_{k=1}^{K}$ such that $\mathbf{0} \preceq \mathbf{\Omega}_k \preceq \mathbf{\Sigma}_k^{-1}$.

**Example 1** (Distributed Scalar Gaussian Information Bottleneck)**.** Consider a scalar instance of the distributed Gaussian Information Bottleneck – that we study in this section – depicted in Figure 4.2a where there are two agents and no side information, i.e., $K = 2$, $\mathbf{Y}_0 = \emptyset$, $n_x = 1$ and $n_1 = n_2 = 1$. The relevance-complexity region of this model is given by (4.25) (wherein with the substitution $\mathbf{H}_0 = \mathbf{0}$). In particular, each encoder observation $Y_k$ is the output of a Gaussian channel with SNR $\rho_k$, i.e., $Y_k = \sqrt{\rho_k}X + N_k$, where $X \sim \mathcal{N}(0, 1)$, $N_k \sim \mathcal{N}(0, 1)$, $k = 1, 2$. Furthermore, the model we consider here is symmetric, i.e., $\rho_1 = \rho_2 = \rho$ and $R_1 = R_2 = R$. For this model, the optimal relevance-complexity pairs $(\Delta^\star, R)$ can be computed from

$$\Delta^\star(R,\rho) = \frac{1}{2}\log\left(1 + 2\rho\exp(-4R)\left[\exp(4R) + \rho - \sqrt{\rho^2 + (1+\rho)\exp(4R)}\right]\right) \ . \quad (4.26)$$




(a) System model

(b) Relevance vs. complexity trade-off.

Figure 4.2: Distributed Scalar Gaussian Information Bottleneck.

The Centralized IB (C-IB) upper bound is given by the pairs $(\Delta_{\text{cIB}}, R)$ achievable if $(Y_1, Y_2)$ are encoded jointly at a single encoder with complexity $2R$, and given by

$$\Delta_{\text{cIB}}(R,\rho) = \frac{1}{2}\log(1 + 2\rho) - \frac{1}{2}\log(1 + 2\rho\exp(-4R)) \ , \quad (4.27)$$

which is an instance of the scalar Gaussian IB problem in [22].

The lower bound is given by the pairs $(\Delta_{\text{ind}}, R)$ achievable if $(Y_1, Y_2)$ are encoded independently at separate encoders, and given by

$$\Delta_{\text{ind}}(R,\rho) = \frac{1}{2}\log(1 + 2\rho - \rho\exp(-2R)) - \frac{1}{2}\log(1 + \rho\exp(-2R)) \ . \quad (4.28)$$

Figure 4.2b shows the optimal relevance-complexity region of tuples $(\Delta^\star, R)$ obtained from (4.26), as well as, the C-IB upper bounds $\Delta_{\text{cIB}}(R,\rho)$ and $\Delta_{\text{cIB}}(\infty,\rho)$, and the lower bound $\Delta_{\text{ind}}(R,\rho)$ for the case in which the channel SNR is 10 dB, i.e., $\rho = 10$. ■

# Chapter 5

# Algorithms

This chapter contains a description of two algorithms and architectures that were developed in [1] for the distributed learning scenario. We state them here for reasons of completeness. In particular, the chapter provides: i) Blahut-Arimoto type iterative algorithms that allow to compute numerically the rate-distortion or relevance-complexity regions of the DM and vector Gaussian CEO problems for the case in which the joint distribution of the data is known perfectly or can be estimated with a high accuracy; and ii) a variational inference type algorithm in which the encoding mappings are parameterized by neural networks and the bound approximated by Monte Carlo sampling and optimized with stochastic gradient descent for the case in which there is only a set of training data is available.

## 5.1 Blahut-Arimoto Type Algorithms for Known Models

### 5.1.1 Discrete Case

Here we develop a BA-type algorithm that allows to compute the convex region $\mathcal{RD}^{\star}_{\mathrm{CEO}}$ for general discrete memoryless sources. To develop the algorithm, we use the Berger-Tung form of the region given in Proposition 11 for $K = 2$. The outline of the proposed method is as follows. First, we rewrite the rate-distortion region $\mathcal{RD}^{\star}_{\mathrm{CEO}}$ in terms of the union of two simpler regions in Proposition 6. The tuples lying on the boundary of each region are parametrically given in Proposition 7. Then, the boundary points of each simpler region are computed numerically via an alternating minimization method derived and detailed in Algorithm 2. Finally, the original rate-distortion region is obtained as the convex hull of the union of the tuples obtained for the two simple regions.

**Equivalent Parameterization**

Define the two regions $\mathcal{RD}^k_{\text{CEO}}$, $k = 1, 2$, as

$$\mathcal{RD}^k_{\text{CEO}} = \{(R_1, R_2, D) \ : \ D \geq D^k_{\text{CEO}}(R_1, R_2)\} \ , \tag{5.1}$$

with

$$\begin{aligned} D^k_{\text{CEO}}(R_1, R_2) := \min \quad & H(X|U_1, U_2, Y_0) \\ \text{s.t.} \quad & R_k \geq I(Y_k; U_k|U_{\bar{k}}, Y_0) \\ & R_{\bar{k}} \geq I(X_{\bar{k}}; U_{\bar{k}}|Y_0) \ , \end{aligned} \tag{5.2}$$

and the minimization is over set of joint measures $P_{U_1,U_2,X,Y_0,Y_1,Y_2}$ that satisfy $U_1 -\!\!\!\circ\!\!\!- Y_1 -\!\!\!\circ\!\!\!- (X, Y_0) -\!\!\!\circ\!\!\!- Y_2 -\!\!\!\circ\!\!\!- U_2$. (We define $\bar{k} := k \pmod 2) + 1$ for $k = 1, 2$.)

As stated in the following proposition, the region $\mathcal{RD}^\star_{\text{CEO}}$ of Theorem 1 coincides with the convex hull of the union of the two regions $\mathcal{RD}^1_{\text{CEO}}$ and $\mathcal{RD}^2_{\text{CEO}}$.

**Proposition 6.** *The region $\mathcal{RD}^\star_{\text{CEO}}$ is given by*

$$\mathcal{RD}^\star_{\text{CEO}} = \text{conv}(\mathcal{RD}^1_{\text{CEO}} \cup \mathcal{RD}^2_{\text{CEO}}) \ . \tag{5.3}$$

*Proof.* An outline of the proof is as follows. Let $P_{U_1,U_2,X,Y_0,Y_1,Y_2}$ and $P_Q$ be such that $(R_1, R_2, D) \in \mathcal{RD}^\star_{\text{CEO}}$. The polytope defined by the rate constraints (A.1), denoted by $\mathcal{V}$, forms a contra-polymatroid with 2! extreme points (vertices) [10,114]. Given a permutation $\pi$ on $\{1, 2\}$, the tuple

$$\tilde{R}_{\pi(1)} = I(Y_{\pi(1)}; U_{\pi(1)}|Y_0) \ , \quad \tilde{R}_{\pi(2)} = I(Y_{\pi(2)}; U_{\pi(2)}|U_{\pi(1)}, Y_0) \ ,$$

defines an extreme point of $\mathcal{V}$ for each permutation. As shown in [10], for every extreme point $(\tilde{R}_1, \tilde{R}_2)$ of $\mathcal{V}$, the point $(\tilde{R}_1, \tilde{R}_2, D)$ is achieved by time-sharing two successive Wyner-Ziv (WZ) strategies. The set of achievable tuples with such successive WZ scheme is characterized by the convex hull of $\mathcal{RD}^{\pi(1)}_{\text{CEO}}$. Convexifying the union of both regions as in (5.3), we obtain the full rate-distortion region $\mathcal{RD}^\star_{\text{CEO}}$. $\square$

The main advantage of Proposition 6 is that it reduces the computation of region $\mathcal{RD}^\star_{\text{CEO}}$ to the computation of the two regions $\mathcal{RD}^k_{\text{CEO}}$, $k = 1, 2$, whose boundary can be efficiently parameterized, leading to an efficient computational method. In what follows, we concentrate on $\mathcal{RD}^1_{\text{CEO}}$. The computation of $\mathcal{RD}^2_{\text{CEO}}$ follows similarly, and is omitted for brevity. Next proposition provides a parameterization of the boundary tuples of the region $\mathcal{RD}^1_{\mathrm{CEO}}$ in terms, each of them, of an optimization problem over the pmfs $\mathbf{P} := \{P_{U_1|Y_1}, P_{U_2|Y_2}\}$.

**Proposition 7.** *For each* $\mathbf{s} := [s_1, s_2]$*,* $s_1 > 0$*,* $s_2 > 0$*, define a tuple* $(R_{1,\mathbf{s}}, R_{2,\mathbf{s}}, D_{\mathbf{s}})$ *parametrically given by*

$$D_{\mathbf{s}} = -s_1 R_{1,\mathbf{s}} - s_2 R_{2,\mathbf{s}} + \min_{\mathbf{P}} F_{\mathbf{s}}(\mathbf{P}) \tag{5.4}$$

$$R_{1,\mathbf{s}} = I(Y_1; U_1^\star | U_2^\star, Y_0) \ , \qquad R_{2,\mathbf{s}} = I(Y_2; U_2^\star | Y_0) \ , \tag{5.5}$$

*where* $F_{\mathbf{s}}(\mathbf{P})$ *is given as follows*

$$F_{\mathbf{s}}(\mathbf{P}) := H(X|U_1, U_2, Y_0) + s_1 I(Y_1; U_1|U_2, Y_0) + s_2 I(Y_2; U_2|Y_0) \ , \tag{5.6}$$

*and;* $\mathbf{P}^\star$ *are the conditional pmfs yielding the minimum in* (5.4) *and* $U_1^\star, U_2^\star$ *are the auxiliary variables induced by* $\mathbf{P}^\star$*. Then, we have:*

1. *Each value of* $\mathbf{s}$ *leads to a tuple* $(R_{1,\mathbf{s}}, R_{2,\mathbf{s}}, D_{\mathbf{s}})$ *on the distortion-rate curve* $D_{\mathbf{s}} = D^1_{\mathrm{CEO}}(R_{1,\mathbf{s}}, R_{2,\mathbf{s}})$*.*

2. *For every point on the distortion-rate curve, there is an* $\mathbf{s}$ *for which* (5.4) *and* (5.5) *hold.*

*Proof.* Suppose that $\mathbf{P}^\star$ yields the minimum in (5.4). For this $\mathbf{P}$, we have $I(Y_1; U_1|U_2, Y_0) = R_{1,\mathbf{s}}$ and $I(Y_2; U_2|Y_0) = R_{2,\mathbf{s}}$. Then, we have

$$\begin{aligned} D_{\mathbf{s}} &= -s_1 R_{1,\mathbf{s}} - s_2 R_{2,\mathbf{s}} + F_{\mathbf{s}}(\mathbf{P}^\star) \\ &= -s_1 R_{1,\mathbf{s}} - s_2 R_{2,\mathbf{s}} + [H(X|U_1^\star, U_2^\star, Y_0) + s_1 R_{1,\mathbf{s}} + s_2 R_{2,\mathbf{s}}] \\ &= H(X|U_1^\star, U_2^\star, Y_0) \geq D^1_{\mathrm{CEO}}(R_{1,\mathbf{s}}, R_{2,\mathbf{s}}) \ . \end{aligned} \tag{5.7}$$

Conversely, if $\mathbf{P}^\star$ is the solution to the minimization in (5.2), then $I(Y_1; U_1^\star|U_2^\star, Y_0) \leq R_1$ and $I(Y_2; U_2^\star|Y_0) \leq R_2$ and for any $\mathbf{s}$,

$$\begin{aligned} D^1_{\mathrm{CEO}}(R_1, R_2) &= H(X|U_1^\star, U_2^\star, Y_0) \\ &\geq H(X|U_1^\star, U_2^\star, Y_0) + s_1(I(Y_1; U_1^\star|U_2^\star, Y_0) - R_1) + s_2(I(Y_2; U_2^\star|Y_0) - R_2) \\ &= D_{\mathbf{s}} + s_1(R_{1,\mathbf{s}} - R_1) + s_2(R_{2,\mathbf{s}} - R_2) \ . \end{aligned}$$

Given $\mathbf{s}$, and hence $(R_{1,\mathbf{s}}, R_{2,\mathbf{s}}, D_{\mathbf{s}})$, letting $(R_1, R_2) = (R_{1,\mathbf{s}}, R_{2,\mathbf{s}})$ yields $D^1_{\mathrm{CEO}}(R_{1,\mathbf{s}}, R_{2,\mathbf{s}}) \geq D_{\mathbf{s}}$, which proves, together with (5.7), statement 1) and 2). □

Next, we show that it is sufficient to run the algorithm for $s_1 \in (0,1]$.

**Lemma 2.** *The range of the parameter $s_1$ can be restricted to* $(0,1]$.

*Proof.* Let $F^\star = \min_{\mathbf{P}} F_{\mathbf{s}}(\mathbf{P})$. If we set $U_1 = \emptyset$, then we have the relation

$$F^\star \leq H(X|U_2,Y_0) + s_2 I(Y_2;U_2|Y_0) \ .$$

For $s_1 > 1$, we have

$$\begin{aligned} F_{\mathbf{s}}(\mathbf{P}) &\overset{(a)}{\geq} (1-s_1)H(X|U_1,U_2,Y_0) + s_1 H(X|U_2,Y_0) + s_2 I(Y_2;U_2|Y_0) \\ &\overset{(b)}{\geq} H(X|U_2,Y_0) + s_2 I(Y_2;U_2|Y_0) \ , \end{aligned}$$

where $(a)$ follows since mutual information is always positive, i.e., $I(Y_1;U_1|X,Y_0) \geq 0$; $(b)$ holds since conditioning reduces entropy and $1-s_1 < 0$. Then,

$$F^\star = H(X|U_2,Y_0) + s_2 I(Y_2;U_2|Y_0) \ , \quad \text{for } s_1 > 1 \ .$$

Hence, we can restrict the range of $s_1$ to $s_1 \in (0,1]$. □

**Computation of $\mathcal{RD}^1_{\mathrm{CEO}}$**

In this section, we derive an algorithm to solve (5.4) for a given parameter value $\mathbf{s}$. To that end, we define a variational bound on $F_{\mathbf{s}}(\mathbf{P})$, and optimize it instead of (5.4). Let $\mathbf{Q}$ be a set of some auxiliary pmfs defined as

$$\mathbf{Q} := \{Q_{U_1} \ , \ Q_{U_2} \ , \ Q_{X|U_1,U_2,Y_0} \ , \ Q_{X|U_1,Y_0} \ , \ Q_{X|U_2,Y_0} \ , \ Q_{Y_0|U_1} \ , \ Q_{Y_0|U_2}\} \ . \tag{5.8}$$

In the following we define the variational cost function $F_{\mathbf{s}}(\mathbf{P},\mathbf{Q})$

$$\begin{aligned} F_{\mathbf{s}}(\mathbf{P},\mathbf{Q}) :=& - s_1 H(X|Y_0) - (s_1+s_2)H(Y_0) \\ &+ \mathbb{E}_{P_{X,Y_0,Y_1,Y_2}}\Big[(1-s_1)\mathbb{E}_{P_{U_1|Y_1}}\mathbb{E}_{P_{U_2|Y_2}}[-\log Q_{X|U_1,U_2,Y_0}] \\ &\qquad + s_1\mathbb{E}_{P_{U_1|Y_1}}[-\log Q_{X|U_1,Y_0}] + s_1\mathbb{E}_{P_{U_2|Y_2}}[-\log Q_{X|U_2,Y_0}] \\ &\qquad + s_1 D_{\mathrm{KL}}(P_{U_1|Y_1}\|Q_{U_1}) + s_2 D_{\mathrm{KL}}(P_{U_2|Y_2}\|Q_{U_2}) \\ &\qquad + s_1\mathbb{E}_{P_{U_1|Y_1}}[-\log Q_{Y_0|U_1}] + s_2\mathbb{E}_{P_{U_2|Y_2}}[-\log Q_{Y_0|U_2}]\Big] \ . \end{aligned} \tag{5.9}$$

The following lemma states that $\mathcal{L}_s(\mathbf{P},\mathbf{Q})$ is an upper bound on $\mathcal{L}_s(\mathbf{P})$ for all distributions $\mathbf{Q}$.

**Lemma 3.** *For fixed* $\mathbf{P}$*, we have*

$$\mathcal{L}_s(\mathbf{P},\mathbf{Q}) \geq \mathcal{L}_s(\mathbf{P}) \ , \quad \textit{for all } \mathbf{Q} \ .$$

*In addition, there exists a* $\mathbf{Q}$ *that achieves the minimum* $\min_{\mathbf{Q}} F_{\mathbf{s}}(\mathbf{P},\mathbf{Q}) = F_{\mathbf{s}}(\mathbf{P})$*, given by*

$$\begin{aligned}&Q_{U_k} = P_{U_k} \ , \quad Q_{X|U_k,Y_0} = P_{X|U_k,Y_0} \ , \quad Q_{Y_0|U_k} = P_{Y_0|U_k} \ , \quad \textit{for } \ k=1,2 \ ,\\ &Q_{X|U_1,U_2,Y_0} = P_{X|U_1,U_2,Y_0} \ .\end{aligned} \tag{5.10}$$

*Proof.* The proof of Lemma 3 is given in Appendix H.1. □

Using the lemma above, the minimization in (5.4) can be written in terms of the variational cost function as follows

$$\min_{\mathbf{P}} F_{\mathbf{s}}(\mathbf{P}) = \min_{\mathbf{P}} \min_{\mathbf{Q}} F_{\mathbf{s}}(\mathbf{P},\mathbf{Q}) \ . \tag{5.11}$$

Motivated by the BA algorithm [19,20], we propose an alternate optimization procedure over the set of pmfs $\mathbf{P}$ and $\mathbf{Q}$ as stated in Algorithm 2. The main idea is that at iteration $t$, for fixed $\mathbf{P}^{(t-1)}$ the optimal $\mathbf{Q}^{(t)}$ minimizing $F_{\mathbf{s}}(\mathbf{P},\mathbf{Q})$ can be found analytically; next, for given $\mathbf{Q}^{(t)}$ the optimal $\mathbf{P}^{(t)}$ that minimizes $F_{\mathbf{s}}(\mathbf{P},\mathbf{Q})$ has also a closed form. So, starting with a random initialization $\mathbf{P}^{(0)}$, the algorithm iterates over distributions $\mathbf{Q}$ and $\mathbf{P}$ minimizing $F_{\mathbf{s}}(\mathbf{P},\mathbf{Q})$ until the convergence, as stated below

$$\mathbf{P}^{(0)} \to \mathbf{Q}^{(1)} \to \mathbf{P}^{(1)} \to \ldots \to \mathbf{P}^{(t)} \to \mathbf{Q}^{(t)} \to \ldots \to \mathbf{P}^{\star} \to \mathbf{Q}^{\star} \ .$$

At each iteration, the optimal values of $\mathbf{P}$ and $\mathbf{Q}$ are found by solving a convex optimization problems. We have the following lemma.

**Lemma 4.** $F_{\mathbf{s}}(\mathbf{P},\mathbf{Q})$ *is convex in* $\mathbf{P}$ *and convex in* $\mathbf{Q}$*.*

*Proof.* The proof of Lemma 4 follows from the log-sum inequality. □

For fixed $\mathbf{P}^{(t-1)}$, the optimal $\mathbf{Q}^{(t)}$ minimizing the variational bound in (5.9) can be found from Lemma 3 and given by (5.10). For fixed $\mathbf{Q}^{(t)}$, the optimal $\mathbf{P}^{(t)}$ minimizing (5.9) can be found by using the next lemma.

**Lemma 5.** *For fixed* $\mathbf{Q}$*, there exists a* $\mathbf{P}$ *that achieves the minimum* $\min_{\mathbf{P}} F_{\mathbf{s}}(\mathbf{P},\mathbf{Q})$*, where* $P_{U_k|Y_k}$ *is given by*

$$p(u_k|y_k) = q(u_k)\frac{\exp[-\psi_k(u_k,y_k)]}{\sum_{u_k} q(u_k)\exp[-\psi_k(u_k,y_k)]} \ , \quad \textit{for } k=1,2 \ , \tag{5.12}$$

*where* $\psi_k(u_k, y_k)$, $k = 1, 2$, *are defined as follows*

$$\psi_k(u_k, y_k) := \frac{1 - s_1}{s_k} \mathbb{E}_{U_{\bar{k}}, Y_0|y_k}[D_{\mathrm{KL}}(P_{X|y_k, U_{\bar{k}}, Y_0} \| Q_{X|u_k, U_{\bar{k}}, Y_0})] \\ + \frac{s_1}{s_k} \mathbb{E}_{Y_0|y_k} D_{\mathrm{KL}}[(P_{X|y_k, Y_0} \| Q_{X|u_k, Y_0})] + D_{\mathrm{KL}}(P_{Y_0|y_k} \| Q_{Y_0|u_k}) \,. \tag{5.13}$$

*Proof.* The proof of Lemma 5 is given in Appendix H.2. □

---

**Algorithm 2** BA-type algorithm to compute $\mathcal{RD}^1_{\mathrm{CEO}}$

---

**input:** pmf $P_{X, Y_0, Y_1, Y_2}$, parameters $1 \geq s_1 > 0$, $s_2 > 0$.

**output:** Optimal $P^\star_{U_1|Y_1}$, $P^\star_{U_2|Y_2}$; triple $(R_{1,\mathbf{s}}, R_{2,\mathbf{s}}, D_{\mathbf{s}})$.

**initialization** Set $t = 0$. Set $\mathbf{P}^{(0)}$ randomly.

**repeat**

  Update the following pmfs for $k = 1, 2$

$$\begin{aligned} p^{(t+1)}(u_k) &= \sum\nolimits_{y_k} p^{(t)}(u_k|y_k) p(y_k), \\ p^{(t+1)}(u_k|y_0) &= \sum\nolimits_{y_k} p^{(t)}(u_k|y_k) p(y_k|y_0), \\ p^{(t+1)}(u_k|x, y_0) &= \sum\nolimits_{y_k} p^{(t)}(u_k|y_k) p(y_k|x, y_0), \\ p^{(t+1)}(x|u_1, u_2, y_0) &= \frac{p^{(t+1)}(u_1|x, y_0) p^{(t+1)}(u_2|x, y_0) p(x, y_0)}{\sum_x p^{(t+1)}(u_1|x, y_0) p^{(t+1)}(u_2|x, y_0) p(x, y_0)} \,. \end{aligned}$$

  Update $\mathbf{Q}^{(t+1)}$ by using (5.10).

  Update $\mathbf{P}^{(t+1)}$ by using (5.12).

  $t \leftarrow t + 1$.

**until** convergence.

---

At each iteration of Algorithm 2, $F_{\mathbf{s}}(\mathbf{P}^{(t)}, \mathbf{Q}^{(t)})$ decreases until eventually it converges. However, since $F_{\mathbf{s}}(\mathbf{P}, \mathbf{Q})$ is convex in each argument but not necessarily jointly convex, Algorithm 2 does not necessarily converge to the global optimum. In particular, next proposition shows that Algorithm 2 converges to a stationary solution of the minimization in (5.4).

**Proposition 8.** *Every limit point of* $\mathbf{P}^{(t)}$ *generated by Algorithm 2 converges to a stationary solution of* (5.4).

*Proof.* Algorithm 2 falls into the class of so-called *Successive Upper-bound Minimization* (SUM) algorithms [115], in which $F_{\mathbf{s}}(\mathbf{P}, \mathbf{Q})$ acts as a globally tight upper bound on $F_{\mathbf{s}}(\mathbf{P})$. Let $\mathbf{Q}^\star(\mathbf{P}) := \arg\min_{\mathbf{Q}} F_{\mathbf{s}}(\mathbf{P}, \mathbf{Q})$. From Lemma 3, $F_{\mathbf{s}}(\mathbf{P}, \mathbf{Q}^\star(\mathbf{P}')) \geq F_{\mathbf{s}}(\mathbf{P}, \mathbf{Q}^\star(\mathbf{P})) = F_{\mathbf{s}}(\mathbf{P})$ for $\mathbf{P}' \neq \mathbf{P}$. It follows that $F_{\mathbf{s}}(\mathbf{P})$ and $F_{\mathbf{s}}(\mathbf{P}, \mathbf{Q}^\star(\mathbf{P}'))$ satisfy [115, Proposition 1] and thus

$F_{\mathbf{s}}(\mathbf{P}, \mathbf{Q}^{\star}(\mathbf{P}'))$ satisfies (A1)–(A4) in [115]. Convergence to a stationary point of (5.4) follows from [115, Theorem 1]. □

**Remark 12.** *Algorithm 2 generates a sequence that is non-increasing. Since this sequence is lower bounded, convergence to a stationary point is guaranteed. This per-se, however, does not necessarily imply that such a point is a stationary solution of the original problem described by* (5.4)*. Instead, this is guaranteed here by showing that the Algorithm 2 is of SUM-type with the function* $F_{\mathbf{s}}(\mathbf{P}, \mathbf{Q})$ *satisfying the necessary conditions [115, (A1)–(A4)].* ■

### 5.1.2 Vector Gaussian Case

Computing the rate-distortion region $\mathcal{RD}^{\star}_{\text{VG-CEO}}$ of the vector Gaussian CEO problem as given by Theorem 4 is a convex optimization problem on $\{\mathbf{\Omega}_k\}_{k=1}^K$ which can be solved using, e.g., the popular generic optimization tool CVX [116]. Alternatively, the region can be computed using an extension of Algorithm 2 to memoryless Gaussian sources as given in the rest of this section.

**Algorithm 3** BA-type algorithm for the Gaussian vector CEO

1: **input:** Covariance $\mathbf{\Sigma}_{(\mathbf{x},\mathbf{y}_0,\mathbf{y}_1,\mathbf{y}_2)}$, parameters$1 \geq s_1 > 0$, $s_2 > 0$.
2: **output:** Optimal pairs $(\mathbf{A}_k^{\star}, \mathbf{\Sigma}_{\mathbf{z}_k^{\star}})$, $k = 1, 2$.
3: **initialization** Set $t = 0$. Set randomly $\mathbf{A}_k^0$ and $\mathbf{\Sigma}_{\mathbf{z}_k^0} \succeq 0$ for $k = 1, 2$.
4: **repeat**
5: For $k = 1, 2$, update the following

$$\begin{aligned} \mathbf{\Sigma}_{\mathbf{u}_k^t} &= \mathbf{A}_k^t \mathbf{\Sigma}_{\mathbf{y}_k} \mathbf{A}_k^{t\,\dagger} + \mathbf{\Sigma}_{\mathbf{z}_k^t} \\ \mathbf{\Sigma}_{\mathbf{u}_k^t|(\mathbf{x},\mathbf{y})} &= \mathbf{A}_k^t \mathbf{\Sigma}_k \mathbf{A}_k^{t\,\dagger} + \mathbf{\Sigma}_{\mathbf{z}_k^t} \ , \end{aligned}$$

and update $\mathbf{\Sigma}_{\mathbf{u}_k^t|(\mathbf{u}_{\bar{k}}^t,\mathbf{y})}$, $\mathbf{\Sigma}_{\mathbf{u}_2^t|\mathbf{y}}$ and $\mathbf{\Sigma}_{\mathbf{y}_k^t|(\mathbf{u}_{\bar{k}}^t,\mathbf{y})}$ from their definitions by using the following

$$\begin{aligned} \mathbf{\Sigma}_{\mathbf{u}_1^t,\mathbf{u}_2^t} &= \mathbf{A}_1^t \mathbf{H}_1 \mathbf{\Sigma}_{\mathbf{x}} \mathbf{H}_2^{\dagger} \mathbf{A}_2^{t\dagger} \\ \mathbf{\Sigma}_{\mathbf{u}_k^t,\mathbf{y}} &= \mathbf{A}_k^t \mathbf{H}_k \mathbf{\Sigma}_{\mathbf{x}} \mathbf{H}_0^{\dagger} \\ \mathbf{\Sigma}_{\mathbf{y}_k,\mathbf{u}_{\bar{k}}^t} &= \mathbf{H}_k \mathbf{\Sigma}_{\mathbf{x}} \mathbf{H}_{\bar{k}}^{\dagger} \mathbf{A}_{\bar{k}}^{t\,\dagger} \ . \end{aligned}$$

6: Compute $\mathbf{\Sigma}_{\mathbf{z}_k^{t+1}}$ as in (5.16a) for $k = 1, 2$.
7: Compute $\mathbf{A}_k^{t+1}$ as (5.16b) for $k = 1, 2$.
8: $t \leftarrow t + 1$.
9: **until** convergence.

For discrete sources with (small) alphabets, the updating rules of $\mathbf{Q}^{(t+1)}$ and $\mathbf{P}^{(t+1)}$ of

Algorithm 2 are relatively easy computationally. However, they become computationally unfeasible for continuous alphabet sources. Here, we leverage on the optimality of Gaussian test channels as shown by Theorem 4 to restrict the optimization of $\mathbf{P}$ to Gaussian distributions, which allows to reduce the search of update rules to those of the associated parameters, namely covariance matrices. In particular, we show that if $P^{(t)}_{\mathbf{U}_k|\mathbf{Y}_k}$, $k=1,2$, is Gaussian and such that

$$\mathbf{U}_k^t = \mathbf{A}_k^t \mathbf{Y}_k + \mathbf{Z}_k^t \;, \tag{5.14}$$

where $\mathbf{Z}_k^t \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Sigma}_{\mathbf{z}_k^t})$, then $P^{(t+1)}_{\mathbf{U}_k|\mathbf{Y}_k}$ is also Gaussian, with

$$\mathbf{U}_k^{t+1} = \mathbf{A}_k^{t+1} \mathbf{Y}_k + \mathbf{Z}_k^{t+1} \;, \tag{5.15}$$

where $\mathbf{Z}_k^{t+1} \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Sigma}_{\mathbf{z}_k^{t+1}})$ and the parameters $\mathbf{A}_k^{t+1}$ and $\mathbf{\Sigma}_{\mathbf{z}_k^{t+1}}$ are given by

$$\mathbf{\Sigma}_{\mathbf{z}_k^{t+1}} = \left( \frac{1}{s_k} \mathbf{\Sigma}^{-1}_{\mathbf{u}_k^t|(\mathbf{x},\mathbf{y}_0)} - \frac{1-s_1}{s_k} \mathbf{\Sigma}^{-1}_{\mathbf{u}_k^t|(\mathbf{u}_{\bar{k}}^t,\mathbf{y}_0)} + \frac{s_k - s_1}{s_k} \mathbf{\Sigma}^{-1}_{\mathbf{u}_k^t|\mathbf{y}_0} \right)^{-1} \tag{5.16a}$$

$$\begin{aligned} \mathbf{A}_k^{t+1} = \; & \mathbf{\Sigma}_{\mathbf{z}_k^{t+1}} \left( \frac{1}{s_k} \mathbf{\Sigma}^{-1}_{\mathbf{u}_k^t|(\mathbf{x},\mathbf{y}_0)} \mathbf{A}_k^t (\mathbf{I} - \mathbf{\Sigma}_{\mathbf{y}_k|(\mathbf{x},\mathbf{y}_0)} \mathbf{\Sigma}^{-1}_{\mathbf{y}_k}) \right) \\ & - \mathbf{\Sigma}_{\mathbf{z}_k^{t+1}} \Bigg( \frac{1-s_1}{s_k} \mathbf{\Sigma}^{-1}_{\mathbf{u}_k^t|(\mathbf{u}_{\bar{k}}^t,\mathbf{y}_0)} \mathbf{A}_k^t (\mathbf{I} - \mathbf{\Sigma}_{\mathbf{y}_k|(\mathbf{u}_{\bar{k}}^t,\mathbf{y}_0)} \mathbf{\Sigma}^{-1}_{\mathbf{y}_k}) \\ & \qquad - \frac{s_k - s_1}{s_k} \mathbf{\Sigma}^{-1}_{\mathbf{u}_k^t|\mathbf{y}_0} \mathbf{A}_k^t (\mathbf{I} - \mathbf{\Sigma}_{\mathbf{y}_k|\mathbf{y}_0} \mathbf{\Sigma}^{-1}_{\mathbf{y}_k}) \Bigg) \; . \end{aligned} \tag{5.16b}$$

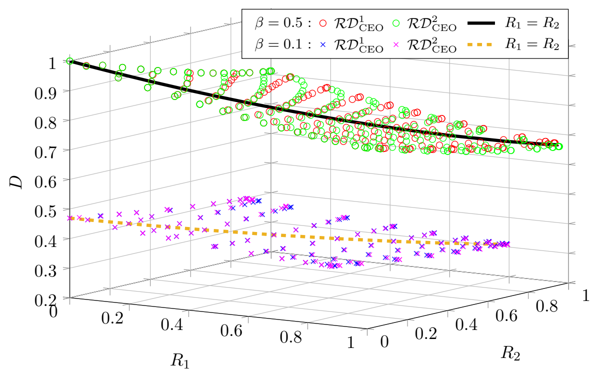
The updating steps are provided in Algorithm 3. The proof of (5.16) can be found in Appendix H.3.

### 5.1.3 Numerical Examples

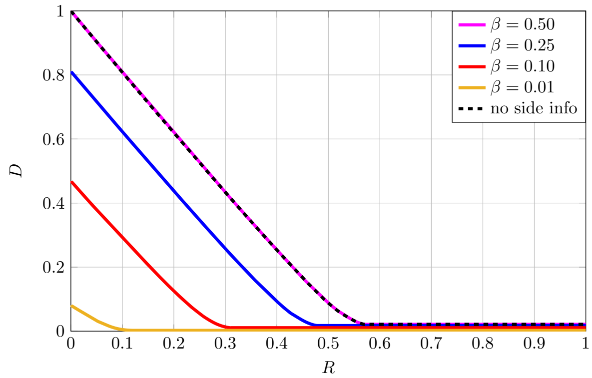
In this section, we discuss two examples, a binary CEO example and a vector Gaussian CEO example.

**Example 2.** Consider the following binary CEO problem. A memoryless binary source $X$, modeled as a Bernoulli-(1/2) random variable, i.e., $X \sim \text{Bern}(1/2)$, is observed remotely at two agents who communicate with a central unit decoder over error-free rate-limited links of capacity $R_1$ and $R_2$, respectively. The decoder wants to estimate the remote source $X$ to within some average fidelity level $D$, where the distortion is measured under the logarithmic loss criterion. The noisy observation $Y_1$ at Agent 1 is modeled as the output of a binary symmetric channel (BSC) with crossover probability $\alpha_1 \in [0,1]$, whose input is

$X$, i.e., $Y_1 = X \oplus S_1$ with $S_1 \sim \text{Bern}(\alpha_1)$. Similarly, the noisy observation $Y_2$ at Agent 2 is modeled as the output of a BSC($\alpha_2$) channel, $\alpha_2 \in [0,1]$, whose has input $X$, i.e., $Y_2 = X \oplus S_2$ with $S_2 \sim \text{Bern}(\alpha_2)$. Also, the central unit decoder observes its own side information $Y_0$ in the form of the output of a BSC($\beta$) channel, $\beta \in [0,1]$, whose input is $X$, i.e., $Y_0 = X \oplus S_0$ with $S_0 \sim \text{Bern}(\beta)$. It is assumed that the binary noises $S_0$, $S_1$ and $S_2$ are independent between them and with the remote source $X$.




(a)




(b)

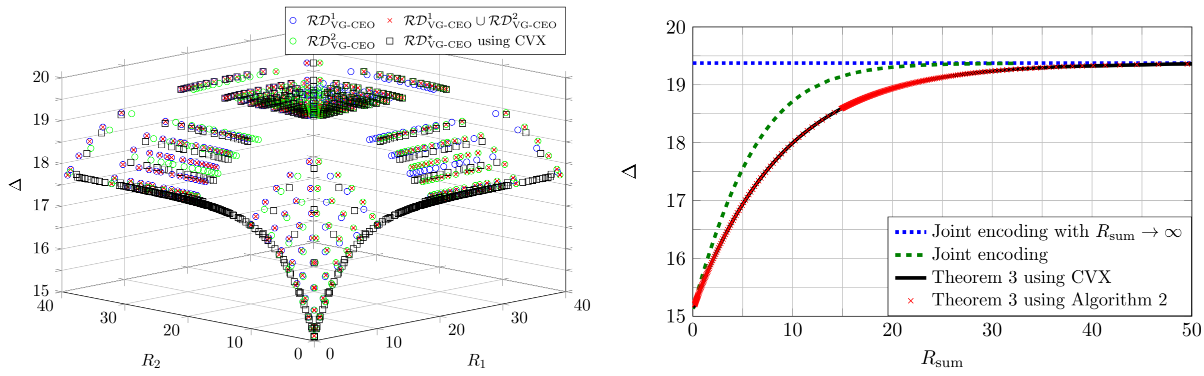
Figure 5.1: Rate-distortion region of the binary CEO network of Example 2, computed using Algorithm 2. (a): set of $(R_1, R_2, D)$ triples such $(R_1, R_2, D) \in \mathcal{RD}^1_{\text{CEO}} \cup \mathcal{RD}^2_{\text{CEO}}$, for $\alpha_1 = \alpha_2 = 0.25$ and $\beta \in \{0.1, 0.25\}$. (b): set of $(R, D)$ pairs such $(R, R, D) \in \mathcal{RD}^1_{\text{CEO}} \cup \mathcal{RD}^2_{\text{CEO}}$, for $\alpha_1 = \alpha_2 = 0.01$ and $\beta \in \{0.01, 0.1, 0.25, 0.5\}$.

We use Algorithm 2 to numerically approximate[1] the set of $(R_1, R_2, D)$ triples such that $(R_1, R_2, D)$ is in the union of the achievable regions $\mathcal{RD}^1_{\text{CEO}}$ and $\mathcal{RD}^2_{\text{CEO}}$ as given by (5.1). The regions are depicted in Figure 5.1a for the values $\alpha_1 = \alpha_2 = 0.25$ and $\beta \in \{0.1, 0.25\}$. Note that for both values of $\beta$, an approximation of the rate-distortion region $\mathcal{RD}_{\text{CEO}}$ is easily found as the convex hull of the union of the shown two regions. For simplicity, Figure 5.1b shows achievable rate-distortion pairs $(R, D)$ in the case in which the rates of the two encoders are constrained to be at most $R$ bits per channel use each, i.e., $R_1 = R_2 = R$, higher quality agents' observations $(Y_1, Y_2)$ corresponding to $\alpha_1 = \alpha_2 = 0.01$ and $\beta \in \{0.01, 0.1, 0.25, 0.5\}$. In this figure, observe that, as expected, smaller values of $\beta$ correspond to higher quality estimate side information $Y_0$ at the decoder; and lead to

[1] We remind the reader that, as already mentioned, Algorithm 2 only converges to stationary points of the rate-distortion region.

smaller distortion values for given rate $R$. The choice $\beta = 0.5$ corresponds to the case of no or independent side information at decoder; and it is easy to check that the associated $(R, D)$ curve coincides with the one obtained through exhaustive search in [10, Figure 3].

■




(a) (b)

Figure 5.2: Rate-information region of the vector Gaussian CEO network of Example 3. Numerical values are $n_x = 3$ and $n_0 = n_1 = n_2 = 4$. (a): set of $(R_1, R_2, \Delta)$ triples such $(R_1, R_2, h(\mathbf{X}) - \Delta) \in \mathcal{RD}^1_{\text{VG-CEO}} \cup \mathcal{RD}^2_{\text{VG-CEO}}$, computed using Algorithm 3. (b): set of $(R_{\text{sum}}, \Delta)$ pairs such $R_{\text{sum}} = R_1 + R_2$ for some $(R_1, R_2)$ for which $(R_1, R_2, h(\mathbf{X}) - \Delta) \in \mathcal{RD}^1_{\text{VG-CEO}} \cup \mathcal{RD}^2_{\text{VG-CEO}}$.

**Example 3.** Consider an instance of the memoryless vector Gaussian CEO problem as described by (4.1) and (4.2) obtained by setting $K = 2$, $n_x = 3$ and $n_0 = n_1 = n_2 = 4$. We use Algorithm 3 to numerically approximate the set of $(R_1, R_2, \Delta)$ triples such $(R_1, R_2, h(\mathbf{X}) - \Delta)$ is in the union of the achievable regions $\mathcal{RD}^1_{\text{VG-CEO}}$ and $\mathcal{RD}^2_{\text{VG-CEO}}$. The result is depicted in Figure 5.2a. The figure also shows the set of $(R_1, R_2, \Delta)$ triples such that $(R_1, R_2, h(\mathbf{X}) - \Delta)$ lies in the region given by Theorem 4 evaluated for the example at hand. Figure 5.2b shows the set of $(R_{\text{sum}}, \Delta)$ pairs such $R_{\text{sum}} := R_1 + R_2$ for some $(R_1, R_2)$ for which $(R_1, R_2, h(\mathbf{X}) - \Delta)$ is in the union of $\mathcal{RD}^1_{\text{VG-CEO}}$ and $\mathcal{RD}^2_{\text{VG-CEO}}$. The region is computed using two different approaches: i) using Algorithm 3 and ii) by directly evaluating the region obtained from Theorem 4 using the CVX optimization tool to find the maximizing covariance matrices $(\mathbf{\Omega}_1, \mathbf{\Omega}_2)$ (note that this problem is convex and so CVX finds the optimal solution). It is worth-noting that Algorithm 3 converges to the optimal solution for the studied vector Gaussian CEO example, as is visible from the figure. For comparisons reasons, the figure also shows the performance of *centralized* or *joint* encoding, i.e., the case both agents observe both $\mathbf{Y}_1$ and $\mathbf{Y}_2$,

$$\Delta(R_{\text{sum}}) = \max_{P_{U|\mathbf{Y}_1,\mathbf{Y}_2} : I(U;\mathbf{Y}_1,\mathbf{Y}_2|\mathbf{Y}_0) \leq R_{\text{sum}}} I(U,\mathbf{Y}_0;\mathbf{X}) \ . \tag{5.17}$$

Finally, we note that the information/sum-rate function (5.17) can be seen an extension of Chechik *et al.* Gaussian Information Bottleneck [21] to the case of side information $\mathbf{Y}_0$ at the decoder. Figure 5.2b shows the loss in terms of information/sum-rate that is incurred by restricting the encoders to operate separately, i.e., distributed Information Bottleneck with side information at decoder. ■

## 5.2 Deep Distributed Representation Learning

Consider the $K$-encoder CEO problem under logarithmic loss that we studied in Chapter 3. In this section, we study the case in which there is no side information, i.e., $Y_0 = \emptyset$. The $K$-encoder CEO source coding problem under logarithmic loss distortion is essentially a distributed learning model, in which the decoder is interested in a soft estimate of $X$ and the inference is done in a distributed manner by $K$ learners (encoders).

Let the logarithmic loss distortion constraint of the CEO problem be replaced by the mutual information constraint

$$I\Big(X^n; \psi^{(n)}\big(\phi_1^{(n)}(Y_1^n), \ldots, \phi_K^{(n)}(Y_K^n)\big)\Big) \geq n\Delta \ . \tag{5.18}$$

In this case, the region $\mathcal{RI}_{\text{DIB}}$ of optimal relevance-complexity tuples $(R_1, \ldots, R_K, \Delta)$ generalizes the Tishby's Information Bottleneck [17] to the distributed case, which is called as Distributed Information Bottleneck (DIB) problem [1]. Since these two problems are equivalent, the region $\mathcal{RI}_{\text{DIB}}$ can be characterized using the relevance-complexity region $\mathcal{RD}^{\star}_{\text{CEO}}$ given in Theorem 1 by substituting therein $\Delta := H(X) - D$. The following corollary states the result.

**Corollary 3.** *The relevance-complexity region $\mathcal{RI}_{\text{DIB}}$ of the distributed learning problem is given by the set of all non-negative relevance-complexity tuples $(R_1, \ldots, R_K, \Delta)$ that satisfy, for all subsets $\mathcal{S} \subseteq \mathcal{K}$,*

$$\Delta \leq \sum_{k \in \mathcal{S}} [R_k - I(Y_k; U_k|X, Q)] + I(X; U_{\mathcal{S}^c}, Q) \ ,$$

*for some auxiliary random variables* $(U_1,\ldots,U_K,Q)$ *with distribution* $P_{U_{\mathcal{K}},Q}(u_{\mathcal{K}},q)$ *such that* $P_{X,Y_{\mathcal{K}},U_{\mathcal{K}},Q}(x,y_{\mathcal{K}},u_{\mathcal{K}},q)$ *factorizes as*

$$P_X(x)\prod_{k=1}^{K}P_{Y_k|X}(y_k|x)\;P_Q(q)\prod_{k=1}^{K}P_{U_k|Y_k,Q}(u_k|y_k,q)\;. \qquad \blacksquare$$

**Remark 13.** *The optimal relevance-complexity tuples* $(R_1,\ldots,R_K,\Delta)$ *of the DIB problem – characterized by Corollary 3 – can be found by solving an optimization problem on* $\{P_{U_k|Y_k,Q}\}_{k=1}^K$ *and* $P_Q$*. Here,* $P_{U_k|Y_k,Q}$ *is the* $k$*-th stochastic encoding that maps the observation* $Y_k$ *to a latent representation* $U_k$ *such that* $U_k$ *captures the relevant information about* $X$ *(similar to the single encoder IB problem), and* $P_Q$ *is the pmf of the time-sharing variable* $Q$ *among* $K$ *encoders. The corresponding optimal decoding mapping is denoted by* $P_{X|U_1,\ldots,U_K,Q}$ *for given* $\{P_{U_k|Y_k,Q}\}_{k=1}^K$ *and* $P_Q$*.* $\blacksquare$

For simplicity, the relevance is maximized under sum-complexity constraint, i.e., $R_{\text{sum}} := \sum_{k=1}^K R_k$. The achievable relevance-complexity region under sum-complexity constraint is defined by

$$\begin{aligned}\mathcal{RI}_{\text{DIB}}^{\text{sum}} := \Big\{(\Delta,R_{\text{sum}}) \in \mathbb{R}_+^2 \;:\; \exists (R_1,\ldots,R_K)\in\mathbb{R}_+^K \quad \text{s.t.}\\ (R_1,\ldots,R_K,\Delta)\in\mathcal{RI}_{\text{DIB}} \quad \text{and} \quad \sum_{k=1}^K R_k = R_{\text{sum}}\Big\}\;.\end{aligned}$$

The region $\mathcal{RI}_{\text{DIB}}^{\text{sum}}$ can be characterized as given in the following proposition.

**Proposition 9.** *[100, Proposition 1] The relevance-complexity region under sum-complexity constraint* $\mathcal{RI}_{\text{DIB}}^{\text{sum}}$ *is given by the convex-hull of all non-negative tuples* $(\Delta,R_{\text{sum}})$ *that satisfy* $\Delta\leq\Delta_{\text{DIB}}^{\text{sum}}(R_{\text{sum}})$ *where*

$$\Delta_{\text{DIB}}^{\text{sum}}(R_{\text{sum}}) := \max_{\mathbf{P}}\;\min\left\{I(X;U_{\mathcal{K}}),R_{\text{sum}}-\sum_{k=1}^K I(Y_k;U_k|X)\right\}, \tag{5.19}$$

*in which the maximization is over the set of conditional pmfs* $\mathbf{P}:=\{P_{U_1|Y_1},\ldots,P_{U_K|Y_K}\}$.

*Proof.* The proof of Proposition 9 is given in Appendix H.4. $\square$

Next proposition provides a parameterization of the boundary tuples $(\Delta_s,R_s)$ of the region $\mathcal{RI}_{\text{DIB}}^{\text{sum}}$ in terms of a parameter $s\geq 0$.

**Proposition 10.** *For each tuple* $(\Delta,R_{\text{sum}})$ *on the boundary of the relevance-complexity region* $\mathcal{RI}_{\text{DIB}}^{\text{sum}}$ *there exists* $s\geq0$ *such that* $(\Delta,R_{\text{sum}})=(\Delta_s,R_s)$*, where*

$$\Delta_s := \frac{1}{1+s}\Big[(1+sK)H(X)+sR_s+\max_{\mathbf{P}}\mathcal{L}_s^{\text{DIB}}(\mathbf{P})\Big] \tag{5.20}$$

$$R_s := I(X; U_{\mathcal{K}}^\star) + \sum_{k=1}^{K} [I(Y_k; U_k^\star) - I(X; U_k^\star)] \;, \tag{5.21}$$

*and* $\mathbf{P}^\star$ *is the set of pmfs that maximize the cost function*

$$\mathcal{L}_s^{\mathrm{DIB}}(\mathbf{P}) := -H(X|U_{\mathcal{K}}) - s \sum_{k=1}^{K} [H(X|U_k) + I(Y_k; U_k)] \;. \tag{5.22}$$

*Proof.* The proof of Proposition 10 is given in Appendix H.5. □

From Proposition 10 it is easy to see that the boundary tuple $(\Delta_s, R_s)$ for a given parameter $s$ can be computed by finding the encoding mappings $\{P_{U_k|Y_k}\}_{k=1}^K$ that maximizes the cost function $\mathcal{L}_s^{\mathrm{DIB}}(\mathbf{P})$ in (5.22). Different boundary tuples of region $\mathcal{RI}_{\mathrm{DIB}}^{\mathrm{sum}}$ can be obtained by finding the encoding mappings maximizing (5.22) for different $s$ values, and computing (5.20) and (5.21) for the resulting solution.

For variational distributions $Q_{U_k}$ on $\mathcal{U}_k$, $k \in \mathcal{K}$ (instead of unknown $P_{U_k}$), a variational stochastic decoder $Q_{X|U_1,\dots,U_K}$ (instead of the unknown optimal decoder $P_{X|U_1,\dots,U_K}$), and $K$ arbitrary decoders $Q_{X|U_k}$, $k \in \mathcal{K}$, let define $\mathbf{Q}$ as follows

$$\mathbf{Q} := \left\{ Q_{X|U_1,\dots,U_K}, Q_{X|U_1}, \dots, Q_{X|U_K}, Q_{X|U_1}, \dots, Q_{X|U_K} \right\} .$$

In the following we define the variational DIB cost function $\mathcal{L}_s^{\mathrm{VDIB}}(\mathbf{P}, \mathbf{Q})$ as

$$\begin{aligned} \mathcal{L}_s^{\mathrm{VDIB}}(\mathbf{P}, \mathbf{Q}) := \mathbb{E}_{P_{X,Y_{\mathcal{K}}}} \Bigg[ & \mathbb{E}_{P_{U_1|Y_1}} \times \cdots \times \mathbb{E}_{P_{U_K|Y_K}} [\log Q_{X|U_{\mathcal{K}}}] \\ & + s \sum_{k=1}^{K} \Big( \mathbb{E}_{P_{U_k|Y_k}} [\log Q_{X|U_k}] - D_{\mathrm{KL}}(P_{U_k|Y_k} \| Q_{U_k}) \Big) \Bigg] . \end{aligned} \tag{5.23}$$

The following lemma states that $\mathcal{L}_s^{\mathrm{VDIB}}(\mathbf{P}, \mathbf{Q})$ is a variational lower bound on the DIB objective $\mathcal{L}_s^{\mathrm{DIB}}(\mathbf{P})$ for all distributions $\mathbf{Q}$.

**Lemma 6.** *For fixed* $\mathbf{P}$*, we have*

$$\mathcal{L}_s^{\mathrm{VDIB}}(\mathbf{P}, \mathbf{Q}) \leq \mathcal{L}_s^{\mathrm{DIB}}(\mathbf{P}) \;, \quad \textit{for all } \mathbf{Q} \;.$$

*In addition, there exists a* $\mathbf{Q}$ *that achieves the maximum* $\max_{\mathbf{Q}} \mathcal{L}_s^{\mathrm{VDIB}}(\mathbf{P}, \mathbf{Q}) = \mathcal{L}_s^{\mathrm{DIB}}(\mathbf{P})$, *and is given by*

$$\begin{aligned} Q_{U_k}^\star = P_{U_k} \;, \qquad Q_{X|U_k}^\star = P_{X|U_k} \;, \qquad k = 1, \dots, K \;, \\ Q_{X|U_1,\dots,U_K}^\star = P_{X|U_1,\dots,U_K} \;, \end{aligned} \tag{5.24}$$

*where* $P_{U_k}$, $P_{X|U_k}$ *and* $P_{X|U_1,\dots,U_K}$ *are computed from* $\mathbf{P}$.

*Proof.* The proof of Lemma 6 is given in Appendix H.6. □

Using Lemma 6, it is easy to see that

$$\max_{\mathbf{P}} \mathcal{L}_s^{\mathrm{DIB}}(\mathbf{P}) = \max_{\mathbf{P}} \max_{\mathbf{Q}} \mathcal{L}_s^{\mathrm{VDIB}}(\mathbf{P}, \mathbf{Q}) \,. \tag{5.25}$$

**Remark 14.** *The variational DIB cost $\mathcal{L}_s^{\mathrm{VDIB}}(\mathbf{P}, \mathbf{Q})$ in (5.23) is composed of the cross-entropy term that is average logarithmic loss of estimating $X$ from all latent representations $U_1, \ldots, U_K$ by using the joint decoder $Q_{X|U_1,\ldots,U_K}$, and a regularization term. The regularization term is consisted of: i) the KL divergence between encoding mapping $P_{U_k|Y_k}$ and the prior $Q_{U_k}$, that also seems in the single encoder case of the variational bound (see (2.33)); and ii) the average logarithmic loss of estimating $X$ from each latent space $U_k$ using the decoder $Q_{X|U_k}$, that does not appear in the single encoder case.* ■

### 5.2.1 Variational Distributed IB Algorithm

In the first part of this chapter, we present the BA-type algorithms which find $\mathbf{P}$, $\mathbf{Q}$ optimizing (5.25) for the cases in which the joint distribution of the data, i.e., $P_{\mathbf{X},\mathbf{Y}_{\mathcal{K}}}$, is known perfectly or can be estimated with a high accuracy. However, this is not the case in general. Instead only a set of training samples $\{(\mathbf{x}_i, \mathbf{y}_{1,i}, \ldots, \mathbf{y}_{K,i})\}_{i=1}^n$ is available.
For this case, we develop a method in which the encoding and decoding mappings are restricted to a family of distributions, whose parameters are the outputs of DNNs. By doing so, the variational bound (5.23) can be written in terms of the parameters of DNNs. Furthermore, the bound can be computed using Monte Carlo sampling and the reparameterization trick [29]. Finally, we use the stochastic gradient descent (SGD) method to train the parameters of DNNs. The proposed method generalizes the variational framework in [30, 78, 117–119] to the distributed case with $K$ learners, and was given in [1].

Let $P_{\theta_k}(\mathbf{u}_k|\mathbf{y}_k)$ denote the encoding mapping from the observation $\mathbf{Y}_k$ to the latent representation $\mathbf{U}_k$, parameterized by a DNN $f_{\theta_k}$ with parameters $\theta_k$. As a common example, the encoder can be chosen as a multivariate Gaussian, i.e., $P_{\theta_k}(\mathbf{u}_k|\mathbf{y}_k) = \mathcal{N}(\mathbf{u}_k; \boldsymbol{\mu}_{\theta_k}, \boldsymbol{\Sigma}_{\theta_k})$. That is the DNN $f_{\theta_k}$ maps the observation $\mathbf{y}_k$ to the parameters of the multivariate Gaussian, namely the mean $\boldsymbol{\mu}_{\theta_k}$ and the covariance $\boldsymbol{\Sigma}_{\theta_k}$, i.e., $(\boldsymbol{\mu}_{\theta_k}, \boldsymbol{\Sigma}_{\theta_k}) = f_\theta(\mathbf{y}_k)$. Similarly, let $Q_{\phi_{\mathcal{K}}}(\mathbf{x}|\mathbf{u}_{\mathcal{K}})$ denote the decoding mapping from all latent representations $\mathbf{U}_1, \ldots, \mathbf{U}_K$ to the target variable $\mathbf{X}$, parameterized by a DNN $g_{\phi_{\mathcal{K}}}$ with parameters $\phi_{\mathcal{K}}$; and let $Q_{\phi_k}(\mathbf{x}|\mathbf{u}_k)$ denote the regularizing decoding mapping from the $k$-th latent representations $\mathbf{U}_k$ to the target variable $\mathbf{X}$, parameterized by a DNN $g_{\phi_k}$ with parameters $\phi_k$, $k = 1, \ldots, K$. Furthermore, let $Q_{\psi_k}(\mathbf{u}_k)$, $k = 1, \ldots, K$, denote the prior of the latent space, which does not depend on a DNN.

By restricting the coders' mappings to a family of distributions as mentioned above, the optimization of the variational DIB cost in (5.25) can be written as follows

$$\max_{\mathbf{P}} \max_{\mathbf{Q}} \mathcal{L}_s^{\mathrm{VDIB}}(\mathbf{P}, \mathbf{Q}) \geq \max_{\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\psi}} \mathcal{L}_s^{\mathrm{NN}}(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\psi}) \;, \tag{5.26}$$

where $\boldsymbol{\theta} := [\theta_1, \ldots, \theta_K]$, $\boldsymbol{\phi} := [\phi_1, \ldots, \phi_K, \phi_{\mathcal{K}}]$, $\boldsymbol{\psi} := [\psi_1, \ldots, \psi_K]$ denote the parameters of encoding DNNs, decoding DNNs, prior distributions, respectively; and the cost $\mathcal{L}_s^{\mathrm{NN}}(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\psi})$ is given as

$$\begin{aligned} \mathcal{L}_s^{\mathrm{NN}}(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\psi}) := \mathbb{E}_{P_{\mathbf{X},\mathbf{Y}_{\mathcal{K}}}} \Bigg[ & \mathbb{E}_{P_{\theta_1}(\mathbf{U}_1|\mathbf{Y}_1)} \times \cdots \times \mathbb{E}_{P_{\theta_K}(\mathbf{U}_K|\mathbf{Y}_K)} [\log Q_{\phi_{\mathcal{K}}}(\mathbf{X}|\mathbf{U}_{\mathcal{K}})] \\ & + s \sum_{k=1}^{K} \Big( \mathbb{E}_{P_{\theta_k}(\mathbf{U}_k|\mathbf{Y}_k)} [\log Q_{\phi_k}(\mathbf{X}|\mathbf{U}_k)] - D_{\mathrm{KL}}(P_{\theta_k}(\mathbf{U}_k|\mathbf{Y}_k) \| Q_{\psi_k}(\mathbf{U}_k)) \Big) \Bigg] \;. \end{aligned} \tag{5.27}$$

Furthermore, the cross-entropy terms in (5.27) can be computed using Monte Carlo sampling and the reparameterization trick [29]. In particular, $P_{\theta_k}(\mathbf{u}_k|\mathbf{y}_k)$ can be sampled by first sampling a random variable $\mathbf{Z}_k$ with distribution $P_{\mathbf{Z}_k}(\mathbf{z}_k)$, i.e., $P_{\mathbf{Z}_k} = \mathcal{N}(\mathbf{0}, \mathbf{I})$, then transforming the samples using some function $\tilde{f}_{\theta_k} : \mathcal{Y}_k \times \mathcal{Z}_k \to \mathcal{U}_k$ parameterized by $\theta_k$, i.e., $\mathbf{u}_k = \tilde{f}_{\theta_k}(\mathbf{y}_k, \mathbf{z}_k) \sim P_{\theta_k}(\mathbf{u}_k|\mathbf{y}_k)$. The reparameterization trick reduces the original optimization to estimating $\theta_k$ of the deterministic function $\tilde{f}_{\theta_k}$; hence, it allows us to compute estimates of the gradient using backpropagation [29]. Thus, we have the empirical DIB cost for the $i$-th sample in the training dataset as follows

$$\begin{aligned} \mathcal{L}_{s,i}^{\mathrm{emp}}(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\psi}) = \frac{1}{m} \sum_{j=1}^{m} & \left[ \log Q_{\phi_{\mathcal{K}}}(\mathbf{x}_i | \mathbf{u}_{1,i,j}, \ldots, \mathbf{u}_{K,i,j}) + s \sum_{k=1}^{K} \log Q_{\phi_k}(\mathbf{x}_i|\mathbf{u}_{k,i,j}) \right] \\ & - s \sum_{k=1}^{K} D_{\mathrm{KL}}(P_{\theta_k}(\mathbf{U}_k|\mathbf{y}_k) \| Q_{\psi_k}(\mathbf{U}_k)) \;. \end{aligned} \tag{5.28}$$

where $m$ is the number of samples for the Monte Carlo sampling.

Finally, we train DNNs to maximize the empirical DIB cost over the parameters $\boldsymbol{\theta}, \boldsymbol{\phi}$ as

$$\max_{\boldsymbol{\theta}, \boldsymbol{\phi}} \; \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_{s,i}^{\mathrm{emp}}(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\psi}) \;. \tag{5.29}$$

For the training step, we use the SGD or Adam optimization tool [83]. The training procedure is detailed in Algorithm 4, so-called variational distributed Information Bottleneck (D-VIB).

**Algorithm 4** D-VIB algorithm for the distributed IB problem [1, Algorithm 3]

1: **input:** Training dataset $\mathcal{D} := \{(\mathbf{x}_i, \mathbf{y}_{1,i}, \ldots, \mathbf{y}_{K,i})\}_{i=1}^{n}$, parameter $s \geq 0$.
2: **output:** $\boldsymbol{\theta}^\star, \boldsymbol{\phi}^\star$ and optimal pairs $(\Delta_s, R_s)$.
3: **initialization** Initialize $\boldsymbol{\theta}, \boldsymbol{\phi}$.
4: **repeat**
5: Randomly select $b$ mini-batch samples $\{(\mathbf{y}_{1,i}, \ldots, \mathbf{y}_{K,i})\}_{i=1}^{b}$ and the corresponding $\{\mathbf{x}_i\}_{i=1}^{b}$ from $\mathcal{D}$.
6: Draw $m$ random i.i.d samples $\{\mathbf{z}_{k,j}\}_{j=1}^{m}$ from $P_{\mathbf{Z}_k}$, $k = 1, \ldots, K$.
7: Compute $m$ samples $\mathbf{u}_{k,i,j} = \tilde{f}_{\theta_k}(\mathbf{y}_{k,i}, \mathbf{z}_{k,j})$
8: For the selected mini-batch, compute gradients of the empirical cost (5.29).
9: Update $\boldsymbol{\theta}, \boldsymbol{\phi}$ using the estimated gradient (e.g. with SGD or Adam).
10: **until** convergence of $\boldsymbol{\theta}, \boldsymbol{\phi}$.

Once our model is trained, with the convergence of the DNN parameters to $\boldsymbol{\theta}^\star, \boldsymbol{\phi}^\star$, for new observations $\mathbf{Y}_1, \ldots, \mathbf{Y}_K$, the target variable $\mathbf{X}$ can be inferred by sampling from the encoders $P_{\theta_k^\star}(\mathbf{U}_k|\mathbf{Y}_k)$ and then estimating from the decoder $Q_{\phi_{\mathcal{K}}^\star}(\mathbf{X}|\mathbf{U}_1, \ldots, \mathbf{U}_K)$.

Now we investigate the choice of parametric distributions $P_{\theta_k}(\mathbf{u}_k|\mathbf{y}_k)$, $Q_{\phi_k}(\mathbf{x}|\mathbf{u}_k)$, $Q_{\phi_{\mathcal{K}}}(\mathbf{x}|\mathbf{u}_{\mathcal{K}})$ and $Q_{\psi_k}(\mathbf{u}_k)$ for the two applications: i) classification, and ii) vector Gaussian model. Nonetheless, the parametric families of distributions should be chosen to be expressive enough to approximate the optimal encoders maximizing (5.22) and the optimal decoders and priors in (5.24) such that the gap between the variational DIB cost (5.23) and the original DIB cost (5.22) is minimized.

### D-VIB Algorithm for Classification

Let us consider a distributed classification problem in which the observations $\mathbf{Y}_1, \ldots, \mathbf{Y}_K$ have arbitrary distribution and $X$ has a discrete distribution on some finite set $\mathcal{X}$ of class labels. For this problem, the choice of the parametric distributions can be the following:

- The decoder $Q_{\phi_{\mathcal{K}}}(x|\mathbf{u}_{\mathcal{K}})$ and decoders used for regularization $Q_{\phi_k}(x|\mathbf{u}_k)$ can be general categorical distributions parameterized by a DNN with a softmax operation in the last layer, which outputs the probabilities of dimension $|\mathcal{X}|$.
- The encoders can be chosen as multivariate Gaussian, i.e. $P_{\theta_k}(\mathbf{u}_k|\mathbf{y}_k) = \mathcal{N}(\mathbf{u}_k; \boldsymbol{\mu}_{\theta_k}, \boldsymbol{\Sigma}_{\theta_k})$.

- The priors of the latent space $Q_{\psi_k}(\mathbf{u}_k)$ can be chosen as multivariate Gaussian (e.g., $\mathcal{N}(\mathbf{0},\mathbf{I})$) such that the KL divergence $D_{\mathrm{KL}}(P_{\theta_k}(\mathbf{U}_k|\mathbf{Y}_k)\|Q_{\psi_k}(\mathbf{U}_k))$ has a closed form solution and is easy to compute [29, 30]; or more expressive parameterizations can also be considered [120, 121].

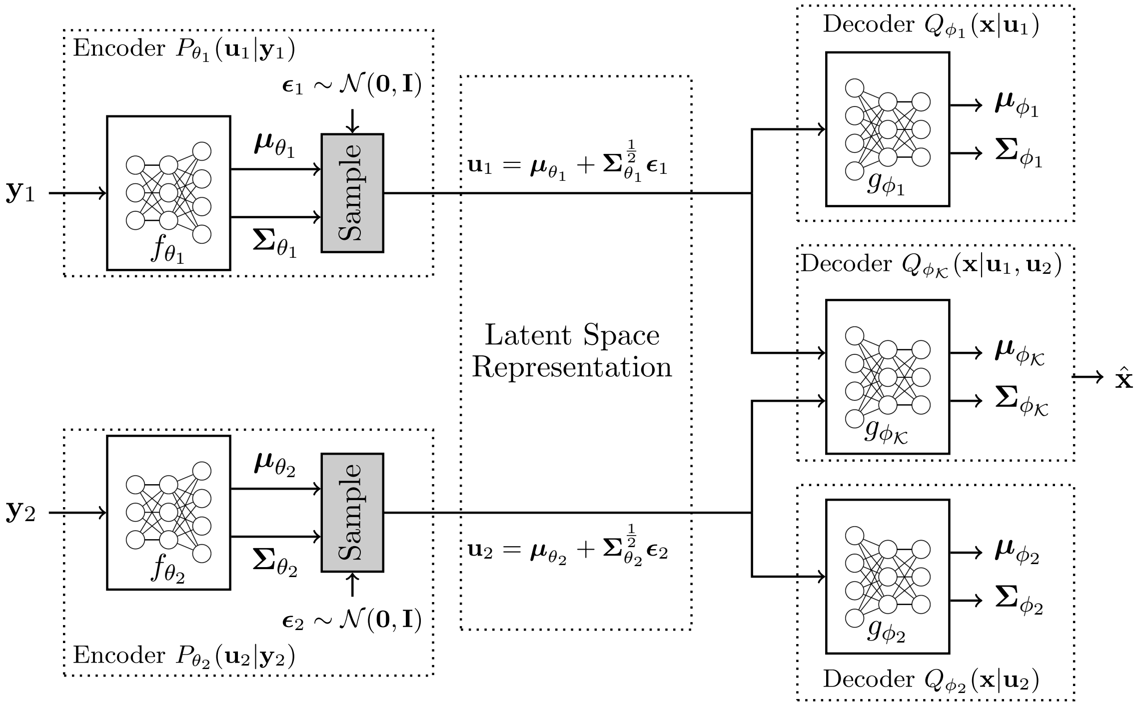



Figure 5.3: An example of distributed supervised learning.

#### D-VIB Algorithm for Vector Gaussian Model

One of the main results of this thesis is that the optimal test channels are Gaussian for the vector Gaussian model (see Theorem 4). Due to this, if the underlying data model is multivariate vector Gaussian, then the optimal distributions $\mathbf{P}$ and $\mathbf{Q}$ are also multivariate Gaussian. Hence, we consider the following parameterization, for $k \in \mathcal{K}$,

$$P_{\theta_k}(\mathbf{u}_k|\mathbf{y}_k) = \mathcal{N}(\mathbf{u}_k; \boldsymbol{\mu}_{\theta_k}, \boldsymbol{\Sigma}_{\theta_k}) \tag{5.30a}$$

$$Q_{\phi_\mathcal{K}}(\mathbf{x}|\mathbf{u}_\mathcal{K}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{\phi_\mathcal{K}}, \boldsymbol{\Sigma}_{\phi_\mathcal{K}}) \tag{5.30b}$$

$$Q_{\phi_k}(\mathbf{x}|\mathbf{u}_k) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{\phi_k}, \boldsymbol{\Sigma}_{\phi_k}) \tag{5.30c}$$

$$Q_{\psi_k}(\mathbf{u}_k) = \mathcal{N}(\mathbf{0},\mathbf{I})\ , \tag{5.30d}$$

where $\boldsymbol{\mu}_{\theta_k}, \boldsymbol{\Sigma}_{\theta_k}$ are the outputs of a DNN $f_{\theta_k}$ that encodes the input $\mathbf{Y}_k$ into a $n_{u_k}$-dimensional Gaussian distribution; $\boldsymbol{\mu}_{\phi_\mathcal{K}}, \boldsymbol{\Sigma}_{\phi_\mathcal{K}}$ are the outputs of a DNN $g_{\phi_\mathcal{K}}$ with inputs

$\mathbf{U}_1, \ldots, \mathbf{U}_K$, sampled from $\mathcal{N}(\mathbf{u}_k; \boldsymbol{\mu}_{\theta_k}, \boldsymbol{\Sigma}_{\theta_k})$; and $\boldsymbol{\mu}_{\phi_k}, \boldsymbol{\Sigma}_{\phi_k}$ are the outputs of a DNN $g_{\phi_k}$ with the input $\mathbf{U}_k$, $k = 1, \ldots, K$.

### 5.2.2 Experimental Results

In this section, numerical results on the synthetic and real datasets are provided to support the efficiency of the D-VIB Algorithm 4. We evaluate the relevance-complexity trade-offs achieved by the BA-type Algorithm 3 and D-VIB Algorithm 4. The resulting relevance-complexity pairs are compared to the optimal relevance-complexity trade-offs and an upper bound, which is denoted by *Centralized IB* (C-IB). The C-IB bound is given by the pairs $(\Delta_s, R_{\mathrm{sum}})$ achievable if $(Y_1, \ldots, Y_K)$ are encoded jointly at a single encoder with complexity $R_{\mathrm{sum}} = R_1 + \cdots + R_K$, and can be obtained by solving the centralized IB problem as follows

$$\Delta_{\mathrm{cIB}}(R_{\mathrm{sum}}) = \max_{P_{U|Y_1,\ldots,Y_K} : I(U;Y_1,\ldots,Y_K) \leq R_{\mathrm{sum}}} I(U;X) \,. \tag{5.31}$$

In the following experiments, the D-VIB Algorithm 4 is implemented by Adam optimizer [29] over 150 epochs and minibatch size of 64. The learning rate is initialized with 0.001 and decreased gradually every 30 epochs with a decay rate of 0.5, i.e., learning rate at epoch $n_{\mathrm{epoch}}$ is given by $0.001 \cdot 0.5^{\lfloor n_{\mathrm{epoch}}/30 \rfloor}$.

#### Regression for Vector Gaussian Data Model

Here we consider a real valued vector Gaussian data model as in [1, Section VI-A]. Specifically, $K = 2$ encoders observe independently corrupted Gaussian noisy versions of a $n_x$-dimensional vector Gaussian source $\mathbf{X} \sim \mathcal{N}(\mathbf{x}; \mathbf{0}, \mathbf{I})$, as $\mathbf{Y}_k = \mathbf{H}_k \mathbf{X} + \mathbf{N}_k$, where $\mathbf{H}_k \in \mathbb{R}^{n_k \times n_x}$ represents the channel connecting the source to the $k$-th encoder and $\mathbf{N}_k \in \mathbb{R}^{n_k}$ is the noise at this encoder, i.e., $\mathbf{N}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $k = 1, 2$.

The optimal complexity-relevance trade-off for this model is characterized as in (4.25) (wherein $\mathbf{H}_0 = \mathbf{0}$), and can be computed using two different approaches: i) using Algorithm 3 and ii) by directly evaluating the region obtained from Theorem 4 using the CVX optimization tool to find the maximizing covariance matrices $(\boldsymbol{\Omega}_1, \boldsymbol{\Omega}_2)$ (note that this problem is convex and so CVX finds the optimal solution). Furthermore, the C-IB upper bound in (5.31) can be computed analytically (see (2.14)) since it is an instance of Gaussian Information Bottleneck problem.

A synthetic dataset of $n$ i.i.d. samples $\{(\mathbf{x}_i, \mathbf{y}_{1,i}, \mathbf{y}_{2,i})\}_{i=1}^n$ is generated from the aforementioned vector Gaussian model. Then, the proposed BA-type and D-VIB algorithms are applied on the generated dataset for regression of the Gaussian target variable $\mathbf{X}$. For the case in which the covariance matrix $\mathbf{\Sigma}_{\mathbf{x},\mathbf{y}_1,\mathbf{y}_2}$ of the data model is known, Algorithm 3 is used to compute the relevance-complexity pairs for different values of $s$. For the case in which the covariance matrix $\mathbf{\Sigma}_{\mathbf{x},\mathbf{y}_1,\mathbf{y}_2}$ is not known, Algorithm 4 is used to train the DNNs determining the encoders and decoders for different value of $s$. The encoders and decoders are parameterized with multivariate Gaussian as in (5.30). We use the following network architecture: Encoder $k$, $k = 1, 2$, is modeled with DNNs with 3 hidden dense layers of 512 neurons with rectified linear unit (ReLU) activations; which is followed by a dense layer without nonlinear activation to generate the outputs of Encoder $k$, i.e., $\boldsymbol{\mu}_{\theta_k}$ and $\mathbf{\Sigma}_{\theta_k}$ of size 512 and $512 \times 512$. Each decoder is modeled with DNNs with 2 hidden dense layers of 512 neurons with ReLU activations. The output of decoder 1, 2 and $\mathcal{K}$ is processed, each, by a fully connected layer without nonlinear activation to generate $\boldsymbol{\mu}_{\phi_k}$ and $\mathbf{\Sigma}_{\phi_k}$, and $\boldsymbol{\mu}_{\phi_\mathcal{K}}$ and $\mathbf{\Sigma}_{\phi_\mathcal{K}}$, of size 2 and $2 \times 2$.

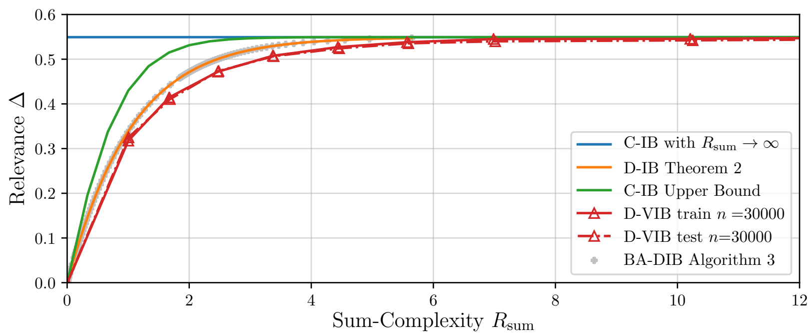



Figure 5.4: Relevance vs. sum-complexity trade-off for vector Gaussian data model with $K = 2$ encoders, $n_x = 1$, $n_1 = n_2 = 3$, and achievable pairs with the BA-type and D-VIB algorithms for $n = 40000$. Figure is taken from [1].

Figure 5.4 shows the tuples $(\Delta_s, R_{\text{sum}})$ resulting from the application of the BA-type Algorithm 3. It is worth-noting that Algorithm 3 converges to the optimal solution obtained directly by evaluation the region from (4.25). To apply the D-VIB algorithm, a synthetic dataset of 40000 i.i.d. samples is generated, which is split into a training set of 30000 samples and a test set of 10000 samples. Figure 5.4 also shows the relevance-complexity pairs resulting form the application of the D-VIB algorithm for different values of $s$ in the range (0, 10] calculated as in Proposition 10. For comparisons reasons, Figure 5.4 also shows the performance of *centralized* or *joint* encoding, i.e., the C-IB bounds $\Delta_{\mathrm{cIB}}(R_{\mathrm{sum}})$ and $\Delta_{\mathrm{cIB}}(\infty)$.

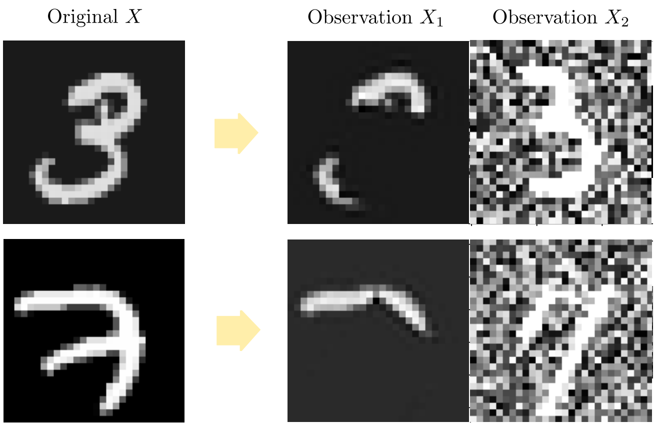



Figure 5.5: Two-view handwritten MNIST dataset. Figure is taken from [1].

| | DNN Layers |
|---|---|
| Encoder $k$ | conv. $[5, 5, 32]$ – ReLU |
| | maxpool $[2, 2, 2]$ |
| | conv. $[5, 5, 64]$ – ReLU |
| | maxpool $[2, 2, 2]$ |
| | dense $[1024]$ – ReLU |
| | dropout 0.4 |
| | dense $[256]$ – ReLU |
| Latent space $k$ | dense $[256]$ – ReLU |
| Decoder $k$ | dense $[256]$ – ReLU |
| Decoder $\mathcal{K}$ | dense $[256]$ – ReLU |

Table 5.1: DNN architecture for Figure 5.6.

**Classification on the multi-view MNIST dataset**

Here the performance of the D-VIB algorithm is evaluated for a classification task on a multi-view version of the MNIST dataset, consisting of gray-scale images of 70000 handwritten digits with a size of $28 \times 28$ pixels from 0 to 9. In the experiments, we use the dataset composed of two views, generated as in [1, Section VI-B]. To generate the view 1, each image in MNIST is rotated by a random angel uniformly selected from the range $[-\pi/4, \pi/4]$, then the pixels in the middle of the image with a size of $25 \times 25$ are occluded. The view 2 is generated from the same digit as in the view 1 by adding a uniformly distributed random noise in the range of $[0, 3]$ to each pixel, and then each pixel value is truncated to $[0, 1]$. An example of the two-view MNIST dataset is depicted in Figure 5.5. The view 1 and view 2 are made available to Encoder 1 and Encoder 2, respectively. Each image is flattened into a vector of length 784, i.e., $\mathbf{y}_k \in [0, 1]^{784}$, $k = 1, 2$. Finally, 70000 two-view samples $\{\mathbf{x}_i, \mathbf{y}_{1,i}, \mathbf{y}_{2,i}\}_{i=1}^{70000}$ are separated into training and test sets of length $n$ and $70000 - n$, respectively. To understand how difficult the classification task is on each view, the centralized VIB (C-VIB) algorithm [30] is applied by using a standard convolutional neural network (CNN) architecture with dropout, which achieves an accuracy of 99.8% for the original MNIST dataset. The resulting accuracies are 92.3% for view 1 and 79.68% for view 2. Therefore, the classification on view 1 is easier than view 2. In other words, view 1 is less noisy.

Now we apply the D-VIB algorithm to the two-view MNIST dataset generated as explained above. The CNN architecture is summarized in Table 5.1. For Encoder $k$, $k=1,2$, we consider a $n_{u_k}=256$ dimensional multivariate Gaussian distribution parameterization, $\mathcal{N}(\boldsymbol{\mu}_{\theta_k},\boldsymbol{\Sigma}_{\theta_k})$, where $\boldsymbol{\mu}_{\theta_k},\boldsymbol{\Sigma}_{\theta_k}$ are the outputs of a DNN $f_{\theta_k}$ consisting of the concatenation of convolutional, dense and maxpool layers with ReLU activations and a dropout. For the last layer of the encoder we use a linear activation. Then, the latent representation $\mathbf{u}_k$, $k=1,2$, is sampled from $\mathcal{N}(\boldsymbol{\mu}_{\theta_k},\boldsymbol{\Sigma}_{\theta_k})$. The prior is chosen as $Q_{\psi_k}(\mathbf{u}_k)=\mathcal{N}(\mathbf{0},\mathbf{I})$. Decoder $k$, $k=1,2$, and Decoder $\mathcal{K}$ takes $\mathbf{u}_k$ and $\mathbf{u}_{\mathcal{K}}$, respectively, as an input. Each decoder is modeled with a DNN ($g_{\phi_k}$ and $g_{\phi_{\mathcal{K}}}$) with 2 hidden dense layers of 256 neurons with ReLU activations. The output of each decoder is processed by a fully connected layer, followed by a softmax, which outputs a normalized vector $\hat{\mathbf{x}}$ of size $|\mathcal{X}|=10$, corresponding to a distribution over the one-hot encoding of the digit labels $\{0,1,\ldots,9\}$ from the $K$ observations, i.e., we have

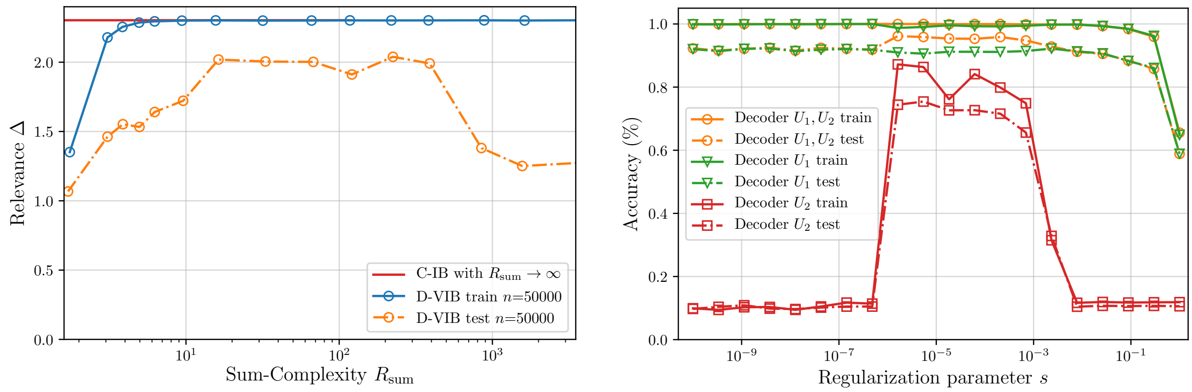
$$\begin{aligned} Q_{\phi_k}(\hat{\mathbf{x}}|\mathbf{u}_k) &= \mathrm{softmax}(g_{\phi_k}(U_k)) \ , \qquad k=1,\ldots,K \ , \\ Q_{\phi_{\mathcal{K}}}(\hat{\mathbf{x}}|\mathbf{u}_{\mathcal{K}}) &= \mathrm{softmax}(g_{\phi_{\mathcal{K}}}(U_1,U_2)) \ , \end{aligned} \tag{5.32}$$

where $\mathrm{softmax}(\mathbf{p})$ for $\mathbf{p}\in\mathbb{R}^d$ is a vector with $i$-th entry is calculated as $[\mathrm{softmax}(\mathbf{p})]_i = \exp(p_i)/\sum_{j=1}^{d}\exp(p_j)$, $i=1,\ldots,d$.

For given parameterization, the log-loss (reconstruction loss) terms are calculated by using the cross-entropy criterion and the KL divergence terms can be computed as in (I.2).

The relevance-complexity pairs obtained from applying the D-VIB Algorithm 4 on the two-view MNIST – consisting of a training set of $n=50000$ samples – is depicted in Figure 5.6a for 15 different values of $s$ in the range $[10^{-10},1]$. For comparisons reasons, the figure also shows the C-IB upper bound for $R_{\mathrm{sum}}\to\infty$ assuming that zero classification error is possible, i.e., $\Delta_{\mathrm{cIB}}(\infty)=\log 10$. During the training phase, it is observed that higher sum-complexity results higher relevance, and that resulting relevance-complexity pairs are very close to the theoretical limit. On the other hand, during the test phase, the achievable relevance decreases for large values of sum-complexity. This is because of the effect of the regularization such that the complexity constraint results in higher generalization.

The accuracies of the D-VIB algorithm achieved by the joint (or main) estimator $Q_{\mathbf{X}|\mathbf{U}_1,\mathbf{U}_2}$, as well as the regularizing decoders $Q_{\mathbf{X}|\mathbf{U}_k}$, $k=1,2$, are depicted in Figure 5.6b




(a) Relevance vs. sum-complexity trade-off.

(b) Accuracy vs. regularization parameter $s$.

Figure 5.6: Distributed representation learning for the two-view MNIST dataset with $K = 2$ encoders, with D-VIB algorithm for $n = 50000$ and $s \in [10^{-10}, 1]$. Figures are taken from [1]

with respect to the regularization parameter $s$. As mentioned previously in this section, view 1 is less noisy. Therefore, the description $\mathbf{U}_1$ from view 1 carries most of the information about the target variable $\mathbf{X}$. While for the range $10^{-6} < s < 10^{-3}$, both descriptions $\mathbf{U}_1$ and $\mathbf{U}_2$ capture the relevant information from the view 1 and view 2, respectively, and that results an increase in the overall performance for $Q_{\mathbf{X}|\mathbf{U}_1,\mathbf{U}_2}$.

| D-VIB | D-VIB-noReg | C-VIB |
|---|---|---|
| **97.24** | 96.72 | 96.68 |

Table 5.2: Accuracy for different algorithms with CNN architectures

In order to understand the advantages of the D-VIB algorithm, now we look at the comparison of accuracy of D-VIB with two different algorithms: i) the C-VIB, where both views are encoded in a centralized manner; and ii) the D-VIB-noReg, where the DIB cost (5.23) is optimized by considering only the divergence terms in the regularizer, without the regularizing decoders $Q_{X|U_k}$, $k = 1, 2$. The D-VIB-noReg can be seen as a naive direct extension of the VIB of [30] to the distributed case. Table 5.2 states the results, where it is seen that the D-VIB has the best accuracy compared to the other algorithms. This justifies that it is better to first partition the data according to its homogeneity, even if the data is available in a centralized manner. The advantage of D-VIB over C-VIB can be explained due to that it is better to learn suitable representations from each group, and optimize the encoding and decoding mappings jointly.

# Chapter 6

# Application to Unsupervised Clustering

Clustering consists of partitioning a given dataset into various groups (clusters) based on some similarity metric, such as the Euclidean distance, $L_1$ norm, $L_2$ norm, $L_\infty$ norm, the popular logarithmic loss measure, or others. The principle is that each cluster should contain elements of the data that are closer to each other than to any other element outside that cluster, in the sense of the defined similarity measure. If the joint distribution of the clusters and data is not known, one should operate blindly in doing so, i.e., using only the data elements at hand; and the approach is called unsupervised clustering [122, 123]. Unsupervised clustering is perhaps one of the most important tasks of unsupervised machine learning algorithms currently, due to a variety of application needs and connections with other problems.

Clustering can be formulated as follows. Consider a dataset that is composed of $N$ samples $\{\mathbf{x}_i\}_{i=1}^N$, which we wish to partition into $|\mathcal{C}| \geq 1$ clusters. Let $\mathcal{C} = \{1, \ldots, |\mathcal{C}|\}$ be the set of all possible clusters and $C$ designate a categorical random variable that lies in $\mathcal{C}$ and stands for the index of the actual cluster. If $\mathbf{X}$ is a random variable that models elements of the dataset, given that $\mathbf{X} = \mathbf{x}_i$ induces a probability distribution on $C$, which the learner should learn, thus mathematically, the problem is that of estimating the values of the unknown conditional probability $P_{C|\mathbf{X}}(\cdot|\mathbf{x}_i)$ for all elements $\mathbf{x}_i$ of the dataset. The estimates are sometimes referred to as the assignment probabilities.

Examples of unsupervised clustering algorithms include the very popular $K$-means [124] and Expectation Maximization (EM) [125]. The $K$-means algorithm partitions the data in a manner that the Euclidean distance among the members of each cluster is minimized. With the EM algorithm, the underlying assumption is that the data comprise a mixture of Gaussian samples, namely a Gaussian Mixture Model (GMM); and one estimates the parameters of each component of the GMM while simultaneously associating each data sample with one of those components. Although they offer some advantages in the context of clustering, these algorithms suffer from some strong limitations. For example, it is well known that the $K$-means is highly sensitive to both the order of the data and scaling; and the obtained accuracy depends strongly on the initial seeds (in addition to that, it does not predict the number of clusters or $K$-value). The EM algorithm suffers mainly from slow convergence, especially for high-dimensional data.

Recently, a new approach has emerged that seeks to perform inference on a transformed domain (generally referred to as latent space), not the data itself. The rationale is that because the latent space often has fewer dimensions, it is more convenient computationally to perform inference (clustering) on it rather than on the high-dimensional data directly. A key aspect then is how to design a latent space that is amenable to accurate low-complexity unsupervised clustering, i.e., one that preserves only those features of the observed high-dimensional data that are useful for clustering while removing all redundant or non-relevant information. Along this line of work, we can mention [126], which utilized Principal Component Analysis (PCA) [127, 128] for dimensionality reduction followed by $K$-means for clustering the obtained reduced dimension data; or [129], which used a combination of PCA and the EM algorithm. Other works that used alternatives for the linear PCA include kernel PCA [130], which employs PCA in a non-linear fashion to maximize variance in the data.

Tishby's Information Bottleneck (IB) method [17] formulates the problem of finding a good representation $\mathbf{U}$ that strikes the right balance between capturing all information about the categorical variable $C$ that is contained in the observation $\mathbf{X}$ and using the most concise representation for it. The IB problem can be written as the following Lagrangian optimization

$$\min_{P_{\mathbf{U}|\mathbf{X}}} \; I(\mathbf{X};\mathbf{U}) - sI(C;\mathbf{U}) \;, \tag{6.1}$$

where $s$ is a Lagrange-type parameter, which controls the trade-off between accuracy and regularization. In [32, 131], a text clustering algorithm is introduced for the case in which the joint probability distribution of the input data is known. This text clustering algorithm uses the IB method with an annealing procedure, where the parameter $s$ is increased gradually. When $s \to 0$, the representation $\mathbf{U}$ is designed with the most compact form, i.e., $|\mathcal{U}| = 1$, which corresponds to the maximum compression. By gradually increasing the parameter $s$, the emphasis on the relevance term $I(C; \mathbf{U})$ increases, and at a critical value of $s$, the optimization focuses on not only the compression, but also the relevance term. To fulfill the demand on the relevance term, this results in the cardinality of $\mathbf{U}$ bifurcating. This is referred as a phase transition of the system. The further increases in the value of $s$ will cause other phase transitions, hence additional splits of $|\mathcal{U}|$ until it reaches the desired level, e.g., $|\mathcal{U}| = |\mathcal{C}|$.

However, in the real-world applications of clustering with large-scale datasets, the joint probability distributions of the datasets are unknown. In practice, the usage of Deep Neural Networks (DNN) for unsupervised clustering of high-dimensional data on a lower dimensional latent space has attracted considerable attention, especially with the advent of Autoencoder (AE) learning and the development of powerful tools to train them using standard backpropagation techniques [29, 132]. Advanced forms include Variational Autoencoders (VAE) [29, 132], which are generative variants of AE that regularize the structure of the latent space, and the more general Variational Information Bottleneck (VIB) of [30], which is a technique that is based on the Information Bottleneck method and seeks a better trade-off between accuracy and regularization than VAE via the introduction of a Lagrange-type parameter $s$, which controls that trade-off and whose optimization is similar to deterministic annealing [32] or stochastic relaxation.

In this chapter, we develop an unsupervised generative clustering framework that combines VIB and the Gaussian Mixture Model. Specifically, in our approach, we use the Variational Information Bottleneck method and model the latent space as a mixture of Gaussians. We derive a bound on the cost function of our model that generalizes the Evidence Lower Bound (ELBO) and provide a variational inference type algorithm that allows computing it. In the algorithm, the coders' mappings are parameterized using Neural Networks (NN), and the bound is approximated by Markov sampling and optimized with stochastic gradient descent. Furthermore, we show how tuning the hyperparameter $s$ appropriately by gradually increasing its value with iterations (number of epochs) results in a better accuracy. Furthermore, the application of our algorithm to the unsupervised clustering of various datasets, including the MNIST [46], REUTERS [47], and STL-10 [48], allows a better clustering accuracy than previous state-of-the-art algorithms. For instance, we show that our algorithm performs better than the Variational Deep Embedding (VaDE) algorithm of [31], which is based on VAE and performs clustering by maximizing the ELBO. Our algorithm can be seen as a generalization of the VaDE, whose ELBO can be recovered by setting $s = 1$ in our cost function. In addition, our algorithm also generalizes the VIB of [30], which models the latent space as an isotropic Gaussian, which is generally not expressive enough for the purpose of unsupervised clustering. Other related works, which are of lesser relevance to the contribution of this paper, are the Deep Embedded Clustering (DEC) of [33] and the Improved Deep Embedded Clustering (IDEC) of [133] and [134]. For a detailed survey of clustering with deep learning, the readers may refer to [135].

To the best of our knowledge, our algorithm performs the best in terms of clustering accuracy by using deep neural networks without any prior knowledge regarding the labels (except the usual assumption that the number of classes is known) compared to the state-of-the-art algorithms of the unsupervised learning category. In order to achieve the outperforming accuracy: (i) we derive a cost function that contains the IB hyperparameter $s$ that controls optimal trade-offs between the accuracy and regularization of the model; (ii) we use a lower bound approximation for the KL term in the cost function, that does not depend on the clustering assignment probability (note that the clustering assignment is usually not accurate in the beginning of the training process); and (iii) we tune the hyperparameter $s$ by following an annealing approach that improves both the convergence and the accuracy of the proposed algorithm.

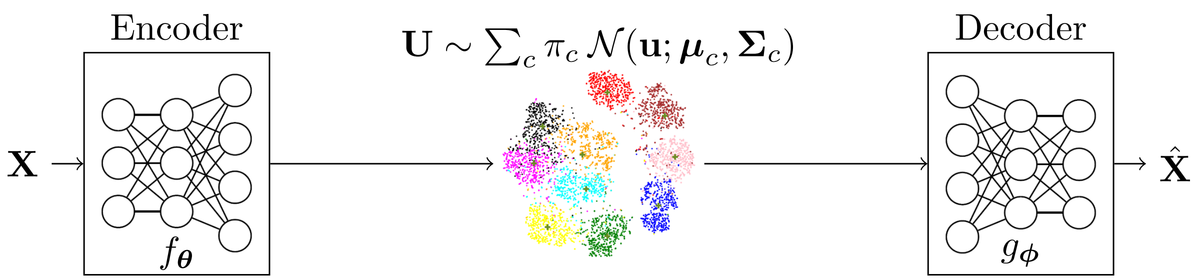



Figure 6.1: Variational Information Bottleneck with Gaussian Mixtures.

## 6.1 Proposed Model

In this section, we explain the proposed model, the so-called Variational Information Bottleneck with Gaussian Mixture Model (VIB-GMM), in which we use the VIB framework and model the latent space as a GMM. The proposed model is depicted in Figure 6.1, where the parameters $\pi_c$, $\boldsymbol{\mu}_c$, $\boldsymbol{\Sigma}_c$, for all values of $c \in \mathcal{C}$, are to be optimized jointly with those of the employed DNNs as instantiation of the coders. Furthermore, the assignment probabilities are estimated based on the values of latent space vectors instead of the observations themselves, i.e., $P_{C|\mathbf{X}} = Q_{C|\mathbf{U}}$. In the rest of this section, we elaborate on the inference and generative network models for our method.

### 6.1.1 Inference Network Model

We assume that observed data $\mathbf{x}$ are generated from a GMM with $|\mathcal{C}|$ components. Then, the latent representation $\mathbf{u}$ is inferred according to the following procedure:

1. One of the components of the GMM is chosen according to a categorical variable $C$.

2. The data $\mathbf{x}$ are generated from the $c$-th component of the GMM, i.e., $P_{\mathbf{X}|C} \sim \mathcal{N}(\mathbf{x}; \tilde{\boldsymbol{\mu}}_c, \tilde{\boldsymbol{\Sigma}}_c)$.

3. Encoder maps $\mathbf{x}$ to a latent representation $\mathbf{u}$ according to $P_{\mathbf{U}|\mathbf{X}} \sim \mathcal{N}(\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta)$.

    3.1. The encoder is modeled with a DNN $f_\theta$, which maps $\mathbf{x}$ to the parameters of a Gaussian distribution, i.e., $[\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta] = f_\theta(\mathbf{x})$.

    3.2. The representation $\mathbf{u}$ is sampled from $\mathcal{N}(\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta)$.

For the inference network, shown in Figure 6.2, the following Markov chain holds

$$C -\!\!\!\!\circ\!\!\!- \mathbf{X} -\!\!\!\!\circ\!\!\!- \mathbf{U}\,. \tag{6.2}$$

$$C \xrightarrow{P_{\mathbf{X}|C}} \mathbf{X} \xrightarrow{P_{\mathbf{U}|\mathbf{X}}} \mathbf{U}$$

Figure 6.2: Inference Network

### 6.1.2 Generative Network Model

Since the encoder extracts useful representations of the dataset and we assume that the dataset is generated from a GMM, we model our latent space also with a mixture of Gaussians. To do so, the categorical variable $C$ is embedded with the latent variable $\mathbf{U}$. The reconstruction of the dataset is generated according to the following procedure:

1. One of the components of the GMM is chosen according to a categorical variable $C$, with a prior distribution $Q_C$.

2. The representation $\mathbf{u}$ is generated from the $c$-th component, i.e., $Q_{\mathbf{U}|C} \sim \mathcal{N}(\mathbf{u}; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$.

3. The decoder maps the latent representation $\mathbf{u}$ to $\hat{\mathbf{x}}$, which is the reconstruction of the source $\mathbf{x}$ by using the mapping $Q_{\mathbf{X}|\mathbf{U}}$.

   3.1. The decoder is modeled with a DNN $g_\phi$ that maps $\mathbf{u}$ to the estimate $\hat{\mathbf{x}}$, i.e., $[\hat{\mathbf{x}}] = g_\phi(\mathbf{u})$.

For the generative network, shown in Figure 6.3, the following Markov chain holds

$$C -\!\!\circ\!\!- \mathbf{U} -\!\!\circ\!\!- \mathbf{X} \,. \tag{6.3}$$

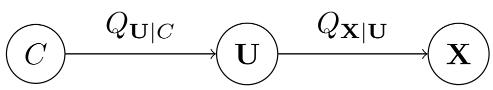



Figure 6.3: Generative Network

## 6.2 Proposed Method

In this section, we present our clustering method. First, we provide a general cost function for the problem of the unsupervised clustering that we study here based on the variational IB framework; and we show that it generalizes the ELBO bound developed in [31]. We then parameterize our model using DNNs whose parameters are optimized jointly with those of the GMM. Furthermore, we discuss the influence of the hyperparameter $s$ that controls optimal trade-offs between accuracy and regularization.

### 6.2.1 Brief Review of Variational Information Bottleneck for Unsupervised Learning

As described in Chapter 6.1, the stochastic encoder $P_{\mathbf{U}|\mathbf{X}}$ maps the observed data $\mathbf{x}$ to a representation $\mathbf{u}$. Similarly, the stochastic decoder $Q_{\mathbf{X}|\mathbf{U}}$ assigns an estimate $\hat{\mathbf{x}}$ of $\mathbf{x}$ based on the vector $\mathbf{u}$. As per the IB method [17], a suitable representation $\mathbf{U}$ should strike the right balance between capturing all information about the categorical variable $C$ that is contained in the observation $\mathbf{X}$ and using the most concise representation for it. This leads to maximizing the following Lagrange problem

$$\mathcal{L}_s(\mathbf{P}) = I(C;\mathbf{U}) - sI(\mathbf{X};\mathbf{U}) \ , \tag{6.4}$$

where $s \geq 0$ designates the Lagrange multiplier and, for convenience, $\mathbf{P}$ denotes the conditional distribution $P_{\mathbf{U}|\mathbf{X}}$.

Instead of (6.4), which is not always computable in our unsupervised clustering setting, we find it convenient to maximize an upper bound of $\mathcal{L}_s(\mathbf{P})$ given by

$$\tilde{\mathcal{L}}_s(\mathbf{P}) := I(\mathbf{X};\mathbf{U}) - sI(\mathbf{X};\mathbf{U}) \overset{(a)}{=} H(\mathbf{X}) - H(\mathbf{X}|\mathbf{U}) - s[H(\mathbf{U}) - H(\mathbf{U}|\mathbf{X})] \ , \tag{6.5}$$

where $(a)$ is due to the definition of mutual information (using the Markov chain $C \multimap \mathbf{X} \multimap \mathbf{U}$, it is easy to see that $\tilde{\mathcal{L}}_s(\mathbf{P}) \geq \mathcal{L}_s(\mathbf{P})$ for all values of $\mathbf{P}$). Noting that $H(\mathbf{X})$ is constant with respect to $P_{\mathbf{U}|\mathbf{X}}$, maximizing $\tilde{\mathcal{L}}_s(\mathbf{P})$ over $\mathbf{P}$ is equivalent to maximizing

$$\mathcal{L}'_s(\mathbf{P}) : = -H(\mathbf{X}|\mathbf{U}) - s[H(\mathbf{U}) - H(\mathbf{U}|\mathbf{X})] \tag{6.6}$$

$$= \mathbb{E}_{P_{\mathbf{X}}}\left[\mathbb{E}_{P_{\mathbf{U}|\mathbf{X}}}[\log P_{\mathbf{X}|\mathbf{U}} + s\log P_{\mathbf{U}} - s\log P_{\mathbf{U}|\mathbf{X}}]\right] \ . \tag{6.7}$$

For a variational distribution $Q_{\mathbf{U}}$ on $\mathcal{U}$ (instead of the unknown $P_{\mathbf{U}}$) and a variational stochastic decoder $Q_{\mathbf{X}|\mathbf{U}}$ (instead of the unknown optimal decoder $P_{\mathbf{X}|\mathbf{U}}$), let $\mathbf{Q} := \{Q_{\mathbf{X}|\mathbf{U}}, Q_{\mathbf{U}}\}$. Furthermore, let

$$\mathcal{L}_s^{\mathrm{VB}}(\mathbf{P},\mathbf{Q}) := \mathbb{E}_{P_{\mathbf{X}}}\left[\mathbb{E}_{P_{\mathbf{U}|\mathbf{X}}}[\log Q_{\mathbf{X}|\mathbf{U}}] - sD_{\mathrm{KL}}(P_{\mathbf{U}|\mathbf{X}}\|Q_{\mathbf{U}})\right] \ . \tag{6.8}$$

**Lemma 7.** *For given $\mathbf{P}$, we have*

$$\mathcal{L}_s^{\mathrm{VB}}(\mathbf{P},\mathbf{Q}) \leq \mathcal{L}'_s(\mathbf{P}), \quad \textit{for all } \mathbf{Q} \ .$$

*In addition, there exists a unique $\mathbf{Q}$ that achieves the maximum $\max_{\mathbf{Q}} \mathcal{L}_s^{\mathrm{VB}}(\mathbf{P},\mathbf{Q}) = \mathcal{L}'_s(\mathbf{P})$, and is given by*

$$Q^*_{\mathbf{X}|\mathbf{U}} = P_{\mathbf{X}|\mathbf{U}} \ , \quad Q^*_{\mathbf{U}} = P_{\mathbf{U}} \ .$$

*Proof.* The proof of Lemma 7 is given in Appendix I.1. □

Using Lemma 7, maximization of (6.6) can be written in term of the variational IB cost as follows

$$\max_{\mathbf{P}} \mathcal{L}'_s(\mathbf{P}) = \max_{\mathbf{P}} \max_{\mathbf{Q}} \mathcal{L}_s^{\mathrm{VB}}(\mathbf{P}, \mathbf{Q}) \,. \tag{6.9}$$

**Remark 15.** *As we already mentioned in the beginning of this chapter, the related work [31] performs unsupervised clustering by combining VAE with GMM. Specifically, it maximizes the following ELBO bound*

$$\mathcal{L}_1^{\mathrm{VaDE}} := \mathbb{E}_{P_{\mathbf{X}}}\left[\mathbb{E}_{P_{\mathbf{U}|\mathbf{X}}}[\log Q_{\mathbf{X}|\mathbf{U}}] - D_{\mathrm{KL}}(P_{C|\mathbf{X}} \| Q_C) - \mathbb{E}_{P_{C|\mathbf{X}}}[D_{\mathrm{KL}}(P_{\mathbf{U}|\mathbf{X}} \| Q_{\mathbf{U}|C})]\right] . \tag{6.10}$$

*Let, for an arbitrary non-negative parameter $s$, $\mathcal{L}_s^{\mathrm{VaDE}}$ be a generalization of the ELBO bound (6.10) of [31] given by*

$$\mathcal{L}_s^{\mathrm{VaDE}} := \mathbb{E}_{P_{\mathbf{X}}}\left[\mathbb{E}_{P_{\mathbf{U}|\mathbf{X}}}[\log Q_{\mathbf{X}|\mathbf{U}}] - sD_{\mathrm{KL}}(P_{C|\mathbf{X}} \| Q_C) - s\mathbb{E}_{P_{C|\mathbf{X}}}[D_{\mathrm{KL}}(P_{\mathbf{U}|\mathbf{X}} \| Q_{\mathbf{U}|C})]\right] . \tag{6.11}$$

*Investigating the RHS of (6.11), we get*

$$\mathcal{L}_s^{\mathrm{VB}}(\mathbf{P}, \mathbf{Q}) = \mathcal{L}_s^{\mathrm{VaDE}} + s\mathbb{E}_{P_{\mathbf{X}}}\left[\mathbb{E}_{P_{\mathbf{U}|\mathbf{X}}}[D_{\mathrm{KL}}(P_{C|\mathbf{X}} \| Q_{C|\mathbf{U}})]\right] , \tag{6.12}$$

*where the equality holds since*

$$\begin{aligned}
\mathcal{L}_s^{\mathrm{VaDE}} &= \mathbb{E}_{P_{\mathbf{X}}}\left[\mathbb{E}_{P_{\mathbf{U}|\mathbf{X}}}[\log Q_{\mathbf{X}|\mathbf{U}}] - sD_{\mathrm{KL}}(P_{C|\mathbf{X}} \| Q_C) - s\mathbb{E}_{P_{C|\mathbf{X}}}[D_{\mathrm{KL}}(P_{\mathbf{U}|\mathbf{X}} \| Q_{\mathbf{U}|C})]\right] && (6.13)\\
&\overset{(a)}{=} \mathbb{E}_{P_{\mathbf{X}}}\left[\mathbb{E}_{P_{\mathbf{U}|\mathbf{X}}}[\log Q_{\mathbf{X}|\mathbf{U}}] - sD_{\mathrm{KL}}(P_{\mathbf{U}|\mathbf{X}} \| Q_{\mathbf{U}}) - s\mathbb{E}_{P_{\mathbf{U}|\mathbf{X}}}\left[D_{\mathrm{KL}}(P_{C|\mathbf{X}} \| Q_{C|\mathbf{U}})\right]\right] && (6.14)\\
&\overset{(b)}{=} \mathcal{L}_s^{\mathrm{VB}}(\mathbf{P}, \mathbf{Q}) - s\mathbb{E}_{P_{\mathbf{X}}}\left[\mathbb{E}_{P_{\mathbf{U}|\mathbf{X}}}\left[D_{\mathrm{KL}}(P_{C|\mathbf{X}} \| Q_{C|\mathbf{U}})\right]\right] , && (6.15)
\end{aligned}$$

*where $(a)$ can be obtained by expanding and rearranging terms under the Markov chain $C -\!\!\circ\!\!- \mathbf{X} -\!\!\circ\!\!- \mathbf{U}$ (for a detailed treatment, please look at Appendix I.2); and $(b)$ follows from the definition of $\mathcal{L}_s^{\mathrm{VB}}(\mathbf{P}, \mathbf{Q})$ in (6.8).*

*Thus, by the non-negativity of relative entropy, it is clear that $\mathcal{L}_s^{\mathrm{VaDE}}$ is a lower bound on $\mathcal{L}_s^{\mathrm{VB}}(\mathbf{P}, \mathbf{Q})$. Furthermore, if the variational distribution $\mathbf{Q}$ is such that the conditional marginal $Q_{C|\mathbf{U}}$ is equal to $P_{C|\mathbf{X}}$, the bound is tight since the relative entropy term is zero in this case.* ■

### 6.2.2 Proposed Algorithm: VIB-GMM

In order to compute (6.9), we parameterize the distributions $P_{\mathbf{U}|\mathbf{X}}$ and $Q_{\mathbf{X}|\mathbf{U}}$ using DNNs. For instance, let the stochastic encoder $P_{\mathbf{U}|\mathbf{X}}$ be a DNN $f_\theta$ and the stochastic decoder $Q_{\mathbf{X}|\mathbf{U}}$ be a DNN $g_{\boldsymbol{\phi}}$. That is

$$\begin{aligned} &P_\theta(\mathbf{u}|\mathbf{x}) = \mathcal{N}(\mathbf{u}; \boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta) \ , \quad \text{where } [\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta] = f_\theta(\mathbf{x}) \ , \\ &Q_\phi(\mathbf{x}|\mathbf{u}) = g_\phi(\mathbf{u}) = [\hat{\mathbf{x}}] \ , \end{aligned} \tag{6.16}$$

where $\theta$ and $\phi$ are the weight and bias parameters of the DNNs. Furthermore, the latent space is modeled as a GMM with $|\mathcal{C}|$ components with parameters $\psi := \{\pi_c, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c\}_{c=1}^{|\mathcal{C}|}$, i.e.,

$$Q_\psi(\mathbf{u}) = \sum_c \pi_c \, \mathcal{N}(\mathbf{u}; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) \ . \tag{6.17}$$

Using the parameterizations above, the optimization of (6.9) can be rewritten as

$$\max_{\theta,\phi,\psi} \ \mathcal{L}_s^{\mathrm{NN}}(\theta, \phi, \psi) \tag{6.18}$$

where the cost function $\mathcal{L}_s^{\mathrm{NN}}(\theta, \phi, \psi)$ given by

$$\mathcal{L}_s^{\mathrm{NN}}(\theta, \phi, \psi) := \mathbb{E}_{P_{\mathbf{X}}}\Big[\mathbb{E}_{P_\theta(\mathbf{U}|\mathbf{X})}[\log Q_\phi(\mathbf{X}|\mathbf{U})] - sD_{\mathrm{KL}}(P_\theta(\mathbf{U}|\mathbf{X}) \| Q_\psi(\mathbf{U}))\Big] \ . \tag{6.19}$$

Then, for a given observations of $N$ samples, i.e., $\{\mathbf{x}_i\}_{i=1}^N$, (6.18) can be approximated in terms of an empirical cost as follows

$$\max_{\theta,\phi,\psi} \ \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_{s,i}^{\mathrm{emp}}(\theta, \phi, \psi) \ , \tag{6.20}$$

where $\mathcal{L}_{s,i}^{\mathrm{emp}}(\theta, \phi, \psi)$ is the empirical cost for the $i$-th observation $\mathbf{x}_i$, and given by

$$\mathcal{L}_{s,i}^{\mathrm{emp}}(\theta, \phi, \psi) = \mathbb{E}_{P_\theta(\mathbf{U}_i|\mathbf{X}_i)}[\log Q_{\boldsymbol{\phi}}(\mathbf{X}_i|\mathbf{U}_i)] - sD_{\mathrm{KL}}(P_\theta(\mathbf{U}_i|\mathbf{X}_i) \| Q_{\boldsymbol{\psi}}(\mathbf{U}_i)) \ . \tag{6.21}$$

Furthermore, the first term of the RHS of (6.21) can be computed using Monte Carlo sampling and the reparameterization trick [29]. In particular, $P_\theta(\mathbf{u}|\mathbf{x})$ can be sampled by first sampling a random variable $\mathbf{Z}$ with distribution $P_{\mathbf{Z}}$, i.e., $P_{\mathbf{Z}} = \mathcal{N}(\mathbf{0}, \mathbf{I})$, then transforming the samples using some function $\tilde{f}_\theta : \mathcal{X} \times \mathcal{Z} \to \mathcal{U}$, i.e., $\mathbf{u} = \tilde{f}_\theta(\mathbf{x}, \mathbf{z})$. Thus,

$$\begin{aligned} \mathbb{E}_{P_\theta(\mathbf{U}_i|\mathbf{X}_i)}[\log Q_\phi(\mathbf{X}_i|\mathbf{U}_i)] &= \frac{1}{M} \sum_{m=1}^{M} \log q_\phi(\mathbf{x}_i|\mathbf{u}_{i,m}) \ , \\ &\text{with } \mathbf{u}_{i,m} = \boldsymbol{\mu}_{\theta,i} + \boldsymbol{\Sigma}_{\theta,i}^{\frac{1}{2}} \cdot \boldsymbol{\epsilon}_m \ , \quad \boldsymbol{\epsilon}_m \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \ , \end{aligned}$$

where $M$ is the number of samples for the Monte Carlo sampling step.

The second term of the RHS of (6.21) is the KL divergence between a single component multivariate Gaussian and a GMM with $|\mathcal{C}|$ components. An exact closed-form solution for the calculation of this term does not exist. However, a variational lower bound approximation [136] of it (see Appendix I.4) can be obtained as

$$D_{\mathrm{KL}}(P_{\boldsymbol{\theta}}(\mathbf{U}_i|\mathbf{X}_i)\|Q_{\boldsymbol{\psi}}(\mathbf{U}_i)) = -\log \sum_{c=1}^{|\mathcal{C}|} \pi_c \exp\left(-D_{\mathrm{KL}}(\mathcal{N}(\boldsymbol{\mu}_{\theta,i}, \boldsymbol{\Sigma}_{\theta,i})\|\mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)\right) \ . \quad (6.22)$$

In particular, in the specific case in which the covariance matrices are diagonal, i.e., $\boldsymbol{\Sigma}_{\theta,i} := \mathrm{diag}(\{\sigma^2_{\theta,i,j}\}_{j=1}^{n_u})$ and $\boldsymbol{\Sigma}_c := \mathrm{diag}(\{\sigma^2_{c,j}\}_{j=1}^{n_u})$, with $n_u$ denoting the latent space dimension, (6.22) can be computed as follows

$$\begin{aligned} &D_{\mathrm{KL}}(P_{\boldsymbol{\theta}}(\mathbf{U}_i|\mathbf{X}_i)\|Q_{\boldsymbol{\psi}}(\mathbf{U}_i)) \\ &\quad = -\log \sum_{c=1}^{|\mathcal{C}|} \pi_c \exp\left( -\frac{1}{2} \sum_{j=1}^{n_u} \left[ \frac{(\mu_{\theta,i,j} - \mu_{c,j})^2}{\sigma^2_{c,j}} + \log \frac{\sigma^2_{c,j}}{\sigma^2_{\theta,i,j}} - 1 + \frac{\sigma^2_{\theta,i,j}}{\sigma^2_{c,j}} \right] \right) , \end{aligned} \quad (6.23)$$

where $\mu_{\theta,i,j}$ and $\sigma^2_{\theta,i,j}$ are the mean and variance of the $i$-th representation in the $j$-th dimension of the latent space. Furthermore, $\mu_{c,j}$ and $\sigma^2_{c,j}$ represent the mean and variance of the $c$-th component of the GMM in the $j$-th dimension of the latent space.

Finally, we train DNNs to maximize the cost function (6.19) over the parameters $\theta, \phi$, as well as those $\psi$ of the GMM. For the training step, we use the ADAM optimization tool [83]. The training procedure is detailed in Algorithm 5.

**Algorithm 5** VIB-GMM algorithm for unsupervised learning.

1: **input:** Dataset $\mathcal{D} := \{\mathbf{x}_i\}_{i=1}^N$, parameter $s \geq 0$.
2: **output:** Optimal DNN weights $\theta^\star, \phi^\star$ and
GMM parameters $\psi^\star = \{\pi_c^\star, \boldsymbol{\mu}_c^\star, \boldsymbol{\Sigma}_c^\star\}_{c=1}^{|\mathcal{C}|}$.
3: **initialization** Initialize $\theta, \phi, \psi$.
4: **repeat**
5: Randomly select $b$ mini-batch samples $\{\mathbf{x}_i\}_{i=1}^b$ from $\mathcal{D}$.
6: Draw $m$ random i.i.d samples $\{\mathbf{z}_j\}_{j=1}^m$ from $P_{\mathbf{Z}}$.
7: Compute $m$ samples $\mathbf{u}_{i,j} = \tilde{f}_\theta(\mathbf{x}_i, \mathbf{z}_j)$
8: For the selected mini-batch, compute gradients of the empirical cost (6.20).
9: Update $\theta, \phi, \psi$ using the estimated gradient (e.g., with SGD or Adam).
10: **until** convergence of $\theta, \phi, \psi$.

Once our model is trained, we assign the given dataset into the clusters. As mentioned in Chapter 6.1, we do the assignment from the latent representations, i.e., $Q_{C|\mathbf{U}} = P_{C|\mathbf{X}}$. Hence, the probability that the observed data $\mathbf{x}_i$ belongs to the $c$-th cluster is computed as follows

$$p(c|\mathbf{x}_i) = q(c|\mathbf{u}_i) = \frac{q_{\psi^\star}(c)q_{\psi^\star}(\mathbf{u}_i|c)}{q_{\psi^\star}(\mathbf{u}_i)} = \frac{\pi_c^\star \mathcal{N}(\mathbf{u}_i; \boldsymbol{\mu}_c^\star, \boldsymbol{\Sigma}_c^\star)}{\sum_c \pi_c^\star \mathcal{N}(\mathbf{u}_i; \boldsymbol{\mu}_c^\star, \boldsymbol{\Sigma}_c^*)} , \tag{6.24}$$

where $^\star$ indicates optimal values of the parameters as found at the end of the training phase. Finally, the right cluster is picked based on the largest assignment probability value.

**Remark 16.** *It is worth mentioning that with the use of the KL approximation as given by* (6.22)*, our algorithm does not require the assumption $P_{C|\mathbf{U}} = Q_{C|\mathbf{U}}$ to hold (which is different from [31]). Furthermore, the algorithm is guaranteed to converge. However, the convergence may be to (only) local minima; and this is due to the problem* (6.18) *being generally non-convex. Related to this aspect, we mention that while without a proper pre-training, the accuracy of the VaDE algorithm may not be satisfactory, in our case, the above assumption is only used in the final assignment after the training phase is completed.* ■

**Remark 17.** *In [78], it is stated that optimizing the original IB problem with the assumption of independent latent representations amounts to disentangled representations. It is noteworthy that with such an assumption, the computational complexity can be reduced from $\mathcal{O}(n_u^2)$ to $\mathcal{O}(n_u)$. Furthermore, as argued in [78], the assumption often results only in some marginal performance loss; and for this reason, it is adopted in many machine learning applications.* ■

#### Effect of the Hyperparameter

As we already mentioned, the hyperparameter $s$ controls the trade-off between the relevance of the representation $\mathbf{U}$ and its complexity. As can be seen from (6.19) for small values of $s$, it is the cross-entropy term that dominates, i.e., the algorithm trains the parameters so as to reproduce $\mathbf{X}$ as accurately as possible. For large values of $s$, however, it is most important for the NN to produce an encoded version of $\mathbf{X}$ whose distribution matches the prior distribution of the latent space, i.e., the term $D_{\mathrm{KL}}(P_{\boldsymbol{\theta}}(\mathbf{U}|\mathbf{X})\|Q_{\boldsymbol{\psi}}(\mathbf{U}))$ is nearly zero.

In the beginning of the training process, the GMM components are randomly selected; and so, starting with a large value of the hyperparameter $s$ is likely to steer the solution towards an irrelevant prior. Hence, for the tuning of the hyperparameter $s$ in practice, it is more efficient to start with a small value of $s$ and gradually increase it with the number of epochs. This has the advantage of avoiding possible local minima, an aspect that is reminiscent of deterministic annealing [32], where $s$ plays the role of the temperature parameter. The experiments that will be reported in the next section show that proceeding in the above-described manner for the selection of the parameter $s$ helps in obtaining higher clustering accuracy and better robustness to the initialization (i.e., no need for a strong pretraining). The pseudocode for annealing is given in Algorithm 6.

**Algorithm 6** Annealing algorithm pseudocode.

1: **input:** Dataset $\mathcal{D} := \{\mathbf{x}_i\}_{i=1}^n$, hyperparameter interval $[s_{\min}, s_{\max}]$.
2: **output:** Optimal DNN weights $\theta^\star, \phi^\star$, GMM parameters $\psi^\star = \{\pi_c^\star, \boldsymbol{\mu}_c^\star, \boldsymbol{\Sigma}_c^\star\}_{c=1}^{|\mathcal{C}|}$, assignment probability $P_{C|\mathbf{X}}$.
3: **initialization** Initialize $\theta, \phi, \psi$.
4: **repeat**
5: Apply VIB-GMM algorithm.
6: Update $\psi, \theta, \phi$.
7: Update $s$, e.g., $s = (1 + \epsilon_s)s_{\text{old}}$.
8: **until** $s$ does not exceed $s_{\max}$.

**Remark 18.** *As we mentioned before, a text clustering algorithm is introduced by Slonim et al. [32, 131], which uses the IB method with an annealing procedure, where the parameter $s$ is increased gradually. In [32], the critical values of $s$ (so-called phase transitions) are observed such that if these values are missed during increasing $s$, the algorithm ends up with the wrong clusters. Therefore, how to choose the step size in the update of $s$ is very important. We note that tuning $s$ is also very critical in our algorithm, such that the step size $\epsilon_s$ in the update of $s$ should be chosen carefully, otherwise phase transitions might be skipped that would cause a non-satisfactory clustering accuracy score. However, the choice of the appropriate step size (typically very small) is rather heuristic; and there exists no concrete method for choosing the right value. The choice of step size can be seen as a trade-off between the amount of computational resource spared for running the algorithm and the degree of confidence about scanning $s$ values not to miss the phase transitions.* ■

## 6.3 Experiments

### 6.3.1 Description of used datasets

In our empirical experiments, we apply our algorithm to the unsupervised clustering of the following datasets.

MNIST: A dataset of gray-scale images of 70000 handwritten digits from 0 to 9 of dimensions $28 \times 28$ pixel each.

STL-10: A dataset of color images collected from 10 categories. Each category consists of 1300 images of size of $96 \times 96$ (pixels) $\times 3$ (RGB code). Hence, the original input dimension $n_x$ is 27648. For this dataset, we use a pretrained convolutional NN model, i.e., ResNet-50 [137] to reduce the dimensionality of the input. This preprocessing reduces the input dimension to 2048. Then, our algorithm and other baselines are used for clustering.

REUTERS10K: A dataset that is composed of 810000 English stories labeled with a category tree. As in [33], 4 root categories (corporate/industrial, government/social, markets, economics) are selected as labels and all documents with multiple labels are discarded. Then, tf-idf features are computed on the 2000 most frequently occurring words. Finally, 10000 samples are taken randomly, which are referred to as REUTERS10K dataset.

### 6.3.2 Network settings and other parameters

We use the following network architecture: the encoder is modeled with DNNs with 3 hidden layers with dimensions $n_x - 500 - 500 - 2000 - n_u$, where $n_x$ is the input dimension and $n_u$ is the dimension of the latent space. The decoder consists of DNNs with dimensions $n_u - 2000 - 500 - 500 - n_x$. All layers are fully connected. For comparison purposes, we chose the architecture of the hidden layers as well as the dimension of the latent space $n_u = 10$ to coincide with those made for the DEC algorithm of [33] and the VaDE algorithm of [31]. All except the last layers of the encoder and decoder are activated with ReLU function. For the last (i.e., latent) layer of the encoder we use a linear activation; and for the last (i.e., output) layer of the decoder we use sigmoid function for MNIST and linear activation for the remaining datasets. The batch size is 100 and the variational bound (6.20) is maximized by the Adam optimizer of [83]. The learning rate is initialized with 0.002 and decreased gradually every 20 epochs with a decay rate of 0.9 until it reaches a small value (0.0005 is our experiments). The reconstruction loss is calculated by using the cross-entropy criterion for MNIST and mean squared error function for the other datasets.

### 6.3.3 Clustering Accuracy

We evaluate the performance of our algorithm in terms of the so-called unsupervised clustering accuracy (ACC), which is a widely used metric in the context of unsupervised learning [135]. For comparison purposes, we also present those of algorithms from the previous state-of-the-art.

| | MNIST | | STL-10 | |
|---|---|---|---|---|
| | Best Run | Average Run | Best Run | Average Run |
| GMM | 44.1 | 40.5 (1.5) | 78.9 | 73.3 (5.1) |
| DEC | | | 80.6† | |
| VaDE | 91.8 | 78.8 (9.1) | 85.3 | 74.1 (6.4) |
| **VIB-GMM** | **95.1** | **83.5** (5.9) | **93.2** | **82.1** (5.6) |

† Values are taken from VaDE [31]

Table 6.1: Comparison of the clustering accuracy of various algorithms. The algorithms are run without pretraining. Each algorithm is run ten times. The values in $(\cdot)$ correspond to the standard deviations of clustering accuracies.

| | MNIST | | REURTERS10K | |
|---|---|---|---|---|
| | Best Run | Average Run | Best Run | Average Run |
| DEC | 84.3‡ | | 72.2‡ | |
| VaDE | 94.2 | 93.2 (1.5) | 79.8 | 79.1 (0.6) |
| **VIB-GMM** | **96.1** | **95.8** (0.1) | **81.6** | **81.2** (0.4) |

‡ Values are taken from DEC [33]

Table 6.2: Comparison of the clustering accuracy of various algorithms. A stacked autoencoder is used to pretrain the DNNs of the encoder and decoder before running algorithms (DNNs are initialized with the same weights and biases of [31]). Each algorithm is run ten times. The values in $(\cdot)$ correspond to the standard deviations of clustering accuracies.

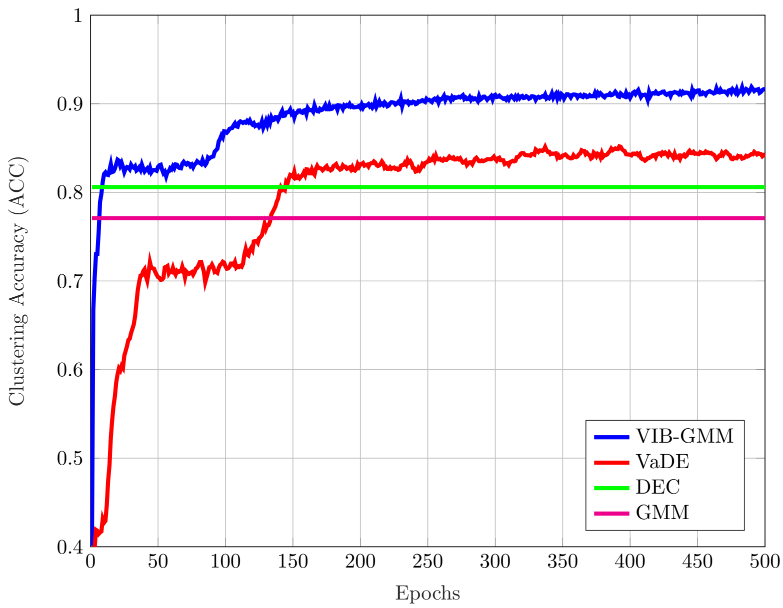



Figure 6.4: Accuracy vs. number of epochs for the STL-10 dataset.

For each of the aforementioned datasets, we run our VIB-GMM algorithm for various values of the hyperparameter $s$ inside an interval $[s_{\min}, s_{\max}]$, starting from the smaller valuer $s_{\min}$ and gradually increasing the value of $s$ every $n_{\text{epoch}}$ epochs. For the MNIST dataset, we set $(s_{\min}, s_{\max}, n_{\text{epoch}}) = (1, 5, 500)$; and for the STL-10 dataset and the REUTERS10K dataset, we choose these parameters to be $(1, 20, 500)$ and $(1, 5, 100)$, respectively. The obtained ACC accuracy results are reported in Table 6.1 and Table 6.2. It is important to note that the reported ACC results are obtained by running each algorithm ten times. For the case in which there is no pretraining[1], Table 6.1 states the accuracies of the best case run and average case run for the MNIST and STL-10 datasets. It is seen that our algorithm outperforms significantly the DEC algorithm of [33], as well as the VaDE algorithm of [31] and GMM for both the best case run and average case run. Besides, in Table 6.1, the values in parentheses correspond to the standard deviations of clustering accuracies. As seen, the standard deviation of our algorithm VIB-GMM is lower than the VaDE; which can be expounded by the robustness of VIB-GMM to non-pretraining. For the case in which there is pretraining, Table 6.2 states the accuracies of the best case run and average case run for the MNIST and REUTERS10K datasets.

[1]In [31] and [33], the DEC and VaDE algorithms are proposed to be used with pretraining; more specifically, the DNNs are initialized with a stacked autoencoder [138].

A stacked autoencoder is used to pretrain the DNNs of the encoder and decoder before running algorithms (DNNs are initialized with the same weights and biases of [31]). It is seen that our algorithm outperforms significantly the DEC algorithm of [33], as well as the VaDE algorithm of [31] and GMM for both the best case run and average case run. The effect of pretraining can be observed comparing Table 6.1 and Table 6.2 for MNIST. Using a stacked autoencoder prior to running the VaDE and VIB-GMM algorithms results in a higher accuracy, as well as a lower standard deviation of accuracies; therefore, supporting the algorithms with a stacked autoencoder is beneficial for a more robust system. Finally, for the STL-10 dataset, Figure 6.4 depicts the evolution of the best case ACC with iterations (number of epochs) for the four compared algorithms.

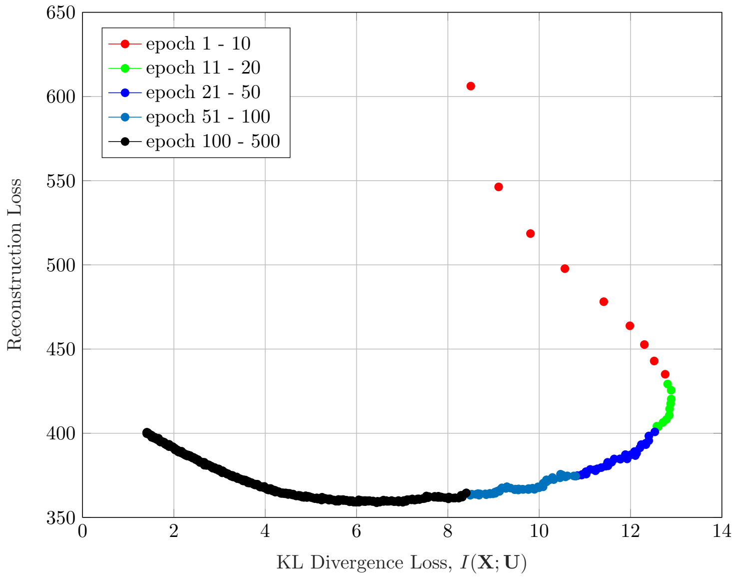
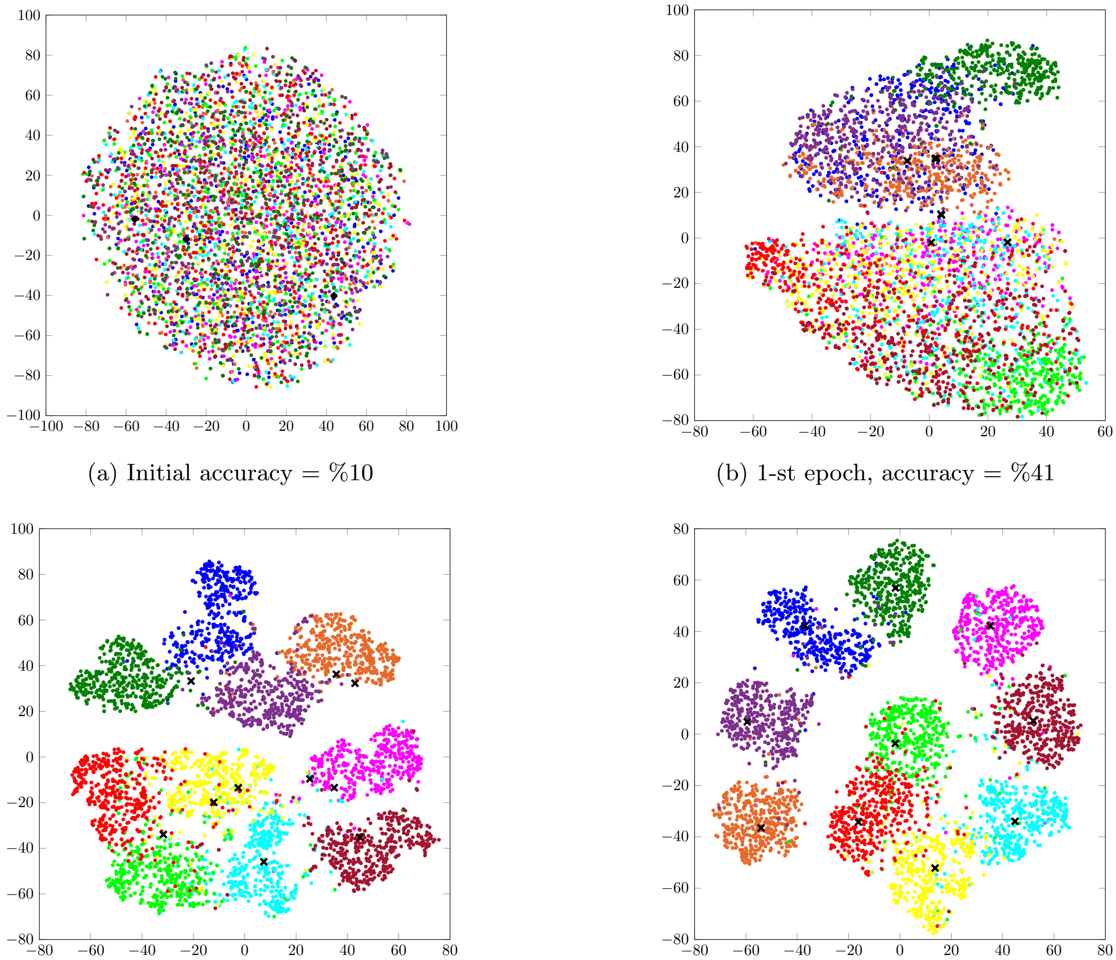



Figure 6.5: Information plane for the STL-10 dataset.

Figure 6.5 shows the evolution of the reconstruction loss of our VIB-GMM algorithm for the STL-10 dataset, as a function of simultaneously varying the values of the hyperparameter $s$ and the number of epochs (recall that, as per the described methodology, we start with $s = s_{\text{min}}$, and we increase its value gradually every $n_{\text{epoch}} = 500$ epochs). As can be seen from the figure, the few first epochs are spent almost entirely on reducing the reconstruction loss (i.e., a fitting phase), and most of the remaining epochs are spent making the found representation more concise (i.e., smaller KL divergence). This is reminiscent of the two-phase (fitting vs. compression) that was observed for supervised learning using VIB in [84].

**Remark 19.** *For a fair comparison, our algorithm VIB-GMM and the VaDE of [31] are run for the same number of epochs, e.g., $n_{epoch}$. In the VaDE algorithm, the cost function (6.11) is optimized for a particular value of hyperparameter $s$. Instead of running $n_{epoch}$ epochs for $s = 1$ as in VaDE, we run $n_{epoch}$ epochs by gradually increasing $s$ to optimize the cost (6.21). In other words, the computational resources are distributed over a range of $s$ values. Therefore, the computational complexity of our algorithm and the VaDE are equivalent.* ■



(a) Initial accuracy = %10

(b) 1-st epoch, accuracy = %41

(c) 5-th epoch, accuracy = %66

(d) Final, accuracy = %91.6

Figure 6.6: Visualization of the latent space before training; and after 1, 5 and 500 epochs.

### 6.3.4 Visualization on the Latent Space

In this section, we investigate the evolution of the unsupervised clustering of the STL-10 dataset on the latent space using our VIB-GMM algorithm. For this purpose, we find it convenient to visualize the latent space through application of the t-SNE algorithm of [139] in order to generate meaningful representations in a two-dimensional space. Figure 6.6 shows 4000 randomly chosen latent representations before the start of the training process and respectively after 1, 5 and 500 epochs. The shown points (with a $\cdot$ marker in the figure) represent latent representations of data samples whose labels are identical. Colors are used to distinguish between clusters. Crosses (with an $\mathbf{x}$ marker in the figure) correspond to the centroids of the clusters. More specifically, Figure 6.6a shows the initial latent space before the training process. If the clustering is performed on the initial representations it allows ACC accuracy of as small as 10%, i.e., as bad as a random assignment. Figure 6.6b shows the latent space after one epoch, from which a partition of some of the points starts to be already visible. With five epochs, that partitioning is significantly sharper and the associated clusters can be recognized easily. Observe, however, that the cluster centers seem still not to have converged. With 500 epochs, the ACC accuracy of our algorithm reaches %91.6 and the clusters and their centroids are neater as visible from Figure 6.6d.

# Chapter 7

# Perspectives

The IB method is connected to many other problems [72], e.g., information combining, the Wyner-Ahlswede-Körner problem, the efficiency of investment information, the privacy funnel problem, and these connections are reviewed in Chapter 2.3.3. The distributed IB problem that we study in this thesis can be instrumental to study the distributed setups of these connected problems. Let us consider the distributed privacy funnel problem. For example, a company, which operates over 2 different regions, needs to share some data – that can be also used to draw some private data – with two different consultants for some analysis. Instead of sharing all data with a single consultant, sharing the data related to each region with different consultants who are experts for different regions may provide better results. The problem is how to share the data with consultants without disclosing the private data, and can be solved by exploring the connections of the distributed IB with the privacy funnel.

This thesis covers the topics related to the problem of the source coding. However, in the information theory it is known that there is a substantial relation – so-called the duality – between the problems of the source and channel coding. This relation has been used to infer solutions from one field (in which there are already known working techniques) to the other one. Now, let consider the CEO problem in a different way, such that the agents are deployed over an area and connected to the cloud (the central processor, or the CEO) via finite capacity backhaul links. This problem is called as the Cloud - Radio Access Networks (C-RAN). The authors in [62, 63] utilize useful connections with the CEO source coding problem under logarithmic loss distortion measure for finding the capacity region of the C-RAN with oblivious relaying (for the converse proof).

Considering the high amount of research which is done recently in machine learning field, the distributed learning may become an important topic in the future. This thesis provides a theoretical background of distributed learning by presenting an information-theoretical connections, as well as, some algorithmic contributions (e.g., the inference type algorithms for classification and clustering). We believe that our contribution can be beneficial to understand the theory behind in the distributed learning area for the future research.

Like for the single-encoder IB problem of [17] and an increasing number of works that followed, including [10, Section III-F], in our approach for the distributed learning problem we adopted we have considered a mathematical formulation that is asymptotic (blocklength $n$ allowed to be large enough). In addition to that it leads to an exact characterization, the result also readily provides a lower bound on the performance in the non-asymptotic setting (e.g., one shot). For the latter setting known approaches (e.g., the functional representation lemma of [140]) would lead to only non-matching inner and outer bounds on the region of optimal trade-off pairs, as this is the case even for the single encoder case [141].

One of the interesting problems left unaddressed in this thesis is the characterization of the optimal input distributions under rate-constrained compression at the relays, where it is known that discrete signaling sometimes outperforms Gaussian signaling for single-user Gaussian C-RAN [60]. One may consider an extension to the frequency selective additive Gaussian noise channel, in parallel to the Gaussian Information Bottleneck [142]; or to the uplink Gaussian inference channel with backhaul links of variable connectivity conditions [143]. Another interesting direction can be to find the worst-case noise for a given input distribution, e.g., Gaussian, for the case in which the compression rate at each relay is constrained. Finally, the processing constraint of continuous waveforms, such as sampling at a given rate [144, 145] with a focus on the logarithmic loss, is another aspect to be mentioned, which in turn boils down to the distributed Information Bottleneck [1, 111].

# Appendices

# Appendix A

# Proof of Theorem 1

## A.1 Direct Part

For the proof of achievability of Theorem 1, we use a slight generalization of Gastpar's inner bound of [146, Theorem 1], which provides an achievable rate region for the multiterminal source coding model with side information, modified to include time-sharing.

**Proposition 11.** *The rate-distortion vector* $(R_1, \ldots, R_K, D)$ *is achievable if*

$$\sum_{k \in \mathcal{S}} R_k \geq I(U_{\mathcal{S}}; Y_{\mathcal{S}} | U_{\mathcal{S}^c}, Y_0, Q) \;, \quad \textit{for } \mathcal{S} \subseteq \mathcal{K} \;, \tag{A.1}$$

$$D \geq \mathbb{E}[d(X, f(U_{\mathcal{K}}, Y_0, Q))] \;,$$

*for some joint measure of the form*

$$P_{X,Y_0,Y_1,Y_2}(x, y_0, y_1, y_2) P_Q(q) \prod_{k=1}^{K} P_{U_k|Y_k,Q}(u_k|y_k, q) \;,$$

*and a reproduction function*

$$f(U_{\mathcal{K}}, Y_0, Q) : \mathcal{U}_1 \times \cdots \times \mathcal{U}_K \times \mathcal{Y}_0 \times \mathcal{Q} \longrightarrow \hat{\mathcal{X}} \;. \qquad \blacksquare$$

The proof of achievability of Theorem 1 simply follows by a specialization of the result of Proposition 11 to the setting in which distortion is measured under logarithmic loss. For instance, we apply Proposition 11 with the reproduction functions chosen as

$$f(U_{\mathcal{K}}, Y_0, Q) = \Pr[X = x | U_{\mathcal{K}}, Y_0, Q] \;.$$

Then, note that with such a choice we have

$$\mathbb{E}[d(X, f(U_{\mathcal{K}}, Y_0, Q))] = H(X | U_{\mathcal{K}}, Y_0, Q) \;.$$

The resulting region can be shown to be equivalent to that given in Theorem 1 using supermodular optimization arguments. The proof is along the lines of that of [10, Lemma 5] and is omitted for brevity.

## A.2 Converse Part

We first state the following lemma, which is an easy extension of that of [10, Lemma 1] to the case in which the decoder also observes statistically dependent side information. The proof of Lemma 8 follows along the lines of that of [10, Lemma 1], and is therefore omitted for brevity.

**Lemma 8.** *Let $T := (\phi_1^{(n)}(Y_1^n), \ldots, \phi_K^{(n)}(Y_K^n))$. Then for the CEO problem of Figure 1.1 under logarithmic loss, we have $n\mathbb{E}[d^{(n)}(X^n, \hat{X}^n)] \geq H(X^n|T, Y_0^n)$.* ■

Let $\mathcal{S}$ be a non-empty set of $\mathcal{K}$ and $J_k := \phi_k^{(n)}(Y_k^n)$ be the message sent by Encoder $k$, $k \in \mathcal{K}$, where $\{\phi_k^{(n)}\}_{k=1}^K$ are the encoding functions corresponding to a scheme that achieves $(R_1, \ldots, R_K, D)$.

Define, for $i = 1, \ldots, n$, the following random variables

$$U_{k,i} := (J_k, Y_k^{i-1}) \; , \qquad Q_i := (X^{i-1}, X_{i+1}^n, Y_0^{i-1}, Y_{0,i+1}^n) \; . \tag{A.2}$$

We can lower bound the distortion $D$ as

$$\begin{aligned} nD &\overset{(a)}{\geq} H(X^n|J_{\mathcal{K}}, Y_0^n) \\ &= \sum_{i=1}^n H(X_i|J_{\mathcal{K}}, X^{i-1}, Y_0^n) \\ &\overset{(b)}{\geq} \sum_{i=1}^n H(X_i|J_{\mathcal{K}}, X^{i-1}, X_{i+1}^n, Y_{\mathcal{K}}^{i-1}, Y_0^n) \\ &= \sum_{i=1}^n H(X_i|J_{\mathcal{K}}, X^{i-1}, X_{i+1}^n, Y_{\mathcal{K}}^{i-1}, Y_0^{i-1}, Y_{0,i}, Y_{0,i+1}^n) \\ &\overset{(c)}{=} \sum_{i=1}^n H(X_i|U_{\mathcal{K},i}, Y_{0,i}, Q_i) \; , \end{aligned} \tag{A.3}$$

where $(a)$ follows due to Lemma 8; $(b)$ holds since conditioning reduces entropy; and $(c)$ follows by substituting using (A.2).

Now, we lower bound the rate term as

$$
\begin{aligned}
&n\sum_{k\in\mathcal{S}} R_k \\
&\geq \sum_{k\in\mathcal{S}} H(J_k) \geq H(J_{\mathcal{S}}) \geq H(J_{\mathcal{S}}|J_{\mathcal{S}^c},Y_0^n) \geq I(J_{\mathcal{S}};X^n,Y_{\mathcal{S}}^n|J_{\mathcal{S}^c},Y_0^n) \\
&= I(J_{\mathcal{S}};X^n|J_{\mathcal{S}^c},Y_0^n) + I(J_{\mathcal{S}};Y_{\mathcal{S}}^n|X^n,J_{\mathcal{S}^c},Y_0^n) \\
&= H(X^n|J_{\mathcal{S}^c},Y_0^n) - H(X^n|J_{\mathcal{K}},Y_0^n) + I(J_{\mathcal{S}};Y_{\mathcal{S}}^n|X^n,J_{\mathcal{S}^c},Y_0^n) \\
&\overset{(a)}{\geq} H(X^n|J_{\mathcal{S}^c},Y_0^n) - nD + I(J_{\mathcal{S}};Y_{\mathcal{S}}^n|X^n,J_{\mathcal{S}^c},Y_0^n) \\
&= \sum_{i=1}^n H(X_i|J_{\mathcal{S}^c},X^{i-1},Y_0^n) - nD + I(J_{\mathcal{S}};Y_{\mathcal{S}}^n|X^n,J_{\mathcal{S}^c},Y_0^n) \\
&\overset{(b)}{\geq} \sum_{i=1}^n H(X_i|J_{\mathcal{S}^c},X^{i-1},X_{i+1}^n,Y_{\mathcal{S}^c}^{i-1},Y_0^n) - nD + I(J_{\mathcal{S}};Y_{\mathcal{S}}^n|X^n,J_{\mathcal{S}^c},Y_0^n) \\
&= \sum_{i=1}^n H(X_i|J_{\mathcal{S}^c},X^{i-1},X_{i+1}^n,Y_{\mathcal{S}^c}^{i-1},Y_0^{i-1},Y_{0,i},Y_{0,i+1}^n) - nD + I(J_{\mathcal{S}};Y_{\mathcal{S}}^n|X^n,J_{\mathcal{S}^c},Y_0^n) \\
&\overset{(c)}{=} \sum_{i=1}^n H(X_i|U_{\mathcal{S}^c,i},Y_{0,i},Q_i) - nD + \Theta \ ,
\end{aligned}
\tag{A.4}
$$

where $(a)$ follows due to Lemma 8; $(b)$ holds since conditioning reduces entropy; and $(c)$ follows by substituting using (A.2) and $\Theta := I(J_{\mathcal{S}};Y_{\mathcal{S}}^n|X^n,J_{\mathcal{S}^c},Y_0^n)$.

To continue with lower-bounding the rate term, we single-letterize the term $\Theta$ as

$$
\begin{aligned}
\Theta &= I(J_{\mathcal{S}};Y_{\mathcal{S}}^n|X^n,J_{\mathcal{S}^c},Y_0^n) \\
&\overset{(a)}{\geq} \sum_{k\in\mathcal{S}} I(J_k;Y_k^n|X^n,Y_0^n) \\
&= \sum_{k\in\mathcal{S}}\sum_{i=1}^n I(J_k;Y_{k,i}|Y_k^{i-1},X^n,Y_0^n) \\
&\overset{(b)}{=} \sum_{k\in\mathcal{S}}\sum_{i=1}^n I(J_k,Y_k^{i-1};Y_{k,i}|X^n,Y_0^n) \\
&= \sum_{k\in\mathcal{S}}\sum_{i=1}^n I(J_k,Y_k^{i-1};Y_{k,i}|X^{i-1},X_i,X_{i+1}^n,Y_0^{i-1},Y_{0,i},Y_{0,i+1}^n) \\
&\overset{(c)}{=} \sum_{k\in\mathcal{S}}\sum_{i=1}^n I(U_{k,i};Y_{k,i}|X_i,Y_{0,i},Q_i) \ ,
\end{aligned}
\tag{A.5}
$$

where $(a)$ follows due to the Markov chain $J_k -\!\!\circ\!\!- Y_k^n -\!\!\circ\!\!- (X^n,Y_0^n) -\!\!\circ\!\!- Y_{\mathcal{S}\setminus k}^n -\!\!\circ\!\!- J_{\mathcal{S}\setminus k}$, $k\in\mathcal{K}$; $(b)$ follows due to the Markov chain $Y_{k,i} -\!\!\circ\!\!- (X^n,Y_0^n) -\!\!\circ\!\!- Y_k^{i-1}$; and $(c)$ follows by substituting using (A.2).

Then, combining (A.4) and (A.5), we get

$$n \sum_{k \in \mathcal{S}} R_k \geq \sum_{i=1}^{n} H(X_i | U_{\mathcal{S}^c,i}, Y_{0,i}, Q_i) - nD + \sum_{k \in \mathcal{S}} \sum_{i=1}^{n} I(U_{k,i}; Y_{k,i} | X_i, Y_{0,i}, Q_i) \ . \tag{A.6}$$

Summarizing, we have from (A.3) and (A.6)

$$\begin{aligned} nD &\geq \sum_{i=1}^{n} H(X_i | U_{\mathcal{K},i}, Y_{0,i}, Q_i) \\ nD + n \sum_{k \in \mathcal{S}} R_k &\geq \sum_{i=1}^{n} H(X_i | U_{\mathcal{S}^c,i}, Y_{0,i}, Q_i) + \sum_{k \in \mathcal{S}} \sum_{i=1}^{n} I(U_{k,i}; Y_{k,i} | X_i, Y_{0,i}, Q_i) \ . \end{aligned}$$

We note that the random variables $U_{\mathcal{K},i}$ satisfy the Markov chain $U_{k,i} -\!\!\circ\!\!- Y_{k,i} -\!\!\circ\!\!- X_i -\!\!\circ\!\!- Y_{\mathcal{K}\setminus k,i} -\!\!\circ\!\!- U_{\mathcal{K}\setminus k,i}$, $k \in \mathcal{K}$. Finally, a standard time-sharing argument completes the proof.

# Appendix B

# Proof of Theorem 2

## B.1 Direct Part

For the proof of achievability of Theorem 2, we use a slight generalization of Gastpar's inner bound of [89, Theorem 2], which provides an achievable rate-distortion region for the multiterminal source coding model of Section 3.2 in the case of general distortion measure, to include time-sharing.

**Proposition 12.** *(Gastpar Inner Bound [89, Theorem 2] with time-sharing) The rate-distortion vector* $(R_1, R_2, D_1, D_2)$ *is achievable if*

$$\begin{aligned} R_1 &\geq I(U_1; Y_1 | U_2, Y_0, Q) \\ R_2 &\geq I(U_2; Y_2 | U_1, Y_0, Q) \\ R_1 + R_2 &\geq I(U_1, U_2; Y_1, Y_2 | Y_0, Q) \\ D_1 &\geq \mathbb{E}[d(X_1, f_1(U_1, U_2, Y_0, Q))] \\ D_2 &\geq \mathbb{E}[d(X_2, f_2(U_1, U_2, Y_0, Q))] \ , \end{aligned}$$

*for some joint measure of the form*

$$P_{Y_0,Y_1,Y_2}(y_0, y_1, y_2) P_Q(q) P_{U_1|Y_1,Q}(u_1|y_1, q) P_{U_2|Y_2,Q}(u_2|y_2, q) \ ,$$

*and reproduction functions*

$$f_k \ : \mathcal{U}_1 \times \mathcal{U}_2 \times \mathcal{Y}_0 \times \mathcal{Q} \longrightarrow \hat{Y}_k \ , \quad \textit{for } k = 1, 2 \ . \qquad \blacksquare$$

The proof of achievability of Theorem 2 simply follows by a specialization of the result of Proposition 12 to the setting in which distortion is measured under logarithmic loss.

For instance, we apply Proposition 12 with the reproduction functions chosen as

$$f_k(U_1, U_2, Y_0, Q) := \Pr[Y_k = y_k | U_1, U_2, Y_0, Q] \ , \quad \text{for } \ k = 1, 2 \ .$$

Then, note that with such a choice we have

$$\mathbb{E}[d(Y_k, f_k(U_1, U_2, Y_0, Q)] = H(Y_k | U_1, U_2, Y_0, Q) \ , \quad \text{for } \ k = 1, 2 \ .$$

## B.2 Converse Part

We first state the following lemma, which is an easy extension of that of [10, Lemma 1] to the case in which the decoder also observes statistically dependent side information. The proof of Lemma 9 follows along the lines of that of [10, Lemma 1], and is therefore omitted for brevity.

**Lemma 9.** *Let $T := (\phi_1^{(n)}(Y_1^n), \phi_2^{(n)}(Y_2^n))$. Then, for the multiterminal source coding problem under logarithmic loss measure we have $n\mathbb{E}[d(Y_k^n, \hat{Y}_k^n)] \geq H(Y_k^n | T, Y_0^n)$ for $k = 1, 2$.* ■

The proof of converse of Theorem 2 follows by Lemma 10 and Lemma 11 below, the proofs of which follow relatively straightforwardly those in the proof of [10, Theorem 12].

**Lemma 10.** *If a rate-distortion quadruple $(R_1, R_2, \tilde{D}_1, D_2)$ is achievable for the model of Section 3.2, then there exist a joint measure*

$$P_{Y_0, Y_1, Y_2}(y_0, y_1, y_2) P_Q(q) P_{U_1|Y_1,Q}(u_1|y_1, q) P_{U_2|Y_2,Q}(u_2|y_2, q) \ , \tag{B.1}$$

*and a $D_1 \leq \tilde{D}_1$ which satisfies*

$$D_1 \geq H(X_1 | U_1, U_2, Y, Q) \tag{B.2a}$$
$$D_2 \geq D_1 + H(X_2 | U_1, U_2, Y, Q) - H(X_1 | U_1, U_2, Y, Q) \ , \tag{B.2b}$$

*and*

$$R_1 \geq H(Y_1 | U_2, Y_0, Q) - D_1 \tag{B.3a}$$
$$R_2 \geq I(U_2; Y_2 | Y_1, Y_0, Q) + H(Y_1 | U_1, Y_0, Q) - D_1 \tag{B.3b}$$
$$R_1 + R_2 \geq I(U_2; Y_2 | Y_1, Y_0, Q) + H(Y_1 | Y_0) - D_1 \ . \tag{B.3c}$$

*Proof.* Let $J_1 := \phi_1^{(n)}(Y_1^n)$ and $J_2 := \phi_2^{(n)}(Y_2^n)$, where the $\phi_1^{(n)}$ and $\phi_2^{(n)}$ are the encoding functions corresponding to a scheme that achieves $(R_1, R_2, \tilde{D}_1, D_2)$. Define

$$D_1 := \frac{1}{n} H(Y_1^n | J_1, J_2, Y_0^n) \ .$$

Also, define, for $i = 1, \ldots, n$, the following random variables

$$U_{1,i} := J_1 \ , \qquad U_{2,i} := (J_2, Y_{2,i+1}^n) \ , \qquad Q_i := (Y_1^{i-1}, Y_{2,i+1}^n, Y_0^{i-1}, Y_{0,i+1}^n) \ . \tag{B.4}$$

First, note that by Lemma 9 we have $n\tilde{D}_1 \geq H(Y_1^n|J_1, J_2, Y_0^n)$; and, so, $D_1 \leq \tilde{D}_1$. Also, we have

$$\begin{aligned}
nD_1 &= \sum_{i=1}^n H(Y_{1,i}|J_1, J_2, Y_1^{i-1}, Y_0^n) \\
&\overset{(a)}{\geq} \sum_{i=1}^n H(Y_{1,i}|J_1, J_2, Y_1^{i-1}, Y_{2,i+1}^n, Y_0^n) \\
&= \sum_{i=1}^n H(Y_{1,i}|J_1, J_2, Y_1^{i-1}, Y_{2,i+1}^n, Y_0^{i-1}, Y_{0,i}, Y_{0,i+1}^n) \\
&\overset{(b)}{=} \sum_{i=1}^n H(X_{1,i}|U_{1,i}, U_{2,i}, Y_{0,i}, Q_i) \ ,
\end{aligned}$$

where $(a)$ holds since conditioning reduces entropy; and $(b)$ follows by substituting using (B.4).

We can lower bound the distortion $D_2$ as

$$\begin{aligned}
nD_2 &\geq H(Y_2^n|J_1, J_2, Y_0^n) \\
&= H(Y_1^n|J_1, J_2, Y_0^n) + [H(Y_2^n|J_1, J_2, Y_0^n) - H(Y_1^n|J_1, J_2, Y_0^n)] \\
&= nD_1 + \Theta \ ,
\end{aligned} \tag{B.5}$$

where $\Theta := H(Y_2^n|J_1, J_2, Y_0^n) - H(Y_1^n|J_1, J_2, Y_0^n)$.

To continue with lower-bounding the distortion $D_2$, we single-letterize the term $\Theta$ as

$$\begin{aligned}
\Theta &= H(Y_2^n|J_1, J_2, Y_0^n) - H(Y_1^n|J_1, J_2, Y_0^n) \\
&= \sum_{i=1}^n H(Y_{2,i}|J_1, J_2, Y_{2,i+1}^n, Y_0^n) - H(Y_{1,i}|J_1, J_2, Y_1^{i-1}, Y_0^n) \\
&= \sum_{i=1}^n I(Y_1^{i-1}; Y_{2,i}|J_1, J_2, Y_{2,i+1}^n, Y_0^n) + H(Y_{2,i}|J_1, J_2, Y_1^{i-1}, Y_{2,i+1}^n, Y_0^n) \\
&\quad - \sum_{i=1}^n I(Y_{2,i+1}^n; Y_{1,i}|J_1, J_2, Y_1^{i-1}, Y_0^n) + H(Y_{1,i}|J_1, J_2, Y_1^{i-1}, Y_{2,i+1}^n, Y_0^n)
\end{aligned}$$

$$\stackrel{(a)}{=} \sum_{i=1}^{n} H(Y_{2,i}|J_1, J_2, Y_1^{i-1}, Y_{2,i+1}^n, Y_0^n) - H(Y_{1,i}|J_1, J_2, Y_1^{i-1}, Y_{2,i+1}^n, Y_0^n) \ , \tag{B.6}$$

where $(a)$ follows by the Csiszár-Körner sum-identity

$$\sum_{i=1}^{n} I(Y_1^{i-1}; Y_{2,i}|J_1, J_2, Y_{2,i+1}^n, Y_0^n) = \sum_{i=1}^{n} I(Y_{2,i+1}^n; Y_{1,i}|J_1, J_2, Y_1^{i-1}, Y_0^n) \ .$$

Then, combining (B.5) and (B.6), we get

$$\begin{aligned}
nD_2 &\geq nD_1 + \sum_{i=1}^{n} H(Y_{2,i}|J_1, J_2, Y_1^{i-1}, Y_{2,i+1}^n, Y_0^n) - H(Y_{1,i}|J_1, J_2, Y_1^{i-1}, Y_{2,i+1}^n, Y_0^n) \\
&= nD_1 + \sum_{i=1}^{n} H(Y_{2,i}|J_1, J_2, Y_1^{i-1}, Y_{2,i+1}^n, Y_0^{i-1}, Y_{0,i}, Y_{0,i+1}^n) \\
&\quad - H(Y_{1,i}|J_1, J_2, Y_1^{i-1}, Y_{2,i+1}^n, Y_0^{i-1}, Y_{0,i}, Y_{0,i+1}^n) \\
&= nD_1 + \sum_{i=1}^{n} H(Y_{2,i}|U_{1,i}, U_{2,i}, Y_{0,i}, Q_i) - H(Y_{1,i}|U_{1,i}, U_{2,i}, Y_{0,i}, Q_i) \ ,
\end{aligned}$$

where the last equality follows by substituting using (B.4).

Rate $R_1$ can be bounded easily as

$$\begin{aligned}
nR_1 &\geq H(J_1) \geq H(J_1|J_2, Y_0^n) \geq I(J_1; Y_1^n|J_2, Y_0^n) = H(Y_1^n|J_2, Y_0^n) - nD_1 \\
&= \sum_{i=1}^{n} H(Y_{1,i}|J_2, Y_{1,i+1}^n, Y_0^n) - nD_1 \\
&\stackrel{(a)}{\geq} \sum_{i=1}^{n} H(Y_{1,i}|J_2, Y_{1,i+1}^n, Y_{2,i+1}^n, Y_0^n) - nD_1 \\
&\stackrel{(b)}{=} \sum_{i=1}^{n} H(Y_{1,i}|J_2, Y_{2,i+1}^n, Y_0^{i-1}, Y_{0,i}, Y_{0,i+1}^n) - nD_1 \\
&\stackrel{(c)}{\geq} \sum_{i=1}^{n} H(Y_{1,i}|J_2, Y_1^{i-1}, Y_{2,i+1}^n, Y_0^{i-1}, Y_{0,i}, Y_{0,i+1}^n) - nD_1 \\
&\stackrel{(d)}{=} \sum_{i=1}^{n} H(Y_{1,i}|U_{2,i}, Y_{0,i}, Q_i) - nD_1 \ ,
\end{aligned}$$

where $(a)$ holds since conditioning reduces entropy; $(b)$ follows since $Y_{1,i} -\!\!\circ\!\!- (J_2, Y_{2,i+1}^n, Y_0^n) -\!\!\circ\!\!- Y_{1,i+1}^n$ forms a Markov chain; $(c)$ holds since conditioning reduces entropy; and $(d)$ follows by substituting using (B.4).

Now, we lower bound the rate $R_2$ as

$$
\begin{aligned}
nR_2 \geq H(J_2) \geq H(J_2|J_1, Y_0^n) &= H(J_2|J_1, Y_1^n, Y_0^n) + I(J_2; Y_1^n|J_1, Y_0^n) \\
&\geq I(J_2; Y_2^n|J_1, Y_1^n, Y_0^n) + I(J_2; Y_1^n|J_1, Y_0^n) \\
&= I(J_2; Y_2^n|J_1, Y_1^n, Y_0^n) + H(Y_1^n|J_1, Y_0^n) - nD_1 \\
&\overset{(a)}{=} I(J_2; Y_2^n|Y_1^n, Y_0^n) + H(Y_1^n|J_1, Y_0^n) - nD_1 \\
&= \sum_{i=1}^{n} I(J_2; Y_{2,i}|Y_1^n, Y_{2,i+1}^n, Y_0^n) + H(Y_{1,i}|J_1, Y_1^{i-1}, Y_0^n) - nD_1 \\
&\overset{(b)}{\geq} \sum_{i=1}^{n} I(J_2; Y_{2,i}|Y_1^n, Y_{2,i+1}^n, Y_0^n) + H(Y_{1,i}|J_1, Y_1^{i-1}, Y_{2,i+1}^n, Y_0^n) - nD_1 \\
&\overset{(c)}{=} \sum_{i=1}^{n} I(J_2, Y_{1,i+1}^n; Y_{2,i}|Y_1^{i-1}, Y_{1,i}, Y_{2,i+1}^n, Y_0^n) + H(Y_{1,i}|J_1, Y_1^{i-1}, Y_{2,i+1}^n, Y_0^n) - nD_1 \\
&= \sum_{i=1}^{n} I(J_2, Y_{1,i+1}^n, Y_{2,i+1}^n; Y_{2,i}|Y_1^{i-1}, Y_{1,i}, Y_{2,i+1}^n, Y_0^n) + H(Y_{1,i}|J_1, Y_1^{i-1}, Y_{2,i+1}^n, Y_0^n) - nD_1 \\
&\overset{(d)}{\geq} \sum_{i=1}^{n} I(J_2, Y_{2,i+1}^n; Y_{2,i}|Y_1^{i-1}, Y_{1,i}, Y_{2,i+1}^n, Y_0^n) + H(Y_{1,i}|J_1, Y_1^{i-1}, Y_{2,i+1}^n, Y_0^n) - nD_1 \\
&\overset{(e)}{=} \sum_{i=1}^{n} I(U_{2,i}; Y_{2,i}|Y_{1,i}, Y_{0,i}, Q_i) + H(Y_{1,i}|U_{1,i}, Y_{0,i}, Q_i) - nD_1 \ ,
\end{aligned}
$$

where $(a)$ holds since $J_1$ is a deterministic function of $Y_1^n$; $(b)$ holds since conditioning reduces the entropy; $(c)$ follows since $Y_{1,i+1}^n \multimap (Y_1^{i-1}, Y_{1,i}, Y_{2,i+1}^n, Y_0^n) \multimap Y_{2,i}$ forms a Markov chain; $(d)$ follows since conditioning reduces the entropy; and $(e)$ follows by substituting using (B.4).

The sum-rate $R_1 + R_2$ can be lower bounded similarly, as

$$
\begin{aligned}
n(R_1 + R_2) &\geq H(J_1) + H(J_2) \\
&\geq H(J_1|J_2, Y_0^n) + H(J_2|Y_0^n) \\
&\geq I(J_1; Y_1^n|J_2, Y_0^n) + I(J_2; Y_1^n, Y_2^n|Y_0^n) \\
&= I(J_1; Y_1^n|J_2, Y_0^n) + I(J_2; Y_1^n|Y_0^n) + I(J_2; Y_2^n|Y_1^n, Y_0^n) \\
&= I(J_1, J_2; Y_1^n|Y_0^n) + I(J_2; Y_2^n|Y_1^n, Y_0^n) \\
&= H(Y_1^n|Y_0^n) - nD_1 + I(J_2; Y_2^n|Y_1^n, Y_0^n) \\
&\overset{(a)}{=} \sum_{i=1}^{n} I(J_2; Y_{2,i}|Y_1^n, Y_{2,i+1}^n, Y_0^n) + H(Y_{1,i}|Y_{0,i}) - nD_1
\end{aligned}
$$

$$
\begin{aligned}
&\stackrel{(b)}{=} \sum_{i=1}^{n} I(J_2, Y_{1,i+1}^n; Y_{2,i}|Y_1^{i-1}, Y_{1,i}, Y_{2,i+1}^n, Y_0^n) + H(Y_{1,i}|Y_{0,i}) - nD_1 \\
&= \sum_{i=1}^{n} I(J_2, Y_{1,i+1}^n, Y_{2,i+1}^n; Y_{2,i}|Y_1^{i-1}, Y_{1,i}, Y_{2,i+1}^n, Y_0^n) + H(Y_{1,i}|Y_{0,i}) - nD_1 \\
&\stackrel{(c)}{\geq} \sum_{i=1}^{n} I(J_2, Y_{2,i+1}^n; Y_{2,i}|Y_1^{i-1}, Y_{1,i}, Y_{2,i+1}^n, Y_0^n) + H(Y_{1,i}|Y_{0,i}) - nD_1 \\
&\stackrel{(d)}{=} \sum_{i=1}^{n} I(U_{2,i}; Y_{2,i}|Y_{1,i}, Y_{0,i}, Q_i) + H(Y_{1,i}|Y_{0,i}) - nD_1 \ ,
\end{aligned}
$$

where $(a)$ follows since the source $(Y_0^n, Y_1^n, Y_2^n)$ is memoryless; $(b)$ follows since $Y_{1,i+1}^n -\!\!\circ\!\!- (Y_1^{i-1}, Y_{1,i}, Y_{2,i+1}^n, Y_0^n) -\!\!\circ\!\!- Y_{2,i}$ forms a Markov chain; $(c)$ holds since conditioning reduces the entropy; and $(d)$ follows by substituting using (B.4).

Summarizing, the distortion pair $(D_1, D_2)$ satisfies

$$
\begin{aligned}
D_1 &\geq \frac{1}{n}\sum_{i=1}^{n} H(X_{1,i}|U_{1,i}, U_{2,i}, Y_{0,i}, Q_i) \\
D_2 &\geq D_1 + \frac{1}{n}\sum_{i=1}^{n} H(Y_{2,i}|U_{1,i}, U_{2,i}, Y_{0,i}, Q_i) - H(Y_{1,i}|U_{1,i}, U_{2,i}, Y_{0,i}, Q_i) \ ,
\end{aligned}
$$

and the rate pair $(R_1, R_2)$ satisfies

$$
\begin{aligned}
R_1 &\geq \frac{1}{n}\sum_{i=1}^{n} H(Y_{1,i}|U_{2,i}, Y_{0,i}, Q_i) - D_1 \\
R_2 &\geq \frac{1}{n}\sum_{i=1}^{n} I(U_{2,i}; Y_{2,i}|Y_{1,i}, Y_{0,i}, Q_i) + H(Y_{1,i}|U_{1,i}, Y_{0,i}, Q_i) - D_1 \\
R_1 + R_2 &\geq \frac{1}{n}\sum_{i=1}^{n} I(U_{2,i}; Y_{2,i}|Y_{1,i}, Y_{0,i}, Q_i) + H(Y_{1,i}|Y_{0,i}) - D_1 \ .
\end{aligned}
$$

It is easy to see that the random variables $(U_{1,i}, U_{2,i}, Q_i)$ satisfy that $U_{1,i} -\!\!\circ\!\!- (Y_{1,i}, Q_i) -\!\!\circ\!\!- (Y_{0,i}, Y_{2,i}, U_{2,i})$ and $U_{2,i} -\!\!\circ\!\!- (X_{2,i}, Q_i) -\!\!\circ\!\!- (Y_{0,i}, Y_{1,i}, U_{1,i})$ form Markov chains. Finally, a standard time-sharing argument proves Lemma 10. □

The rest of the proof of converse of Theorem 2 follows using the following lemma, the proof of which is along the lines of that of [10, Lemma 9] and is omitted for brevity.

**Lemma 11.** *Let a rate-distortion quadruple $(R_1, R_2, D_1, D_2)$ be given. If there exists a joint measure of the form (B.1) such that (B.2) and (B.3) are satisfied, then the rate-distortion quadruple $(R_1, R_2, D_1, D_2)$ is in the region described by Theorem 2.* ■

# Appendix C

# Proof of Proposition 3

We start with the proof of the direct part. Let a non-negative tuple $(R_1, \ldots, R_K, E) \in \mathcal{R}_{\mathrm{HT}}$ be given. Since $\mathcal{R}_{\mathrm{HT}} = \overline{\mathcal{R}^\star}$, then there must exist a series of non-negative tuples $\{(R_1^{(m)}, \ldots, R_K^{(m)}, E^{(m)})\}_{m\in\mathbb{N}}$ such that

$$(R_1^{(m)}, \ldots, R_K^{(m)}, E^{(m)}) \in \mathcal{R}^\star \ , \quad \text{for all } m \in \mathbb{N}, \quad \text{and} \tag{C.1a}$$

$$\lim_{m\to\infty} (R_1^{(m)}, \ldots, R_K^{(m)}, E^{(m)}) = (R_1, \ldots, R_K, E) \ . \tag{C.1b}$$

Fix $\delta' > 0$. Then, $\exists\ m_0 \in \mathbb{N}$ such that for all $m \geq m_0$, we have

$$R_k \geq R_k^{(m)} - \delta' \ , \quad \text{for } k = 1, \ldots, K \ , \tag{C.2a}$$

$$E \leq E^{(m)} + \delta' \ . \tag{C.2b}$$

For $m \geq m_0$, there exist a series $\{n_m\}_{m\in\mathbb{N}}$ and functions $\{\breve{\phi}_k^{(n_m)}\}_{k\in\mathcal{K}}$ such that

$$R_k^{(m)} \geq \frac{1}{n_m} \log|\breve{\phi}_k^{(n_m)}| \ , \quad \text{for } k = 1, \ldots, K \ , \tag{C.3a}$$

$$E^{(m)} \leq \frac{1}{n_m} I(\{\breve{\phi}_k^{(n_m)}(Y_k^{n_m})\}_{k\in\mathcal{K}}; X^{n_m}|Y_0^{n_m}) \ . \tag{C.3b}$$

Combining (C.2) and (C.3) we get that for all $m \geq m_0$,

$$R_k \geq \frac{1}{n_m} \log|\breve{\phi}_k^{(n_m)}(Y_k^{n_m})| - \delta' \ , \quad \text{for } k = 1, \ldots, K \ , \tag{C.4a}$$

$$E \leq \frac{1}{n_m} I(\{\breve{\phi}_k^{(n_m)}(Y_k^{n_m})\}_{k\in\mathcal{K}}; X^{n_m}|Y_0^{n_m}) + \delta' \ . \tag{C.4b}$$

The second inequality of (C.4) implies that

$$H(X^{n_m}|\{\breve{\phi}_k^{(n_m)}(Y_k^{n_m})\}_{k\in\mathcal{K}}, Y_0^{n_m}) \leq n_m(H(X|Y_0) - E) + n_m\delta' \ . \tag{C.5}$$

Now, consider the $K$-encoder CEO source coding problem of Figure 3.1; and let the encoding function $\phi_k^{(n_m)}$ at Encoder $k \in \mathcal{K}$ be such that $\phi_k^{(n_m)} := \breve{\phi}_k^{(n_m)}$. Also, let the decoding function at the decoder be

$$\psi^{(n_m)} : \{1, \ldots, M_1^{(n_m)}\} \times \ldots \times \{1, \ldots, M_K^{(n_m)}\} \times \mathcal{Y}_0^{n_m} \longrightarrow \mathcal{X}^{n_m} \tag{C.6}$$

$$(m_1, \ldots, m_K, y_0^{n_m}) \longrightarrow p(x^{n_m}|m_1, \ldots, m_K, y_0^{n_m}) \; . \tag{C.7}$$

With such a choice, the achieved average logarithmic loss distortion is

$$\mathbb{E}[d^{(n_m)}(X^{n_m}, \psi^{(n_m)}(\{\phi_k^{(n_m)}(Y_k^{n_m})\}_{k\in\mathcal{K}}, Y_0^{n_m}))] = \frac{1}{n_m} H(X^{n_m}|\{\phi_k^{(n_m)}(Y_k^{n_m})\}_{k\in\mathcal{K}}, Y_0^{n_m}) \; . \tag{C.8}$$

Combined with (C.5), the last equality implies that

$$\mathbb{E}[d^{(n_m)}(X^{n_m}, \psi^{(n_m)}(\{\phi_k^{(n_m)}(Y_k^{(n_m)})\}_{k\in\mathcal{K}}, Y_0^{n_m}))] \le n_m(H(X|Y_0) - E) + \delta' \; . \tag{C.9}$$

Finally, substituting $\breve{\phi}_k^{(n_m)}$ with $\phi_k^{(n_m)}$ in (C.4), and observing that $\delta'$ can be chosen arbitrarily small in the obtained set of inequalities as well as in (C.9), it follows that $(R_1, \ldots, R_K, H(X|Y_0) - E) \in \mathcal{RD}^{\star}_{\mathrm{CEO}}$.

We now show the reverse implication. Let a non-negative tuple $(R_1, \ldots, R_K, H(X|Y_0) - E) \in \mathcal{RD}^{\star}_{\mathrm{CEO}}$ be given. Then, there exist encoding functions $\{\phi^{(n)}\}_{k\in\mathcal{K}}$ and a decoding function $\psi^{(n)}$ such that

$$R_k \ge \frac{1}{n} \log |\phi_k^{(n)}(Y_k^n)| \; , \quad \text{for } k = 1, \ldots, K \; , \tag{C.10a}$$

$$H(X|Y_0) - E \ge \mathbb{E}[d^{(n)}(X^n, \psi^{(n)}(\{\phi_k^{(n)}(X_k^{(n)})\}_{k\in\mathcal{K}}, Y_0^n))] \; . \tag{C.10b}$$

Using Lemma 8 (see the proof of converse of Theorem 1 in Appendix A), the RHS of the second inequality of (C.10) can be lower-bounded as

$$\mathbb{E}[d^{(n)}(X^n, \psi^{(n)}(\{\phi_k^{(n)}(X_k^{(n)})\}_{k\in\mathcal{K}}, Y_0^n))] \ge \frac{1}{n} H(X^n|\{\phi_k^{(n)}(X_k^{(n)})\}_{k\in\mathcal{K}}, Y_0^n) \; . \tag{C.11}$$

Combining the second inequality of (C.10) and (C.11), we get

$$H(X^n|\psi^{(n)}(\{\phi_k^{(n)}(X_k^{(n)})\}_{k\in\mathcal{K}}, Y_0^n)) \le n(H(X|Y_0) - E) \; , \tag{C.12}$$

from which it holds that

$$I(\{\phi_k^{(n)}(X_k^{(n)})\}_{k\in\mathcal{K}}; X^n|Y_0^n) = nH(X|Y_0) - H(X^n|\psi^{(n)}(\{\phi_k^{(n)}(X_k^{(n)})\}_{k\in\mathcal{K}}, Y_0^n)) \quad \text{(C.13a)}$$

$$\geq nE \ , \quad \text{(C.13b)}$$

where the equality follows since $(X^n, Y_0^n)$ is memoryless and the inequality follows by using (C.12).

Now, using the first inequality of (C.10) and (C.13), it follows that $(R_1, \ldots, R_K, E) \in \mathcal{R}^\star\Big(n, \{\phi_k^{(n)}\}_{k\in\mathcal{K}}\Big)$. Finally, using Proposition 2, it follows that $(R_1, \ldots, R_K, E) \in \mathcal{R}_{\mathrm{HT}}$; and this concludes the proof of the reverse part and the proposition.

# Appendix D

# Proof of Proposition 4

First let us define the rate-information region $\mathcal{RI}^\star_{\mathrm{CEO}}$ for discrete memoryless vector sources as the closure of all rate-information tuples $(R_1, \ldots, R_K, \Delta)$ for which there exist a blocklength $n$, encoding functions $\{\phi_k^{(n)}\}_{k=1}^K$ and a decoding function $\psi^{(n)}$ such that

$$\begin{aligned} R_k &\geq \frac{1}{n} \log M_k^{(n)} \ , \quad \text{for } \ k = 1, \ldots, K \ , \\ \Delta &\leq \frac{1}{n} I(\mathbf{X}^n; \psi^{(n)}(\phi_1^{(n)}(\mathbf{Y}_1^n), \ldots, \phi_K^{(n)}(\mathbf{Y}_K^n), \mathbf{Y}_0^n)) \ . \end{aligned}$$

It is easy to see that a characterization of $\mathcal{RI}^\star_{\mathrm{CEO}}$ can be obtained by using Theorem 1 and substituting distortion levels $D$ therein with $\Delta := H(\mathbf{X}) - D$. More specifically, the region $\mathcal{RI}^\star_{\mathrm{CEO}}$ is given as in the following theorem.

**Proposition 13.** *The rate-information region $\mathcal{RI}^\star_{\mathrm{CEO}}$ of the vector DM CEO problem under logarithmic loss is given by the set of all non-negative tuples $(R_1, \ldots, R_K, \Delta)$ that satisfy, for all subsets $\mathcal{S} \subseteq \mathcal{K}$,*

$$\sum_{k \in \mathcal{S}} R_k \geq \sum_{k \in \mathcal{S}} I(\mathbf{Y}_k; U_k | \mathbf{X}, \mathbf{Y}_0, Q) - I(\mathbf{X}; U_{\mathcal{S}^c}, \mathbf{Y}_0, Q) + \Delta \ ,$$

*for some joint measure of the form $P_{\mathbf{Y}_0, \mathbf{Y}_\mathcal{K}, \mathbf{X}}(\mathbf{y}_0, \mathbf{y}_\mathcal{K}, \mathbf{x}) P_Q(q) \prod_{k=1}^K P_{U_k|\mathbf{Y}_k, Q}(u_k|y_k, q)$.* ■

The region $\mathcal{RI}^\star_{\mathrm{CEO}}$ involves mutual information terms only (not entropies); and, so, using a standard discretization argument, it can be easily shown that a characterization of this region in the case of continuous alphabets is also given by Proposition 13.

Let us now return to the vector Gaussian CEO problem under logarithmic loss that we study in this section. First, we state the following lemma, whose proof is easy and is omitted for brevity.

**Lemma 12.** *$(R_1, \ldots, R_K, D) \in \mathcal{RD}^{\star}_{\text{VG-CEO}}$ if and only if $(R_1, \ldots, R_K, h(\mathbf{X}) - D) \in \mathcal{RI}^{\star}_{\text{CEO}}$.* ■

For vector Gaussian sources, the region $\mathcal{RD}^{\star}_{\text{VG-CEO}}$ can be characterized using Proposition 13 and Lemma 12. This completes the proof of first equality $\mathcal{RD}^{\star}_{\text{VG-CEO}} = \widetilde{\mathcal{RD}}^{\text{I}}_{\text{CEO}}$.

To complete the proof of Proposition 4, we need to show that two regions are equivalent, i.e., $\widetilde{\mathcal{RD}}^{\text{I}}_{\text{CEO}} = \widetilde{\mathcal{RD}}^{\text{II}}_{\text{CEO}}$. To do that, it is sufficient to show that, for fixed conditional distributions $\{p(u_k|\mathbf{y}_k, q)\}_{k=1}^{K}$, the extreme points of the polytope $\mathcal{P}_D$ defined by (4.5) are *dominated* by points that are in $\widetilde{\mathcal{RD}}^{\text{II}}_{\text{CEO}}$ that achieves distortion at most $D$. This is shown in the proof of Proposition 5 in Appendix F.

# Appendix E

# Proof of Converse of Theorem 4

The proof of the converse of Theorem 4 relies on deriving an outer bound on the region $\widetilde{\mathcal{RD}}_{\mathrm{CEO}}^{\mathrm{I}}$ given by Proposition 4. In doing so, we use the technique of [11, Theorem 8] which relies on the de Bruijn identity and the properties of Fisher information; and extend the argument to account for the time-sharing variable $Q$ and side information $\mathbf{Y}_0$.

We first state the following lemma.

**Lemma 13.** *[11, 147] Let $(\mathbf{X}, \mathbf{Y})$ be a pair of random vectors with pmf $p(\mathbf{x}, \mathbf{y})$. We have*

$$\log|(\pi e)\mathbf{J}^{-1}(\mathbf{X}|\mathbf{Y})| \leq h(\mathbf{X}|\mathbf{Y}) \leq \log|(\pi e)\mathrm{mmse}(\mathbf{X}|\mathbf{Y})| \ ,$$

*where the conditional Fisher information matrix is defined as*

$$\mathbf{J}(\mathbf{X}|\mathbf{Y}) := \mathbb{E}[\nabla \log p(\mathbf{X}|\mathbf{Y}) \nabla \log p(\mathbf{X}|\mathbf{Y})^{\dagger}] \ ,$$

*and the minimum mean squared error (MMSE) matrix is*

$$\mathrm{mmse}(\mathbf{X}|\mathbf{Y}) := \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}|\mathbf{Y}])(\mathbf{X} - \mathbb{E}[\mathbf{X}|\mathbf{Y}])^{\dagger}] \ . \qquad \blacksquare$$

Now, we derive an outer bound on (4.5) as follows. For each $q \in \mathcal{Q}$ and fixed pmf $\prod_{k=1}^{K} p(u_k|\mathbf{y}_k, q)$, choose $\{\mathbf{\Omega}_{k,q}\}_{k=1}^{K}$ satisfying $\mathbf{0} \preceq \mathbf{\Omega}_{k,q} \preceq \mathbf{\Sigma}_k^{-1}$ such that

$$\mathrm{mmse}(\mathbf{Y}_k|\mathbf{X}, U_{k,q}, \mathbf{Y}_0, q) = \mathbf{\Sigma}_k - \mathbf{\Sigma}_k \mathbf{\Omega}_{k,q} \mathbf{\Sigma}_k \ . \tag{E.1}$$

Such $\mathbf{\Omega}_{k,q}$ always exists since, for all $q \in \mathcal{Q}$, $k \in \mathcal{K}$, we have

$$\mathbf{0} \preceq \mathrm{mmse}(\mathbf{Y}_k|\mathbf{X}, U_{k,q}, \mathbf{Y}_0, q) \preceq \mathbf{\Sigma}_{\mathbf{y}_k|(\mathbf{x},\mathbf{y}_0)} = \mathbf{\Sigma}_{\mathbf{n}_k|\mathbf{n}_0} = \mathbf{\Sigma}_k \ .$$

Then, for $k \in \mathcal{K}$ and $q \in \mathcal{Q}$, we have

$$\begin{aligned} I(\mathbf{Y}_k; U_k|\mathbf{X}, \mathbf{Y}_0, Q = q) &= h(\mathbf{Y}_k|\mathbf{X}, \mathbf{Y}_0, Q = q) - h(\mathbf{Y}_k|\mathbf{X}, U_{k,q}, \mathbf{Y}_0, Q = q) \\ &\overset{(a)}{\geq} \log|(\pi e)\mathbf{\Sigma}_k| - \log|(\pi e)\mathrm{mmse}(\mathbf{Y}_k|\mathbf{X}, U_{k,q}, \mathbf{Y}_0, Q = q)| \\ &\overset{(b)}{=} -\log|\mathbf{I} - \mathbf{\Omega}_{k,q}\mathbf{\Sigma}_k| \ , \end{aligned} \tag{E.2}$$

where $(a)$ is due to Lemma 13; and $(b)$ is due to (E.1).

For convenience, the matrix $\mathbf{\Lambda}_{\bar{\mathcal{S}},q}$ is defined as follows

$$\mathbf{\Lambda}_{\bar{\mathcal{S}},q} := \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathrm{diag}(\{\mathbf{\Sigma}_k - \mathbf{\Sigma}_k\mathbf{\Omega}_{k,q}\mathbf{\Sigma}_k\}_{k\in\mathcal{S}^c}) \end{bmatrix} . \tag{E.3}$$

Then, for $q \in \mathcal{Q}$ and $\mathcal{S} \subseteq \mathcal{K}$, we have

$$\begin{aligned} h(\mathbf{X}|U_{\mathcal{S}^c,q}, \mathbf{Y}_0, Q = q) &\overset{(a)}{\geq} \log|(\pi e)\mathbf{J}^{-1}(\mathbf{X}|\mathbf{U}_{\mathcal{S}^c,q}, \mathbf{Y}_0, q)| \\ &\overset{(b)}{=} \log\left|(\pi e)\left(\mathbf{\Sigma}_{\mathbf{x}}^{-1} + \mathbf{H}_{\bar{\mathcal{S}}}^{\dagger}\mathbf{\Sigma}_{\mathbf{n}_{\bar{\mathcal{S}}}}^{-1}(\mathbf{I} - \mathbf{\Lambda}_{\bar{\mathcal{S}},q}\mathbf{\Sigma}_{\mathbf{n}_{\bar{\mathcal{S}}}}^{-1})\mathbf{H}_{\bar{\mathcal{S}}}\right)^{-1}\right| \ , \end{aligned} \tag{E.4}$$

where $(a)$ follows from Lemma 13; and for $(b)$, we use the connection of the MMSE and the Fisher information to show the following equality

$$\mathbf{J}(\mathbf{X}|U_{\mathcal{S}^c,q}, \mathbf{Y}_0, q) = \mathbf{\Sigma}_{\mathbf{x}}^{-1} + \mathbf{H}_{\bar{\mathcal{S}}}^{\dagger}\mathbf{\Sigma}_{\mathbf{n}_{\bar{\mathcal{S}}}}^{-1}(\mathbf{I} - \mathbf{\Lambda}_{\bar{\mathcal{S}},q}\mathbf{\Sigma}_{\mathbf{n}_{\bar{\mathcal{S}}}}^{-1})\mathbf{H}_{\bar{\mathcal{S}}} \ . \tag{E.5}$$

In order to proof (E.5), we use de Bruijn identity to relate the Fisher information with the MMSE as given in the following lemma.

**Lemma 14.** *[11, 148] Let $(\mathbf{V}_1, \mathbf{V}_2)$ be a random vector with finite second moments and $\mathbf{Z} \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Sigma}_{\mathbf{z}})$ independent of $(\mathbf{V}_1, \mathbf{V}_2)$. Then*

$$\mathrm{mmse}(\mathbf{V}_2|\mathbf{V}_1, \mathbf{V}_2 + \mathbf{Z}) = \mathbf{\Sigma}_{\mathbf{z}} - \mathbf{\Sigma}_{\mathbf{z}}\mathbf{J}(\mathbf{V}_2 + \mathbf{Z}|\mathbf{V}_1)\mathbf{\Sigma}_{\mathbf{z}} \ . \qquad \blacksquare$$

From MMSE estimation of Gaussian random vectors, for $\mathcal{S} \subseteq \mathcal{K}$, we have

$$\mathbf{X} = \mathbb{E}[\mathbf{X}|\mathbf{Y}_{\bar{\mathcal{S}}}] + \mathbf{W}_{\bar{\mathcal{S}}} = \mathbf{G}_{\bar{\mathcal{S}}}\mathbf{Y}_{\bar{\mathcal{S}}} + \mathbf{W}_{\bar{\mathcal{S}}} \ , \tag{E.6}$$

where $\mathbf{G}_{\bar{\mathcal{S}}} := \mathbf{\Sigma}_{\mathbf{w}_{\bar{\mathcal{S}}}}\mathbf{H}_{\bar{\mathcal{S}}}^{\dagger}\mathbf{\Sigma}_{\mathbf{n}_{\bar{\mathcal{S}}}}^{-1}$, and $\mathbf{W}_{\bar{\mathcal{S}}} \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Sigma}_{\mathbf{w}_{\bar{\mathcal{S}}}})$ is a Gaussian vector that is independent of $\mathbf{Y}_{\bar{\mathcal{S}}}$ and

$$\mathbf{\Sigma}_{\mathbf{w}_{\bar{\mathcal{S}}}}^{-1} := \mathbf{\Sigma}_{\mathbf{x}}^{-1} + \mathbf{H}_{\bar{\mathcal{S}}}^{\dagger}\mathbf{\Sigma}_{\mathbf{n}_{\bar{\mathcal{S}}}}^{-1}\mathbf{H}_{\bar{\mathcal{S}}} \ . \tag{E.7}$$

Now we show that the cross-terms of $\mathrm{mmse}\left(\mathbf{Y}_{\mathcal{S}^c}|\mathbf{X}, U_{\mathcal{S}^c,q}, \mathbf{Y}_0, q\right)$ are zero (similarly to [11, Appendix V]). For $i \in \mathcal{S}^c$ and $j \neq i$, we have

$$\begin{aligned}
&\mathbb{E}\big[(Y_i - \mathbb{E}[Y_i|\mathbf{X}, U_{\mathcal{S}^c,q}, \mathbf{Y}_0, q])(Y_j - \mathbb{E}[Y_j|\mathbf{X}, U_{\mathcal{S}^c,q}, \mathbf{Y}_0, q])^\dagger\big] \\
&\quad\overset{(a)}{=} \mathbb{E}\Big[\mathbb{E}\big[(Y_i - \mathbb{E}[Y_i|\mathbf{X}, U_{\mathcal{S}^c,q}, \mathbf{Y}_0, q])(Y_j - \mathbb{E}[Y_j|\mathbf{X}, U_{\mathcal{S}^c,q}, \mathbf{Y}_0, q])^\dagger|\mathbf{X}, \mathbf{Y}_0\big]\Big] \\
&\quad\overset{(b)}{=} \mathbb{E}\Big[\mathbb{E}\big[(Y_i - \mathbb{E}[Y_i|\mathbf{X}, U_{\mathcal{S}^c,q}, \mathbf{Y}_0, q])|\mathbf{X}, \mathbf{Y}_0\big]\mathbb{E}\big[(Y_j - \mathbb{E}[Y_j|\mathbf{X}, U_{\mathcal{S}^c,q}, \mathbf{Y}_0, q])^\dagger|\mathbf{X}, \mathbf{Y}_0\big]\Big] \\
&\quad= \mathbf{0} \;, \qquad\qquad (\text{E.8})
\end{aligned}$$

where $(a)$ is due to the law of total expectation; $(b)$ is due to the Markov chain $\mathbf{Y}_k -\!\!\circ\!\!- (\mathbf{X}, \mathbf{Y}_0) -\!\!\circ\!\!- \mathbf{Y}_{\mathcal{K}\setminus k}$.

Then, for $k \in \mathcal{K}$ and $q \in \mathcal{Q}$, we have

$$\begin{aligned}
&\mathrm{mmse}\big(\mathbf{G}_{\bar{\mathcal{S}}}\mathbf{Y}_{\bar{\mathcal{S}}}\big|\mathbf{X}, U_{\mathcal{S}^c,q}, \mathbf{Y}_0, q\big) \\
&\quad= \mathbf{G}_{\bar{\mathcal{S}}}\, \mathrm{mmse}\left(\mathbf{Y}_{\bar{\mathcal{S}}}|\mathbf{X}, U_{\mathcal{S}^c,q}, \mathbf{Y}_0, q\right) \mathbf{G}_{\bar{\mathcal{S}}}^\dagger \\
&\quad\overset{(a)}{=} \mathbf{G}_{\bar{\mathcal{S}}} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathrm{diag}(\{\mathrm{mmse}(\mathbf{Y}_k|\mathbf{X}, U_{\mathcal{S}^c,q}, \mathbf{Y}_0, q)\}_{k\in\mathcal{S}^c}) \end{bmatrix} \mathbf{G}_{\bar{\mathcal{S}}}^\dagger \\
&\quad\overset{(b)}{=} \mathbf{G}_{\bar{\mathcal{S}}}\mathbf{\Lambda}_{\bar{\mathcal{S}},q}\mathbf{G}_{\bar{\mathcal{S}}}^\dagger \;, \qquad\qquad (\text{E.9})
\end{aligned}$$

where $(a)$ follows since the cross-terms are zero as shown in (E.8); and $(b)$ follows due to (E.1) and the definition of $\mathbf{\Lambda}_{\bar{\mathcal{S}},q}$ given in (E.3).

Finally, we obtain the equality (E.5) by applying Lemma 14 and noting (E.6) as follows

$$\begin{aligned}
\mathbf{J}(\mathbf{X}|U_{\mathcal{S}^c,q}, \mathbf{Y}_0, q) &\overset{(a)}{=} \mathbf{\Sigma}_{\mathbf{w}_{\bar{\mathcal{S}}}}^{-1} - \mathbf{\Sigma}_{\mathbf{w}_{\bar{\mathcal{S}}}}^{-1}\, \mathrm{mmse}\big(\mathbf{G}_{\bar{\mathcal{S}}}\mathbf{Y}_{\bar{\mathcal{S}}}\big|\mathbf{X}, U_{\mathcal{S}^c,q}, \mathbf{Y}_0, q\big)\mathbf{\Sigma}_{\mathbf{w}_{\bar{\mathcal{S}}}}^{-1} \\
&\overset{(b)}{=} \mathbf{\Sigma}_{\mathbf{w}_{\bar{\mathcal{S}}}}^{-1} - \mathbf{\Sigma}_{\mathbf{w}_{\bar{\mathcal{S}}}}^{-1}\mathbf{G}_{\bar{\mathcal{S}}}\mathbf{\Lambda}_{\bar{\mathcal{S}},q}\mathbf{G}_{\bar{\mathcal{S}}}^\dagger\mathbf{\Sigma}_{\mathbf{w}_{\bar{\mathcal{S}}}}^{-1} \\
&\overset{(c)}{=} \mathbf{\Sigma}_{\mathbf{x}}^{-1} + \mathbf{H}_{\bar{\mathcal{S}}}^\dagger\mathbf{\Sigma}_{\mathbf{n}_{\bar{\mathcal{S}}}}^{-1}\mathbf{H}_{\bar{\mathcal{S}}} - \mathbf{H}_{\bar{\mathcal{S}}}^\dagger\mathbf{\Sigma}_{\mathbf{n}_{\bar{\mathcal{S}}}}^{-1}\mathbf{\Lambda}_{\bar{\mathcal{S}},q}\mathbf{\Sigma}_{\mathbf{n}_{\bar{\mathcal{S}}}}^{-1}\mathbf{H}_{\bar{\mathcal{S}}} \\
&= \mathbf{\Sigma}_{\mathbf{x}}^{-1} + \mathbf{H}_{\bar{\mathcal{S}}}^\dagger\mathbf{\Sigma}_{\mathbf{n}_{\bar{\mathcal{S}}}}^{-1}\big(\mathbf{I} - \mathbf{\Lambda}_{\bar{\mathcal{S}},q}\mathbf{\Sigma}_{\mathbf{n}_{\bar{\mathcal{S}}}}^{-1}\big)\mathbf{H}_{\bar{\mathcal{S}}} \;,
\end{aligned}$$

where $(a)$ is due to Lemma 14; $(b)$ is due to (E.9); and $(c)$ follows due to the definitions of $\mathbf{\Sigma}_{\mathbf{w}_{\bar{\mathcal{S}}}}^{-1}$ and $\mathbf{G}_{\bar{\mathcal{S}}}$.

Next, we average (E.2) and (E.4) over the time-sharing $Q$ and letting $\mathbf{\Omega}_k := \sum_{q\in\mathcal{Q}} p(q)\mathbf{\Omega}_{k,q}$, we obtain the lower bound

$$\begin{aligned} I(\mathbf{Y}_k;\mathbf{U}_k|\mathbf{X},\mathbf{Y}_0,Q) &= \sum_{q\in\mathcal{Q}} p(q) I(\mathbf{Y}_k;\mathbf{U}_k|\mathbf{X},\mathbf{Y}_0,Q=q) \\ &\overset{(a)}{\geq} -\sum_{q\in\mathcal{Q}} p(q)\log|\mathbf{I}-\mathbf{\Omega}_{k,q}\mathbf{\Sigma}_k| \\ &\overset{(b)}{\geq} -\log|\mathbf{I}-\sum_{q\in\mathcal{Q}} p(q)\mathbf{\Omega}_{k,q}\mathbf{\Sigma}_k| \\ &= -\log|\mathbf{I}-\mathbf{\Omega}_k\mathbf{\Sigma}_k| \,, \end{aligned} \tag{E.10}$$

where $(a)$ follows from (E.2); and $(b)$ follows from the concavity of the log-determinant function and Jensen's Inequality.

Besides, we can derive the following lower bound

$$\begin{aligned} h(\mathbf{X}|U_{\mathcal{S}^c},\mathbf{Y}_0,Q) &= \sum_{q\in\mathcal{Q}} p(q) h(\mathbf{X}|U_{\mathcal{S}^c,q},\mathbf{Y}_0,Q=q) \\ &\overset{(a)}{\geq} \sum_{q\in\mathcal{Q}} p(q)\log\left|(\pi e)\left(\mathbf{\Sigma}_{\mathbf{x}}^{-1}+\mathbf{H}_{\bar{\mathcal{S}}}^{\dagger}\mathbf{\Sigma}_{\mathbf{n}_{\bar{\mathcal{S}}}}^{-1}(\mathbf{I}-\mathbf{\Lambda}_{\bar{\mathcal{S}},q}\mathbf{\Sigma}_{\mathbf{n}_{\bar{\mathcal{S}}}}^{-1})\mathbf{H}_{\bar{\mathcal{S}}}\right)^{-1}\right| \\ &\overset{(b)}{\geq} \log\left|(\pi e)\left(\mathbf{\Sigma}_{\mathbf{x}}^{-1}+\mathbf{H}_{\bar{\mathcal{S}}}^{\dagger}\mathbf{\Sigma}_{\mathbf{n}_{\bar{\mathcal{S}}}}^{-1}(\mathbf{I}-\mathbf{\Lambda}_{\bar{\mathcal{S}}}\mathbf{\Sigma}_{\mathbf{n}_{\bar{\mathcal{S}}}}^{-1})\mathbf{H}_{\bar{\mathcal{S}}}\right)^{-1}\right| \,, \end{aligned} \tag{E.11}$$

where $(a)$ is due to (E.4); and $(b)$ is due to the concavity of the log-determinant function and Jensen's inequality and the definition of $\mathbf{\Lambda}_{\bar{\mathcal{S}}}$ given in (4.7).

Finally, the outer bound on $\widetilde{\mathcal{RD}}_{\mathrm{CEO}}^{\mathrm{I}}$ is obtained by applying (E.10) and (E.11) in (4.5), noting that $\mathbf{\Omega}_k = \sum_{q\in\mathcal{Q}} p(q)\mathbf{\Omega}_{k,q} \preceq \mathbf{\Sigma}_k^{-1}$ since $\mathbf{0} \preceq \mathbf{\Omega}_{k,q} \preceq \mathbf{\Sigma}_k^{-1}$, and taking the union over $\mathbf{\Omega}_k$ satisfying $\mathbf{0} \preceq \mathbf{\Omega}_k \preceq \mathbf{\Sigma}_k^{-1}$.

# Appendix F

# Proof of Proposition 5 (Extension to $K$ Encoders)

For the proof of Proposition 5, it is sufficient to show that, for fixed Gaussian distributions $\{p(u_k|\mathbf{y}_k)\}_{k=1}^K$, the extreme points of the polytope $\mathcal{P}_D$ defined by (4.5) are *dominated* by points that are in $\widetilde{\mathcal{RD}}_{\text{CEO}}^{\text{II}}$ and which are achievable using Gaussian conditional distributions $\{p(v_k|\mathbf{y}_k, q')\}_{k=1}^K$. The proof is similar to [10, Appendix C, Lemma 6].

First, we characterize the extreme points of $\mathcal{P}_D$. Let the function $f : 2^K \to \mathbb{R}$ be such that for all $\mathcal{S} \subseteq \mathcal{K}$,

$$f(\mathcal{S}) = I(\mathbf{Y}_\mathcal{S}; U_\mathcal{S}|U_{\mathcal{S}^c}, \mathbf{Y}_0, Q) + h(\mathbf{X}|U_1, \ldots, U_K, \mathbf{Y}_0, Q) - D\ . \tag{F.1}$$

It is easy to see that $f(\cdot)$ and the function $S \to [f(\mathcal{S})]^+ := \max\{f(\mathcal{S}), 0\}$ are supermodular functions. Also, for all subsets $\mathcal{S} \subseteq \mathcal{K}$, we have

$$\begin{aligned}
f(\mathcal{S}) &= I(\mathbf{Y}_\mathcal{S}; U_\mathcal{S}|U_{\mathcal{S}^c}, \mathbf{Y}_0, Q) + h(\mathbf{X}|U_1, \ldots, U_K, \mathbf{Y}_0, Q) - D \\
&\overset{(a)}{=} I(\mathbf{Y}_\mathcal{S}, \mathbf{X}; U_\mathcal{S}|U_{\mathcal{S}^c}, \mathbf{Y}_0, Q) + h(\mathbf{X}|U_1, \ldots, U_K, \mathbf{Y}_0, Q) - D \\
&= I(\mathbf{Y}_\mathcal{S}; U_\mathcal{S}|\mathbf{X}, U_{\mathcal{S}^c}, \mathbf{Y}_0, Q) + I(\mathbf{X}; U_\mathcal{S}|U_{\mathcal{S}^c}, \mathbf{Y}_0, Q) + h(\mathbf{X}|U_1, \ldots, U_K, \mathbf{Y}_0, Q) - D \\
&= I(\mathbf{Y}_\mathcal{S}; U_\mathcal{S}|\mathbf{X}, U_{\mathcal{S}^c}, \mathbf{Y}_0, Q) + h(\mathbf{X}|U_{\mathcal{S}^c}, \mathbf{Y}_0, Q) - h(\mathbf{X}|U_\mathcal{S}, U_{\mathcal{S}^c}, \mathbf{Y}_0, Q) \\
&\quad + h(\mathbf{X}|U_1, \ldots, U_K, \mathbf{Y}_0, Q) - D \\
&\overset{(b)}{=} \sum_{k\in\mathcal{S}} I(\mathbf{Y}_k; U_k|\mathbf{X}, \mathbf{Y}_0, Q) + h(\mathbf{X}|U_{\mathcal{S}^c}, \mathbf{Y}_0, Q) - D\ ,
\end{aligned} \tag{F.2}$$

where $(a)$ follows using the Markov chain $U_\mathcal{S} -\!\!\circ\!\!- \mathbf{Y}_\mathcal{S} -\!\!\circ\!\!- \mathbf{X}$; and $(b)$ follows by using the chain rule and the Markov chain $(U_k, \mathbf{Y}_k) -\!\!\circ\!\!- (\mathbf{X}, \mathbf{Y}_0, Q) -\!\!\circ\!\!- (U_{\mathcal{K}\setminus k}, \mathbf{Y}_{\mathcal{K}\setminus k})$. Then, by construction, we have that $\mathcal{P}_D$ is given by the set of $(R_1, \ldots, R_K)$ that satisfy for all subsets $\mathcal{S} \subseteq \mathcal{K}$,

$$\sum_{k \in \mathcal{S}} R_k \geq [f(\mathcal{S})]^+ .$$

Proceeding along the lines of [103, Appendix B], we have that for a linear ordering $i_1 \prec i_2 \prec \cdots \prec i_K$ on the set $\mathcal{K}$, an extreme point of $\mathcal{P}_D$ can be computed as follows

$$\tilde{R}_{i_k} = [f(\{i_1, i_2, \ldots, i_k\})]^+ - [f(\{i_1, i_2, \ldots, i_{k-1}\})]^+ , \quad \text{for } k = 1, \ldots, K .$$

All the $K!$ extreme points of $\mathcal{P}_D$ can be enumerated by looking over all linear orderings $i_1 \prec i_2 \prec \cdots \prec i_K$ of $\mathcal{K}$. Each ordering of $\mathcal{K}$ is analyzed in the same manner and, therefore, for notational simplicity, the only ordering we consider is the natural ordering, i.e., $i_k = k$, in the rest of the proof. Then, by construction, we have

$$\begin{aligned}\tilde{R}_k = &\Big[\sum_{i=1}^{k} I(\mathbf{Y}_i; U_i|\mathbf{X}, \mathbf{Y}_0, Q) + h(\mathbf{X}|U_{k+1}^K, \mathbf{Y}_0, Q) - D\Big]^+ \\ &- \Big[\sum_{i=1}^{k-1} I(\mathbf{Y}_i; U_i|\mathbf{X}, \mathbf{Y}_0, Q) + h(\mathbf{X}|U_{k}^K, \mathbf{Y}_0, Q) - D\Big]^+ .\end{aligned} \tag{F.3}$$

Let $j$ be the first index for which $f(\{1, 2, \ldots, j\}) > 0$. Then it follows from (F.3) that

$$\begin{aligned}\tilde{R}_j &= \sum_{k=1}^{j} I(\mathbf{Y}_k; U_k|\mathbf{X}, \mathbf{Y}_0, Q) + h(\mathbf{X}|U_{j+1}^K, \mathbf{Y}_0, Q) - D \\ &= I(\mathbf{Y}_j; U_j|\mathbf{X}, \mathbf{Y}_0, Q) + \sum_{k=1}^{j-1} I(\mathbf{X}_k; U_k|\mathbf{X}, \mathbf{Y}_0, Q) + h(\mathbf{X}|U_{j+1}^K, \mathbf{Y}_0, Q) - D \\ &\quad + h(\mathbf{X}|U_j^K, \mathbf{Y}_0, Q) - h(\mathbf{X}|U_j, U_{j+1}^K, \mathbf{Y}_0, Q) \\ &\overset{(a)}{=} f(\{1, 2, \ldots, j-1\}) + I(\mathbf{Y}_j; U_j|\mathbf{X}, U_{j+1}^K, \mathbf{Y}_0, Q) + I(\mathbf{X}; U_j|U_{j+1}^K, \mathbf{Y}_0, Q) \\ &= f(\{1, 2, \ldots, j-1\}) + I(\mathbf{Y}_j, \mathbf{X}; U_j|U_{j+1}^K, \mathbf{Y}_0, Q) \\ &\overset{(b)}{=} f(\{1, 2, \ldots, j-1\}) + I(\mathbf{Y}_j; U_j|U_{j+1}^K, \mathbf{Y}_0, Q) \\ &= (1-\theta) I(\mathbf{Y}_j; U_j|U_{j+1}^K, \mathbf{Y}_0, Q) ,\end{aligned}$$

where $(a)$ follows due to the Markov chain $U_j -\!\!\circ\!\!- \mathbf{Y}_j -\!\!\circ\!\!- \mathbf{X} -\!\!\circ\!\!- U_{\mathcal{K}\setminus j}$ and (F.2); $(b)$ follows due to the Markov chain $U_j -\!\!\circ\!\!- \mathbf{Y}_j -\!\!\circ\!\!- \mathbf{X}$; and $\theta \in (0, 1]$ is defined as

$$\theta := \frac{-f(\{1, 2, \ldots, j-1\})}{I(\mathbf{Y}_j; U_j|U_{j+1}^K, \mathbf{Y}_0, Q)} = \frac{D - h(\mathbf{X}|U_{\mathcal{K}}, \mathbf{Y}_0, Q) - I(\mathbf{Y}_1^{j-1}; U_1^{j-1}|U_j^K, \mathbf{Y}_0, Q)}{I(\mathbf{Y}_j; U_j|U_{j+1}^K, \mathbf{Y}_0, Q)} . \tag{F.4}$$

Furthermore, for all indices $k > j$, we have

$$\begin{aligned}
\tilde{R}_k &= f(\{1,2,\dots,k\}) - f(\{1,2,\dots,k-1\}) \\
&= I(\mathbf{Y}_k; U_k|\mathbf{X}, \mathbf{Y}_0, Q) + I(\mathbf{X}; U_k|U_{k+1}^K, \mathbf{Y}_0, Q) \\
&\overset{(a)}{=} I(\mathbf{Y}_k; U_k|\mathbf{X}, U_{k+1}^K, \mathbf{Y}_0, Q) + I(\mathbf{X}; U_k|U_{k+1}^K, \mathbf{Y}_0, Q) \\
&= I(\mathbf{Y}_k, \mathbf{X}; U_k|U_{k+1}^K, \mathbf{Y}_0, Q) \\
&\overset{(b)}{=} I(\mathbf{Y}_k; U_k|U_{k+1}^K, \mathbf{Y}_0, Q) \;,
\end{aligned}$$

where $(a)$ follows due to the Markov chain $U_k -\!\!\circ\!\!- \mathbf{Y}_k -\!\!\circ\!\!- \mathbf{X} -\!\!\circ\!\!- U_{\mathcal{K}\setminus k}$; and $(b)$ follows due to the Markov chain $U_k -\!\!\circ\!\!- \mathbf{Y}_k -\!\!\circ\!\!- \mathbf{X}$.

Therefore, for the natural ordering, the extreme point $(\tilde{R}_1, \dots, \tilde{R}_K)$ is given as

$$\begin{aligned}
(\tilde{R}_1, \dots, \tilde{R}_K) = \Big(&0, \dots, 0, (1-\theta) I(\mathbf{Y}_j; U_j|U_{j+1}^K, \mathbf{Y}_0, Q), I(\mathbf{Y}_{j+1}; U_{j+1}|U_{j+2}^K, \mathbf{Y}_0, Q), \\
&\dots, I(\mathbf{Y}_K; U_K|\mathbf{Y}_0, Q)\Big) \;.
\end{aligned}$$

Next, we show that $(\tilde{R}_1, \dots, \tilde{R}_K) \in \mathcal{P}_D$ is dominated by a point $(R_1, \dots, R_K, \bar{D}) \in \widetilde{\mathcal{RD}}_{\mathrm{CEO}}^{\mathrm{II}}$ that achieves a distortion $\bar{D} \leq D$.

We consider an instance of the CEO setup in which for a fraction $\theta \in (0,1]$ of the time the decoder recovers $U_{j+1}^n, \dots, U_K^n$ while encoders $k = 1, \dots, j$ are inactive; and for the remaining fraction $(1-\theta)$ of the time the decoder recovers $U_j^n, \dots, U_K^n$ while encoders $k = 1, \dots, j-1$ are inactive. Then, the source $X$ is decoded. Formally, we consider a pmf $p(q') \prod_{k=1}^K p(v_k|\mathbf{y}_k, q')$ for the CEO setup as follows. Let $B$ denote a Bernoulli random variable with parameter $\theta$, i.e., $B = 1$ with probability $\theta$ and $B = 0$ with probability $1-\theta$. We let $\theta$ as in (F.4) and $Q' := (B, Q)$. Then, let the tuple of random variables be distributed as follows

$$(Q', V_{\mathcal{K}}) = \begin{cases} \big((1,Q), \emptyset, \dots, \emptyset, U_{j+1}, \dots, U_K\big) \;, & \text{if } B = 1 \;, \\ \big((0,Q), \emptyset, \dots, \emptyset, U_j, \dots, U_K\big) \;, & \text{if } B = 0 \;. \end{cases} \tag{F.5}$$

Using Definition 6, we have $(R_1, \dots, R_K, \bar{D}) \in \widetilde{\mathcal{RD}}_{\mathrm{CEO}}^{\mathrm{II}}$, where

$$\begin{aligned}
R_k &= I(\mathbf{Y}_k; V_k|V_{k+1}, \dots, V_K, \mathbf{Y}_0, Q') \;, \quad \text{for } k = 1, \dots, K \;, \\
\bar{D} &= h(\mathbf{X}|V_1, \dots, V_K, \mathbf{Y}_0, Q') \;.
\end{aligned}$$

Then, for $k = 1, \dots, j-1$, we have

$$R_k = I(\mathbf{Y}_k; V_k | V_{k+1}, \dots, V_K, \mathbf{Y}_0, Q') \overset{(a)}{=} 0 = \tilde{R}_k \ , \tag{F.6}$$

where $(a)$ follows since $V_k = \emptyset$ for $k < j$ independently of $B$.

For $k = j$, we have

$$\begin{aligned} R_j &= I(\mathbf{Y}_j; V_j | V_{j+1}, \dots, V_K, \mathbf{Y}_0, Q') \\ &= \theta I(\mathbf{Y}_j; U_j | U_{j+1}, \dots, U_K, \mathbf{Y}_0, Q, B=1) \\ &\quad + (1-\theta) I(\mathbf{Y}_j; U_j | U_{j+1}, \dots, U_K, \mathbf{Y}_0, Q, B=0) \\ &\overset{(a)}{=} (1-\theta) I(\mathbf{Y}_j; U_j | U_{j+1}, \dots, U_K, \mathbf{Y}_0, Q) = \tilde{R}_j \ , \end{aligned} \tag{F.7}$$

where $(a)$ follows since $V_j = \emptyset$ for $B = 0$ and $V_j = U_j$ for $B = 1$.

For $k = j+1, \dots, K$, we have

$$\begin{aligned} R_k &= I(\mathbf{Y}_k; V_k | V_{k+1}, \dots, V_K, \mathbf{Y}_0, Q') \\ &= \theta I(\mathbf{Y}_j; U_j | U_{j+1}, \dots, U_K, \mathbf{Y}_0, Q, B=1) \\ &\quad + (1-\theta) I(\mathbf{Y}_j; U_j | U_{j+1}, \dots, U_K, \mathbf{Y}_0, Q, B=0) \\ &\overset{(a)}{=} I(\mathbf{Y}_j; U_j | U_{j+1}, \dots, U_K, \mathbf{Y}_0, Q) = \tilde{R}_k \ , \end{aligned} \tag{F.8}$$

where $(a)$ is due to $V_j = U_j$ for $k > j$ independently of $B$.

Besides, the distortion $\bar{D}$ satisfies

$$\begin{aligned} \bar{D} &= h(\mathbf{X} | V_1, \dots, V_K, \mathbf{Y}_0, Q') \\ &= \theta h(\mathbf{X} | U_{j+1}, \dots, U_K, \mathbf{Y}_0, Q, B=1) + (1-\theta) h(\mathbf{X} | U_j, \dots, U_K, \mathbf{Y}_0, Q, B=0) \\ &= h(\mathbf{X} | U_j^K, \mathbf{Y}_0, Q) + \theta I(\mathbf{X}; U_j | U_{j+1}^K, \mathbf{Y}_0, Q) \\ &\overset{(a)}{=} h(\mathbf{X} | U_j^K, \mathbf{Y}_0, Q) \\ &\quad + \frac{D - h(\mathbf{X} | U_{\mathcal{K}}, \mathbf{Y}_0, Q) - I(\mathbf{Y}_1^{j-1}; U_1^{j-1} | U_j^K, \mathbf{Y}_0, Q)}{I(\mathbf{Y}_j, \mathbf{X}; U_j | U_{j+1}^K, \mathbf{Y}_0, Q)} I(\mathbf{X}; U_j | U_{j+1}^K, \mathbf{Y}_0, Q) \\ &= h(\mathbf{X} | U_j^K, \mathbf{Y}_0, Q) \\ &\quad + \frac{D - h(\mathbf{X} | U_{\mathcal{K}}, \mathbf{Y}_0, Q) - I(\mathbf{Y}_1^{j-1}; U_1^{j-1} | U_j^K, \mathbf{Y}_0, Q)}{I(\mathbf{X}; U_j | U_{j+1}^K, \mathbf{Y}_0, Q) + I(\mathbf{Y}_j; U_j | \mathbf{X}, U_{j+1}^K, \mathbf{Y}_0, Q)} I(\mathbf{X}; U_j | U_{j+1}^K, \mathbf{Y}_0, Q) \\ &\leq D + h(\mathbf{X} | U_j^K, \mathbf{Y}_0, Q) - h(\mathbf{X} | U_{\mathcal{K}}, \mathbf{Y}_0, Q) - I(\mathbf{Y}_1^{j-1}; U_1^{j-1} | U_j^K, \mathbf{Y}_0, Q) \end{aligned}$$

$$
\begin{aligned}
&= D + I(\mathbf{X}; U_1^{j-1} | U_j^K, \mathbf{Y}_0, Q) - I(\mathbf{Y}_1^{j-1}; U_1^{j-1} | U_j^K, \mathbf{Y}_0, Q) \\
&\stackrel{(b)}{=} D + I(\mathbf{X}; U_1^{j-1} | U_j^K, \mathbf{Y}_0, Q) - I(\mathbf{Y}_1^{j-1}, \mathbf{X}; U_1^{j-1} | U_j^K, \mathbf{Y}_0, Q) \\
&= D - I(\mathbf{Y}_1^{j-1}; U_1^{j-1} | \mathbf{X}, U_j^K, \mathbf{Y}_0, Q) \leq D \,, 
\end{aligned} \tag{F.9}
$$

where $(a)$ follows from (F.4) and due to the Markov chain $U_j -\!\!\circ\!\!- \mathbf{Y}_j -\!\!\circ\!\!- \mathbf{X}$; and $(b)$ follows due to the Markov chain $U_{\mathcal{S}} -\!\!\circ\!\!- \mathbf{Y}_{\mathcal{S}} -\!\!\circ\!\!- \mathbf{X}$ for all subsets $\mathcal{S} \subseteq \mathcal{K}$.

Summarizing, using (F.6), (F.7), (F.8) and (F.9), it follows that the extreme point $(\tilde{R}_1, \tilde{R}_2, \ldots, \tilde{R}_K) \in \mathcal{P}_D$ is dominated by the point $(R_1, \ldots, R_K, D) \in \widetilde{\mathcal{RD}}_{\mathrm{CEO}}^{\mathrm{II}}$ satisfying $\bar{D} \leq D$. Similarly, by considering all possible orderings each extreme point of $\mathcal{P}_D$ can be shown to be dominated by a point which lies in $\widetilde{\mathcal{RD}}_{\mathrm{CEO}}^{\mathrm{II}}$. The proof is terminated by observing that, for all extreme points, $V_k$ is set either equal $U_k^{\mathrm{G}}$ (which is Gaussian distributed conditionally on $\mathbf{Y}_k$) or a constant.

# Appendix G

# Proof of Theorem 5

We first present the following lemma, which essentially states that Theorem 4 provides an outer bound on $\mathcal{RD}^{\text{det}}_{\text{VG-CEO}}$.

**Lemma 15.** *If* $(R_1, \dots, R_K, D) \in \mathcal{RD}^{\text{det}}_{\text{VG-CEO}}$*, then* $(R_1, \dots, R_K, \log(\pi e)^{n_x} D) \in \widetilde{\mathcal{RD}}^{\text{I}}_{\text{CEO}}$.

*Proof.* Let a tuple $(R_1, \dots, R_K, D) \in \mathcal{RD}^{\text{det}}_{\text{VG-CEO}}$ be given. Then, there exist a blocklength $n$, $K$ encoding functions $\{\breve{\phi}^{(n)}_k\}_{k=1}^K$ and a decoding function $\breve{\psi}^{(n)}$ such that

$$
\begin{aligned}
R_k &\geq \frac{1}{n} \log M_k^{(n)} \;, \quad \text{for } \; k = 1, \dots, K \;, \\
D &\geq \left| \frac{1}{n} \sum_{i=1}^n \text{mmse}(\mathbf{X}_i | \breve{\phi}^{(n)}_1(\mathbf{Y}^n_1), \dots, \breve{\phi}^{(n)}_K(\mathbf{Y}^n_K), \mathbf{Y}^n_0) \right| \;. 
\end{aligned} \tag{G.1}
$$

We need to show that there exist $(U_1, \dots, U_K, Q)$ such that

$$
\sum_{k \in \mathcal{S}} R_k + \log(\pi e)^{n_x} D \geq \sum_{k \in \mathcal{S}} I(\mathbf{Y}_k; U_k | \mathbf{X}, \mathbf{Y}_0, Q) + h(\mathbf{X} | U_{\mathcal{S}^c}, \mathbf{Y}_0, Q) \;, \quad \text{for } \mathcal{S} \subseteq \mathcal{K} \;. \tag{G.2}
$$

Let us define

$$
\bar{\Delta}^{(n)} := \frac{1}{n} h(\mathbf{X}^n | \breve{\phi}^{(n)}_1(\mathbf{Y}^n_1), \dots, \breve{\phi}^{(n)}_K(\mathbf{Y}^n_K), \mathbf{Y}^n_0) \;.
$$

It is easy to justify that expected distortion $\bar{\Delta}^{(n)}$ is achievable under logarithmic loss (see Proposition 4). Then, following straightforwardly the lines in the proof of Theorem 1 (see (A.6)), we have

$$
\sum_{k \in \mathcal{S}} R_k \geq \sum_{k \in \mathcal{S}} \frac{1}{n} \sum_{i=1}^n I(\mathbf{Y}_{k,i}; U_{k,i} | \mathbf{X}_i, \mathbf{Y}_{0,i}, Q_i) + \frac{1}{n} \sum_{i=1}^n h(\mathbf{X}_i | U_{\mathcal{S}^c,i}, \mathbf{Y}_{0,i}, Q_i) - \bar{\Delta}^{(n)} \;. \tag{G.3}
$$

Next, we upper bound $\bar{\Delta}^{(n)}$ in terms of $D$ as follows

$$\begin{aligned}
\bar{\Delta}^{(n)} &= \frac{1}{n} h(\mathbf{X}^n | \breve{\phi}_1^{(n)}(\mathbf{Y}_1^n), \ldots, \breve{\phi}_K^{(n)}(\mathbf{Y}_K^n), \mathbf{Y}_0^n) \\
&= \frac{1}{n} \sum_{i=1}^{n} h(\mathbf{X}_i | \mathbf{X}_{i+1}^n, \breve{\phi}_1^{(n)}(\mathbf{Y}_1^n), \ldots, \breve{\phi}_K^{(n)}(\mathbf{Y}_K^n), \mathbf{Y}_0^n) \\
&= \frac{1}{n} \sum_{i=1}^{n} h(\mathbf{X}_i - \mathbb{E}[\mathbf{X}_i | J_{\mathcal{K}}] \big| \mathbf{X}_{i+1}^n, \breve{\phi}_1^{(n)}(\mathbf{Y}_1^n), \ldots, \breve{\phi}_K^{(n)}(\mathbf{Y}_K^n), \mathbf{Y}_0^n) \\
&\overset{(a)}{\leq} \frac{1}{n} \sum_{i=1}^{n} h(\mathbf{X}_i - \mathbb{E}[\mathbf{X}_i | \breve{\phi}_1^{(n)}(\mathbf{Y}_1^n), \ldots, \breve{\phi}_K^{(n)}(\mathbf{Y}_K^n), \mathbf{Y}_0^n) \\
&\overset{(b)}{\leq} \frac{1}{n} \sum_{i=1}^{n} \log(\pi e)^{n_x} \left| \mathrm{mmse}(\mathbf{X}_i | \breve{\phi}_1^{(n)}(\mathbf{Y}_1^n), \ldots, \breve{\phi}_K^{(n)}(\mathbf{Y}_K^n), \mathbf{Y}_0^n) \right| \\
&\overset{(c)}{\leq} \log(\pi e)^{n_x} \left| \frac{1}{n} \sum_{i=1}^{n} \mathrm{mmse}(\mathbf{X}_i | \breve{\phi}_1^{(n)}(\mathbf{Y}_1^n), \ldots, \breve{\phi}_K^{(n)}(\mathbf{Y}_K^n), \mathbf{Y}_0^n) \right| \\
&\overset{(d)}{\leq} \log(\pi e)^{n_x} D \ , \qquad \text{(G.4)}
\end{aligned}$$

where $(a)$ holds since conditioning reduces entropy; $(b)$ is due to the maximal differential entropy lemma; $(c)$ is due to the convexity of the log-determinant function and Jensen's inequality; and $(d)$ is due to (G.1).

Combining (G.4) with (G.3), and using standard arguments for single-letterization, we get (G.2); and this completes the proof of the lemma. □

The proof of Theorem 5 is as follows. By Lemma 15 and Proposition 5, there must exist Gaussian test channels $(V_1^{\mathrm{G}}, \ldots, V_K^{\mathrm{G}})$ and a time-sharing random variable $Q'$, with joint distribution that factorizes as

$$P_{\mathbf{X},\mathbf{Y}_0}(\mathbf{x}, \mathbf{y}_0) \prod_{k=1}^{K} P_{\mathbf{Y}_k|\mathbf{X},\mathbf{Y}_0}(\mathbf{y}_k|\mathbf{x}, \mathbf{y}_0) \ P_Q'(q') \prod_{k=1}^{K} P_{V_k|\mathbf{Y}_k,Q'}(v_k|\mathbf{y}_k, q') \ ,$$

such that the following holds

$$\sum_{k \in \mathcal{S}} R_k \geq I(\mathbf{Y}_{\mathcal{S}}; V_{\mathcal{S}}^{\mathrm{G}} | V_{\mathcal{S}^c}^{\mathrm{G}}, \mathbf{Y}_0, Q') \ , \quad \text{for } \mathcal{S} \subseteq \mathcal{K} \ , \qquad \text{(G.5)}$$

$$\log(\pi e)^{n_x} D \geq h(\mathbf{X} | V_1^{\mathrm{G}}, \ldots, V_K^{\mathrm{G}}, \mathbf{Y}_0, Q') \ . \qquad \text{(G.6)}$$

This is clearly achievable by the Berger-Tung coding scheme with Gaussian test channels and time-sharing $Q'$, since the achievable error matrix under quadratic distortion has determinant that satisfies

$$\log \Big( (\pi e)^{n_x} |\mathrm{mmse}(\mathbf{X} | V_1^{\mathrm{G}}, \ldots, V_K^{\mathrm{G}}, \mathbf{Y}_0, Q')| \Big) = h(\mathbf{X} | V_1^{G}, \ldots, V_K^{G}, \mathbf{Y}_0, Q') \ .$$

The above shows that the rate-distortion region of the quadratic vector Gaussian CEO problem with determinant constraint is given by (G.6), i.e., $\widetilde{\mathcal{RD}}^{\mathrm{II}}_{\mathrm{CEO}}$ (with distortion parameter $\log(\pi e)^{n_x} D$). Recalling that $\widetilde{\mathcal{RD}}^{\mathrm{II}}_{\mathrm{CEO}} = \widetilde{\mathcal{RD}}^{\mathrm{I}}_{\mathrm{CEO}} = \mathcal{RD}^{\star}_{\text{VG-CEO}}$, and substituting in Theorem 4 using distortion level $\log(\pi e)^{n_x} D$ completes the proof.

# Appendix H

# Proofs for Chapter 5

## H.1 Proof of Lemma 3

First, we rewrite $\mathcal{L}_s(\mathbf{P})$ in (5.6). To that end, the second term of the RHS of (5.6) can be proceeded as

$$
\begin{aligned}
I(Y_1;U_1|U_2,Y_0) &\overset{(a)}{=} I(X,Y_1;U_1|U_2,Y_0) \\
&= I(X;U_1|U_2,Y_0) + I(Y_1;U_1|U_2,Y_0,X) \\
&\overset{(b)}{=} I(X;U_1|U_2,Y_0) + I(Y_1;U_1|X,Y_0) \\
&= I(X;U_1|U_2,Y_0) + I(Y_1,X;U_1|Y_0) - I(X;U_1|Y_0) \\
&\overset{(c)}{=} I(X;U_1|U_2,Y_0) + I(Y_1;U_1|Y_0) - I(X;U_1|Y_0) \\
&= H(X|U_2,Y_0) - H(X|U_1,U_2,Y_0) + H(U_1|Y_0) - H(U_1|Y_0,Y_1) \\
&\quad - H(X|Y_0) + H(X|U_1,Y_0) \\
&= H(X|U_2,Y_0) - H(X|U_1,U_2,Y_0) + H(U_1) - H(Y_0) + H(Y_0|U_1) \\
&\quad - H(U_1|Y_0,Y_1) - H(X|Y_0) + H(X|U_1,Y_0)\ ,
\end{aligned} \tag{H.1}
$$

where $(a)$, $(b)$ and $(c)$ follows due to the Markov chain $U_1 -\!\!\circ\!\!- Y_1 -\!\!\circ\!\!- (X,Y_0) -\!\!\circ\!\!- Y_2 -\!\!\circ\!\!- U_2$. Besides, the third term of the RHS of (5.6) can be written as

$$
\begin{aligned}
I(Y_2;U_2|Y_0) &= H(U_2|Y_0) - H(U_2|Y_0,Y_2) \\
&\overset{(a)}{=} H(U_2|Y_0) - H(U_2|Y_2) \\
&= H(U_2) - H(Y_0) + H(Y_0|U_2) - H(U_2|Y_2)\ ,
\end{aligned} \tag{H.2}
$$

where $(a)$ follows due to the Markov chain $U_1 -\!\!\circ\!\!- Y_1 -\!\!\circ\!\!- (X,Y_0) -\!\!\circ\!\!- Y_2 -\!\!\circ\!\!- U_2$.

By applying (H.1) and (H.2) in (5.6), we have

$$
\begin{aligned}
F_{\mathbf{s}}(\mathbf{P}) =& - s_1 H(X|Y_0) - (s_1+s_2)H(Y_0) + (1-s_1)H(X|U_1,U_2,Y_0) \\
&+ s_1 H(X|U_1,Y_0) + s_1 H(X|U_2,Y_0) + s_1 H(U_1) - s_1 H(U_1|Y_1) \\
&+ s_2 H(U_2) - s_2 H(U_2|Y_2) + s_1 H(Y_0|U_1) + s_2 H(Y_0|U_2) \\
=& - s_1 H(X|Y_0) - (s_1+s_2)H(Y_0) \\
&- (1-s_1) \sum_{u_1 u_2 x y_0} p(u_1,u_2,x,y_0) \log p(x|u_1,u_2,y_0) \\
&- s_1 \sum_{u_1 x y_0} p(u_1,x,y_0) \log p(x|u_1,y_0) - s_1 \sum_{u_2 x y_0} p(u_2,x,y_0) \log p(x|u_2,y_0) \\
&- s_1 \sum_{u_1} p(u_1) \log p(u_1) + s_1 \sum_{u_1 y_1} p(u_1,y_1) \log p(u_1|y_1) \\
&- s_2 \sum_{u_2} p(u_2) \log p(u_2) + s_2 \sum_{u_2 y_2} p(u_2,y_2) \log p(u_2|y_2) \\
&- s_1 \sum_{u_1 y_0} p(u_1,y_0) \log p(y_0|u_1) - s_2 \sum_{u_2 y_0} p(u_2,y_0) \log p(y_0|u_2) \;, \qquad \text{(H.3)}
\end{aligned}
$$

Then, marginalizing (H.3) over variables $X, Y_0, Y_1, Y_2$, and using the Markov chain $U_1 -\!\!\circ\!\!- Y_1 -\!\!\circ\!\!- (X,Y_0) -\!\!\circ\!\!- Y_2 -\!\!\circ\!\!- U_2$, it is easy to see that $F_{\mathbf{s}}(\mathbf{P})$ can be written as

$$
\begin{aligned}
F_{\mathbf{s}}(\mathbf{P}) =& - s_1 H(X|Y_0) - (s_1+s_2)H(Y_0) \\
&+ \mathbb{E}_{P_{X,Y_0,Y_1,Y_2}} \Big[ (1-s_1) \mathbb{E}_{P_{U_1|Y_1}} \mathbb{E}_{P_{U_2|Y_2}} [-\log P_{X|U_1,U_2,Y_0}] \\
&\qquad + s_1 \mathbb{E}_{P_{U_1|Y_1}} [-\log P_{X|U_1,Y_0}] + s_1 \mathbb{E}_{P_{U_2|Y_2}} [-\log P_{X|U_2,Y_0}] \\
&\qquad + s_1 D_{\mathrm{KL}}(P_{U_1|Y_1} \| P_{U_1}) + s_2 D_{\mathrm{KL}}(P_{U_2|Y_2} \| P_{U_2}) \\
&\qquad + s_1 \mathbb{E}_{P_{U_1|Y_1}} [-\log P_{Y_0|U_1}] + s_2 \mathbb{E}_{P_{U_2|Y_2}} [-\log P_{Y_0|U_2}] \Big] \,. \qquad \text{(H.4)}
\end{aligned}
$$

Hence, we have

$$
\begin{aligned}
F_{\mathbf{s}}(\mathbf{P},\mathbf{Q}) - F_{\mathbf{s}}(\mathbf{P}) =& (1-s_1) \mathbb{E}_{U_1,U_2,Y_0} [D_{\mathrm{KL}}(P_{X|U_1,U_2,Y_0} \| Q_{X|U_1,U_2,Y_0})] \\
&+ s_1 \mathbb{E}_{U_1,Y_0} [D_{\mathrm{KL}}(P_{X|U_1,Y_0} \| Q_{X|U_1,Y_0})] + s_1 \mathbb{E}_{U_2,Y_0} [D_{\mathrm{KL}}(P_{X|U_2,Y_0} \| Q_{X|U_2,Y_0})] \\
&+ s_1 D_{\mathrm{KL}}(P_{U_1} \| Q_{U_1}) + s_2 D_{\mathrm{KL}}(P_{U_2} \| Q_{U_2}) \\
&+ s_1 \mathbb{E}_{U_1} [D_{\mathrm{KL}}(P_{Y_0|U_1} \| Q_{Y_0|U_1})] + s_2 \mathbb{E}_{U_2} [D_{\mathrm{KL}}(P_{Y_0|U_2} \| Q_{Y_0|U_2})] \\
\geq& \, 0 \;,
\end{aligned}
$$

where it holds with equality if and only if (5.10) is satisfied. Note that we have the relation $1 - s_1 \geq 0$ due to Lemma 2. This completes the proof.

## H.2 Proof of Lemma 5

We have that $F_{\mathbf{s}}(\mathbf{P},\mathbf{Q})$ is convex in $\mathbf{P}$ from Lemma 4. For a given $\mathbf{Q}$ and $\mathbf{s}$, in order to minimize $F_{\mathbf{s}}(\mathbf{P},\mathbf{Q})$ (given in (H.3)) over the convex set of pmfs $\mathbf{P}$, let us define the Lagrangian as

$$\mathcal{L}_{\mathbf{s}}(\mathbf{P},\mathbf{Q},\boldsymbol{\lambda}) := F_{\mathbf{s}}(\mathbf{P},\mathbf{Q}) + \sum_{y_1}\lambda_1(y_1)[1-\sum_{u_1}p(u_1|y_1)] + \sum_{y_2}\lambda_2(y_2)[1-\sum_{u_2}p(u_2|y_2)]\ ,$$

where $\lambda_1(y_1)\geq 0$ and $\lambda_2(y_2)\geq 0$ are the Lagrange multipliers corresponding the constrains $\sum_{u_k}p(u_k|y_k)=1$, $y_k\in\mathcal{Y}_k$, $k=1,2$, of the pmfs $P_{U_1|Y_1}$ and $P_{U_2|Y_2}$, respectively. Due to the convexity of $F_{\mathbf{s}}(\mathbf{P},\mathbf{Q})$, the Karush-Kuhn-Tucker (KKT) conditions are necessary and sufficient for optimality. By applying the KKT conditions

$$\frac{\partial\mathcal{L}_{\mathbf{s}}(\mathbf{P},\mathbf{Q},\boldsymbol{\lambda})}{\partial p(u_1|y_1)}=0\ ,\qquad \frac{\partial\mathcal{L}_{\mathbf{s}}(\mathbf{P},\mathbf{Q},\boldsymbol{\lambda})}{\partial p(u_2|y_2)}=0\ ,$$

and arranging terms, we obtain

$$\begin{aligned}
&\log p(u_k|y_k)\\
&= \log q(u_k) + \frac{1-s_1}{s_k}\sum_{u_{\bar{k}}xy_0}p(x,y_0|y_k)p(u_{\bar{k}}|x,y_0)\log q(x|u_k,u_{\bar{k}},y_0)\\
&\quad + \frac{s_1}{s_k}\sum_{xy_0}p(x,y_0|y_k)\log q(x|u_k,y_0) + \sum_{y_0}p(y_0|y_k)\log q(y_0|u_k) + \frac{\lambda_k(y_k)}{s_kp(y_k)} - 1\\
&= \log q(u_k) + \frac{1-s_1}{s_k}\sum_{u_{\bar{k}}y_0}p(u_{\bar{k}},y_0|y_k)\sum_x p(x|y_k,u_{\bar{k}},y_0)\log q(x|u_k,u_{\bar{k}},y_0)\\
&\quad + \frac{s_1}{s_k}\sum_{y_0}p(y_0|y_k)\sum_x p(x|y_k,y_0)\log q(x|u_k,y_0) + \sum_{y_0}p(y_0|y_k)\log q(y_0|u_k) + \frac{\lambda_k(y_k)}{s_kp(y_k)} - 1\\
&= \log q(u_k) - \frac{1-s_1}{s_k}\sum_{u_{\bar{k}}y_0}p(u_{\bar{k}},y_0|y_k)\sum_x p(x|y_k,u_{\bar{k}},y_0)\log\frac{p(x|y_k,u_{\bar{k}},y_0)}{q(x|u_k,u_{\bar{k}},y_0)}\frac{1}{p(x|y_k,u_{\bar{k}},y_0)}\\
&\quad - \frac{s_1}{s_k}\sum_{y_0}p(y_0|y_k)\sum_x p(x|y_k,y_0)\log\frac{p(x|y_k,y_0)}{q(x|u_k,y_0)}\frac{1}{p(x|y_k,y_0)}\\
&\quad - \sum_{y_0}p(y_0|y_k)\log\frac{p(y_0|y_k)}{q(y_0|u_k)}\frac{1}{p(y_0|y_k)} + \frac{\lambda_k(y_k)}{s_kp(y_k)} - 1\\
&= \log q(u_k) - \psi_k(u_k,y_k) + \tilde{\lambda}_k(y_k)\ ,
\end{aligned} \tag{H.5}$$

where $\psi_k(u_k,y_k)$, $k=1,2$, are given by (5.13), and $\tilde{\lambda}_k(y_k)$ contains all terms independent of $u_k$ for $k=1,2$. Then, we proceeded by rearranging (H.5) as follows

$$p(u_k|y_k) = e^{\tilde{\lambda}_k(y_k)}q(u_k)e^{-\psi_k(u_k,y_k)}\ ,\quad \text{for } k=1,2\ . \tag{H.6}$$

Finally, the Lagrange multipliers $\lambda_k(y_k)$ satisfying the KKT conditions are obtained by finding $\tilde{\lambda}_k(y_k)$ such that $\sum_{u_k} p(u_k|y_k) = 1$, $k = 1, 2$. Substituting in (H.6), $p(u_k|y_k)$ can be found as in (5.12).

## H.3 Derivation of the Update Rules of Algorithm 3

In this section, we derive the update rules in Algorithm 3 and show that the Gaussian distribution is invariant to the update rules in Algorithm 2, in line with Theorem 4.

First, we recall that if $(\mathbf{X}_1, \mathbf{X}_2)$ are jointly Gaussian, then

$$P_{\mathbf{X}_2|\mathbf{X}_1} \sim \mathcal{CN}(\boldsymbol{\mu}_{\mathbf{x}_2|\mathbf{x}_1}, \boldsymbol{\Sigma}_{\mathbf{x}_2|\mathbf{x}_1}) \ ,$$

where $\boldsymbol{\mu}_{\mathbf{x}_2|\mathbf{x}_1} := \mathbf{K}_{\mathbf{x}_2|\mathbf{x}_1}\mathbf{x}_1$, $\mathbf{K}_{\mathbf{x}_2|\mathbf{x}_1} := \boldsymbol{\Sigma}_{\mathbf{x}_2,\mathbf{x}_1}\boldsymbol{\Sigma}_{\mathbf{x}_1}^{-1}$.

Then, for $\mathbf{Q}^{(t+1)}$ computed as in (5.10) from $\mathbf{P}^{(t)}$, which is a set of Gaussian distributions, we have

$$Q_{\mathbf{X}|\mathbf{U}_1,\mathbf{U}_2,\mathbf{Y}_0} \sim \mathcal{CN}(\boldsymbol{\mu}_{\mathbf{x}|\mathbf{u}_1,\mathbf{u}_2,\mathbf{y}_0}, \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{u}_1,\mathbf{u}_2,\mathbf{y}_0}) \ , \qquad Q_{\mathbf{X}|\mathbf{U}_k,\mathbf{Y}_0} \sim \mathcal{CN}(\boldsymbol{\mu}_{\mathbf{x}|\mathbf{u}_k,\mathbf{y}_0}, \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{u}_k,\mathbf{y}_0}) \ ,$$

$$Q_{\mathbf{Y}_0|\mathbf{U}_k} \sim \mathcal{CN}(\boldsymbol{\mu}_{\mathbf{y}_0|\mathbf{u}_k}, \boldsymbol{\Sigma}_{\mathbf{y}_0|\mathbf{u}_k}) \ , \qquad Q_{\mathbf{U}_k} \sim \mathcal{CN}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{u}_k}) \ .$$

Next, we look at the update $\mathbf{P}^{(t+1)}$ as in (5.12) from given $\mathbf{Q}^{(t+1)}$. To compute $\psi_k(\mathbf{u}_k^t, \mathbf{y}_k)$, first, we note that

$$\begin{aligned} &E_{\mathbf{U}_{\bar{k}},\mathbf{Y}_0|\mathbf{y}_k}[D_{\mathrm{KL}}(P_{\mathbf{X}|\mathbf{y}_k,\mathbf{U}_{\bar{k}},\mathbf{Y}_0}\|Q_{\mathbf{X}|\mathbf{u}_k,\mathbf{U}_{\bar{k}},\mathbf{Y}_0})] \\ &\qquad = D_{\mathrm{KL}}(P_{\mathbf{U}_{\bar{k}},\mathbf{X},\mathbf{Y}_0|\mathbf{y}_k}\|Q_{\mathbf{U}_{\bar{k}},\mathbf{X},\mathbf{Y}_0|\mathbf{u}_k}) - D_{\mathrm{KL}}(P_{\mathbf{U}_{\bar{k}},\mathbf{Y}_0|\mathbf{y}_k}\|Q_{\mathbf{U}_{\bar{k}},\mathbf{Y}_0|\mathbf{u}_k}) \end{aligned} \tag{H.7a}$$

$$\begin{aligned} &\mathbb{E}_{\mathbf{Y}_0|\mathbf{y}_k}[D_{\mathrm{KL}}(P_{\mathbf{X}|\mathbf{y}_k,\mathbf{Y}_0}\|Q_{\mathbf{X}|\mathbf{u}_k,\mathbf{Y}_0})] \\ &\qquad = D_{\mathrm{KL}}(P_{\mathbf{X},\mathbf{Y}_0|\mathbf{y}_k}\|Q_{\mathbf{X},\mathbf{Y}_0|\mathbf{u}_k}) - D_{\mathrm{KL}}(P_{\mathbf{Y}_0|\mathbf{y}_k}\|Q_{\mathbf{Y}_0|\mathbf{u}_k}) \ , \end{aligned} \tag{H.7b}$$

and that for two multivariate Gaussian distributions, i.e., $P_{\mathbf{X}_1} \sim \mathcal{CN}(\boldsymbol{\mu}_{\mathbf{x}_1}, \boldsymbol{\Sigma}_{\mathbf{x}_1})$ and $P_{\mathbf{X}_2} \sim \mathcal{CN}(\boldsymbol{\mu}_{\mathbf{x}_2}, \boldsymbol{\Sigma}_{\mathbf{x}_2})$ in $\mathbb{C}^N$,

$$D_{\mathrm{KL}}(P_{\mathbf{X}_1}\|P_{\mathbf{X}_2}) = (\boldsymbol{\mu}_{\mathbf{x}_1} - \boldsymbol{\mu}_{\mathbf{x}_2})^\dagger \boldsymbol{\Sigma}_{\mathbf{x}_2}^{-1}(\boldsymbol{\mu}_{\mathbf{x}_1} - \boldsymbol{\mu}_{\mathbf{x}_2}) + \log|\boldsymbol{\Sigma}_{\mathbf{x}_2}\boldsymbol{\Sigma}_{\mathbf{x}_1}^{-1}| + \mathrm{tr}(\boldsymbol{\Sigma}_{\mathbf{x}_2}^{-1}\boldsymbol{\Sigma}_{\mathbf{x}_1}) - N \ . \tag{H.8}$$

Applying (H.7) and (H.8) in (5.13) and noting that all involved distributions are Gaussian, it follows that $\psi_k(\mathbf{u}_k^t, \mathbf{y}_k)$ is a quadratic form. Then, since $q^{(t)}(\mathbf{u}_k)$ is also Gaussian, the product $\log(q^{(t)}(\mathbf{u}_k)\exp(-\psi_k(\mathbf{u}_k^t, \mathbf{y}_k)))$ is also a quadratic form, and identifying constant, first and second order terms, we can write

$$\log p^{(t+1)}(\mathbf{u}_k|\mathbf{y}_k) = -(\mathbf{u}_k - \boldsymbol{\mu}_{\mathbf{u}_k^{t+1}|\mathbf{y}_k})^\dagger \boldsymbol{\Sigma}_{\mathbf{z}_k^{t+1}}^{-1}(\mathbf{u}_k - \boldsymbol{\mu}_{\mathbf{u}_k^{t+1}|\mathbf{y}_k}) + Z(\mathbf{y}_k) \;,$$

where

$$\begin{aligned}\boldsymbol{\Sigma}_{\mathbf{z}_k^{t+1}}^{-1} &= \boldsymbol{\Sigma}_{\mathbf{u}_k^t}^{-1} + \frac{1-s_1}{s_k}\mathbf{K}^\dagger_{(\mathbf{u}_{\bar{k}}^t,\mathbf{x},\mathbf{y}_0)|\mathbf{u}_k^t}\boldsymbol{\Sigma}^{-1}_{(\mathbf{u}_{\bar{k}}^t,\mathbf{x},\mathbf{y}_0)|\mathbf{u}_k^t}\mathbf{K}_{(\mathbf{u}_{\bar{k}}^t,\mathbf{x},\mathbf{y}_0)|\mathbf{u}_k^t}\\ &\quad - \frac{1-s_1}{s_k}\mathbf{K}^\dagger_{(\mathbf{u}_{\bar{k}}^t,\mathbf{y}_0)|\mathbf{u}_k^t}\boldsymbol{\Sigma}^{-1}_{(\mathbf{u}_{\bar{k}}^t,\mathbf{y}_0)|\mathbf{u}_k^t}\mathbf{K}_{(\mathbf{u}_{\bar{k}}^t,\mathbf{y}_0)|\mathbf{u}_k^t}\\ &\quad + \frac{s_1}{s_k}\mathbf{K}^\dagger_{(\mathbf{x},\mathbf{y}_0)|\mathbf{u}_k^t}\boldsymbol{\Sigma}^{-1}_{(\mathbf{x},\mathbf{y}_0)|\mathbf{u}_k^t}\mathbf{K}_{(\mathbf{x},\mathbf{y}_0)|\mathbf{u}_k^t} + \frac{s_k-s_1}{s_k}\mathbf{K}^\dagger_{\mathbf{y}_0|\mathbf{u}_k^t}\boldsymbol{\Sigma}^{-1}_{\mathbf{y}_0|\mathbf{u}_k^t}\mathbf{K}_{\mathbf{y}_0|\mathbf{u}_k^t}\end{aligned} \tag{H.9}$$

$$\begin{aligned}\boldsymbol{\mu}_{\mathbf{u}_k^{t+1}|\mathbf{y}_k} = \boldsymbol{\Sigma}_{\mathbf{z}_k^{t+1}}\Bigg(&\frac{1-s_1}{s_k}\mathbf{K}^\dagger_{(\mathbf{u}_{\bar{k}}^t,\mathbf{x},\mathbf{y}_0)|\mathbf{u}_k^t}\boldsymbol{\Sigma}^{-1}_{(\mathbf{u}_{\bar{k}}^t,\mathbf{x},\mathbf{y}_0)|\mathbf{u}_k^t}\mathbf{K}_{(\mathbf{u}_{\bar{k}}^t,\mathbf{x},\mathbf{y}_0)|\mathbf{y}_k}\\ &- \frac{1-s_1}{s_k}\mathbf{K}^\dagger_{(\mathbf{u}_{\bar{k}}^t,\mathbf{y}_0)|\mathbf{u}_k^t}\boldsymbol{\Sigma}^{-1}_{(\mathbf{u}_{\bar{k}}^t,\mathbf{y}_0)|\mathbf{u}_k^t}\mathbf{K}_{(\mathbf{u}_{\bar{k}}^t,\mathbf{y}_0)|\mathbf{y}_k}\\ &+ \frac{s_1}{s_k}\mathbf{K}^\dagger_{(\mathbf{x},\mathbf{y}_0)|\mathbf{u}_k^t}\boldsymbol{\Sigma}^{-1}_{(\mathbf{x},\mathbf{y}_0)|\mathbf{u}_k^t}\mathbf{K}_{(\mathbf{x},\mathbf{y}_0)|\mathbf{y}_k} + \frac{s_k-s_1}{s_k}\mathbf{K}^\dagger_{\mathbf{y}_0|\mathbf{u}_k^t}\boldsymbol{\Sigma}^{-1}_{\mathbf{y}_0|\mathbf{u}_k^t}\mathbf{K}_{\mathbf{y}_0|\mathbf{y}_k}\Bigg)\mathbf{y}_k \;.\end{aligned} \tag{H.10}$$

This shows that $p^{(t+1)}(\mathbf{u}_k|\mathbf{y}_k)$ is a Gaussian distribution and that $\mathbf{U}_k^{t+1}$ is distributed as $\mathbf{U}_k^{t+1}{\sim}\mathcal{CN}(\boldsymbol{\mu}_{\mathbf{u}_k^{t+1}|\mathbf{y}_k}, \boldsymbol{\Sigma}_{\mathbf{z}_k^{t+1}})$.

Next, we simplify (H.9) to obtain the update rule (5.16a). From the matrix inversion lemma, similarly to [21], for $(\mathbf{X}_1, \mathbf{X}_2)$ jointly Gaussian we have

$$\boldsymbol{\Sigma}^{-1}_{\mathbf{x}_2|\mathbf{x}_1} = \boldsymbol{\Sigma}^{-1}_{\mathbf{x}_2} + \mathbf{K}^\dagger_{\mathbf{x}_1|\mathbf{x}_2}\boldsymbol{\Sigma}^{-1}_{\mathbf{x}_1|\mathbf{x}_2}\mathbf{K}_{\mathbf{x}_1|\mathbf{x}_2} \;. \tag{H.11}$$

Applying (H.11) in (H.9), we have

$$\begin{aligned}\boldsymbol{\Sigma}_{\mathbf{z}_k^{t+1}}^{-1} &= \boldsymbol{\Sigma}_{\mathbf{u}_k^t}^{-1} + \frac{1-s_1}{s_k}\left(\boldsymbol{\Sigma}^{-1}_{\mathbf{u}_k^t|(\mathbf{u}_{\bar{k}}^t,\mathbf{x},\mathbf{y}_0)} - \boldsymbol{\Sigma}^{-1}_{\mathbf{u}_k^t}\right) - \frac{1-s_1}{s_k}\left(\boldsymbol{\Sigma}^{-1}_{\mathbf{u}_k^t|(\mathbf{u}_{\bar{k}}^t,\mathbf{y}_0)} - \boldsymbol{\Sigma}^{-1}_{\mathbf{u}_k^t}\right)\\ &\quad + \frac{s_1}{s_k}\left(\boldsymbol{\Sigma}^{-1}_{\mathbf{u}_k^t|(\mathbf{x},\mathbf{y}_0)} - \boldsymbol{\Sigma}^{-1}_{\mathbf{u}_k^t}\right) + \frac{s_k-s_1}{s_k}\left(\boldsymbol{\Sigma}^{-1}_{\mathbf{u}_k^t|\mathbf{y}_0} - \boldsymbol{\Sigma}^{-1}_{\mathbf{u}_k^t}\right)\\ &\overset{(a)}{=} \frac{1}{s_k}\boldsymbol{\Sigma}^{-1}_{\mathbf{u}_k^t|(\mathbf{x},\mathbf{y}_0)} - \frac{1-s_1}{s_k}\boldsymbol{\Sigma}^{-1}_{\mathbf{u}_k^t|(\mathbf{u}_{\bar{k}}^t,\mathbf{y}_0)} + \frac{s_k-s_1}{s_k}\boldsymbol{\Sigma}^{-1}_{\mathbf{u}_k^t|\mathbf{y}_0} \;,\end{aligned}$$

where $(a)$ is due to the Markov chain $\mathbf{U}_1 -\!\!\circ\!\!- \mathbf{X} -\!\!\circ\!\!- \mathbf{U}_2$. We obtain (5.16a) by taking the inverse of both sides of $(a)$.

Also from the matrix inversion lemma [21], for $(\mathbf{X}_1, \mathbf{X}_2)$ jointly Gaussian we have

$$\mathbf{\Sigma}_{\mathbf{x}_1}^{-1}\mathbf{\Sigma}_{\mathbf{x}_1,\mathbf{x}_2}\mathbf{\Sigma}_{\mathbf{x}_2|\mathbf{x}_1}^{-1} = \mathbf{\Sigma}_{\mathbf{x}_1|\mathbf{x}_2}^{-1}\mathbf{\Sigma}_{\mathbf{x}_1,\mathbf{x}_2}\mathbf{\Sigma}_{\mathbf{x}_2}^{-1}\,. \tag{H.12}$$

Now, we simplify (H.10) to obtain the update rule (5.16b) as follows

$$\begin{aligned}
\boldsymbol{\mu}_{\mathbf{u}_k^{t+1}|\mathbf{y}_k} &= \mathbf{\Sigma}_{\mathbf{z}_k^{t+1}}\Bigg(\frac{1-s_1}{s_k}\mathbf{\Sigma}_{\mathbf{u}_k^t}^{-1}\mathbf{\Sigma}_{\mathbf{u}_k^t,(\mathbf{u}_{\bar{k}}^t,\mathbf{x},\mathbf{y}_0)}\mathbf{\Sigma}_{(\mathbf{u}_{\bar{k}}^t,\mathbf{x},\mathbf{y}_0)|\mathbf{u}_k^t}^{-1}\mathbf{\Sigma}_{(\mathbf{u}_{\bar{k}}^t,\mathbf{x},\mathbf{y}_0),\mathbf{y}_k}\mathbf{\Sigma}_{\mathbf{y}_k}^{-1}\\
&\quad-\frac{1-s_1}{s_k}\mathbf{\Sigma}_{\mathbf{u}_k^t}^{-1}\mathbf{\Sigma}_{\mathbf{u}_k^t,(\mathbf{u}_{\bar{k}}^t,\mathbf{y}_0)}\mathbf{\Sigma}_{(\mathbf{u}_{\bar{k}}^t,\mathbf{y}_0)|\mathbf{u}_k^t}^{-1}\mathbf{\Sigma}_{(\mathbf{u}_{\bar{k}}^t,\mathbf{y}_0),\mathbf{y}_k}\mathbf{\Sigma}_{\mathbf{y}_k}^{-1}\\
&\quad+\frac{s_1}{s_k}\mathbf{\Sigma}_{\mathbf{u}_k^t}^{-1}\mathbf{\Sigma}_{\mathbf{u}_k^t,(\mathbf{x},\mathbf{y}_0)}\mathbf{\Sigma}_{(\mathbf{x},\mathbf{y}_0)|\mathbf{u}_k^t}^{-1}\mathbf{\Sigma}_{(\mathbf{x},\mathbf{y}_0),\mathbf{y}_k}\mathbf{\Sigma}_{\mathbf{y}_k}^{-1}\\
&\quad+\frac{s_k-s_1}{s_k}\mathbf{\Sigma}_{\mathbf{u}_k^t}^{-1}\mathbf{\Sigma}_{\mathbf{u}_k^t,\mathbf{y}_0}\mathbf{\Sigma}_{\mathbf{y}_0|\mathbf{u}_k^t}^{-1}\mathbf{\Sigma}_{\mathbf{y}_0,\mathbf{y}_k}\mathbf{\Sigma}_{\mathbf{y}_k}^{-1}\Bigg)\mathbf{y}_k\\
&\stackrel{(a)}{=} \mathbf{\Sigma}_{\mathbf{z}_k^{t+1}}\Bigg(\frac{1-s_1}{s_k}\mathbf{\Sigma}_{\mathbf{u}_k^t|(\mathbf{u}_{\bar{k}}^t,\mathbf{x},\mathbf{y}_0)}^{-1}\mathbf{\Sigma}_{\mathbf{u}_k^t,(\mathbf{u}_{\bar{k}}^t,\mathbf{x},\mathbf{y}_0)}\mathbf{\Sigma}_{(\mathbf{u}_{\bar{k}}^t,\mathbf{x},\mathbf{y}_0)}^{-1}\mathbf{\Sigma}_{(\mathbf{u}_{\bar{k}}^t,\mathbf{x},\mathbf{y}_0),\mathbf{y}_k}\mathbf{\Sigma}_{\mathbf{y}_k}^{-1}\\
&\quad-\frac{1-s_1}{s_k}\mathbf{\Sigma}_{\mathbf{u}_k^t|(\mathbf{u}_{\bar{k}}^t,\mathbf{y}_0)}^{-1}\mathbf{\Sigma}_{\mathbf{u}_k^t,(\mathbf{u}_{\bar{k}}^t,\mathbf{y}_0)}\mathbf{\Sigma}_{(\mathbf{u}_{\bar{k}}^t,\mathbf{y}_0)}^{-1}\mathbf{\Sigma}_{(\mathbf{u}_{\bar{k}}^t,\mathbf{y}_0),\mathbf{y}_k}\mathbf{\Sigma}_{\mathbf{y}_k}^{-1}\\
&\quad+\frac{s_1}{s_k}\mathbf{\Sigma}_{\mathbf{u}_k^t|(\mathbf{x},\mathbf{y}_0)}^{-1}\mathbf{\Sigma}_{\mathbf{u}_k^t,(\mathbf{x},\mathbf{y}_0)}\mathbf{\Sigma}_{(\mathbf{x},\mathbf{y}_0)}^{-1}\mathbf{\Sigma}_{(\mathbf{x},\mathbf{y}_0),\mathbf{y}_k}\mathbf{\Sigma}_{\mathbf{y}_k}^{-1}\\
&\quad+\frac{s_k-s_1}{s_k}\mathbf{\Sigma}_{\mathbf{u}_k^t|\mathbf{y}_0}^{-1}\mathbf{\Sigma}_{\mathbf{u}_k^t,\mathbf{y}_0}\mathbf{\Sigma}_{\mathbf{y}_0}^{-1}\mathbf{\Sigma}_{\mathbf{y}_0,\mathbf{y}_k}\mathbf{\Sigma}_{\mathbf{y}_k}^{-1}\Bigg)\mathbf{y}_k\\
&\stackrel{(b)}{=} \mathbf{\Sigma}_{\mathbf{z}_k^{t+1}}\Bigg(\frac{1-s_1}{s_k}\mathbf{\Sigma}_{\mathbf{u}_k^t|(\mathbf{u}_{\bar{k}}^t,\mathbf{x},\mathbf{y}_0)}^{-1}\mathbf{A}_k^t\mathbf{\Sigma}_{\mathbf{y}_k,(\mathbf{u}_{\bar{k}}^t,\mathbf{x},\mathbf{y}_0)}\mathbf{\Sigma}_{(\mathbf{u}_{\bar{k}}^t,\mathbf{x},\mathbf{y}_0)}^{-1}\mathbf{\Sigma}_{(\mathbf{u}_{\bar{k}}^t,\mathbf{x},\mathbf{y}_0),\mathbf{y}_k}\mathbf{\Sigma}_{\mathbf{y}_k}^{-1}\\
&\quad-\frac{1-s_1}{s_k}\mathbf{\Sigma}_{\mathbf{u}_k^t|(\mathbf{u}_{\bar{k}}^t,\mathbf{y}_0)}^{-1}\mathbf{A}_k^t\mathbf{\Sigma}_{\mathbf{y}_k,(\mathbf{u}_{\bar{k}}^t,\mathbf{y}_0)}\mathbf{\Sigma}_{(\mathbf{u}_{\bar{k}}^t,\mathbf{y}_0)}^{-1}\mathbf{\Sigma}_{(\mathbf{u}_{\bar{k}}^t,\mathbf{y}_0),\mathbf{y}_k}\mathbf{\Sigma}_{\mathbf{y}_k}^{-1}\\
&\quad+\frac{s_1}{s_k}\mathbf{\Sigma}_{\mathbf{u}_k^t|(\mathbf{x},\mathbf{y}_0)}^{-1}\mathbf{A}_k^t\mathbf{\Sigma}_{\mathbf{y}_k,(\mathbf{x},\mathbf{y}_0)}\mathbf{\Sigma}_{(\mathbf{x},\mathbf{y}_0)}^{-1}\mathbf{\Sigma}_{(\mathbf{x},\mathbf{y}_0),\mathbf{y}_k}\mathbf{\Sigma}_{\mathbf{y}_k}^{-1}\\
&\quad+\frac{s_k-s_1}{s_k}\mathbf{\Sigma}_{\mathbf{u}_k^t|\mathbf{y}_0}^{-1}\mathbf{A}_k^t\mathbf{\Sigma}_{\mathbf{y}_k,\mathbf{y}_0}\mathbf{\Sigma}_{\mathbf{y}_0}^{-1}\mathbf{\Sigma}_{\mathbf{y}_0,\mathbf{y}_k}\mathbf{\Sigma}_{\mathbf{y}_k}^{-1}\Bigg)\mathbf{y}_k\\
&\stackrel{(c)}{=} \mathbf{\Sigma}_{\mathbf{z}_k^{t+1}}\Bigg(\frac{1-s_1}{s_k}\mathbf{\Sigma}_{\mathbf{u}_k^t|(\mathbf{u}_{\bar{k}}^t,\mathbf{x},\mathbf{y}_0)}^{-1}\mathbf{A}_k^t(\mathbf{\Sigma}_{\mathbf{y}_k}-\mathbf{\Sigma}_{\mathbf{y}_k|(\mathbf{u}_{\bar{k}}^t,\mathbf{x},\mathbf{y}_0)})\mathbf{\Sigma}_{\mathbf{y}_k}^{-1}\\
&\quad-\frac{1-s_1}{s_k}\mathbf{\Sigma}_{\mathbf{u}_k^t|(\mathbf{u}_{\bar{k}}^t,\mathbf{y}_0)}^{-1}\mathbf{A}_k^t(\mathbf{\Sigma}_{\mathbf{y}_k}-\mathbf{\Sigma}_{\mathbf{y}_k|(\mathbf{u}_{\bar{k}}^t,\mathbf{y}_0)})\mathbf{\Sigma}_{\mathbf{y}_k}^{-1}\\
&\quad+\frac{s_1}{s_k}\mathbf{\Sigma}_{\mathbf{u}_k^t|(\mathbf{x},\mathbf{y}_0)}^{-1}\mathbf{A}_k^t(\mathbf{\Sigma}_{\mathbf{y}_k}-\mathbf{\Sigma}_{\mathbf{y}_k|(\mathbf{x},\mathbf{y}_0)})\mathbf{\Sigma}_{\mathbf{y}_k}^{-1}\\
&\quad+\frac{s_k-s_1}{s_k}\mathbf{\Sigma}_{\mathbf{u}_k^t|\mathbf{y}_0}^{-1}\mathbf{A}_k^t(\mathbf{\Sigma}_{\mathbf{y}_k}-\mathbf{\Sigma}_{\mathbf{y}_k|\mathbf{y}_0})\mathbf{\Sigma}_{\mathbf{y}_k}^{-1}\Bigg)\mathbf{y}_k
\end{aligned}$$

$$\begin{aligned}
&\overset{(d)}{=} \mathbf{\Sigma}_{\mathbf{z}_k^{t+1}} \Bigg( \frac{1}{s_k} \mathbf{\Sigma}^{-1}_{\mathbf{u}_k^t|(\mathbf{x},\mathbf{y}_0)} \mathbf{A}_k^t (\mathbf{I} - \mathbf{\Sigma}_{\mathbf{y}_k|(\mathbf{x},\mathbf{y}_0)} \mathbf{\Sigma}^{-1}_{\mathbf{y}_k}) \\
&\qquad - \frac{1-s_1}{s_k} \mathbf{\Sigma}^{-1}_{\mathbf{u}_k^t|(\mathbf{u}_{\bar{k}}^t,\mathbf{y}_0)} \mathbf{A}_k^t (\mathbf{I} - \mathbf{\Sigma}_{\mathbf{y}_k|(\mathbf{u}_{\bar{k}}^t,\mathbf{y}_0)} \mathbf{\Sigma}^{-1}_{\mathbf{y}_k}) \\
&\qquad + \frac{s_k - s_1}{s_k} \mathbf{\Sigma}^{-1}_{\mathbf{u}_k^t|\mathbf{y}_0} \mathbf{A}_k^t (\mathbf{I} - \mathbf{\Sigma}_{\mathbf{y}_k|\mathbf{y}_0} \mathbf{\Sigma}^{-1}_{\mathbf{y}_k}) \Bigg) \mathbf{y}_k \ ,
\end{aligned}$$

where $(a)$ follows from (H.12); $(b)$ follows from the relation $\mathbf{\Sigma}_{\mathbf{u}_k,\mathbf{y}_0} = \mathbf{A}_k \mathbf{\Sigma}_{\mathbf{y}_k,\mathbf{y}_0}$; $(c)$ is due the definition of $\mathbf{\Sigma}_{\mathbf{x}_1|\mathbf{x}_2}$; and $(d)$ is due to the Markov chain $\mathbf{U}_1 -\!\!\circ\!\!- \mathbf{X} -\!\!\circ\!\!- \mathbf{U}_2$. Equation (5.16b) follows by noting that $\boldsymbol{\mu}_{\mathbf{u}_k^{t+1}|\mathbf{y}_k} = \mathbf{A}_k^{t+1}\mathbf{y}_k$.

## H.4 Proof of Proposition 9

For simplicity of exposition, the proof is given for the case $K = 2$ encoders. The proof for $K > 2$ follows similarly. By the definition of $\mathcal{RI}_{\text{DIB}}^{\text{sum}}$, the tuple $(\Delta, R_{\text{sum}})$ is achievable if there exists some random variables $X, Y_1, Y_2, U_1, U_2$ with joint distribution $P_X(x)\prod_{k=1}^K P_{Y_k|X}(y_k|x)\prod_{k=1}^K P_{U_k|Y_k}(u_k|y_k)$ satisfying

$$\Delta \leq I(X;U_1,U_2) \tag{H.13a}$$
$$\Delta \leq R_1 - I(Y_1;U_1|X) + I(X;U_2) \tag{H.13b}$$
$$\Delta \leq R_2 - I(Y_2;U_2|X) + I(X;U_1) \tag{H.13c}$$
$$\Delta \leq R_1 + R_2 - I(Y_1;U_1|X) - I(Y_2;U_2|X) \tag{H.13d}$$
$$R_1 + R_2 \leq R_{\text{sum}} \ . \tag{H.13e}$$

The application of the Fourier-Motzkin elimination to project out $R_1$ and $R_2$ reduces (H.13) to the following system of inequalities

$$\Delta \leq I(X;U_1,U_2) \tag{H.14a}$$
$$2\Delta \leq R_{\text{sum}} - I(Y_1;U_1|X) - I(Y_2;U_2|X) + I(X;U_1) + I(X;U_2) \tag{H.14b}$$
$$\Delta \leq R_{\text{sum}} - I(Y_1;U_1|X) - I(Y_2;U_2|X) \ . \tag{H.14c}$$

We note that we have $I(X;U_1,U_2) \leq I(X;U_1) + I(X;U_2)$ due to the Markov chain $U_1 -\!\!\circ\!\!- Y_1 -\!\!\circ\!\!- X -\!\!\circ\!\!- Y_2 -\!\!\circ\!\!- U_2$. Therefore, inequality (H.14b) is redundant as it is implied by (H.14a) and (H.14c). This completes the proof.

## H.5 Proof of Proposition 10

Suppose that $\mathbf{P}^\star$ yields the maximum in (5.20). Then,

$$\begin{aligned}
\Delta_s &= \frac{1}{1+s}\Big[(1+sK)H(X) + sR_s + \mathcal{L}_s^{\mathrm{DIB}}(\mathbf{P}^\star)\Big] \\
&\overset{(a)}{=} \frac{1}{1+s}\Big[(1+sK)H(X) + sR_s - H(X|U_{\mathcal{K}}^\star) + s\sum_{k=1}^{K}[-H(X|U_k^\star) - I(Y_k;U_k^\star)]\Big] \\
&= \frac{1}{1+s}\Big[sR_s + H(X) - H(X|U_{\mathcal{K}}^\star) + s\sum_{k=1}^{K}[H(X) - H(X|U_k^\star) - I(Y_k;U_k^\star)]\Big] \\
&= \frac{1}{1+s}\Big[sR_s + I(X;U_{\mathcal{K}}^\star) + s\sum_{k=1}^{K}[I(X;U_k^\star) - I(Y_k;U_k^\star)]\Big] \\
&\overset{(b)}{=} \frac{1}{1+s}\Big[sR_s + I(X;U_{\mathcal{K}}^\star) + s\Big(I(X;U_{\mathcal{K}}^\star) - R_s\Big)\Big] \\
&= \frac{1}{1+s}\Big[sR_s + I(X;U_{\mathcal{K}}^\star) + s\Big(I(X;U_{\mathcal{K}}^\star) - R_s\Big)\Big] \\
&= I(X;U_{\mathcal{K}}^\star) \\
&\overset{(c)}{\geq} \Delta_{\mathrm{DIB}}^{\mathrm{sum}}(R_s)\ ,
\end{aligned} \tag{H.15}$$

where $(a)$ follows from the definition of $\mathcal{L}_s^{\mathrm{DIB}}(\mathbf{P})$ in (5.22); $(b)$ is due to the definition of $R_s$ in (5.21); $(c)$ follows follows from (5.19).

Conversely, if $\mathbf{P}^\star$ is the solution which maximize $\Delta_{\mathrm{DIB}}^{\mathrm{sum}}(R_{\mathrm{sum}})$ in (5.19) such that $\Delta_{\mathrm{DIB}}^{\mathrm{sum}}(R_{\mathrm{sum}}) = \Delta_s$, then the following will be held

$$\Delta_s \leq I(X;U_{\mathcal{K}}^\star) \tag{H.16a}$$

$$\Delta_s \leq R_{\mathrm{sum}} - \sum_{k=1}^{K} I(Y_k;U_k^\star|X)\ . \tag{H.16b}$$

Besides, for any $s \geq 0$, we have

$$\begin{aligned}
\Delta_{\mathrm{DIB}}^{\mathrm{sum}}(R_{\mathrm{sum}}) &= \Delta_s \\
&\overset{(a)}{\leq} \Delta_s + \Big(I(X;U_{\mathcal{K}}^\star) - \Delta_s\Big) + s\Big(R_{\mathrm{sum}} - \sum_{k=1}^{K} I(Y_k;U_k^\star|X) - \Delta_s\Big) \\
&= I(X;U_{\mathcal{K}}^\star) - s\Delta_s + sR_{\mathrm{sum}} - s\sum_{k=1}^{K} I(Y_k;U_k^\star|X) \\
&\overset{(b)}{=} I(X;U_{\mathcal{K}}^\star) - s\Delta_s + sR_{\mathrm{sum}} - s\sum_{k=1}^{K}[I(Y_k;U_k^\star) - I(X;U_k^\star)]
\end{aligned}$$

$$= (1+sK)H(X) - s\Delta_s + sR_{\text{sum}} - H(X|U^\star_{\mathcal{K}}) - s\sum_{k=1}^{K}[H(X|U^\star_k) + I(Y_k;U^\star_k)]$$
$$\overset{(c)}{\leq} (1+sK)H(X) - s\Delta_s + sR_{\text{sum}} + \mathcal{L}^\star_s$$
$$\overset{(d)}{\leq} (1+sK)H(X) - s\Delta_s + sR_{\text{sum}} + (1+s)\Delta_s - (1+sK)H(X) - sR_s$$
$$= \Delta_s + s(R_{\text{sum}} - R_s)\ , \tag{H.17}$$

where $(a)$ due to the inequalities (H.16); $(b)$ follows since we have $I(Y_k;U_k|X) = I(Y_k,X;U_k) - I(X;U_k) = I(Y_k;U_k) - I(X;U_k)$ due to the Markov chain $U_k -\!\!\circ\!\!- Y_k -\!\!\circ\!\!- X -\!\!\circ\!\!- Y_{\mathcal{K}\setminus k} -\!\!\circ\!\!- U_{\mathcal{K}\setminus k}$; $(c)$ follows since $\mathcal{L}^\star_s$ is the value maximizing (5.22) over all possible $\mathbf{P}$ values (not necessarily $\mathbf{P}^\star$ maximizing $\Delta^{\text{sum}}_{\text{DIB}}(R_{\text{sum}})$); and $(d)$ is due to (5.20).

Finally, (H.17) is valid for any $R_{\text{sum}} \geq 0$ and $s \geq 0$. For a given $s$, letting $R_{\text{sum}} = R_s$, (H.17) yields $\Delta^{\text{sum}}_{\text{DIB}}(R_s) \leq \Delta_s$. Together with (H.15), this completes the proof.

## H.6 Proof of Lemma 6

First, we expand $\mathcal{L}^{\text{DIB}}_s(\mathbf{P})$ in (5.22) as follows

$$\begin{aligned}\mathcal{L}^{\text{DIB}}_s(\mathbf{P}) &= -H(X|U_{\mathcal{K}}) - s\sum_{k=1}^{K}[H(X|U_k) + H(U_k) - H(U_k|Y_k)]\\ &= \sum_{u_{\mathcal{K}}}\sum_{x} p(u_{\mathcal{K}},x)\log p(x|u_{\mathcal{K}}) + s\sum_{k=1}^{K}\sum_{u_k}\sum_{x} p(u_k,x)\log p(x|u_k)\\ &\quad + s\sum_{k=1}^{K}\sum_{u_k} p(u_k)\log p(u_k) - s\sum_{k=1}^{K}\sum_{u_k}\sum_{y_k} p(u_k,y_k)\log p(u_k|y_k)\ .\end{aligned} \tag{H.18}$$

Then, $\mathcal{L}^{\text{VDIB}}_s(\mathbf{P},\mathbf{Q})$ is defined as follows

$$\begin{aligned}\mathcal{L}^{\text{VDIB}}_s(\mathbf{P},\mathbf{Q}) &= \sum_{u_{\mathcal{K}}}\sum_{x} p(u_{\mathcal{K}},x)\log q(x|u_{\mathcal{K}}) + s\sum_{k=1}^{K}\sum_{u_k}\sum_{x} p(u_k,x)\log q(x|u_k)\\ &\quad + s\sum_{k=1}^{K}\sum_{u_k} p(u_k)\log q(u_k) - s\sum_{k=1}^{K}\sum_{u_k}\sum_{y_k} p(u_k,y_k)\log p(u_k|y_k)\ .\end{aligned} \tag{H.19}$$

Hence, from (H.18) and (H.19) we have the following relation

$$\begin{aligned}\mathcal{L}_s^{\mathrm{DIB}}(\mathbf{P}) - \mathcal{L}_s^{\mathrm{VDIB}}(\mathbf{P},\mathbf{Q}) =& \mathbb{E}_{P_{U_{\mathcal{K}}}}[D_{\mathrm{KL}}(P_{X|U_{\mathcal{K}}}\|Q_{X|U_{\mathcal{K}}}] \\ &+ s\sum_{k=1}^{K}\Big(\mathbb{E}_{P_{U_k}}[D_{\mathrm{KL}}(P_{X|U_k}\|Q_{X|U_k}] + D_{\mathrm{KL}}(P_{U_k}\|Q_{U_k})\Big) \\ \geq& 0\ ,\end{aligned}$$

where it holds with an equality if and only if $Q_{X|U_{\mathcal{K}}} = P_{X|U_{\mathcal{K}}}$, $Q_{X|U_k} = P_{X|U_k}$, $Q_{U_k} = P_{U_k}$, $k = 1,\dots,K$. We note that $s \geq 0$.

Now, we will complete the proof by showing that (H.19) is equal to (5.23). To do so, we proceed (H.19) as follows

$$\begin{aligned}\mathcal{L}_s^{\mathrm{VDIB}}(\mathbf{P},\mathbf{Q}) =& \sum_{u_{\mathcal{K}}}\sum_{x}\sum_{y_{\mathcal{K}}} p(u_{\mathcal{K}},x,y_{\mathcal{K}})\log q(x|u_{\mathcal{K}}) \\ &+ s\sum_{k=1}^{K}\sum_{u_k}\sum_{x}\sum_{y_{\mathcal{K}}} p(u_k,x,y_{\mathcal{K}})\log q(x|u_k) \\ &- s\sum_{k=1}^{K}\sum_{u_k}\sum_{x}\sum_{y_{\mathcal{K}}} p(u_k,x,y_{\mathcal{K}})\log\frac{p(u_k|y_k)}{q(u_k)} \\ \stackrel{(a)}{=}& \sum_{x}\sum_{y_{\mathcal{K}}} p(x,y_{\mathcal{K}})\sum_{u_{\mathcal{K}}} p(u_1|y_1)\times\cdots\times p(u_K|y_K)\log q(x|u_{\mathcal{K}}) \\ &+ s\sum_{x}\sum_{y_{\mathcal{K}}} p(x,y_{\mathcal{K}})\sum_{k=1}^{K}\sum_{u_k} p(u_k|y_k)\log q(x|u_k) \\ &+ s\sum_{x}\sum_{y_{\mathcal{K}}} p(x,y_{\mathcal{K}})\sum_{k=1}^{K}\sum_{u_k} p(u_k|y_k)\log\frac{p(u_k|y_k)}{q(u_k)} \\ =& \mathbb{E}_{P_{X,Y_{\mathcal{K}}}}\Bigg[\mathbb{E}_{P_{U_1|Y_1}}\times\cdots\times\mathbb{E}_{P_{U_K|Y_K}}[\log Q_{X|U_{\mathcal{K}}}] \\ &+ s\sum_{k=1}^{K}\Big(\mathbb{E}_{P_{U_k|Y_k}}[\log Q_{X|U_k}] - D_{\mathrm{KL}}(P_{U_k|Y_k}\|Q_{U_k})\Big)\Bigg]\ ,\end{aligned}$$

where $(a)$ follows due to the Markov chain $U_k -\!\!\circ\!\!- Y_k -\!\!\circ\!\!- X -\!\!\circ\!\!- Y_{\mathcal{K}\setminus k} -\!\!\circ\!\!- U_{\mathcal{K}\setminus k}$. This completes the proof.

# Appendix I

# Supplementary Material for Chapter 6

## I.1 Proof of Lemma 7

First, we expand $\mathcal{L}'_s(\mathbf{P})$ as follows

$$
\begin{aligned}
\mathcal{L}'_s(\mathbf{P}) &= - H(\mathbf{X}|\mathbf{U}) - sI(\mathbf{X};\mathbf{U}) \\
&= - H(\mathbf{X}|\mathbf{U}) - s[H(\mathbf{U}) - H(\mathbf{U}|\mathbf{X})] \\
&= \iint_{\mathbf{ux}} p(\mathbf{u},\mathbf{x}) \log p(\mathbf{x}|\mathbf{u}) \, d\mathbf{u} \, d\mathbf{x} \\
&\quad + s \int_{\mathbf{u}} p(\mathbf{u}) \log p(\mathbf{u}) \, d\mathbf{u} - s \iint_{\mathbf{ux}} p(\mathbf{u},\mathbf{x}) \log p(\mathbf{u}|\mathbf{x}) \, d\mathbf{u} \, d\mathbf{x}.
\end{aligned}
$$

Then, $\mathcal{L}_s^{\mathrm{VB}}(\mathbf{P},\mathbf{Q})$ is defined as follows

$$
\begin{aligned}
\mathcal{L}_s^{\mathrm{VB}}(\mathbf{P},\mathbf{Q}) := & \iint_{\mathbf{ux}} p(\mathbf{u},\mathbf{x}) \log q(\mathbf{x}|\mathbf{u}) \, d\mathbf{u} \, d\mathbf{x} \\
& + s \int_{\mathbf{u}} p(\mathbf{u}) \log q(\mathbf{u}) \, d\mathbf{u} - s \iint_{\mathbf{ux}} p(\mathbf{u},\mathbf{x}) \log p(\mathbf{u}|\mathbf{x}) \, d\mathbf{u} \, d\mathbf{x}.
\end{aligned} \tag{I.1}
$$

Hence, we have the following relation

$$
\mathcal{L}'_s(\mathbf{P}) - \mathcal{L}_s^{\mathrm{VB}}(\mathbf{P},\mathbf{Q}) = \mathbb{E}_{P_{\mathbf{X}}}[D_{\mathrm{KL}}(P_{\mathbf{X}|\mathbf{U}} \| Q_{\mathbf{X}|\mathbf{U}})] + sD_{\mathrm{KL}}(P_{\mathbf{U}} \| Q_{\mathbf{U}}) \geq 0
$$

where equality holds under equalities $Q_{\mathbf{X}|\mathbf{U}} = P_{\mathbf{X}|\mathbf{U}}$ and $Q_{\mathbf{U}} = P_{\mathbf{U}}$. We note that $s \geq 0$.

Now, we complete the proof by showing that (I.1) is equal to (6.8). To do so, we proceed (I.1) as follows

$$\begin{aligned}\mathcal{L}_s^{\mathrm{VB}}(\mathbf{P},\mathbf{Q}) &= \int_{\mathbf{x}} p(\mathbf{x}) \int_{\mathbf{u}} p(\mathbf{u}|\mathbf{x}) \log q(\mathbf{x}|\mathbf{u})\, d\mathbf{u}\, d\mathbf{x} \\ &\quad + s \int_{\mathbf{x}} p(\mathbf{x}) \int_{\mathbf{u}} p(\mathbf{u}|\mathbf{x}) \log q(\mathbf{u})\, d\mathbf{u} - s \int_{\mathbf{x}} p(\mathbf{x}) \int_{\mathbf{u}} p(\mathbf{u}|\mathbf{x}) \log p(\mathbf{u}|\mathbf{x})\, d\mathbf{u}\, d\mathbf{x} \\ &= \mathbb{E}_{P_{\mathbf{X}}}\Big[\mathbb{E}_{P_{\mathbf{U}|\mathbf{X}}}[\log Q_{\mathbf{X}|\mathbf{U}}] - s D_{\mathrm{KL}}(P_{\mathbf{U}|\mathbf{X}} \| Q_{\mathbf{U}})\Big].\end{aligned}$$

## I.2 Alternative Expression $\mathcal{L}_s^{\mathrm{VaDE}}$

Here, we show that (6.13) is equal to (6.14).

To do so, we start with (6.14) and proceed as follows

$$\begin{aligned}\mathcal{L}_s^{\mathrm{VaDE}} &= \mathbb{E}_{P_{\mathbf{X}}}\Big[\mathbb{E}_{P_{\mathbf{U}|\mathbf{X}}}[\log Q_{\mathbf{X}|\mathbf{U}}] - s D_{\mathrm{KL}}(P_{\mathbf{U}|\mathbf{X}} \| Q_{\mathbf{U}}) - s \mathbb{E}_{P_{\mathbf{U}|\mathbf{X}}}\big[D_{\mathrm{KL}}(P_{C|\mathbf{X}} \| Q_{C|\mathbf{U}})\big]\Big] \\ &= \mathbb{E}_{P_X}\left[\mathbb{E}_{P_{\mathbf{U}|\mathbf{X}}}[\log Q_{\mathbf{X}|\mathbf{U}}]\right] - s \int_{\mathbf{x}} p(\mathbf{x}) \int_{\mathbf{u}} p(\mathbf{u}|\mathbf{x}) \log \frac{p(\mathbf{u}|\mathbf{x})}{q(\mathbf{u})}\, d\mathbf{u}\, d\mathbf{x} \\ &\quad - s \int_{\mathbf{x}} p(\mathbf{x}) \int_{\mathbf{u}} p(\mathbf{u}|\mathbf{x}) \sum_c p(c|\mathbf{x}) \log \frac{p(c|\mathbf{x})}{q(c|\mathbf{u})}\, d\mathbf{u}\, d\mathbf{x} \\ &\overset{(a)}{=} \mathbb{E}_{P_X}\left[\mathbb{E}_{P_{\mathbf{U}|\mathbf{X}}}[\log Q_{\mathbf{X}|\mathbf{U}}]\right] - s \iint_{\mathbf{ux}} p(\mathbf{x}) p(\mathbf{u}|\mathbf{x}) \log \frac{p(\mathbf{u}|\mathbf{x})}{q(\mathbf{u})}\, d\mathbf{u}\, d\mathbf{x} \\ &\quad - s \iint_{\mathbf{ux}} \sum_c p(\mathbf{x}) p(\mathbf{u}|c,\mathbf{x}) p(c|\mathbf{x}) \log \frac{p(c|\mathbf{x})}{q(c|\mathbf{u})}\, d\mathbf{u}\, d\mathbf{x} \\ &= \mathbb{E}_{P_X}\left[\mathbb{E}_{P_{\mathbf{U}|\mathbf{X}}}[\log Q_{\mathbf{X}|\mathbf{U}}]\right] - s \iint_{\mathbf{ux}} \sum_c p(\mathbf{u},c,\mathbf{x}) \log \frac{p(\mathbf{u}|\mathbf{x}) p(c|\mathbf{x})}{q(\mathbf{u}) q(c|\mathbf{u})}\, d\mathbf{u}\, d\mathbf{x} \\ &= \mathbb{E}_{P_X}\left[\mathbb{E}_{P_{\mathbf{U}|\mathbf{X}}}[\log Q_{\mathbf{X}|\mathbf{U}}]\right] - s \iint_{\mathbf{ux}} \sum_c p(\mathbf{u},c,\mathbf{x}) \log \frac{p(c|\mathbf{x})}{q(c)} \frac{p(\mathbf{u}|\mathbf{x})}{q(\mathbf{u}|c)}\, d\mathbf{u}\, d\mathbf{x} \\ &= \mathbb{E}_{P_X}\left[\mathbb{E}_{P_{\mathbf{U}|\mathbf{X}}}[\log Q_{\mathbf{X}|\mathbf{U}}]\right] - s \int_{\mathbf{x}} \sum_c p(c,\mathbf{x}) \log \frac{p(c|\mathbf{x})}{q(c)}\, d\mathbf{x} \\ &\quad - s \iint_{\mathbf{ux}} \sum_c p(\mathbf{x}) p(c|\mathbf{x}) p(\mathbf{u}|c,\mathbf{x}) \log \frac{p(\mathbf{u}|\mathbf{x})}{q(\mathbf{u}|c)}\, d\mathbf{u}\, d\mathbf{x} \\ &\overset{(b)}{=} \mathbb{E}_{P_{\mathbf{X}}}\Big[\mathbb{E}_{P_{\mathbf{U}|\mathbf{X}}}[\log Q_{\mathbf{X}|\mathbf{U}}] - s D_{\mathrm{KL}}(P_{C|\mathbf{X}} \| Q_C) - s \mathbb{E}_{P_{C|\mathbf{X}}}[D_{\mathrm{KL}}(P_{\mathbf{U}|\mathbf{X}} \| Q_{\mathbf{U}|C})]\Big],\end{aligned}$$

where $(a)$ and $(b)$ follow due to the Markov chain $C \multimap\!\!- \mathbf{X} \multimap\!\!- \mathbf{U}$.

## I.3 KL Divergence Between Multivariate Gaussian Distributions

The KL divergence between two multivariate Gaussian distributions $P_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $P_2 \sim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ in $\mathbb{R}^J$ is

$$D_{\mathrm{KL}}(P_1 \| P_2) = \frac{1}{2}\Big((\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \log|\boldsymbol{\Sigma}_2| - \log|\boldsymbol{\Sigma}_1| - J + \mathrm{tr}(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1)\Big). \tag{I.2}$$

For the case in which $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ covariance matrices are diagonal, i.e., $\boldsymbol{\Sigma}_1 := \mathrm{diag}(\{\sigma_{1,j}^2\}_{j=1}^J)$ and $\boldsymbol{\Sigma}_2 := \mathrm{diag}(\{\sigma_{2,j}^2\}_{j=1}^J)$, (I.2) boils down to the following

$$D_{\mathrm{KL}}(P_1 \| P_2) = \frac{1}{2}\Big(\sum_{j=1}^{J} \frac{(\mu_{1,j} - \mu_{2,j})^2}{\sigma_{2,j}^2} + \log\frac{\sigma_{2,j}^2}{\sigma_{1,j}^2} - 1 + \frac{\sigma_{1,j}^2}{\sigma_{2,j}^2}\Big). \tag{I.3}$$

## I.4 KL Divergence Between Gaussian Mixture Models

An exact close form for the calculation of the KL divergence between two Gaussian mixture models does not exist. In this paper, we use a variational lower bound approximation for calculations of KL between two Gaussian mixture models. Let $f$ and $g$ be GMMs and the marginal densities of $x$ under $f$ and $g$ are

$$\begin{aligned} f(x) &= \sum_{m=1}^{M} \omega_m \mathcal{N}(x; \mu_m^{\mathrm{f}}, \Sigma_m^{\mathrm{f}}) = \sum_{m=1}^{M} \omega_m f_m(x) \\ g(x) &= \sum_{C=1}^{C} \pi_c \mathcal{N}(x; \mu_c^{\mathrm{g}}, \Sigma_c^{\mathrm{g}}) = \sum_{c=1}^{C} \pi_c g_c(x). \end{aligned}$$

The KL divergence between two Gaussian mixtures f an g can be approximated as follows

$$D_{\mathrm{vKL}}(f \| g) := \sum_{m=1}^{M} \omega_m \log \frac{\sum_{m' \in \mathcal{M} \setminus m} \omega_{m'} \exp\left(-D_{\mathrm{KL}}(f_m \| f_{m'})\right)}{\sum_{c=1}^{C} \pi_c \exp\left(-D_{\mathrm{KL}}(f_m \| g_c)\right)}. \tag{I.4}$$

In this paper, we are interested, in particular, $M = 1$. Hence, (I.4) simplifies to

$$D_{\mathrm{vKL}}(f \| g) = -\log \sum_{c=1}^{C} \pi_c \exp\left(-D_{\mathrm{KL}}(f \| g_c)\right) \tag{I.5}$$

where $D_{\mathrm{KL}}(\cdot \| \cdot)$ is the KL divergence between single component multivariate Gaussian distribution, defined as in (I.2).

# Bibliography

[1] Inaki Estella Aguerri and Abdellatif Zaidi, “Distributed variational representation learning,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[2] Toby Berger, Zhen Zhang, and Harish Viswanathan, “The CEO problem,” *IEEE Transactions on Information Theory*, vol. 42, no. 3, pp. 887 – 902, May 1996.

[3] Yasutada Oohama, “Rate-distortion theory for Gaussian multiterminal source coding systems with several side informations at the decoder,” *IEEE Transactions on Information Theory*, vol. 51, no. 7, pp. 2577 – 2593, July 2005.

[4] Vinod Prabhakaran, David Tse, and Kannan Ramachandran, “Rate region of the quadratic Gaussian CEO problem,” in *Proceedings of IEEE International Symposium on Information Theory*, June – July 2004, p. 117.

[5] Jun Chen and Jia Wang, “On the vector Gaussian CEO problem,” in *Proceedings of IEEE International Symposium on Information Theory*, July – August 2011, pp. 2050 – 2054.

[6] Jia Wang and Jun Chen, “On the vector Gaussian $L$-terminal CEO problem,” in *Proceedings of IEEE International Symposium on Information Theory*, July 2012, pp. 571 – 575.

[7] Tie Liu and Pramod Viswanath, “An extremal inequality motivated by multiterminal information-theoretic problems,” *IEEE Transactions on Information Theory*, vol. 53, no. 5, pp. 1839 – 1851, May 2007.

[8] Yinfei Xu and Qiao Wang, “Rate region of the vector Gaussian CEO problem with the trace distortion constraint,” *IEEE Transactions on Information Theory*, vol. 62, no. 4, pp. 1823 – 1835, April 2016.

[9] Thomas A. Courtade and Richard D. Wesel, "Multiterminal source coding with an entropy-based distortion measure," in *Proceedings of IEEE International Symposium on Information Theory*, July – August 2011, pp. 2040 – 2044.

[10] Thomas A. Courtade and Tsachy Weissman, "Multiterminal source coding under logarithmic loss," *IEEE Transactions on Information Theory*, vol. 60, no. 1, pp. 740 – 761, January 2014.

[11] Ersen Ekrem and Sennur Ulukus, "An outer bound for the vector Gaussian CEO problem," *IEEE Transactions on Information Theory*, vol. 60, no. 11, pp. 6870 – 6887, November 2014.

[12] Saurabha Tavildar and Pramod Viswanath, "On the sum-rate of the vector Gaussian CEO problem," in *Proceedings of* 39*-th Asilomar Conference on Signals, Systems, and Computers*, October – November 2005, pp. 3 – 7.

[13] Hanan Weingarten, Yossef Steinberg, and Shlomo Shamai (Shitz), "The capacity region of the gaussian multiple-input multiple-output broadcast channel," *IEEE Transactions on Information Theory*, vol. 52, no. 9, pp. 3936 – 3964, September 2006.

[14] Daniel Perez Palomar, John M. Cioffi, and Miguel Angel Lagunas, "Joint Tx-Rx beamforming design for multicarrier MIMO channels: A unified framework for convex optimization," *IEEE Transactions on Signal Processing*, vol. 51, no. 9, pp. 2381 – 2401, September 2003.

[15] Anna Scaglione, Petre Stoica, Sergio Barbarossa, Georgios B. Giannakis, and Hemanth Sampath, "Optimal designs for space-time linear precoders and decoders," *IEEE Transactions on Signal Processing*, vol. 50, no. 5, pp. 1051 – 1064, May 2002.

[16] Md. Saifur Rahman and Aaron B. Wagner, "On the optimality of binning for distributed hypothesis testing," *IEEE Transactions on Information Theory*, vol. 58, no. 10, pp. 6282 – 6303, October 2012.

[17] Naftali Tishby, Fernando C. Pereira, and William Bialek, "The information bottleneck method," in *Proceedings of the* 37*-th Annual Allerton Conference on Communication, Control and Computing*, 1999, pp. 368 – 377.

[18] Peter Harremoes and Naftali Tishby, "The information bottleneck revisited or how to choose a good distortion measure," in *Proceedings of IEEE International Symposium on Information Theory*, June 2007, pp. 566 – 570.

[19] Richard E. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Transactions on Information Theory*, vol. IT-18, no. 4, pp. 460 – 473, July 1972.

[20] Suguru Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Transactions on Information Theory*, vol. IT-18, no. 1, pp. 14 – 20, January 1972.

[21] Gal Chechik, Amir Globerson, Naftali Tishby, and Yair Weiss, "Information bottleneck for Gaussian variables," *Journal of Machine Learning Research*, vol. 6, pp. 165 – 188, January 2005.

[22] Andreas Winkelbauer and Gerald Matz, "Rate-information-optimal Gaussian channel output compression," in *Proceedings of the* 48*-th Annual Conference on Information Sciences and Systems*, August 2014.

[23] Samuel Cheng, Vladimir Stankovic, and Zixiang Xiong, "Computing the channel capacity and rate-distortion function with two-sided state information," *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4418 – 4425, December 2005.

[24] Mung Chiang and Stephen Boyd, "Geometric programming duals of channel capacity and rate distortion," *IEEE Transactions on Information Theory*, vol. 50, no. 2, pp. 245 – 258, February 2004.

[25] Frederic Dupuis, Wei Yu, and Frans M. J. Willems, "Blahut-Arimoto algorithms for computing channel capacity and rate-distortion with side information," in *Proceedings of IEEE International Symposium on Information Theory*, June – July 2004, p. 181.

[26] Mohammad Rezaeian and Alex Grant, "A generalization of Arimoto-Blahut algorithm," in *Proceedings of IEEE International Symposium on Information Theory*, June – July 2004, p. 180.

[27] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner, "$\beta$-vae: Learning basic vi-

sual concepts with a constrained variational framework," in *Proceedings of the* 5*-th International Conference on Learning Representations*, 2017.

[28] Alexander A. Alemi, Ben Poole, Ian Fischer, Ian Fischer, Joshua V. Dillon, Rif A. Saurous, and Kevin Murphy, "Fixing a broken ELBO," in *Proceedings of the* 35*-th International Conference on Machine Learning*, 2018.

[29] Diederik P. Kingma and Max Welling, "Auto-encoding variational bayes," in *Proceedings of the* 2*-nd International Conference on Learning Representations*, 2014.

[30] Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy, "Deep variational information bottleneck," in *Proceedings of the* 5*-th International Conference on Learning Representations*, 2017.

[31] Zhuxi Jiang, Yin Zheng, Huachun Tan, Bangsheng Tang, and Hanning Zhou, "Variational deep embedding: An unsupervised and generative approach to clustering," in *Proceedings of the* 26*-th International Joint Conference on Artificial Intelligence*, 2017, pp. 1965 – 1972.

[32] Noam Slonim, *The Information Bottleneck: Theory and Applications.* PhD dissertation, Hebrew University, 2002.

[33] Junyuan Xie, Ross Girshick, and Ali Farhadi, "Unsupervised deep embedding for clustering analysis," in *Proceedings of the* 33*-rd International Conference on Machine Learning*, 2016, pp. 478 – 487.

[34] Hans S. Witsenhausen, "Indirect rate distortion problems," *IEEE Transactions on Information Theory*, vol. IT-26, no. 5, pp. 518 – 521, September 1980.

[35] Yossef Steinberg, "Coding and common reconstruction," *IEEE Transactions on Information Theory*, vol. 55, no. 11, pp. 4995 – 5010, November 2009.

[36] Ilan Sutskover, Shlomo Shamai (Shitz), and Jacob Ziv, "Extremes of information combining," *IEEE Transactions on Information Theory*, vol. 51, no. 4, pp. 1313 – 1325, April 2005.

[37] Ingmar Land and Johannes Huber, "Information combining," *Foundations and Trends in Communication and Information Theory*, vol. 3, no. 3, pp. 227 – 330, November 2006.

[38] Ingmar Land, Simon Huettinger, Peter A. Hoeher, and Johannes B. Huber, "Bounds on information combining," *IEEE Transactions on Information Theory*, vol. 51, no. 2, pp. 612 – 619, February 2005.

[39] Aaron D. Wyner, "On source coding with side information at the decoder," *IEEE Transactions on Information Theory*, vol. 21, no. 3, pp. 294 – 300, May 1975.

[40] Rudolf Ahlswede and Janos Korner, "Source coding with side information and a converse for degraded broadcast channels," *IEEE Transactions on Information Theory*, vol. 21, no. 6, pp. 629 – 637, November 1975.

[41] Elza Erkip and Thomas Cover, "The efficiency of investment information," *IEEE Transactions on Information Theory*, vol. 44, no. 3, pp. 1026 – 1040, May 1998.

[42] Ali Makhdoumi, Salman Salamatian, Nadia Fawaz, and Muriel Medard, "From the information bottleneck to the privacy funnel," in *Proceedings of IEEE Information Theory Workshop*, November 2014, pp. 501 – 505.

[43] Yoshua Bengio, Aaron Courville, and Pascal Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798 – 1828, August 2013.

[44] Chang Xu, Dacheng Tao, and Chao Xu, "A survey on multi-view learning," *arXiv: 1304.5634*, 2013.

[45] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes, "On deep multi-view representation learning," in *Proceedings of the* 32*-nd International Conference on Machine Learning*, 2015.

[46] Yann Lecun, Leon Bottou, Yoshua Bengio, and Patrick Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, vol. 86, no. 11, 1998, pp. 2278 – 2324.

[47] David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li, "A new benchmark collection for text categorization research," *The Journal of Machine Learning Research*, vol. 5, pp. 361 – 397, 2004.

[48] Adam Coates, Andrew Ng, and Honglak Lee, "An analysis of single-layer networks in unsupervised feature learning," in *Proceedings of the* 14*-th International Conference on Artificial Intelligence and Statistics*, 2011, pp. 215 – 223.

[49] Georg Pichler, Pablo Piantanida, and Gerald Matz, "Distributed information-theoretic biclustering," in *Proceedings of IEEE International Symposium on Information Theory*, July 2016, pp. 1083 – 1087.

[50] Georg Pichler, Pablo Piantanida, and Gerald Matz, "A multiple description CEO problem with log-loss distortion," in *Proceedings of IEEE International Symposium on Information Theory*, June 2017, pp. 111 – 115.

[51] Jiantao Jiao, Thomas A. Courtade, Kartik Venkat, and Tsachy Weissman, "Justification of logarithmic loss via the benefit of side information," *IEEE Transactions on Information Theory*, vol. 61, no. 10, pp. 5357 – 5365, October 2015.

[52] Albert No and Tsachy Weissman, "Universality of logarithmic loss in lossy compression," in *Proceedings of IEEE International Symposium on Information Theory*, June 2015, pp. 2166 – 2170.

[53] Yanina Shkel, Maxim Raginsky, and Sergio Verdu, "Universal lossy compression under logarithmic loss," in *Proceedings of IEEE International Symposium on Information Theory*, June 2017, pp. 1157 – 1161.

[54] Nicolo Cesa-Bianchi and Gabor Lugosi, *Prediction, Learning and Games.* Cambridge University Press, 2006.

[55] Thomas Andre, Marc Antonini, Michel Barlaud, and Robert M. Gray, "Entropy-based distortion measure for image coding," in *Proceedings of IEEE International Conference on Image Processing*, October 2006, pp. 1157 – 1160.

[56] Kittipong Kittichokechai, Yeow-Khiang Chia, Tobias J. Oechtering, Mikael Skoglund, and Tsachy Weissman, "Secure source coding with a public helper," *IEEE Transactions on Information Theory*, vol. 62, no. 7, pp. 3930 – 3949, July 2016.

[57] Amichai Painsky and Gregory Wornell, "On the universality of the logistic loss function," in *Proceedings of IEEE International Symposium on Information Theory*, June 2018, pp. 936 – 940.

[58] Cheuk Ting Li, Xiugang Wu, Ayfer Ozgur, and Abbas El Gamal, "Minimax learning for remote prediction," in *Proceedings of IEEE International Symposium on Information Theory*, June 2018, pp. 541 – 545.

[59] Chao Tian and Jun Chen, "Remote vector Gaussian source coding with decoder side information under mutual information and distortion constraints," *IEEE Transactions on Information Theory*, vol. 55, no. 10, pp. 4676 – 4680, October 2009.

[60] Amichai Sanderovich, Shlomo Shamai (Shitz), Yossef Steinberg, and Gerhard Kramer, "Communication via decentralized processing," *IEEE Transactions on Information Theory*, vol. 54, no. 7, pp. 3008 – 3023, July 2008.

[61] Osvaldo Simeone, Elza Erkip, and Shlomo Shamai (Shitz), "On codebook information for interference relay channels with out-of-band relaying," *IEEE Transactions on Information Theory*, vol. 57, no. 5, pp. 2880 – 2888, May 2011.

[62] Inaki Estella Aguerri, Abdellatif Zaidi, Giuseppe Caire, and Shlomo Shamai (Shitz), "On the capacity of cloud radio access networks with oblivious relaying," in *Proceedings of IEEE International Symposium on Information Theory*, June 2017, pp. 2068 – 2072.

[63] Inaki Estella Aguerri, Abdellatif Zaidi, Giuseppe Caire, and Shlomo Shamai (Shitz), "On the capacity of cloud radio access networks with oblivious relaying," *IEEE Transactions on Information Theory*, vol. 65, no. 7, pp. 4575 – 4596, July 2019.

[64] Flavio P. Calmon, Ali Makhdoumi, Muriel Medard, Mayank Varia, Mark Christiansen, and Ken R. Duffy, "Principal inertia components and applications," *IEEE Transactions on Information Theory*, vol. 63, no. 8, pp. 5011 – 5038, July 2017.

[65] Rudolf Ahlswede and Imre Csiszar, "Hypothesis testing with communication constraints," *IEEE Transactions on Information Theory*, vol. IT - 32, no. 4, pp. 533 – 542, July 1986.

[66] Te Sun Han, "Hypothesis testing with multiterminal data compression," *IEEE Transactions on Information Theory*, vol. IT - 33, no. 6, pp. 759 – 772, November 1987.

[67] Chao Tian and Jun Chen, "Successive refinement for hypothesis testing and lossless one-helper problem," *IEEE Transactions on Information Theory*, vol. 54, no. 10, pp. 4666 – 4681, October 2008.

[68] Sadaf Salehkalaibar, Michele Wigger, and Roy Timo, "On hypothesis testing against conditional independence with multiple decision centers," *IEEE Transactions on Communications*, vol. 66, no. 6, pp. 2409 – 2420, June 2018.

[69] Ran Gilad-Bachrach, Amir Navot, and Naftali Tishby, "An information theoretic tradeoff between complexity and accuracy," in *Proceedings of Conference on Learning Theory*, 2003, pp. 595 – 609.

[70] Andreas Winkelbauer, Stefan Farthofer, and Gerald Matz, "The rate-information trade-off for Gaussian vector channels," in *Proceedings of IEEE International Symposium on Information Theory*, June 2014, pp. 2849 – 2853.

[71] Michael Meidlinger, Andreas Winkelbauer, and Gerald Matz, "On the relation between the Gaussian information bottleneck and MSE-optimal rate-distortion quantization," in *Proceedings of IEEE Workshop on Statistical Signal Processing*, June 2014, pp. 89 – 92.

[72] Abdellatif Zaidi, Inaki Estella Aguerri, and Shlomo Shamai (Shitz), "On the information bottleneck problems: Models, connections, applications and information theoretic views," *Entropy*, vol. 22, no. 2, p. 151, January 2020.

[73] Aaron D. Wyner and Jacob Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Transactions on Information Theory*, vol. IT – 22, no. 1, pp. 1 – 10, January 1976.

[74] Meryem Benammar and Abdellatif Zaidi, "Rate-distortion of a Heegard-Berger problem with common reconstruction constraint," in *Proceedings of International Zurich Seminar on Communications*, 2016, pp. 150 – 154.

[75] Meryem Benammar and Abdellatif Zaidi, "Rate-distortion function for a Heegard-Berger problem with two sources and degraded reconstruction sets," *IEEE Transactions on Information Theory*, vol. 62, no. 9, pp. 5080 – 5092, September 2016.

[76] Flavio du Pin Calmon and Nadia Fawaz, "Privacy against statistical inference," in *Proceedings of the* 50*-th Annual Allerton Conference on Communication, Control and Computing*, October 2012, pp. 1401 – 1408.

[77] Shahab Asoodeh, Mario Diaz, Fady Alajaji, and Tamas Linder, "Information extraction under privacy constraints," *Information*, vol. 7, no. 15, March 2016.

[78] Alessandro Achille and Stefano Soatto, "Information dropout: Learning optimal representations through noisy computation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 12, pp. 2897 – 2905, December 2018.

[79] Satosi Watanabe, "Information theoretical analysis of multivariate correlation," *IBM Journal of Research and Development*, vol. 4, no. 1, pp. 66 – 82, January 1960.

[80] Ricky T. Q. Chen, Xuechen Li, Roger Grosse, and David Duvenaud, "Isolating sources of disentanglement in VAEs," in *Proceedings of the* 32*-nd Conference on Neural Information Processing Systems*, 2018.

[81] Ohad Shamir, Sivan Sabato, and Naftali Tishby, "Learning and generalization with the information bottleneck," in *Proceedings of the* 19*-th International Conference on Algorithmic Learning Theory*, October 2008, pp. 92 – 107.

[82] Naftali Tishby and Noga Zaslavsky, "Deep learning and the information bottleneck principle," in *Proceedings of IEEE Information Theory Workshop*, April 2015.

[83] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *Proceedings of the* 3*-rd International Conference on Learning Representations*, 2015.

[84] Ravid Schwartz-Ziv and Naftali Tishby, "Opening the black box of deep neural networks via information," *arXiv: 1703.00810*, 2017.

[85] Andrew Michael Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan Daniel Tracey, and David Daniel Cox, "On the information bottleneck theory of deep learning," in *Proceedings of the* 6*-th International Conference on Learning Representations*, 2018.

[86] Lewandowsky and Gerhard Bauch, "Information-optimum LDPC decoders based on the information bottleneck method," *IEEE Access*, vol. 6, pp. 4054 – 4071, 2018.

[87] Michael Meidlinger, Alexios Balatsoukas-Stimming, Andreas Burg, and Gerald Matz, "Quantized message passing for LDPC codes," in *Proceedings of* 49*-th Asilomar Conference on Signals, Systems, and Computers*, November 2015, pp. 1606 – 1610.

[88] J. Korner and K. Marton, “How to encode the modulo-two sum of binary sources,” *IEEE Transactions on Information Theory*, vol. 25, no. 02, pp. 219 – 221, March 1979.

[89] Michael Gastpar, “The Wyner-Ziv problem with multiple sources,” *IEEE Transactions on Information Theory*, vol. 50, no. 11, pp. 2762 – 2768, November 2004.

[90] Daniel Russo and James Zou, “How much does your data exploration overfit? Controlling bias via information usage,” *arXiv: 1511.05219*, 2015.

[91] Aolin Xu and Maxim Raginsky, “Information-theoretic analysis of generalization capability of learning algorithms,” in *Proceedings of the* 31*-st Conference on Neural Information Processing Systems*, 2017, pp. 2524 – 2533.

[92] Amir R. Asadi, Emmanuel Abbe, and Sergio Verdu, “Chaining mutual information and tightening generalization bounds,” in *Proceedings of the* 32*-nd Conference on Neural Information Processing Systems*, 2018.

[93] Toby Berger, “Decentralized estimation and decision theory,” in *Proceedings of IEEE Spring Workshop on Information Theory*, 1979.

[94] Hossam M. H. Shalaby and Adrian Papamarcou, “Multiterminal detection with zero-rate data compression,” *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 254 – 267, March 1992.

[95] Wenwen Zhao and Lifeng Lai, “Distributed testing with zero-rate compression,” in *Proceedings of IEEE International Symposium on Information Theory*, June 2015, pp. 2792 – 2796.

[96] Pierre Escamilla, Michele Wigger, and Abdellatif Zaidi, “Distributed hypothesis testing with concurrent detections,” in *Proceedings of IEEE International Symposium on Information Theory*, June 2018, pp. 166 – 170.

[97] Pierre Escamilla, Michele Wigger, and Abdellatif Zaidi, “Distributed hypothesis testing with collaborative detection,” in *Proceedings of the* 56*-th Annual Allerton Conference on Communication, Control, and Computing*, October 2018, pp. 512 – 518.

[98] Jiachun Liao, Lalitha Sankar, Flavio P. Calmon, and Vincent Y. F. Tan, “Hypothesis testing under maximal leakage privacy constraints,” in *Proceedings of IEEE International Symposium on Information Theory*, June 2017, pp. 779 – 783.

[99] Sreejith Sreekumar, Asaf Cohen, and Deniz Gunduz, "Distributed hypothesis testing with a privacy constraint," in *Proceedings of IEEE Information Theory Workshop*, November 2018.

[100] Abdellatif Zaidi and Inaki Estella Aguerri, "Optimal rate-exponent region for a class of hypothesis testing against conditional independence problems," in *Proceedings of IEEE Information Theory Workshop*, August 2019.

[101] Toby Berger, *Multiterminal source coding.* The Information Theory Approach to Communications, CSIM Courses and Lectures, 1978, vol. 229.

[102] S. Y. Tung, *Multiterminal source coding.* PhD dissertation, Cornell University, 1978.

[103] Yuhan Zhou, Yinfei Xu, Wei Yu, and Jun Chen, "On the optimal fronthaul compression and decoding strategies for uplink cloud radio access networks," *IEEE Transactions on Information Theory*, vol. 62, no. 12, pp. 7402 – 7418, December 2016.

[104] Thomas A. Courtade, "Gaussian multiterminal source coding through the lens of logarithmic loss," in *Information Theory and Applications Workshop*, 2015.

[105] Thomas A. Courtade, "A strong entropy power inequality," *IEEE Transactions on Information Theory*, vol. 64, no. 4, pp. 2173 – 2192, April 2018.

[106] Aaron B. Wagner, Saurabha Tavildar, and Pramod Viswanath, "Rate region of the quadratic Gaussian two-encoder source-coding problem," *IEEE Transactions on Information Theory*, vol. 54, no. 5, pp. 1938 – 1961, May 2008.

[107] Thomas A. Courtade and Jiantao Jiao, "An extremal inequality for long Markov chains," in *Proceedings of the* 52*-nd Annual Allerton Conference on Communication, Control and Computing*, September 2014, pp. 763 – 770.

[108] Y. Oohama, "The rate-distortion function for the quadratic gaussian ceo problem," *IEEE Transactions on Information Theory*, vol. 44, no. 3, pp. 1057 – 1070, May 1998.

[109] Saurabha Tavildar, Pramod Viswanath, and Aaron B. Wagner, "The gaussian many-help-one distributed source coding problem," *IEEE Transactions on Information Theory*, vol. 56, no. 1, pp. 564 – 581, January 2010.

[110] Md. Saifur Rahman and Aaron B. Wagner, “Rate region of the vector gaussian one-helper source-coding problem,” *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2708 – 2728, May 2015.

[111] Inaki Estella Aguerri and Abdellatif Zaidi, “Distributed information bottleneck method for discrete and Gaussian sources,” in *Proceedings of International Zurich Seminar on Information and Communication*, February 2018.

[112] Noam Slonim and Naftali Tishby, “The power of word clusters for text classification,” in *Proceedings of* 23*-rd European Colloquium on Information Retrieval Research*, 2001, pp. 191 – 200.

[113] Yoram Baram, Ran El-Yaniv, and Kobi Luz, “Online choice of active learning algorithms,” *Journal of Machine Learning Research*, vol. 5, pp. 255 – 291, March 2004.

[114] Jun Chen and Toby Berger, “Successive Wyner-Ziv coding scheme and its application to the quadratic Gaussian CEO problem,” *IEEE Transactions on Information Theory*, vol. 54, no. 4, pp. 1586 – 1603, April 2008.

[115] Meisam Razaviyayn, Mingyi Hong, and Zhi-Quan Luo, “A unified convergence analysis of block successive minimization methods for nonsmooth optimization,” *SIAM Journal on Optimization*, vol. 23, no. 2, pp. 1126 – 1153, June 2013.

[116] Michael Grant and Stephen Boyd, “CVX: Matlab software for disciplined convex programming,” http://cvxr.com/cvx, March 2014.

[117] Matthew Chalk, Olivier Marre, and Gasper Tkacik, “Relevant sparse codes with variational information bottleneck,” in *Proceedings of the* 30*-th Conference on Neural Information Processing Systems*, 2016.

[118] Xue Bin Peng, Angjoo Kanazawa, Sam Toyer, Pieter Abbeel, and Sergey Levine, “Variational discriminator bottleneck: Improving imitation learning, inverse rl, and gans by constraining information flow,” in *Proceedings of the* 7*-th International Conference on Learning Representations*, 2019.

[119] Bin Dai, Chen Zhu, and David P. Wipf, “Compressing neural networks using the variational information bottleneck,” in *Proceedings of the* 35*-th International Conference on Machine Learning*, 2018.

[120] Diederik P. Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling, "Improved variational inference with inverse autoregressive flow," in *Proceedings of* 30*-st Conference on Neural Information Processing Systems*, 2016.

[121] George Papamakarios, Theo Pavlakou, and Iain Murray, "Masked autoregressive flow for density estimation," in *Proceedings of* 31*-st Conference on Neural Information Processing Systems*, 2017.

[122] D. Sculley, "Web-scale $K$-means clustering," in *Proceedings of the* 19*-th International Conference on World Wide Web*, April 2010, pp. 1177 – 1178.

[123] Zhexue Huang, "Extensions to the $K$-means algorithm for clustering large datasets with categorical values," *Data Mining and Knowledge Discovery*, vol. 2, no. 3, pp. 283 – 304, September 1998.

[124] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A $k$-means clustering algorithm," *Journal of the Royal Statistical Society*, vol. 28, pp. 100 – 108, 1979.

[125] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, pp. 1 – 38, 1977.

[126] Chris Ding and Xiaofeng He, "$K$-means clustering via principal component analysis," in *Proceedings of the* 21*-st International Conference on Machine Learning*, 2004.

[127] Karl Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine*, vol. 2, no. 11, pp. 559 – 572, November 1901.

[128] Svante Wold, Kim Esbensen, and Paul Geladi, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, pp. 37 – 52, August 1987.

[129] Sam Roweis, "EM algorithms for PCA and SPCA," in *Advances in Neural Information Processing Systems* 10, 1997, pp. 626 – 632.

[130] Thomas Hofmann, Bernhard Schölkopf, and Alexander J. Smola, "Kernel methods in machine learning," *The Annals of Statistics*, vol. 36, pp. 1171 – 1220, June 2008.

[131] N. Slonim and N. Tishby, "Document clustering using word clusters via the information bottleneck method," in *Proceedings of the* 23*-rd Annual International ACM*

*SIGIR Conference on Research and Development in Information Retrieval*, July 2000, pp. 208 – 215.

[132] Danilo J. Rezende, Shakir Mohamed, and Daan Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *Proceedings of the* 31*-st International Conference on Machine Learning*, 2014, pp. 1278 – 1286.

[133] Xifeng Guo, Long Gao, Xinwang Liu, and Jianping Yin, "Improved deep embedded clustering with local structure preservation," in *Proceedings of the* 26*-th International Joint Conference on Artificial Intelligence*, 2017, pp. 1753 – 1759.

[134] Nat Dilokthanakul, Pedro A. M. Mediano, Marta Garnelo, Matthew C.H. Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahani, "Deep unsupervised clustering with Gaussian mixture variational autoencoders," *arXiv: 1611.02648*, 2017.

[135] Erxue Min, Xifeng Guo, Qiang Liu, Gen Zhang, Jianjing Cui, and Jun Long, "A survey of clustering with deep learning: From the perspective of network architecture," *IEEE Access*, vol. 6, pp. 39 501 – 39 514, 2018.

[136] John R. Hershey and Peder A. Olsen, "Approximating the Kullback Leibler divergence between Gaussian mixture models," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, April 2007, pp. 317 – 320.

[137] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770 – 778.

[138] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *The Journal of Machine Learning Research*, vol. 11, pp. 3371 – 3408, December 2010.

[139] Laurens van der Maaten and Geoffrey Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research 9*, pp. 2579 – 2605, November 2008.

[140] Cheuk Ting Li and Abbas El Gamal, "Strong functional representation lemma and applications to coding theorems," *IEEE Transactions on Information Theory*, vol. 64, no. 11, pp. 6967 – 6978, November 2018.

[141] Cheuk Ting Li, Xiugang Wu, Ayfer Ozgur, and Abbas El Gama, “Minimax learning for remote prediction,” *arXiv: 1806.00071*, 2018.

[142] Adi Homri, Michael Peleg, and Shlomo Shamai (Shitz), “Oblivious fronthaul-constrained relay for a Gaussian channel,” *IEEE Transactions on Communications*, vol. 66, no. 11, pp. 5112 – 5123, November 2018.

[143] Roy Karasik, Osvaldo Simeone, and Shlomo Shamai (Shitz), “Robust uplink communications over fading channels with variable backhaul connectivity,” *IEEE Transactions on Wireless Communications*, vol. 12, no. 11, pp. 5788 – 5799, November 2013.

[144] Yuxin Chen, Andrea J. Goldsmith, and Yonina C. Eldar, “Channel capacity under sub-nyquist nonuniform sampling,” *IEEE Transactions on Information Theory*, vol. 60, no. 8, pp. 4739 – 4756, August 2014.

[145] Alon Kipnis, Yonina C. Eldar, and Andrea J. Goldsmith, “Analog-to-digital compression: A new paradigm for converting signals to bits,” *IEEE Signal Processing Magazine*, vol. 35, no. 3, pp. 16 – 39, May 2018.

[146] Michael Gastpar, “On Wyner-Ziv networks,” in *Proceedings of* 37*-th Asilomar Conference on Signals, Systems, and Computers*, November 2003, pp. 855 – 859.

[147] Amir Dembo, Thomas M. Cover, and Joy A. Thomas, “Information theoretic inequalities,” *IEEE Transactions on Information Theory*, vol. 37, no. 6, pp. 1501 – 1518, November 1991.

[148] Daniel P. Palomar and Sergio Verdu, “Gradient of mutual information in linear vector gaussian channels,” *IEEE Transactions on Information Theory*, vol. 52, no. 1, pp. 141 – 154, January 2006.

# Publications


[Y1] Yiğit Uğur, Iñaki Estella Aguerri, and Abdellatif Zaidi, "Vector Gaussian CEO problem under logarithmic loss and applications," accepted for publication in *IEEE Transactions on Information Theory*, January 2020.

[Y2] Yiğit Uğur, Iñaki Estella Aguerri, and Abdellatif Zaidi, "A generalization of Blahut-Arimoto algorithm to compute rate-distortion regions of multiterminal source coding under logarithmic loss," in *Proceedings of IEEE Information Theory Workshop*, November 2017, pp. 349 – 353.

[Y3] Yiğit Uğur, Iñaki Estella Aguerri, and Abdellatif Zaidi, "Vector Gaussian CEO problem under logarithmic loss," in *Proceedings of IEEE Information Theory Workshop*, November 2018, pp. 515 – 519.

[Y4] Yiğit Uğur, George Arvanitakis, and Abdellatif Zaidi, "Variational information bottleneck for unsupervised clustering: Deep Gaussian mixture embedding," *Entropy*, vol. 22, no. 2, p. 213, February 2020.