The Minimum Information Principle in Learning and Neural Data Analysis

Thesis submitted for the degree
"Doctor of Philosophy"
by

Amir Globerson

Submitted to the Senate of the Hebrew University

September 2005

This work was carried out under the supervision of Prof. Naftali Tishby and Prof. Eilon Vaadia

Acknowledgments

First and foremost, I would like to thank my advisor Tali Tishby. Tali's enthusiasm for understanding fundamental principles, and seeing their implications to a wide array of scientific fields has been a continuous source of inspiration. His strong intuition about the structure of solutions to problems have often proved themselves, although it often took me a while before I could formalize what he envisioned. My research has taken a while, and I must acknowledge and thank Tali for being patient and believing that good ideas take time to flourish. I hope this was the case here.

Eilon Vaadia, my second advisor, has been very supportive in our joint quest to apply information theoretic ideas to neuroscience. He has helped me to see things through the eyes of an experimentalist, and to understand where novel methods could be most helpful. I would also like to thank Eilon for his great contribution for making the ICNC what it is, namely a great program, with true enthusiasm for big questions, and a superb group of students.

To my roommates Amir Navot and Ran Gilad-Bacharch, whose razor sharp intellect and knowledge of any subject under the sun, made them always fun to be around and to learn from.

The Machine Learning "corridor" was the best working environment I could hope for: a bunch of ultra-clever people, who were also fun to be around. Specifically, I would like to thank Michael Fink, Koby Crammer, Shai Shalev-Schwartz, Elad Schneidman, Naama Parush, Lavi Shpigelman, Tomer Hertz, Noam Slonim, Michal Rosen-Zvi and Yoseph Barash, for both their professional and personal presence. I would also like to thank Neta Zach and Rony Paz from Eilon's lab for sharing their data, and helping me look at it properly.

Two fellow student collaborators have had a major impact on my work. Gal Chechik has served as a good friend/mentor/collaborator. He has had a great impact on the way I think about things, both professionally and personaly. I've learned more from him than I am able to recall, and cannot thank him enough.

Eran Stark, who is an MD/PhD and student of Moshe Abeles, has also been a close collaborator. My talks with Eran resulted in much of the application of this work to neural data analysis, and working with him has been, and still is, pure pleasure.

I have also had the good fortune to collaborate with Yair Weiss on the Gaussian Information Bottleneck work, which was initially a project in his excellent course on graphical models. A visit to the University of Pennsilvania resulted in a fruitful collaboration with Fernando Pereira, who has sparked my enthusiasm and interest in Natural Language Processing. I would also like to thank John Blitzer from UPenn, who is always fun to work and to hang out with.

I am forever indebted to a group of close friends, who have been there all along, offering support and wise words at rough times. To Neta Martan, Ronen Golan, Ranit Aharonov, Robert Guetig, Udi Biener, Erel Morris, and Tal El-Hay. I am priviliged to have you around.

To Oriti, for the two wonderful years we spent together.

To my brother Eitan, who has infused me with his intellectual curiosity from an early age. To my sister Drorit, who is always there for me, and has been a pillar of support and caring for all these years.

Finally, to my dear parents, Dov and Amiela Globerson. Thank you for your endless love and support. I owe so much of what I am to you. I love you both with all my heart, and wish all our family nothing but happiness in the years to come.

Abstract

As science enters the 21st century, the study of complex systems is increasingly gaining importance. Complex systems are characterized by an extremely large number of units (many millions) which cooperate to produce complex observable behavior. Perhaps the most striking example of such a system is the human brain, a network of 10^{10} nerve cells whose simultaneous activity results in all human functions, from locomotion to emotion. *Understanding* the brain involves several extremely difficult paradigmatic and experimental challenges. At the current point of time, it is impossible to observe the activity of more than a few hundred cells simultaneously. Furthermore, even if one could obtain such a *"Brain TV"* which reports the simultaneous activity of all nerve cells in the brain, it is not clear that such a device would lead to a satisfactory explanation as to how the brain operates.

These conceptual difficulties are not unique to brain research. They are manifested in the study of other complex systems such as biological networks (i.e., metabolic pathways), weather forecasting, and huge document databases such as the world wide web. *Machine learning* is an important field of research which studies fundamental theoretic and algorithmic aspects of extracting rules from empirical measurements of such systems. We make frequent use of machine learning concepts throughout the dissertation.

It is clear that any empirical measurement of complex systems is bound to be partial, in the sense that only a subset of its units may be measured at a given time [1]. Furthermore, experimental procedures typically limit the duration for which one can observe a given unit [2]. We are thus faced with the following fundamental question: what can one say about a system, given such a set of partial observations? An immediate extension of this question is how to choose the observations which would be most beneficial for studying a system. These two questions, and their various extensions are the focus of this dissertation.

One of the central formal tools in the current thesis is Information Theory, introduced by Claude Shannon in the 1940's as a comprehensive mathematical theory of

[1] Perhaps with the exception of the world wide web.

[2] This state of affairs may be considerably enhanced in the future, with the improvement of chronically implanted electrodes [39].

communications. Information Theory allows one to quantify information transmission in systems and is thus an attractive tool for studying information processing systems like the brain. Substantial literature exists on estimating information between neural activity and the external wold, in an effort to understand how the world is encoded in cortical activity.

However, information theoretic methods also face two difficulties when applied to complex system analysis: one is the partial measurement problem described above, and the other is the need to understand which properties of the system are important for its function (for example: what is the information in single neuron responses as compared to information in neuronal correlations). The current dissertation deals with these methodological issues by designing tools for measuring information under partial measurement and in specific properties of the system. The new Minimum Mutual Information (MinMI) principle , developed here, quantifies information in these scenarios by considering all possible systems with a given property, and returning the minimum information in this set of systems. The resulting number captures the information in the given property, since systems with higher information necessarily contain additional, information enhancing properties. Furthermore, it is a lower bound on the information in the system whose partial properties were measured.

The first chapter of the thesis covers previous approaches to the problem of inference from partial measurements. This problem was previously addressed in the framework of the Maximum Entropy (MaxEnt) principle, developed by Maxwell and Jaynes [73], and more recently in the machine learning [34], and neural coding [122] literature. MaxEnt is similar to MinMI, in that it addresses all systems with a given property. However, unlike MinMI, MaxEnt returns the system that maximizes entropy, which is an information theoretic measure of uncertainty. The MinMI principle is the natural extension of MaxEnt methods to information processing systems, since it is targeted at information measurement. The relation between these two principles is discussed in Chapters 2 and 3. Chapter 1 also covers basic concepts in information theory and geometry of distributions, which will be used throughout the thesis.

The MinMI principle is presented in Chapter 2. The general form of its solution is derived, and several algorithms are given for calculating its parameters. The algorithms are based on geometric projections of distributions (*I-projections*), and con-

vergence proofs are given. MinMI may also be used as a classification algorithm. The chapter describes this approach, and provides upper bounds on its generalization error. Furthermore, we prove a theorem showing that the MinMI classifier is optimal in a game theoretic sense, since it minimizes the worst case loss in a game which we define. The proof of the theorems uses concepts from convex duality. We conclude with empirical results on classification tasks, where MinMI outperforms other methods on a subset of the databases.

Chapter 3 discusses the application of MinMI to studying the neural code, and analyzing data obtained in neuro-physiological experiments. We show how MinMI may be used to measure information in properties of the neural response, such as single cell responses. The results demonstrate that MinMI can differentiate between populations where neurons have similar codes and those in which their codes differ. A similar analysis is performed for pairwise responses. Finally, we show that MinMI can quantify the information in the neuronal temporal response profiles. This allows us to detect neurons whose temporal response provides information about the stimulus, increasing the number of informative neurons by 35% for data recorded in the motor cortex of behaving monkeys. MinMI extends current information theoretic methods in neuroscience, by yielding a measure of information in various scenarios which are not covered by currently available methods. We discuss its various advantages over existing approaches.

In Chapters 2-3 we assume that the partial measurements are determined in advance (e.g., the responses of single cells). Interestingly, the method can be extended (under some assumptions) to finding the optimal set of measurements, i.e. those which contain the maximum amount of information. In Chapter 4 we introduce this maximization problem and its algorithmic solution, also based on geometric projections of distributions. The method is named Sufficient Dimensionality Reduction (SDR), due to its close link to the concept of sufficient statistics. We apply SDR to several text analysis tasks, and show how it can be used to extract meaningful features in large datasets. In Chapter 5, SDR is extended to the case where data about the noise structure is available, so that one can learn to disregard features which describe the noise (e.g., illumination conditions in a face recognition task). Experimental results show that the method improves performance in image recognition tasks.

Another key question with respect to information processing is the tradeoff between accurate representation of the external world, and the complexity of this representation. The theoretical aspects of this tradeoff have been studied in the context of information theory, and more recently using the Information Bottleneck approach [135]. In Chapter 6, we present an analytical characterization of this tradeoff for the case of Gaussian variables. The analysis demonstrates how the representation dimension is a natural outcome of this tradeoff: more accurate representations require more dimensions than the less accurate ones, and the dimension is increased in a continuous manner. Finally, we provide an algorithm for finding the most accurate representation for a given complexity level.

We apply our novel tools to a wide array of applications, from studying the neural code to analyzing documents and images, and discuss their advantages in such cases.

Taken together, our results serve to illustrate the utility and importance of information theoretic concepts, and specifically mutual information minimization, in machine learning and in analyzing data measured in complex systems.

Contents

Chapter 1

Introduction

The human brain is one of the most complex machines in the universe. Using a network of 10^{10} nerve cells, the brain generates all aspects of human behavior from moving in the environment to deciphering the laws of nature and composing symphonies or literary masterpieces. We are still far from a reasonable understanding of how the brain achieves all these, although much progress has been made over the last century.

The difficulties encountered by the brain researcher are numerous. The first is the technical difficulty of measuring physical activity in the brain. It is currently possible to record currents from hundreds of cells simultaneously using chronically implanted electrodes [22] and to measure large scale activity of brain regions using Magnetic Resonance Imaging (MRI). However, with all these methods we are still far from a complete characterization of the activity of all neurons in a small brain patch.

The second difficulty is a conceptual one. Even if we had access to a complete description of neural activity, it is far from obvious that this would endow us with an *understanding* of how the brain generates behavior. A useful metaphor comes from statistical physics: imagine we could measure the location of each gas molecule in a container. Would this yield understanding of the properties of the gas? Thus, experimental measurements must be supplemented with appropriate conceptual tools in order to achieve understanding. The search for the basic mechanisms underlying the transformation between neural activity and behavior is often referred to as the problem of studying the *neural code* [116].

A theoretical tool which has proven useful in studying the neural code is information theory. Introduced by Claude Shannon in 1948 as a mathematical theory of communication [123], information theory formally quantified such notions as compression, transmission over noisy channels and information processing. Since the brain may be interpreted as a complex information processing machine whose input is sensual experience and output is behavior, it was only natural for neuro-scientists to be interested in this general theory.

One of the first uses of information theory in neuroscience was in Miller's famous

"Magic Number 7" paper [93], where information theoretic concepts were used to quantify the limits on short term memory. In later years, information theoretic studies were used to study different properties of neural coding from temporal aspects [101] to population coding [113].

While the information theoretic approach has advanced our understanding of neural coding principles, its application is still limited in several important aspects. The first, technical limitation, is the typical need for many repetitions of an experiment in order to calculate information theoretic measures. The second, explanatory limitation, is that although we may know information exists in a given neural activity, it is not always clear what properties of the activity carry that information.

One of the goals of the current dissertation was to help in developing a theory to approach these two problems, and study its various extensions. The current chapter introduces basic concepts to be used throughout the thesis.

1.1 General Setup

We shall consider systems which can be generally divided into two observed random variables, X and Y, whose interrelationship we are interested in studying.

For example, Y can be a stimulus presented to the brain, and X the neural response. Alternatively, Y can be the motor output of the brain. Another setup which will be considered extensively in what follows is the classification problem where Y is a class variable and X is the value of a set of object features. Common applications in this scenario include speech recognition, face recognition, document categorization etc. Thus, in the face recognition example X could be an image of a person [1], and Y the identity of that person.

In what follows, we assume that X and Y are discrete variables, unless otherwise specified (see Chapter 6). Since X and Y are random variables, their dependence can be described by a joint distribution $p(X = x, Y = y)$, which we abbreviate by $p(x, y)$. A central motivation for the approach presented in this thesis is that the distribution $p(x, y)$ cannot be reliably estimated in many cases. For example consider a population of 100 neurons whose individual responses are denoted by X_i $(i = 1 \dots 100)$. The total response is then $X = [X_1, \dots, X_{100}]$, a random variable with 2^{100} possible values. It is clear that the distribution $p(x, y)$ cannot be reliably estimated via any feasible experiment.

Even when $p(x, y)$ can be reliably estimated, it may be advantageous to have a description with more explanatory power than merely stating $|X||Y|$ numbers, where $|X|$ is the number of different values the variable X can take. One such option is to use a parametric model with a small set of parameters, preferably one which is motivated

[1] Represented by the grey level of pixels in the image for example.

by a model of the system. However, in many domains there is no natural parametric model.

1.2 Information Theory

Information theory was introduced by Claude Shannon in 1948 [123] as a mathematical theory of communication. The fundamental problem addressed by Shannon was how to transmit a source over a noisy channel such that it can be reconstructed with acceptable error, while making minimum use of the channel. Shannon showed that such transmission can in fact be carried out, and characterized the performance of the optimal coding schemes.

Shannon's theorems were not constructive in the sense that they only showed the limits on communication but not how they may be achieved in practice. This prompted decades of research into channel coding mechanisms, which is still going on to this day.

One of Shannon's basic insights was that most communication channels can be broken into two components. The first is the source X, which one wants to transmit (e.g., a set of images). The distribution of the source is denoted by $p(x)$ (low $p(x)$ indicates that x will only rarely be sent). The second component is a noisy channel, over which the transmission is carried out. A message X sent over a channel usually undergoes some degradation as a result of noise (e.g., a low-quality telephone line). Thus the output Y of the channel is usually a stochastic function of the input, and its behavior can be described via a distribution $p(y|x)$. Shannon realized that the physical properties of the source and channel are of no importance, given the two distributions $p(x)$ and $p(y|x)$. Thus the entire theory of communication relies on the properties of univariate and bivariate distributions.

Two functionals of distributions play a central role in information theory: the entropy and the mutual information.

Definition 1 *The entropy of a distribution $p(x)$ is defined as* [2] [3]

$$H[p(x)] = -\sum_x p(x) \log p(x) \ . \tag{1.1}$$

It will alternatively be written as $H(X)$ when the distribution $p(x)$ is clear from the context.

The entropy can be shown to be the answer to the following question: imagine a *20 questions* like game, where one player generates a value of X and the other needs

[2] The base used in the log changes the entropy by a multiplicative factor, and is of importance only in defining units. Here we will typically use the natural logarithm except when otherwise noted. Information is measured in *bits* when using base 2 in the log, and in *nats* when using the natural logarithm.

[3] Values of $p(x) = 0$ do not contribute to the sum since $0 \log 0 = 0$ in the limit.

to discover this value by asking yes/no questions. The second player naturally seeks to minimize the number of questions asked. It turns out that the minimum number of questions that can be asked on average [4] is bounded between $H(X)$ and $H(X)+1$ [27] [5]. In a communication setting this scenario can be described as: given a random variable X, the minimum average number of bits needed to encode X is bounded between $H(X)$ and $H(X)+1$. These properties suggest that entropy is the degree of uncertainty one has about the variable X before observing it. Furthermore, Shannon [123] also showed that entropy is, in some sense, the only possible measure of uncertainty. He stated three properties which one should expect *any* measure of uncertainty to satisfy, and showed that entropy is the only measure satisfying them.

Although the definition of entropy forms the basis of Shannon's presentation, the more fundamental measure used in his coding theorems is the mutual information defined next.

Definition 2 *The mutual information between two random variables X, Y with joint distribution $p(x, y)$ is defined as $I[p(x, y)] = \sum_{x,y} p(x, y) \log_2 \frac{p(x,y)}{p(x)p(y)}$. It will alternatively be written as $I(X; Y)$ when the distribution $p(x, y)$ is clear from the context.*

It is easy to prove [27] that the mutual information is non-negative (with equality if and only if X, Y are independent) and that $I(X; Y) = H(X) + H(Y) - H(X, Y) = H(X) - H(X|Y)$. Thus the mutual information measures the reduction in uncertainty about X which results from knowledge of Y, and is therefore a measure of dependence between the two variables. One of the important advantages of mutual information as such a measure is that it does not make any modeling assumptions about $p(x, y)$ (e.g., it does not assume any parametric form). It can therefore capture information in any property of the joint distribution, and not only in its first and second moments for example. It is this aspect of the mutual information that makes it an attractive tool in studying coding systems, such as the brain, where complex properties of the system potentially contribute to its coding power. However, this advantage comes with two disadvantages: one is the need to obtain an estimate of $p(x, y)$, and the other is the need to understand which properties of the system carry the information. We will address these issues throughout the thesis.

Most of the key theorems of information theory involve finding a distribution which minimizes or maximizes information subject to some constraints. The next section present one such theorem, which is closely tied to the current work.

1.2.1 Rate Distortion Theory

Voice communication devices such as cellular phones are designed to transfer speech signals from one end to the other while transmitting as little data as possible, in order

[4] Where averaging is w.r.t the random variable X.

[5] Here we use $\log_2$.

to minimize the related costs (e.g. energy, storage, bandwidth). While the received signal should be audible, it by no means needs to be an identical copy of the sent signal, since the human ear can recognize speech under a wide array of distortions and interferences. The principle of transmission with some allowed distortion is common to many communication systems, and is also known as the problem of "lossy compression" as opposed to "lossless compression".

In his Rate Distortion Theory, Shannon analyzed the limits of such communication. We next give a brief outline of his approach. Many details are left out, since the main purpose is to provide intuition. Denote by $X, \hat{X}$ the sent and received signals (assume a noiseless channel for convenience), and by $d(x, \hat{x})$ some measure of distortion between them. Thus if $X, \hat{X}$ are speech signals, $d(x, \hat{x})$ would be low if they seem similar to the human ear. The communication system is defined as follows: n signals $X_1, \ldots, X_n$ are represented using a sequence of nR bits, which are to be sent to the receiver. The factor R is referred to as the rate of the code. This is done via a mapping $f : \mathcal{X}^n \to \{1, \ldots, 2^{nR}\}$. The nR bits are then decoded at the receiver using a mapping $g : \{1, \ldots, 2^{nR}\} \to \hat{\mathcal{X}}$. The average distortion of this system is defined as $D(f, g) \equiv \sum_{x^n} d(x^n, g(f(x^n)))p(x^n)$.

Shannon showed [27] that if the maximum distortion allowed is D_0, then one can design a system with distortion at most D_0, which uses the following rate

$$R^{(I)}(D_0) \equiv \min_{p(\hat{x}|x):\sum_x p(x)p(\hat{x}|x)d(x,\hat{x}) \leq D_0} I[p(x, \hat{x})] \ . \tag{1.2}$$

He furthermore showed that no system can achieve a lower rate.

The function $R^{(I)}(D_0)$ is known as the Rate Distortion function, since it describes the tradeoff between allowed distortion and the rate of the code. For our purposes, it is important to note that $R^{(I)}(D_0)$ is calculated via constrained minimization of the mutual information functional, where the constraints are given by some linear functional of the distribution p. A related information minimization problem will be described in detail in Chapter 2.

1.2.2 Information and Prediction Error

Mutual information quantifies dependence between variables. A different quantifier is the minimum possible error incurred in predicting Y from X. It is easy to see that the optimal predictor of Y from X is the "Maximum a Posteriori" (MAP) predictor defined as $g(x) = \arg\max_y p(y|x)$. The error incurred by this predictor is called the Bayes error and is denoted by e_p^*.

Since the Bayes error and the mutual information both quantify the dependence between X and Y, it is natural to expect that they are related. The following two theorems give some information about their interrelationships. The first, known as Fano's inequality (see [27]) provides a lower bound on the Bayes error.

Theorem 1 $h(e_p^*) + e_p^* log(|Y| - 1) \geq H(Y|X)$, *where* $h(e_p^*) = -e_p^* \log e_p^* - (1 - e_p^*) \log (1 - e_p^*)$

The second result, proved in [67], and stated below, yields an upper bound on the Bayes error.

Theorem 2 *[67] The Bayes error is upper bounded by the conditional entropy* $e_p^* \leq \frac{1}{2} H(Y|X)$ [6].

The upper and lower bounds are illustrated in Figure 1.1, for the $|Y| = 2$ case. It can be seen that the range of possible prediction errors is small for extremal mutual information values. For a value of 0.5 bits, the prediction error lies in a range of approximately 0.1.

Machine learning algorithms often seek distributions which minimize prediction error subject to constraints on the structure of $p(x, y)$ (for example $p(x, y)$ is limited to some parametric family [36]). Since prediction error is a non-smooth function of $p(x, y)$, it is typically hard to minimize directly. Mutual information on the other hand, is a smooth function of $p(x, y)$ and thus it is sometimes easier to construct machine learning algorithms that attempt to *maximize* it rather that minimize the prediction error [136]. An illustration of this approach appears in the current work (Section 2.3).

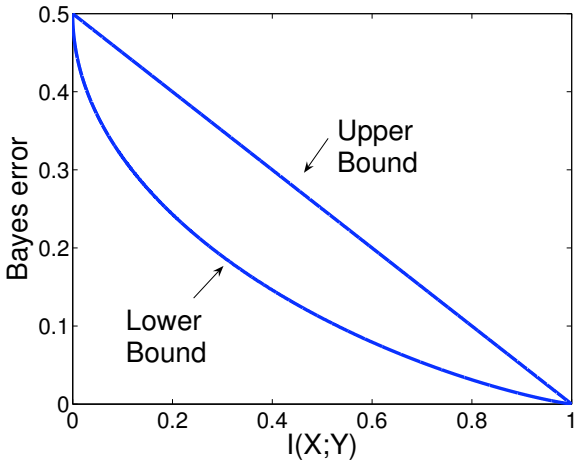


Figure 1.1: Upper and lower bounds on the Bayes error as a function of the Mutual Information. A value of $H(Y) = 0.5$ is used in calculating $H(Y|X)$ from $I(X;Y)$. Information values are calculated in bits (using base 2 in the log).

1.3 Information in Measurements

A key question in what follows is what can one say about a system from *partial* measurements of it. Before explicitly defining this notion, we give a brief example. Consider

[6] Entropy here is measured in bits.

again the network of 100 neurons described in Section 1.1. Now assume we can reliably estimate the individual response of each neuron $p(x_i|y)$ under the different stimuli Y. This constitutes what we shall call a partial measurement of the system.

To give a more formal definition, we begin with distributions over a single variable X. In subsequent chapters, we will discuss the generalization to bivariate distributions. Consider a vector function $\vec{\phi}(x) : X \to \Re^d$. We shall say that the *measurement* of $\vec{\phi}(x)$ under a distribution $\hat{p}(x)$ is the set of d expected values $\langle \vec{\phi}(x) \rangle_{\hat{p}(x)}$, where $\langle\rangle$ is the expectation operator defined as: $\langle f(x) \rangle_{p(x)} = \sum_x p(x) f(x)$. A given measurement value $\vec{a}$ does not generally uniquely specify an underlying distribution, and may in fact be obtained from a large set different distributions (e.g. there are infinitely many distributions over $\Re$ with zero mean). We next define the set of distributions which share a given measurement value as

$$\mathcal{P}_x(\vec{\phi}(x), \vec{a}) \equiv \left\{ \hat{p}(x) : \ \langle \vec{\phi}(x) \rangle_{\hat{p}(x)} = \vec{a} \ \right\} . \tag{1.3}$$

A simple extension of the above set is to the case where expectations are not known with certainty but are rather known to lie within some range of values. This is often the case since expectations are commonly calculated from finite samples, which leaves some uncertainty regarding the true underlying values [7]. The set of distributions which share a given set of measurement *ranges* are defined by

$$\mathcal{P}_x(\vec{\phi}(x), \vec{a}, \vec{\beta}) \equiv \left\{ \hat{p}(x) : \ \vec{a} - \vec{\beta} \leq \langle \vec{\phi}(x) \rangle_{\hat{p}(x)} \leq \vec{a} + \vec{\beta} \ \right\} , \tag{1.4}$$

where $\vec{\beta}$ is an element-wise positive vector reflecting the uncertainty about the measurement value. What can one say about the underlying distribution $p(x)$ given only partial measurement values? A possible approach is described in the next section.

1.3.1 The Maximum Entropy Principle

Suppose one knows that a distribution $p(x)$ has the expected value $\langle \vec{\phi}(x) \rangle_{p(x)} = \vec{a}$, what can be said about the values of $p(x)$ for all x? A simple illustration of this problem, due to Boltzmann, is a dice whose expected outcome is known, but not the individual probabilities of each its faces. The above problem is of course ill posed, since the only thing one can say with certainty about $p(x)$ is that it is in the set $\mathcal{P}_x(\vec{\phi}(x), \vec{a})$. However, we shall see that there are specific distributions in this set which may constitute a reasonable answer.

One of the key approaches to this problem has been the Maximum Entropy principle (MaxEnt) introduced in statistical mechanics in the 19^{th} century and later expanded by Jaynes [73] . The MaxEnt principle simply states that if a distribution $p(x)$ is known

[7] A *likely* range of values may be calculated using concentration bounds such as Chernoff's. See more on this in Section 2.8.1.

to have an expected value $\langle\vec{\phi}(x)\rangle_{p(x)} = \vec{a}$, then the *best* guess at $p(x)$ is the distribution in $\mathcal{P}_x(\vec{\phi}(x), \vec{a})$ with maximum entropy. Formally, the MaxEnt distribution is given by:

$$p_{ME} = \arg \max_{p \in \mathcal{P}_x(\vec{\phi}(x), \vec{a})} H[p(x)] \ . \tag{1.5}$$

The MaxEnt distribution is intuitively the distribution with the highest degree of uncertainty in $\mathcal{P}_x(\vec{\phi}(x), \vec{a})$ and as such reflects no additional knowledge above that given in the observations (it is sometimes referred to as the distribution *least committed* to unseen data). An alternative justification of the principle is related to the asymptotic equipartition theorem, and states that the MaxEnt distribution has the largest number of characteristic samples (see [27] page 266).

A different argument, given in [63], suggests a game theoretic interpretation. Suppose you are to choose a distribution $q(x)$ such that any time x is drawn, you pay $-\log q(x)$. Your goal is to minimize your loss $-\sum_x p(x) \log q(x)$. It can be shown that the MaxEnt distribution minimizes the worst case loss in this game. Formally

$$p_{ME} = \arg \min_{q(x)} \max_{p \in \mathcal{P}_x(\vec{\phi}(x), \vec{a})} -\sum_x p(x) \log q(x) \ . \tag{1.6}$$

We shall return to this interpretation in the next chapter (Section 2.3).

Finally, MaxEnt has been successfully used as a modeling tool in a wide range of applications, from Natural Language Processing [12, 34, 98] and spectral estimation (see Chapter 11 in [27]) to ecological modeling [109].

1.4 Generalizing MaxEnt: *I-projections*

We now define a problem which generalizes MaxEnt, and discuss its algorithmic solutions. We begin with defining the Kullback Liebler (KL) divergence between two distributions (see e.g. [27]) [8]:

$$D_{KL}[p(x)|q(x)] \equiv \sum_x p(x) \log \frac{p(x)}{q(x)} \ . \tag{1.7}$$

The KL divergence is a non-negative, non-symmetric measure of similarity between two distributions over a random variable X. Although it is not a metric, it does have some metric properties, such as being zero if and only if the two distributions are identical [9].

Given a distribution $q(x)$ and a set of distributions $\mathcal{F}$, we may ask what is the distribution in $\mathcal{F}$ that is *closest* to $q(x)$, in the KL divergence sense. Since the divergence is not symmetric this can be defined in two ways. We shall focus on one possibility,

[8] We will mostly use the natural logarithm in what follows.

[9] When x is continuous, $p(x), q(x)$ may differ over a countable set of points.

and following Csiszar [30], define the *I-projection* of $q(x)$ on $\mathcal{F}$ as [10]

$$IPR(q, \mathcal{F}) \equiv \arg\min_{p \in \mathcal{F}} D_{KL}[p|q] \ . \tag{1.8}$$

In what follows we shall use the notation $p^*(x) = IPR(q, \mathcal{F})$ for brevity. When $q(x)$ is the uniform distribution, the KL divergence is the negative entropy of $p(x)$ (up to an additive constant), and thus *I-projection* is equivalent to MaxEnt in the set $\mathcal{F}$. The distribution $q(x)$ can be interpreted as supplying some prior knowledge about the distribution of X. When none is given, $q(x)$ is taken to be uniform. Thus *I-projection* offers a geometric interpretation of MaxEnt and extends it (see [34] for the use of *I-projections* in Natural Language Processing).

An interesting and useful property of the *I-projection* is the so called Pythagorean property. When $\mathcal{F}$ is a convex set, it can be shown [30] that the following inequality is satisfied for all $p(x) \in \mathcal{F}$ and all distributions $q(x)$

$$D_{KL}[p|q] \geq D_{KL}[p|p^*] + D_{KL}[p^*|q] \ .$$

Furthermore, when $\mathcal{F}$ is defined via expectation constraints as in Equation 1.3, the above becomes an equality. We shall make repeated use of this property in proving convergence of algorithms later.

When the set of distributions $\mathcal{F}$ is given by expectation constraints $\mathcal{P}_x(\vec{\phi}(x), \vec{a})$ the general form of the distribution $p^*(x)$ may be found explicitly. Using Lagrange multipliers, we obtain

$$p^*(x) = \frac{q(x)}{Z} e^{\vec{\phi}(x) \cdot \vec{\psi}^*} \ , \tag{1.9}$$

where $Z = \sum_x q(x) e^{\vec{\phi}(x) \cdot \vec{\psi}^*}$ is the partition function, and $\vec{\psi}^*$ should be chosen such that $p^*(x)$ satisfies the expectation constraints (i.e., $p^* \in \mathcal{P}_x(\vec{\phi}(x), \vec{a})$). The parameters $\vec{\psi}^*$ do not have a closed form solution, and need to be solved using iterative procedures as shown in the next section.

1.4.1 Calculating *I-projections*

There are several approaches to finding the parameters $\vec{\psi}^*$. To describe the first, we note that the optimization problem in Equation 1.8 is convex. This is due to the fact that the KL divergence is a convex functional in $q(x)$ for fixed $p(x)$ (see [27] page 30), and that the set $\mathcal{P}_x(\vec{\phi}(x), \vec{a})$ is convex. The theory of convex duality (see Appendix A.1.2) asserts that any convex problem has a convex dual which is equivalent to the original (primal) problem [11]. For the current problem, this dual can be shown to be

[10] In some contexts the *I-projection* is called the minimum relative entropy problem [71].

[11] An equivalent dual requires that the primal problem be strictly feasible (Slater's condition). If the expected values are obtained from an empirical distribution, the empirical distribution is then itself a feasible point, although not strictly so, since some of the probabilities may be zero. However, there always exists an exponential distribution with the same expected values [34], which is strictly feasible, since it has no zero values.

the following *unconstrained* problem (see e.g. [26])

$$\vec{\psi}^* = \arg\min_{\vec{\psi}} \ \vec{a} \cdot \vec{\psi} + \log Z \ . \tag{1.10}$$

This convex unconstrained minimization problem can be solved using any general purpose optimization method (such as the Conjugate Gradient or Broyden-Fletcher-Goldfarb-Shanno (BFGS) procedures; see [88] for a comparison of these methods for the MaxEnt problem). The solution is guaranteed to be unique due to convexity.

The above dual problem can also be interpreted in terms of maximum likelihood. Suppose that the value $\vec{a}$ was obtained by averaging $\vec{\phi}(x)$ over a sample $x_1, \ldots, x_n$ generated by a distribution of the exponential form 1.9. Then the expression in Equation 1.10 is the negative-likelihood of the sample, and thus its minimizer is the maximum likelihood parameter.

An alternative approach to finding $\vec{\psi}^*$ is an elegant procedure introduced by Darroch and Rattcliff in [32]. Their *Generalized Iterative Scaling* (GIS), is an iterative algorithm which generates a sequence of parameters $\vec{\psi}$ which converges to the optimal one. The idea is simple: given the parameter $\vec{\psi}$ after the t^{th} iteration, construct an exponential distribution as in Equation 1.9. Now calculate the expected value of $\vec{\phi}(x)$ according to this distribution, and use the ratio between it and the desired expected values to update $\vec{\psi}$. The algorithm is depicted in Figure 1.2. Note that the algorithm requires the functions $\vec{\phi}(x)$ to be non-negative and to sum up to some constant M for all x values. These two requirements can be easily fulfilled for any general function $\vec{\phi}(x)$ using the following procedure: add a constant to make all functions positive, and construct an additional function $\phi_{d+1}(x)$ which complements the original d functions such that $\sum_i \phi_i(x) = M$. The description of the algorithm shows how to obtain the distribution p^*. It is straightforward to derive the corresponding updates of the parameter $\vec{\psi}$.

Input: Distribution $q(x)$. A set of d positive functions $\vec{\phi}(x)$ such that $\vec{\phi}(x) \geq 0$ for all x, and $\sum_i \phi_i(x) = M$ for all x. A set of expected values $\vec{a}$.

Output: $IPR(q(x), \mathcal{P}_x(\vec{\phi}(x), \vec{a}))$

Initialize: $p_0(x) = q(x)$

Iterate:

- Set $p_{t+1}(x) = \frac{1}{Z_t} p_t(x) \exp\left[\frac{1}{M} \sum_{i=1}^{d} \phi_i(x) \log \frac{a_i}{\langle \phi_i(x) \rangle_{p_t(x)}}\right]$, where Z_t is a normalization constant
- Halt on small enough change in $p_t(x)$

Figure 1.2: The Generalized Iterative Scaling Algorithm.

1.4.2 *I-projections* for Uncertain Expectations

As mentioned in Section 1.3, it is often the case that the expected value of $\vec{\phi}(x)$ is known up to some uncertainty $\vec{\beta}$. In this case, one is interested in *I-projecting* on the set of distributions defined in Equation 1.4, a problem which was addressed in a recent work [41].

It can be shown that the *I-projection* in this case has the same exponential form as in Equation 1.9. However, since the constraints are different from the standard $\vec{\beta} = \vec{0}$ case, the resulting parameters $\vec{\psi}^*$ will also be different.

The dual problem in this case can be shown to be the following:

$$\vec{\psi}^* = \arg\min_{\vec{\psi}} \ \vec{a} \cdot \vec{\psi} + \log Z + |\vec{\psi}| \cdot \vec{\beta} \ . \tag{1.11}$$

It thus turns out that the dual in the uncertain expectations case has an L1 regularization element in addition to elements which appear in the original dual. This problem is again convex and unconstrained, and may be minimized using standard optimization machinery [12].

1.5 The Machine Learning Perspective

Machine learning is a wide field of research whose goal is to build algorithms that learn rules from examples, and to understand the theoretical properties of these algorithms. As such, it is intimately connected to several disciplines such as statistics, combinatorics, artificial intelligence, and cognitive sciences (see [64, 40, 138] for a thorough introduction). A common machine learning task is the *supervised learning* scenario. Here, one is given a set of pairs $(x_i, y_i), i = 1 \ldots n$, where X is some set of features and Y is a class variable (e.g. X is an image of a face, and Y is the identity of a person). The goal is then to learn a function $\hat{y} = f(x)$ such that it can predict the value of Y for values of X which did not appear in the training set (e.g. a new image of one of the people in the training set). In *unsupervised learning*, one is given a dataset $x_1, \ldots, x_n$ without explicit class structure, and the goal is generally to construct a *simple* model which approximates the statistical behavior of X.

Much progress has been made over the last 40 years in designing supervised learning algorithms that can understand speech [114], language [89], and images [9] among many other tasks. Such algorithms have also been recently used in analyzing biomedical data, with applications such as predicting cancer types from gene expression profiles [115]. All the above applications, involving very different domains, often employ similar or

[12] Although the absolute value function is not differentiable, it may be optimized using a simple trick: introduce two non negative parameter sets $\psi^+ \geq 0$ and $\psi^- \geq 0$ such that $\psi \equiv \psi^+ - \psi^-$. Then it can be shown that one of them will be zero at the optimum. Thus $|\psi| = \psi^+ + \psi^-$, and we have a differentiable problem, with non-negativity constraints, which can be solved using constrained conjugate gradient for example. See further details in Section A.1.3.

related algorithms. The reason is that the basic rule learning problems that underlie all of them are similar, and can thus be approached using a similar toolbox. On the other hand, it is commonly accepted that a *good* machine learning algorithm must work with an appropriate representation of the data. Thus, for example, representing an image by the grey level of its pixels is not a very good idea since images which appear similar to the human eye may have very different pixel representation because of translation effects for example.

Indeed, all machine learning algorithms will fail miserably if given an inadequate representation of their inputs. In some sense, this state of affairs is theoretically insoluble, as manifested in the so called *No Free Lunch* theorems [142].

The problem of finding the *correct* representation of a given data set has attracted considerable interest in the machine learning community, and will also be dealt with in the current dissertation. Although, as mentioned above, no comprehensive solution can be found to this problem, real world data often have structure that allows some progress. For instance, in complex systems much information can be gained from considering contributions of single units, disregarding their correlations and higher order statistics (this will become relevant in Chapters 2-3). We shall be interested in ways of using such features to learn about a system.

Information theoretic tools have been widely used in machine learning, in different contexts. One concept which has had a very large impact on Natural Language Processing has been that of maximum entropy (MaxEnt) models, covered in the previous sections. MaxEnt models were successfully used to model distributions in several domains of NLP such as language modeling and translation (see e.g. [34]). Information theoretic concepts were also exploited in many studies of efficient compression of data, which aims to preserve relevant information. One such model, which we analyze here, is the Information Bottleneck method [135], which suggests a way of clustering data in a maximally informative manner.

The mutual information function has often been used as a measure of independence between variables, in order to build representation of data which contain independent components. A well known example of this approach is the Independent Component Analysis (ICA) algorithm [10].

The current dissertation will use both machine learning and information theoretic concepts in order to study methods of meaningful feature extraction.

1.6 Outline and Novel Contributions

The problem of learning from partial measurements as presented in the introduction occupies a large part of the thesis. The first novel contribution to this question is to treat it in the context of input-output systems, where mutual information, rather than entropy is the appropriate information theoretic measure.

In Chapter 2 we introduce the minimum mutual information (MinMI) principle, as an extension of the MaxEnt principle to such systems. The MinMI principle considers the distribution which minimizes mutual information, rather than maximizing entropy, subject to a set of partial measurements. We characterize the solution to the MinMI problem, and provide several algorithms for obtaining it. We also show how this distribution may be used in classification and provide generalization bounds and game theoretic interpretations of its performance. Finally, the performance of the MinMI classifier is demonstrated on classification problems, where it is shown to outperform other methods in some cases.

Since the brain may be interpreted, on some level, as an information processing system, we can use MinMI to study its properties of the neural code given partial measurements. Chapter 3 discusses the application of MinMI to studying the neural code, and analyzing data obtained in neuro-physiological experiments. We show how MinMI may be used to measure information in properties of the neural response, such as single cell responses. The results demonstrate that MinMI can differentiate between populations where neurons have similar codes and those in which their codes differ. A similar analysis is performed for pairwise responses. Finally, we show that MinMI can quantify the information in the neuronal temporal response profile. This allows us to detect neurons whose temporal response provides information about the stimulus, increasing the number of informative neurons by 35%, for data recorded in the motor cortex of behaving monkeys. MinMI extends current information theoretic methods in neuroscience, by yielding a measure of information in various scenarios which are not covered by present methods. We discuss its various advantages over existing approaches.

When using information in measurements, a natural question that arises is which measurements are more informative. In other words, what properties of the system provide more insight into its information processing capabilities? Chapter 4 formalizes this question and introduces its formal and algorithmic solution. The resulting novel method, which we name Sufficient Dimensionality Reduction (SDR), is described, and an algorithm for finding the optimal features is provided. The algorithm uses the notion of *I-projections* described in the current chapter, and is proved to converge. SDR is then applied to several text analysis tasks and is shown to find useful features of the data. Chapter 5 presents an extension of SDR to cases where the structure of the noise is known, so that one can avoid finding irrelevant features.

The final chapter tackles the question of feature dimension. When analyzing data, one typically aspires to find a small set of features which characterize it. Ideally, this should be done without compromising the accuracy of the description. This tradeoff between dimensionality and accuracy is formally treated in Chapter 6. We use the information bottleneck formalism [135] to show that feature dimensionality arises naturally when describing feature extraction in an information theoretic context, where a tradeoff between information minimization and maximization is performed.

Taken together, our results serve to illustrate the utility of information theoretic concepts, and specifically mutual information minimization, in machine learning and in analyzing data measured in complex systems. The material in this thesis is partly covered in the following publications [13] [24, 53, 54, 55, 56, 57].

[13] In [24], A.G. had a contribution equal to that of the first author.

Chapter 2

The Minimum Information Principle

Some of the greatest challenges to science today involve complex systems, such as the brain and gene regulatory networks. Such systems are characterized by a very large number of interacting units that potentially cooperate in complex ways to produce ordered behavior. Some of the more interesting systems may be viewed, to a certain degree, as input-output systems. The brain, for example, receives multiple inputs from the environment and processes them to generate behavior. Such processing is often accompanied by highly complex communication between the different cortical areas [117]. In order to obtain insight into such systems, this processing of information needs to be quantified. An attractive mathematical tool in this context is *Information Theory*, discussed in the previous chapter. Information theory has been used in neuroscience ever since its introduction [93], yielding insights into design principles in neural coding [4, 11, 85], and offering new methods for analyzing intricate data obtained in neurophysiological experiments [116]. Such experimental works employ information theory by calculating the mutual information between aspects of the external world (e.g. motor activity [65] or a visual stimulus [14]) and aspects of the neuronal response (e.g. spike counts [49] and precise spike times [31] among others).

Empirical studies of complex systems in general and information theoretic analyses in particular are fundamentally limited by the fact that the space of possible system states is extremely large. Thus any measurement of the system is bound to be partial and reveal only a subset of its possible states.

For example, it is not practical to fully characterize the statistics of a 100 ms spike train of even a single neuron, because of its high dimensionality (2^{100}) and the relatively limited number of experimental trials. The problem of partial measurements is even more acute for multiple neuron recording due to two reasons. First, the dimension of the response space grows exponentially with the number of neurons. Second, neurons are often not recorded simultaneously but rather over several recording sessions, so that

their joint statistics are not accessible.

Here, we present a new principle and framework for extending information theoretic analysis to handle partial measurements of complex systems. At the basis of our approach is the assumption that the partial measurements hold for the *true* underlying system, whose complete characterization cannot be accessed. We next consider all *hypothetical* systems that are consistent with the observed partial measurements. Clearly, there is a large set of such systems, each with its own value of mutual information between input and output. Our goal is to find the value of information that can be attributed *only* to the given measurements. Intuitively, the systems with relatively high mutual information in the hypothetical set have some additional structure which cannot be inferred based on the given partial measurements. However, the system with minimum information in this set cannot be simplified further (in the mutual information sense) and its information can thus be taken to reflect the information available in the given measurements.

Our minimum information (MinMI) principle thus states that given a set of measurements of a system, the mutual information available in these measurements is the minimum mutual information between input and output in any system consistent with the given measurements. An immediate implication of the above construction is that this minimum information is a lower bound on the mutual information in the true underlying system.

Another conceptual tool which has previously been used to tackle partial measurements is the Maximum Entropy (MaxEnt) principle [73, 122] discussed in the previous chapter. The MinMI principle is more appropriate for handling input-output systems, since in the latter, mutual information, rather than entropy is the measure of interest. This is further strengthened by the fact that MinMI offers a bound on the information in the true system, whereas MaxEnt does not. We shall point to the main practical differences between the approaches in the text. One important technical difference which will become evident is that since MinMI minimizes a *difference* between entropies, it results in a substantially different solution. The MinMI formalism also extends the well known, and frequently used data processing inequality, and allows the estimation of information in scenarios where this was not previously possible.

In what follows, we formalize this approach, and show how the minimum mutual information can be calculated. We shall also be interested in using the MinMI formalism for constructing an algorithm that predicts Y from X (e.g. predict movement from neural response, or class variables from features in machine learning applications). A bound on the error incurred by such an algorithm will be given, along with a game theoretic interpretation of its performance. We shall also discuss the relation between the MinMI classification algorithm and two well known approaches to classification: generative and discriminative modeling.

In this chapter we demonstrate the applicability of the approach to machine learning

applications in various domains. The next chapter discusses its applications to studying the neural code. Some of the material in this chapter was published in [55].

2.1 Problem Formulation

We now define the minimum information problem, and characterize its solution.

To simplify presentation, we assume we have n independently identically distributed (IID) samples (x_i, y_i) $(i = 1, \ldots, n)$ drawn from an underlying distribution $p(x, y)$. We use those to calculate our partial measurements. The partial measurements may be obtained in other ways, for example by measuring different properties from different samples, as in neurons recorded on different days.

We assume that the marginal of Y can be reliably estimated from the data. This is often the case when Y is a stimulus or some behavioral condition whose frequency is governed by the experimentalist, or when it is a class variable and there are relatively few classes. The empirical marginal of Y is

$$\bar{p}(y) \equiv \frac{1}{n} \sum_i \delta_{y_i, y} \ . \tag{2.1}$$

Let $\vec{\phi}(x) : X \to \Re^d$ be a given function of X. The conditional means of $\vec{\phi}(x)$ are then

$$\vec{a}(y) \equiv \frac{1}{n\bar{p}(y)} \sum_{i:y_i=y} \vec{\phi}(x_i) \ . \tag{2.2}$$

In what follows we assume that the expected values are exact, i.e., $\vec{a}(y) = \langle \vec{\phi}(x) \rangle_{p(y|x)}$, where $p(x, y)$ is the *true* underlying distribution. This occurs when $n \to \infty$. For finite sample sizes, there will be a divergence between these two values, which can be controlled using concentration bounds such as Chernoff's. In Section 2.8.1 we show how MinMI may be extended to handle imprecise measurements.

We now consider the distribution which has minimum mutual information while agreeing with the sample on both the expected values of $\vec{\phi}(x)$, and the marginal $\bar{p}(y)$. Define the set of distributions agreeing with the sample by

$$\mathcal{P}_{x|y}\big(\vec{\phi}(x), \vec{a}(y), \bar{p}(y)\big) \equiv \left\{ \hat{p}(x, y) : \begin{array}{cc} \langle \vec{\phi}(x) \rangle_{\hat{p}(x|y)} = \vec{a}(y) & \forall y \\ \hat{p}(y) = \bar{p}(y) & \end{array} \right\} . \tag{2.3}$$

The information minimizing distribution is then given by

$$p_{MI}(x, y) \equiv \arg \min_{\hat{p}(x,y) \in \mathcal{P}_{x|y}\left(\vec{\phi}(x), \vec{a}(y), \bar{p}(y)\right)} I[\hat{p}(x, y)] \ , \tag{2.4}$$

where $I[\hat{p}(x, y)]$ denotes the mutual information between X and Y under the distribution $\hat{p}(x, y)$. Note that since the marginal $\hat{p}(y)$ is constrained, we are actually optimizing over $\hat{p}(x|y)$.

This minimization problem is convex since the mutual information is a convex function of $p(x|y)$ for a fixed $p(y)$ [27] and the set of constraints is also convex. It thus has no local minima.

We also define

$$I_{min}\left[\vec{\phi}(x),\vec{a}(y),\bar{p}(y)\right] \equiv \min_{\hat{p}(x,y)\in\mathcal{P}_{x|y}\left(\vec{\phi}(x),\vec{a}(y),\bar{p}(y)\right)} I[\hat{p}(x,y)] \ , \tag{2.5}$$

as the minimum mutual information in any distribution agreeing with the constraints. It is clear from the definition that this information is a lower bound on the information in the *true* distribution $p(x,y)$ since the latter is also in the set $\mathcal{P}_{x|y}\big(\vec{\phi}(x),\vec{a}(y),\bar{p}(y)\big)$.

Using Lagrange multipliers to solve the constrained optimization in Equation 2.4 we obtain the following characterization of the solution

$$p_{MI}(x|y) = p_{MI}(x)e^{\vec{\phi}(x)\cdot\vec{\psi}(y)+\gamma(y)} \ , \tag{2.6}$$

where $\vec{\psi}(y)$ are the Lagrange multipliers corresponding to the constraints, and $\gamma(y)$ is set to normalize the distribution. Note that this does not provide an analytic characterization of $p_{MI}(x|y)$ since $p_{MI}(x)$ itself depends on $p_{MI}(x|y)$ through the marginalization

$$p_{MI}(x) = \sum_y p_{MI}(x|y)\bar{p}(y) \ . \tag{2.7}$$

The minimum mutual information has the following simple expression

$$I_{min}\left[\vec{\phi}(x),\vec{a}(y),\bar{p}(y)\right] \equiv I[p_{MI}(x,y)] = \langle\vec{\psi}(y)\cdot\vec{a}(y)+\gamma(y)\rangle_{\bar{p}(y)} \ , \tag{2.8}$$

where the operator $\langle\rangle_{\bar{p}(y)}$ denotes expectation with respect to $\bar{p}(y)$.

2.1.1 Using $p_{MI}(x,y)$ to Predict Y from X

Since $p_{MI}(x,y)$ is in some sense a model of $p(x,y)$, it seems reasonable to use it in order to predict Y from X. This is indeed a common strategy in machine learning where a model $\hat{p}(y|x)$ is constructed from data, and is used to predict Y [97]. We shall later want to analyze the prediction power of our classifier (Section 2.5).

Principally, in order to obtain $p_{MI}(y|x)$ from $p_{MI}(x|y)$ one needs to use Bayes law, i.e., multiply by $\bar{p}(y)$ and divide by $p_{MI}(x)$. However, $p_{MI}(x)$ may well be zero for a large number of x values (see Section 2.2), resulting in $p_{MI}(y|x)$ being undefined. We thus consider the expression that would have been obtained for $p_{MI}(x)\neq 0$, but note that it may not be a *legal* distribution.

$$f_{MI}(y|x) = \bar{p}(y)e^{\vec{\phi}(x)\cdot\vec{\psi}(y)+\gamma(y)} \ . \tag{2.9}$$

Note that this distribution has a form similar to logistic regression as in [79]. However, there are two main differences between $p_{MI}(y|x)$ and the standard logistic regression.

One is that $p_{MI}(y|x)$ does not have a normalization function dependent on x. This is a common property of information minimizing distributions, and is also seen in Rate Distortion Theory [27] and the Information Bottleneck method [135]. The second difference is that the optimal parameters of $p_{MI}(y|x)$ are not those obtained via (conditional) maximum likelihood, but rather those which satisfy the conditions in Equation 2.6. This constitutes another difference between our formalism and that of MaxEnt, which is known to be equivalent to Maximum Likelihood estimation in exponential models [34].

The Information in k^{th} Order Marginals: $I^{(k)}$

Models of distributions over large sets of variables often focus on the marginal properties of subsets of these variables. Furthermore, maximum likelihood estimation over Markov fields is known to be equivalent to matching the empirical marginals of the cliques in the graph [76]. We now define the MinMI version of the marginal matching problem.

Denote by $X \equiv (X_1, \ldots, X_N)$ an N dimensional feature vector, and by $\{X_C\}$ a set of subsets of variables of X (e.g., all singletons or pairs of X_i). Assume we are given the empirical conditional marginals $p(X_C|Y)$. In our notation, this is equivalent to choosing the following functions to measure

$$\phi_{x_C}(\hat{x}) = \delta_{\hat{x}_C, x_C} \ , \tag{2.10}$$

which (with some abuse of notation) are the indicator functions for a specific assignment to the variables in x_C (the number of functions is the total number of assignments to variables in the sets x_C). The function defined above can be seen to have the expected value $p(x_C|y)$, i.e., the conditional marginal of the set x_C.

The MinMI distribution in this case would have the following form

$$p_{MI}(x|y) = p_{MI}(x) e^{\sum_{x_C} \psi(x_C, y) + \gamma(y)} \ . \tag{2.11}$$

In what follows, and especially in the next chapter, we will often be interested in measuring all the k^{th} order marginals of a distribution and calculating the respective minimum information. We shall denote the minimum information under the set of all k^{th} order marginals by $I^{(k)}$. For example, the information available from the marginals $p(x_i|y)$ will be denoted $I^{(1)}$, and that from $p(x_i, x_j|y)$ by $I^{(2)}$.

The $I^{(k)}$ measure allows us to write the *full* information $I(X_1, \ldots, X_N; Y)$ as a sum of positive terms which reflect interactions on different orders (see [122] for a different, MaxEnt based factorization). Define $\Delta_k \equiv I^{(k)} - I^{(k-1)}$, and $I^{(0)} \equiv 0$. The difference Δ_k is always positive, and measures the information available in k^{th} order marginals not available in orders lower than k. The full information can then be written as the sum $I(X_1, \ldots, X_N; Y) = \sum_k \Delta_k$. It is thus decomposed into the contributions of the different statistical orders. It is important to note that when the true distribution is

conditionally independent, Δ_k will generally not be zero, since conditional independence increases information with respect to the minimum information scenario (see Section 3.3.1). In Section 3.1 we introduce measures which study the departure from conditional independence.

2.2 Duality and Sparsity

The constrained information minimization in Equation 2.4 is a convex optimization problem, and therefore has an equivalent convex dual (see Section A.1.2 for a brief introduction to duality). The dual for a similar problem - finding the Rate Distortion function, was recently shown to be a geometric program [26].

Using similar duality transformations to those in [26], we obtain the following geometric program (in convex form), which is equivalent to the MinMI problem in Equation 2.4,

$$\begin{array}{ll} \text{maximize} & \langle \vec{\psi}(y) \cdot \vec{a}(y) + \gamma(y) \rangle_{\bar{p}(y)} \\ \text{subject to} & \log \sum_y \bar{p}(y) e^{\vec{\phi}(x) \cdot \vec{\psi}(y) + \gamma(y)} \leq 0 \quad \forall x \end{array} , \tag{2.12}$$

where the optimization is over the variables $(\vec{\psi}(y), \gamma(y))$. The duality proof is given in Appendix A.1.3. Note that the definition of $f_{MI}(y|x)$ (Equation 2.9) implies that the constraint can be written as $\sum_y f_{MI}(y|x) \leq 1$. In the dual problem, optimization is over the variables $(\vec{\psi}(y), \gamma(y))$, and there are $|X|$ constraints. By convex duality (see A.1.2), the maximum of Equation 2.12 is equal to the minimum information obtained in Equation 2.4. Furthermore, the value of the maximized dual function in Equation 2.12 at any feasible point yields a lower bound on the minimum information. We shall use this property later in designing algorithms for calculating I_{min}.

It is interesting to study when the dual constraint is not achieved with equality. The duality proof implies that this will happen only if $\lambda_{xy} > 0$ for some y (see Appendix A.1.3 for the definition of $\lambda_{xy} > 0$), which by the Kuhn-Tucker conditions ([18], page 243) implies that $p_{MI}(x|y) = 0$. Due to the structure of $p_{MI}(x|y)$ (Equation 2.6), it follows that $p_{MI}(x) = 0$ (ignoring the anomalous case of infinite parameters). Thus, we conclude that if the dual constraint is not achieved with equality, then $p_{MI}(x) = 0$. To see why this implies that $p_{MI}(x)$ is a sparse distribution, note that there are $|X|$ inequalities in the dual problem, but only $|Y|(d+1)$ variables (where d is the dimension of $\vec{\phi}(x)$). In the general case, not all of these can be satisfied with equality, implying that if $|X| >> |Y|(d+1)$, most of $p_{MI}(x)$ will be set to zero [1].

The structure of $p_{MI}(x)$ is illustrated in Figure 2.1, which also demonstrates that indeed $\sum_y f_{MI}(y|x) < 1$ implies $p_{MI}(x) = 0$, as claimed above.

[1] It will be interesting to obtain a more quantitative estimate of this sparsity ratio. There are cases when all constraints are satisfied with equality, for example when all $\vec{\psi}(y)$ are zero. However, these are not likely to be the optimal parameters in the general case.

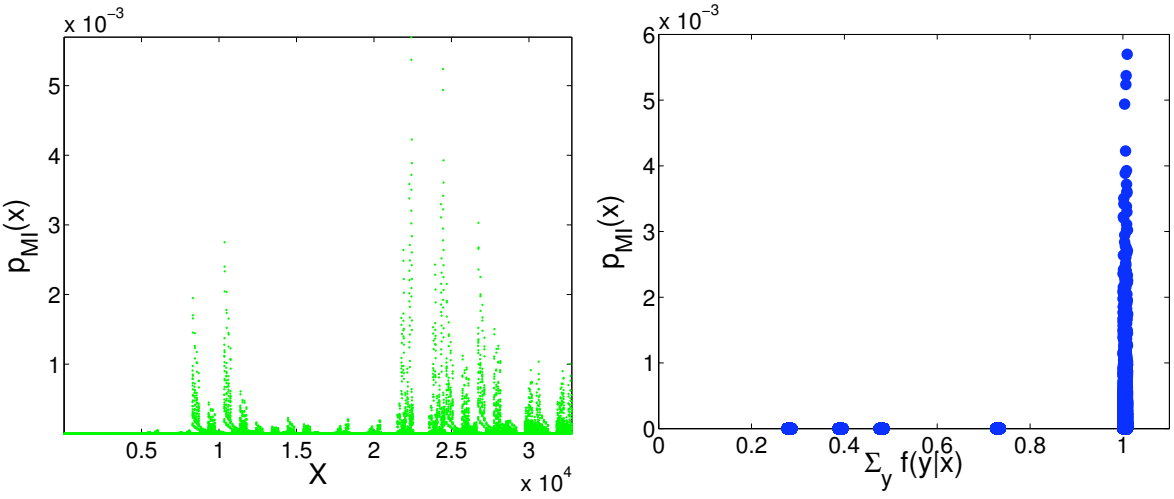


Figure 2.1: Sparsity of the distribution $p_{MI}(x)$ and strict inequality in the dual constraints. The minimum information distribution was calculated for first order constraints drawn randomly for $X = X_1, \ldots, X_{15}$. Each X_i was binary and Y was also binary. Left figure shows the values of $p_{MI}(x)$. Right figure shows the values of $p_{MI}(x)$ as a function of $\sum_y f_{MI}(y|x)$. It can be seen that whenever $\sum_y f_{MI}(y|x) < 1$, we have $p_{MI}(x) = 0$. Note that there is a high overlap for points where $\sum_y f_{MI}(y|x) < 1$ since apparently only a discrete set of values is obtained for this sum.

In Section 2.7 we discuss algorithmic solutions to both the primal and the dual problems.

2.3 A Game Theoretic Interpretation

In [63] Grünwlad gives a game theoretic interpretation of the MaxEnt principle (see Section 1.3.1). We now describe a similar interpretation which applies to the MinMI principle. The proof of the result is similar to that in [63] and is given in Appendix A.1.4.

Proposition 1 *Let $\mathcal{A}$ be the set of all functions $f(y|x)$ such that $\sum_y f(y|x) \leq 1$ for all x. Then the minimum information function $f_{MI}(y|x)$ satisfies*

$$f_{MI}(y|x) = \arg \min_{f(y|x) \in \mathcal{A}} \max_{\hat{p}(x,y) \in \mathcal{P}_{x|y}\left(\vec{\phi}(x), \vec{a}(y), \bar{p}(y)\right)} -\langle \log f(y|x) \rangle_{\hat{p}(x,y)} \ .$$

The above proposition implies that $f_{MI}(y|x)$ is obtained by playing the following game: Nature chooses a distribution $\hat{p}(x,y)$ from $\mathcal{P}_{x|y}\left(\vec{\phi}(x), \vec{a}(y), \bar{p}(y)\right)$. The player, who does not know $\hat{p}(x,y)$ then chooses a predictor $f(y|x)$ aimed at predicting Y from X. The loss incurred in choosing $f(y|x)$ is given by $-\langle \log f(y|x) \rangle_{\hat{p}(x,y)}$. The proposition states that $f_{MI}(y|x)$ corresponds to the strategy which minimizes the worst case loss incurred in this game.

To see how the above argument is related to classification error, we focus on the binary class case, and take the class variable to be $y = \pm 1$. In this case, a classifier based on $f(y|x)$ will decide $y = 1$ if $f(y = 1|x) \geq \frac{a_x}{2}$, where $a_x = \sum_y f(y|x)$. The zero-one loss is thus

$$c_{zo}(x, y, f) = \Theta\big[- y\big(f(y = 1|x) - \frac{a_x}{2}\big)\big] \ , \tag{2.13}$$

where Θ is the step function [2], and y is the *true* label for x. It can be seen that for $f(y|x) \in \mathcal{A}$, the zero-one loss is bounded from above by the loss function $-\log_2 f(y|x)$

$$c_{zo}(x, y, f) \leq -\log_2 f(y|x) \ . \tag{2.14}$$

as illustrated in Figure 2.2.

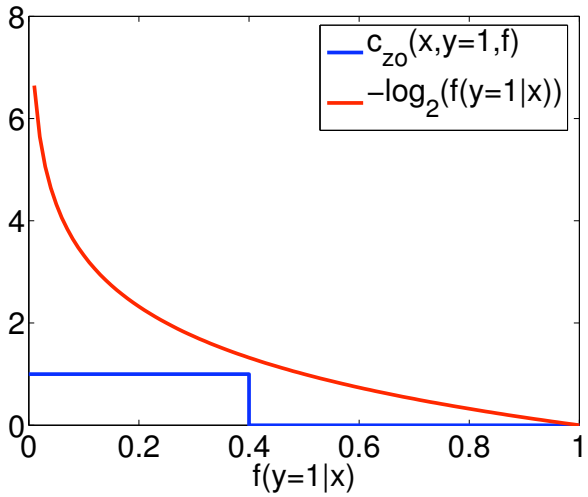


Figure 2.2: Demonstration of the log bound on the zero one loss in Equation 2.14. The bound is illustrated for $y = 1$. A similar picture is obtained for $y = -1$. Here $a_x = 0.8$ so that the decision boundary is at $f(y = 1|x) = 0.4$.

The classification error incurred by $f(y|x)$ is thus bounded from above by the expected loss

$$\langle c_{zo} \rangle_{\hat{p}(x,y)} \leq \langle -\log_2 f(y|x) \rangle_{\hat{p}(x,y)} \ . \tag{2.15}$$

Note that for the information minimizing distribution $p_{MI}(y|x)$ the above loss is the familiar logistic loss.

We thus have the following elegant formulation of MinMI: the predictor $f_{MI}(y|x)$ is the one which minimizes the worst case upper bound on classification error.

Interestingly, there have been recent works [70, 80] which minimize the worst case error itself (rather than a bound, as done here). However, this was done only for the case of first and second order moment constraints.

[2] The step function is 0 for $x < 0$ and 1 for $x \geq 0$.

2.3.1 Relation to Minimum Description Length

The minimum information principle is intuitively related to a parsimonious description of the system. A seemingly related approach is the Minimum Description Length (MDL) principle which suggests that a *good* model of data is one which can be described using the minimum number of bits (see e.g. [5]). We now discuss the relation between these two approaches, using the game theoretic results above.

The Kraft-McMillan theorem states that any code for an alphabet $\{x_1, \ldots x_n\}$ with codeword lengths of $\{l_1, \ldots l_n\}$ is a uniquely decodable binary prefix code if and only if $\sum_i 2^{-l_i} \leq 1$. In the previous section we posed the constraint $\sum_y f(y|x) \leq 1$ over the predictor functions. This suggests that $-\log f(y|x)$ may be interpreted as a code length for the variable Y. Indeed, if we define $l_y \equiv \lceil -log_2 f(y|x) \rceil$ then

$$\sum_y 2^{-l_y} = \sum_y 2^{-\lceil -\log f(y|x) \rceil} \leq \sum_y 2^{\log f(y|x)} = \sum_y f(y|x) \leq 1 \ . \tag{2.16}$$

Thus l_y may be interpreted as the lengths of a (prefix) code for Y. The loss function in the game of Proposition 1 satisfies

$$\langle l_y \rangle_{\hat{p}(x,y)} - 1 \leq -\langle \log_2 f(y|x) \rangle_{\hat{p}(x,y)} \leq \langle l_y \rangle_{\hat{p}(x,y)} \ . \tag{2.17}$$

When $|Y|$ is large (as in block coding) we expect the $\log_2 f(y|x)$ to be large so the difference of one becomes insignificant and $-\langle \log_2 f(y|x) \rangle_{\hat{p}(x,y)} \approx \langle l_y \rangle_{\hat{p}(x,y)}$.

Comparison of the result in Equation 2.17 with Proposition 1 implies that the loss in the minimax game may be interpreted as the length of a code. We can now give a *description length* interpretation of Proposition 1: For each value of x, describe a code for Y such that its expected length is minimal in the worst case.

This proposition is indeed reminiscent of those proved for MDL. However, we do not know of a similar result in the literature, for the current setting.

2.4 MinMI and Joint Typicality

The rationale for the MaxEnt principle, as given by Boltzmann, Jaynes and others, is based on the fact that samples with atypical empirical histograms - hence with lower empirical entropy - are exponentially (in the sample size) unlikely to occur. Thus we can assert by a histogram counting argument that out of all histograms consistent with observed expectation values, those with maximum entropy are the most likely to be observed among all consistent histograms in the absence of any other knowledge.

When dealing with classification or regression problems, the issue is predictions of Y from X, and it is the notion of *joint typicality* of the two sequences that replaces the simple typicality and Asymptotic Equipartition Property (AEP) in the MaxEnt case. Here we are asking for the most uncommitted distribution of x, *given* that we know the marginal distribution of y, $p(y)$, together with a set of empirical conditional

expectations. For this case a similar histogram counting argument is supplied through the notion of joint typicality, as stated e.g. in [27] page 359.

Let $Y^n = Y_1, Y_2, ..., Y_n$ be drawn IID from $p(y)$. Then for any sequence $x^n = x_1, x_2, ..., x_n$, the probability that (x^n, Y^n) are jointly drawn IID from $p(x, y)$ is $\simeq 2^{-nI(X;Y)}$, via the standard AEP property. In other words, if we partition all the possible empirical histograms of x^n into equivalent classes according their (empirical) mutual information with Y^n, $I(X;Y)$, the relative volume of such a class is exponential in its mutual information and proportional to $2^{-nI(X;Y)}$.

Without any other constraints the (overwhelmingly) largest joint-histogram of x^n and Y^n is the one with $I(X;Y) = 0$, i.e., independent X and Y. Otherwise, with additional empirical constraints on the joint distribution, the overwhelming large fraction among the x^n histograms is occupied by the one with the minimal empirical mutual information. This is the distribution selected by our proposed MinMI procedure.

2.5 Generalization Bounds

The Minimum Information principle suggests a parsimonious description of the data, and therefor one would expect it to have generalization capabilities. We discuss several generalization related results below. To simplify the discussion, we focus on the binary class case. Denote by $p(x, y)$ the *true* distribution underlying the data. Also, denote by e_p^* the optimal Bayes error associated with $p(x, y)$ (see Section 1.2.2), and by e_{MI} the generalization error when using $f_{MI}(y|x)$ for classification.

The Bayes error e_p^* is the minimum classification error one could hope for, when predicting Y from X under $p(x, y)$. In Section 1.2.2 we quote a result bounding the Bayes error via the mutual information [67]

$$e_p^* \leq \frac{1}{2} \left(H(Y) - I[p(x, y)] \right) \ . \tag{2.18}$$

In what follows we assume, as before, that the empirical constraints $\vec{a}(y)$ and $\bar{p}(y)$ correspond to their true values, i.e., $p(x, y) \in \mathcal{P}_{x|y}\big(\vec{\phi}(x), \vec{a}(y), \bar{p}(y)\big)$. Since $p(x, y) \in \mathcal{P}_{x|y}\big(\vec{\phi}(x), \vec{a}(y), \bar{p}(y)\big)$, its information must be larger than or equal to that of $p_{MI}(x, y)$, and thus

$$e_p^* \leq \frac{1}{2} \left(H(Y) - I[p_{MI}(x, y)] \right) \ . \tag{2.19}$$

We thus have a model free bound on the Bayes error of the unknown distribution $p(x, y)$. An obvious shortcoming of the above bound is that it does not relate to the classification error under when using the classifier $f_{MI}(y|x)$ as the class predictor. Denote this error by e_{MI}. Then using Equation 2.15 with $f(y|x) = f_{MI}(y|x)$ we have

$$e_{MI} \leq -\sum p(x, y) \log_2 f_{MI}(y|x) \ . \tag{2.20}$$

But the special form of $f_{MI}(y|x)$ implies that we can replace expectation over $p(x,y)$ with expectation over $p_{MI}(x,y)$. We can then replace $\log_2 f_{MI}(y|x)$ with $\log_2 p_{MI}(y|x)$, since $p_{MI}(y|x) = f_{MI}(y|x)$ when $p_{MI}(x) > 0$. Thus

$$-\sum p(x,y)\log_2 f_{MI}(y|x) = -\sum p_{MI}(x,y)\log_2 f_{MI}(y|x) = -\sum p_{MI}(x,y)\log_2 p_{MI}(y|x) \ .$$

The right hand side is the conditional entropy $H[p_{MI}(y|x)]$, which implies the following bound

Proposition 2 *The generalization error of the classifier based on $f_{MI}(y|x)$ satisfies*

$$e_{MI} \leq H(Y) - I[p_{MI}(x,y)] \ . \tag{2.21}$$

Note that the bound on the optimal Bayes error of the true distribution is tighter than the above bound by a factor of 2. It will be interesting to see whether these bounds can be improved.

2.6 Relation to Other Methods

The current section studies the relations between MinMI and other methods in the literature. We will be interested in two aspects of MinMI. The first is that it calculates a lower bound on the information in a system, where the bound is given by I_{min}. The other is that the function $f_{MI}(y|x)$ can be used to predict Y from X. Below we present different approaches to these two tasks, and also discuss the relation to rate distortion theory.

2.6.1 Information Estimation

The information theoretic literature in neuroscience describes several approaches to calculating mutual information. All suggest ways of handling the inherent small sample problem encountered in experimental data. To stress that this is always a problem, consider a recording of $100ms$ from a *single* cell. Assuming a single spike can be fired at most in each millisecond, the possible set of responses has 2^{100} elements. Again it is impractical to estimate response probabilities over this space. Thus every method which tackles the information estimation problem must consider this difficulty.

The Data Processing Inequality

A very common approach to the problem mentioned above is to consider not the entire response space X but rather some function of it $f(X)$. For example one can characterize a $100ms$ spike train not by its spike firing times, but rather by the total number of spikes that were fired (i.e., the spike count). There are two advantages to this approach. The first is that the distribution $p(f(x), y)$ is much more practical to estimate, since $|f(X)|$ (i.e., the number of different values $f(x)$ can take) is often considerably less

than $|X|$. The second is that $f(x)$ is often some physiologically meaningful property of the response (e.g. spike count), so that the question "What is the information in $f(X)$ about Y?" is meaningful on its own.

There still remains the question of how $I[p(f(x),y)]$ is related to $I[p(x,y)]$. It would seem like a problem if the former may be larger than the latter. However, as intuitively expected, this is not the case. The theorem which guarantees this is known as the data processing inequality (see [27], page 32) which states that $I[p(f(x),y)] \leq I[p(x,y)]$ with equality if and only if $X \to f(X) \to Y$ forms a Markov chain.

We would now like to claim that the above approach constitutes a specific example of the MinMI principle. The data processing approach assumes that we can estimate the distribution $p(f(x),y)$. This is equivalent to having access to the expected values of the functions $\phi_k(x) = \delta_{f(x),k}$, since $\langle \phi_k(x) \rangle_{p(x|y)} = p(f(x) = k|y)$. The following proposition states that the minimum information subject to constraints on the expected values of $\phi_k(x)$ is in fact that calculated via the data processing inequality.

Proposition 3 $I_{min}[\vec{\phi}(x), p(f(x)|y), \bar{p}(y)] = I[p(f(x),y)]$

Proof: *We prove the claim by explicitly defining the distribution $q(x,y)$ which minimizes the information, and showing that its information is equal to that of $p(f(x),y)$. Denote by n_k the number of value of x such that $f(x) = k$. Now consider the conditional distribution $q(x|y) \equiv \frac{p(f(x)|y)}{n_{f(x)}}$, and the joint distribution $q(x,y) \equiv q(x|y)\bar{p}(y)$. Note that $q(x) = \sum_y \bar{p}(y) \frac{p(f(x)|y)}{n_{f(x)}} = \frac{p(f(x))}{n_{f(x)}}$*

It is easy to see that $q(x,y) \in \mathcal{P}\big(\vec{\phi}(x), p(f(x)|y), \bar{p}(y)\big)$, i.e., it has the desired expected values of $\vec{\phi}(x)$, and is thus in the set we are minimizing over. Next, observe that the information $I[q(x,y)]$ is equal to $I[p(f(x),y)]$ since

$$\begin{aligned} I[q(x,y)] &= \sum_y \bar{p}(y) \sum_x q(x|y) \log \frac{q(x|y)}{q(x)} = \sum_y \bar{p}(y) \sum_k n_k \frac{p(f(x)=k|y)}{n_k} \log \frac{p(f(x)=k|y)}{p(f(x)=k)} \\ &= \sum_y \bar{p}(y) \sum_k p(f(x)=k|y) \log \frac{p(f(x)=k|y)}{p(f(x)=k)} = I[p(f(x),y)] \ . \end{aligned}$$

Finally, the data processing inequality guarantees that $I[p(f(x),y)] \leq I[r(x,y)]$ for any distribution $r(x,y) \in \mathcal{P}\big(\vec{\phi}(x), p(f(x)|y), \bar{p}(y))\big)$, and thus

$$I[p(f(x),y)] \leq I_{min}[\vec{\phi}(x), p(f(x)|y), \bar{p}(y)] \ . \tag{2.22}$$

Since we have found a distribution in $\mathcal{P}\big(\vec{\phi}(x), p(f(x)|y), \bar{p}(y)\big)$ where this is an equality, it must be that $I[p(f(x),y)] = I_{min}[\vec{\phi}(x), p(f(x)|y), \bar{p}(y)]$.

We can thus conclude that use of the data processing inequality can be understood as minimizing information subject to knowledge of the *quantized* distribution $p(f(x)|y)$.

Information via Classification

Mutual information is closely related to prediction error as described in Section 1.2.2. It thus seems natural that an estimate of prediction error should lead to an estimate of the mutual information. Indeed, assume one has access to some prediction function $\hat{y} = f(x)$. Then, using the data processing inequality described above, we can deduce that $I(\hat{Y}; Y) \leq I(X; Y)$. The distribution $p(\hat{y}, y)$ is simply the error matrix of the classifier $f(x)$, where the element $p(\hat{y}|y)$ is the probability that the classifier would predict class $\hat{y}$ given that the real class is y. This classification approach is used in a recent work by [95] and in [116], among others. While these methods may be efficient in estimating the true value of the information, it is not clear what property (e.g. statistical interaction order) of the stimulus-response statistics generates this information. MinMI on the other hand, provides an estimate of the information that is directly related to some given statistical property. Note also that since the classification approach is an instance of the data processing inequality, it is related to the MinMI procedure, as described in the section above.

Maximum Entropy and Other Approaches

An approach which is closer in spirit to ours appears in [122, 90]. As in the current work, these consider the case when partial measurements are given (e.g. first or second order marginals) and consider models of the data which fit those. The approach they take is that of Maximum Entropy, as described in Section 1.3.1. Namely, they consider the distribution over the entire response space which maximizes entropy. The entropy in this distribution is taken as a measure of the correlation strength in the given statistics.

The above two approaches differ from ours in that they do not consider mutual information directly, but rather only properties (e.g. entropy) of the response itself. They can thus not be used directly to obtain a bound on information.

In [21], the information in a *binary event* was addressed. The authors introduced a method for calculating the information content of events such as single spikes, pairs of spikes etc. Their approach differs from ours in that the response is characterized by the event alone, whereas MinMI considers distributions over the entire response space. It will be interesting to further study the relations between these two approaches.

2.6.2 Rate Distortion Theory

In contradistinction with other information theoretic methods in neuroscience, MinMI does not aim to estimate the underlying distribution directly, but rather uses the distribution as a variable to be optimized over. This in fact is the mathematical structure of the central coding theorems in information theory [28], where information is either minimized (as in the Rate Distortion theorem) or maximized (as in the Channel Capacity theorem) under some constraints on the joint distribution $p(x, y)$. Our approach

is most closely related to Rate Distortion Theory (RDT) (see Section 1.2.1) which sets the achievable limits on "lossy" compression, i.e., compression which results in some distortion of the original signal. The RDT compression bound is obtained by minimizing information with respect for a fixed $p(y)$, and a constraint on the expected distortion. In MinMI, we also fix $p(y)$ but introduce additional constraints on $p(x|y)$ via its expected values. This can be understood as searching for a distribution $p(x|y)$ as in RDT, but with the *single* distortion constraint replaced by *multiple* constraints on expected values of several functions.

2.6.3 Classification Algorithms

As seen previously, MinMI provides a model for $f_{MI}(y|x)$ which can be used as a class predictor. The literature on building classification algorithms is vast and will not be covered here. The algorithms closest to ours are those which build some probabilistic model $p(y|x)$ of the data, which is then used as a predictor. It is customary to divide such methods into two classes: Generative and Discriminative. Generative methods approximate the joint distribution $p(x, y)$ and use Bayes rule to obtain $p(y|x)$. Often $p(y)$ is assumed to be known and the class conditional distributions $p(x|y)$ are estimated separately. On the other hand, Discriminative models approximate $p(y|x)$ directly. The latter have the clear advantage of solving the classification problem directly. In what follows, we show how our approach is related to both these schemes.

Maximum Entropy of the Joint Distribution

The joint entropy of X and Y is related to the mutual information via

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \ . \tag{2.23}$$

Thus, if both marginals are assumed to be known, the problems of Maximum Entropy and Minimum Mutual Information coincide. The joint distribution which optimizes the above problem has the following form

$$p_{ME}(x,y) = \frac{1}{Z} e^{\vec{\phi}(x)\cdot\vec{\psi}(y)+A(x)+B(y)} \ . \tag{2.24}$$

where $A(x), B(y)$ are free parameters which are adjusted so that $p_{ME}(x, y)$ has the desired marginals over X and Y.

The resulting conditional model is then

$$p_{ME}(y|x) = \frac{1}{Z_x} e^{\vec{\phi}(x)\cdot\vec{\psi}(y)+A(x)+B(y)} \ . \tag{2.25}$$

When the marginal $p(x)$ is not known, but $p(y)$ is, maximizing the joint entropy is equivalent to maximizing $H(X|Y)$, which is equivalent to maximizing $H(X|Y = y)$ for each value of y independently. Note that under this approach, changing the values

of $\vec{a}(y)$ for a given value of y will not change $p(x|y)$ for other values of y. This does not seem to be a desirable property, as it does not consider the class structure of the problem. MinMI, on the other hand, considers all values of $\vec{a}(y)$ simultaneously and therefore does not have the property mentioned above. One example of maximizing joint entropy is the Naive Bayes model which results from maximizing $H(X|Y)$ subject to a constraint on conditional singleton marginals, and the class marginals [97].

Conditional Random Fields and Logistic Regression

Conditional Random Fields (CRF) are models of the conditional distribution [79]

$$p_\lambda(y|x) = \frac{1}{Z_\lambda(x)} e^{\sum_{k=1}^{d} \lambda_k f_k(x,y)} \ , \tag{2.26}$$

where $Z_\lambda(x)$ is a partition function dependent on x. In CRFs, both Y and X are typically multivariate. When Y is a single categorical variable, CRFs become the standard logistic regression model [64].

The d functions $f_k(x, y)$ are assumed to be known in advance, or are chosen from some large set of possible functions. This becomes similar to our setting if one chooses functions

$$f_{i,y_j}(x, y) = \delta_{y,y_j}\phi_i(x) \ . \tag{2.27}$$

In fact, the MinMI formalism could be equally applied to general functions of X and Y as in CRFs. We focus on functions of X for ease of presentation.

CRFs are commonly trained by choosing the parameters λ_i which maximize the conditional maximum likelihood [79] given by

$$\sum_{x,y} \bar{p}(x, y) \log p_\lambda(y|x) = -\langle \log Z_\lambda(x) \rangle_{\bar{p}(x)} + \sum_{k=1}^{d} \lambda_k \langle f_k \rangle_{\bar{p}(x,y)} \ ,$$

where $\bar{p}(x, y)$ is the empirical distribution.

This target function can be seen to depend on the empirical expected values of f_k but also on the empirical marginal $\bar{p}(x)$. This is of course true for all conditional logistic regression models, and differentiates them from MinMI, which has access only to the expected values of $\vec{\phi}(x)$. It thus seems logical that MinMI may outperform these models for small sample sizes, where expected values are reliable, but $\bar{p}(x)$ is not. This can indeed be seen in the experimental evaluation in Section 2.9.2.

2.7 MinMI Algorithms

In order to find the information minimizing distribution $p_{MI}(x|y)$, the optimization problem in Equation 2.4 or its dual in Equation 2.12 need to be solved. This section describes several algorithmic approaches to calculating $p_{MI}(x|y)$. When $|X|$ is small

enough to allow $O(|X|)$ operations, exact algorithms can be used. For the large $|X|$ case we present an approximate algorithm, which uses the primal and dual problems to obtain upper and lower bounds on I_{min}.

2.7.1 Solving the Primal Problem

The characterization of $p_{MI}(x|y)$ is similar to that of the Rate Distortion channel [27] or the related Information Bottleneck distribution in [135]. There are iterative procedures for finding the optimal distributions in these cases, although usually as a function of the Lagrange multipliers (i.e., $\vec{\psi}(y)$) rather than of the value of the constraints. In what follows we outline an algorithm which finds $p_{MI}(x|y)$ for any set of empirical constraints.

The basic building block of the algorithm is the *I-projection* [30] described in Section 1.4. Recall that when we *I-project* a distribution $q(x)$ on a set of expectation constraints of a function $\vec{\phi}(x)$, the projection has the form

$$p^*(x) = \frac{1}{Z_\lambda} q(x) e^{\vec{\phi}(x)\cdot\vec{\lambda}} \ . \tag{2.28}$$

The similarity between the form of the projection in Equation 2.28 and the characterization of $p_{MI}(x|y)$ in Equation 2.6, implies that $p_{MI}(x|y)$ is an *I-projection* of $p_{MI}(x)$ on the set $\mathcal{P}_x(\vec{\phi}(x), \vec{a}(y))$ (see Equation 1.3). The fact that $p_{MI}(x)$ is dependent on $p_{MI}(x|y)$ through marginalization implies an iterative algorithm where marginalization and projection are performed. This procedure is described in Figure 4.1. It can be shown to converge using the Pythagorean property of the *I-projection*. The convergence proof is given in Appendix A.1.1.

The above algorithm cannot be implemented in a straightforward manner when $|X|$ is large, since it involves an explicit representation of $p_t(x)$. Section 2.7.3 addresses a possible approach to this scenario.

2.7.2 Solving the Dual Problem

The dual problem as given in Equation 2.12 is a geometric program and as such can be solved efficiently using interior point algorithms [26]. When $|X|$ is too large to allow $O(|X|)$ operations, such algorithms are no longer practical. However, oracle based algorithms such as the Ellipsoid algorithm or Cutting Plane Methods [59] are still applicable. The above algorithms require an oracle which specifies if a given point is feasible, and if not, specifies a constraint which it violates. For the constraints in Equation 2.12 this amounts to finding the x maximizing the constrained function

$$\begin{aligned} x_{max} &\equiv \arg\max_x f(x) \\ f(x) &\equiv \sum_y \bar{p}(y) e^{\vec{\phi}(x)\cdot\vec{\psi}(y)+\gamma(y)} \ . \end{aligned}$$

Input: Set of functions $\vec{\phi}(x)$ and their class conditional empirical means $\vec{a}(y)$.

Output: The distribution $p_{MI}(x|y)$.

Initialization:

- Initialize $p_0(x)$ randomly.

Iterate:

- For all y set $p_{t+1}(x|y)$ to be the *I-projection* of $p_t(x)$ on $\mathcal{P}_x(\vec{\phi}(x), \vec{a}(y))$

$$p_{t+1}(x|y) = p_t(x) e^{\vec{\phi}(x)\cdot\vec{\psi}_{t+1}(y)+\gamma_{t+1}(y)} \ . \tag{2.29}$$

- Set $p_{t+1}(x) = \sum_y p_{t+1}(x|y)\bar{p}(y)$
- Halt on convergence.

Figure 2.3: An algorithm for solving the primal problem.

The point $(\gamma(y), \vec{\psi}(y))$ is then feasible if $f(x_{max}) \leq 1$. Since $f(x)$ may be interpreted as an unnormalized distribution over x, finding x_{max} is equivalent to finding its maximum probability assignment. This is known as the MAP problem in the AI literature. When $|X|$ is large, it typically cannot be solved exactly, but can be tackled using random sampling techniques as in [106] [3].

2.7.3 An Approximate Primal Algorithm with Dual Bounds

Although the dual algorithm presented above may be applied to the large $|X|$ case, we have found it converges very slowly. In this section we present an approximation of the primal algorithm which can be used in the large $|X|$ case. To derive an approximate algorithm for the large system case, we first note that after t iterations of the iterative algorithm, the distribution $p_t(x)$ is a mixture of the form

$$p_t(x) = \sum_{k=1}^{|S|^t} \frac{c_k}{Z_k} e^{\vec{\phi}(x)\cdot\vec{\psi}_k(y)} \ , \tag{2.30}$$

where every iteration increases the number of components by a factor of $|S|$. For the approximate algorithm, we limit the number of elements in this mixture to some constant c by clustering its components after each iteration using a *K-means* algorithm

[3] We found it useful to start with the maximum likelihood (ML) assignment as an initial guess, followed by Gibbs sampling.

[40] [4]. The resulting mixture is represented using its mixing probabilities c_k and parameters $\vec{\psi}_k(y)$ (resulting in $O(c)$ parameters). We denote the resulting approximate distribution $\hat{p}_t(x)$.

For the $I^{(1)}$ case (see Section 2.1.1) it is straightforward to calculate the *I-projection* of $\hat{p}_t(x)$ on the relevant constraints. Thus we can calculate the parameters at time $t+1$, namely $(\vec{\psi}_{t+1}(y), \gamma_{t+1}(y))$, and the corresponding distribution at iteration $t+1$

$$p_{t+1}(x|y) = \hat{p}_t(x) e^{\vec{\phi}(x)\cdot\vec{\psi}_{t+1}(y)+\gamma_{t+1}(y)} \ ,$$

where the parameters are calculated to satisfy the expectation and normalization constraints.

For the higher order cases, such as $I^{(2)}$, the marginals of the mixture do not have a closed form solution, and require approximate methods such as Monte Carlo [76] or loopy belief propagation [147]. For the applications presented here, we used the approximate algorithm only for the $I^{(1)}$ case.

To measure the progress of the algorithm, we would have liked to calculate $I[p_{t+1}(x,y)]$. However, this is typically impractical for the large $|X|$ case, since mixtures do not have a closed form expression for information. The following proposition gives an easily calculable upper bound on $I[p_{t+1}(x,y)]$

Proposition 4 *The information in the distribution at time $t+1$ satisfies*

$$I[p_{t+1}(x,y)] \leq \langle \vec{\psi}_{t+1}(y)\cdot\vec{a}(y) + \gamma_{t+1}(y)\rangle_{\bar{p}(y)} \ . \tag{2.31}$$

Proof:

$$\begin{aligned}
I[p_{t+1}(x,y)] &= \sum_y \bar{p}(y) \sum_x p_{t+1}(x|y) \log \frac{p_{t+1}(x|y)}{p_{t+1}(x)} \\
&= \sum_y \bar{p}(y) \sum_x p_{t+1}(x|y) \log \frac{\hat{p}_t(x)}{p_{t+1}(x)} \\
&\quad + \sum_y \bar{p}(y) \sum_x p_{t+1}(x|y) \left(\vec{\phi}(x)\cdot\vec{\psi}_{t+1}(y) + \gamma_{t+1}(y)\right) \\
&= \sum_x p_{t+1}(x) \log \frac{\hat{p}_t(x)}{p_{t+1}(x)} \\
&\quad + \sum_y \bar{p}(y)\gamma_{t+1}(y) + \sum_y \bar{p}(y)\vec{\psi}_{t+1}(y)\cdot\sum_x p_{t+1}(x|y)\vec{\phi}(x) \\
&= -D_{KL}[p_{t+1}(x)|\hat{p}_t(x)] + \langle\vec{\psi}_{t+1}(y)\cdot\vec{a}(y) + \gamma_{t+1}(y)\rangle_{\bar{p}(y)} \\
&\leq \langle\vec{\psi}_{t+1}(y)\cdot\vec{a}(y) + \gamma_{t+1}(y)\rangle_{\bar{p}(y)} \ ,
\end{aligned}$$

where in the last two steps we used the fact that $p_{t+1}(x|y)$ satisfies the expectation constraints and the non-negativity of the KL divergence.

[4] For the $I^{(1)}$ case, we cluster the vectors $e^{\vec{\psi}(x_i,y)}$ using an $L2$ norm based algorithm, where each vector is weighted by c_k.

The proposition shows that we can bound $I[p_{t+1}(x,y)]$ using a simple function of the current dual parameters. Although this bound may not be decreasing, we have found that it does decrease in the general case, when c (the number of mixture components) is large enough. Furthermore, as the algorithm converges $D_{KL}[p_{t+1}(x)|\hat{p}_t(x)]$ will typically decrease, and thus a tighter bound will be obtained.

To obtain a lower bound on I_{min} we use the duality result. Since any dual feasible point provides a lower bound on the optimal value, we use the current set of dual parameters $(\vec{\psi}_{t+1}(y), \gamma_{t+1}(y))$ (which are not necessarily feasible) to obtain a dual feasible point.

In principle we could find the dual feasible point that is closest to our current set of parameters in the Euclidean sense. An algorithm for projecting a point on an intersection of constraints was introduced by Boyle and Dykstra [19]. However, it is hard to apply it to the case where the number of constraints is large. We therefore choose a less accurate approach described in Figure 2.5. We start with the current set of parameters and perform alternating projections on the set of constraints. This procedure is guaranteed to yield a feasible point [20] although possibly not the one closest to the point we started from. When the initial point *is* feasible, the algorithm will not change its value.

The projection algorithm involves checking if a point $(\gamma(y), \vec{\psi}(y))$ violates any of the dual constraints. As mentioned in Section 2.7.2 there are methods which approximate this check for the large $|X|$ case. The other element of the projection algorithm is a Euclidean projection on a set of points $(\vec{\psi}(y), \gamma(y))$ which satisfy

$$\log \sum_y p(y) e^{\vec{\phi}(x)\cdot\vec{\psi}(y)+\gamma(y)} \leq 0 \ . \tag{2.32}$$

Since this set is convex in $(\gamma(y), \vec{\psi}(y))$ and so is the Euclidean distance, we have a simple convex problem with $O(d \times |Y|)$ parameters, which can be solved easily using an interior point algorithm for example.

The iterations of the algorithm are repeated until the upper and lower bound are sufficiently close (in the results reported here, we stop on a 1% difference). Figure 2.4 shows the bounds obtained by the algorithm, and illustrates their convergence to a single value.

2.8 Extensions

The MinMI principle can be extended in several ways to accommodate different variations on the setup introduced above. We discuss two interesting extensions below.

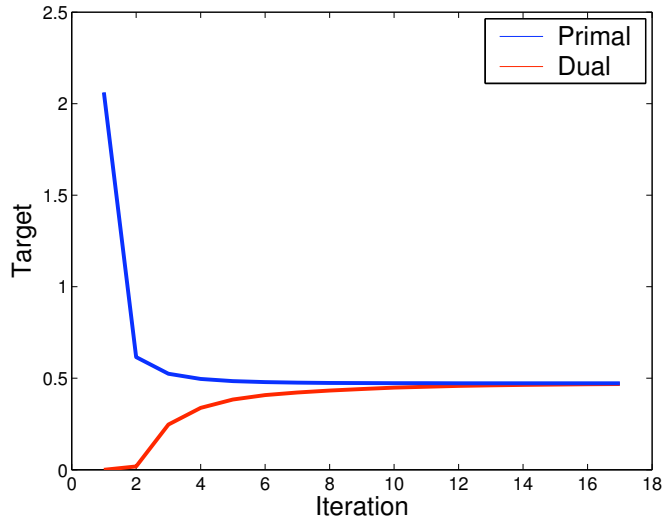


Figure 2.4: Lower and upper bounds on the minimum information obtained using the algorithm of Section 2.7.3. Results are for the $I^{(1)}$ case. We used a binary Y and $X = (X_1, \ldots, X_5)$. The distributions $p(x_i|y)$ were generated randomly.

2.8.1 Uncertainty in Expectation Values

In the presentation of MinMI, we assumed that the expected values $\vec{a}(y)$ were known exactly. In other words, the true underlying distribution was assumed to have these expected values for $\vec{\phi}(x)$. Since these expected values are often calculated from finite samples, they cannot be expected to be exact, but rather to converge to their true values for large enough samples. This concentration around the true means may be quantified using measure concentration theorems such as Chernoff or Markov [36]. This will result in statements such as: "The expected value of $\vec{\phi}(x)$ is expected to be in the range $\vec{a}(y) \pm \vec{\beta}(y)$ at a certainty of 0.95" [5].

In this new setup, we replace the exact knowledge of $\vec{a}(y)$ with some range of possible values. The straightforward extension of MinMI to this case is to consider all distributions which have expected values lying in the given range, and returning the one with minimum information. Define

$$\mathcal{P}_{x|y}\big(\vec{\phi}(x), \vec{a}(y), \vec{\beta}(y), \bar{p}(y)\big) \equiv \Big\{\hat{p}(x,y) : \begin{array}{c} \vec{a}(y) - \vec{\beta}(y) \leq \langle \vec{\phi}(x) \rangle_{\hat{p}(x|y)} \leq \vec{a}(y) + \vec{\beta}(y) \quad \forall y \\ \hat{p}(y) = \bar{p}(y) \end{array} \Big\} . \tag{2.33}$$

Then the minimizing information is defined via

$$I_{min} \equiv \min_{\hat{p}(x,y) \in \mathcal{P}_{x|y}\left(\vec{\phi}(x), \vec{a}(y), \vec{\beta}(y), \bar{p}(y)\right)} I[\hat{p}(x,y)] \ . \tag{2.34}$$

[5] There is still a probability of 0.05 that the values are outside this range. But this level of certainty may be decreased arbitrarily, of course at the cost of increasing the range of uncertainty.

Input: Set of dual parameters $(\vec{\psi}_0(y), \gamma_0(y))$ (possibly infeasible).

Output: A dual feasible point.

Iterate:

- Given current dual parameters $(\vec{\psi}_t(y), \gamma_t(y))$, check for a violated constraint, i.e., look for a value of x such $\log \sum_y \bar{p}(y) e^{\vec{\phi}(x)\cdot\vec{\psi}_t(y)+\gamma_t(y)} > 0$. If non exists, current point is feasible, return it.
- Set $(\vec{\psi}_{t+1}(y), \gamma_{t+1}(y))$ to be the point in the set $\log \sum_y \bar{p}(y) e^{\vec{\phi}(x)\cdot\vec{\psi}(y)+\gamma(y)} \leq 0$ which minimizes the Euclidean distance to $(\vec{\psi}_t(y), \gamma_t(y))$.

Figure 2.5: An algorithm for obtaining dual feasible points from a possibly non-feasible dual set of parameters.

It is easy to see that the constraints on $\hat{p}(x, y)$ are still linear, and therefore the problem is still convex. To obtain the form of the minimizing distribution we follow a procedure similar to [41]. In principle one needs two sets of Lagrange multipliers $\vec{\psi}(y)^+, \vec{\psi}(y)^- \geq 0$ to enforce the two inequality constraints on expectations. However, as shown in Appendix A.1.3, these turn out to be equivalent to one set of multipliers $\vec{\psi}(y)$ as in the original problem. Thus, the information minimization problem in Equation 2.34 turns out to have the exponential form given in Equation 2.6. Of course, the expectation values are different and thus will result in a different solution for the values of $\vec{\psi}(y)$.

An interesting difference is in the dual problem for the current case, which turns out to be

$$\begin{array}{ll} \text{Maximize} & \langle \vec{\psi}(y) \cdot \vec{a}(y) + \gamma(y) - |\vec{\psi}(y)| \cdot \vec{\beta}(y) \rangle_{\bar{p}(y)} \\ \text{Subject to} & \log \sum_y \bar{p}(y) e^{\vec{\phi}(x)\cdot\vec{\psi}(y)+\gamma(y)} \leq 0 \quad \forall x \end{array} . \tag{2.35}$$

The only difference between this and the original dual in Equation 2.12 is the addition of the L1 regularization term $|\vec{\psi}(y)| \cdot \vec{\beta}(y)$.

The optimization algorithm in this case turns out to be identical to that of the equality constraints, with one difference: *I-projection* is on the set $\mathcal{P}_x(\vec{\phi}(x), \vec{a}, \vec{\beta})$ instead of $\mathcal{P}_x(\vec{\phi}(x), \vec{a})$. Such a projection can be easily calculated, as described in Section 1.4.2.

An illustration of I_{min} values in the above case is shown in Figure 2.6. It can be seen that, as expected, information drops off as the range of possible expectation values increases.

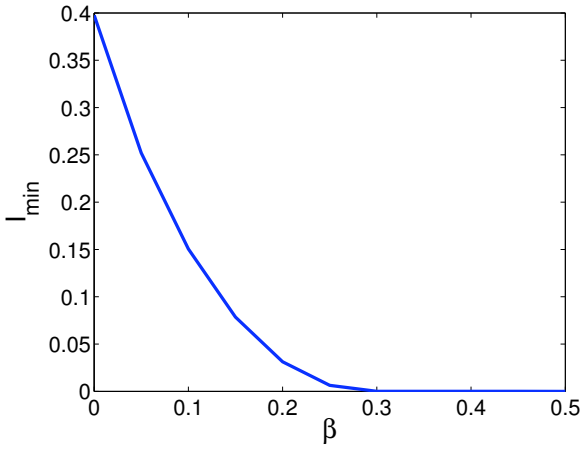


Figure 2.6: Minimum information value for uncertain expectation values, as a function of the uncertainty range $\beta(y)$, which in this case is uniform for all y values. In the simulation we calculated $I^{(1)}$ for $X_1, \ldots, X_{10}$ binary variables with random expectation values. As expected, information drops as β increases since the constraint set increases. Eventually the information drops to zero, since the constraint set includes distributions where $p(x|y)$ is identical for all y values, yielding zero information at the minimum.

2.8.2 Entropy Regularization

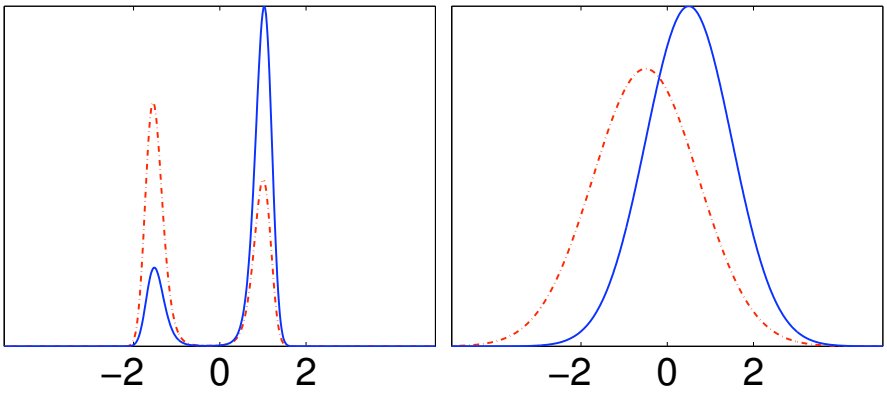
The MinMI algorithm yields a lower bound on the information in the underlying distribution. However, as seen in Section 2.2 the information minimizing distribution is highly sparse, and it is not likely to be the true underlying distribution. There are two possible ways around this situation. One is to add more constraints (i.e., increase the dimensionality of $\vec{\phi}(x)$ so that the constrained set of distribution is smaller). The other is to enforce smoothness constraints on the distribution. Since smooth distributions are closely related to high entropy distributions, we may consider the following target function to be minimized subject to expectation constraints

$$f(\lambda) \equiv I(X;Y) - \lambda H(X) = (1-\lambda)H(X) - H(X|Y) \ , \tag{2.36}$$

where $\lambda \geq 0$ is some tradeoff parameter. Note that the minimum here will no longer yield a bound on the true information, but should yield a smoother distribution. It also has the advantage of creating a continuous parameterization of the range between minimum information ($\lambda = 0$) and maximum entropy ($\lambda = 1$).

The above function is still convex (since $I(X;Y)$ is convex and $H(X)$ is concave), and thus the resulting problem may be solved via convex optimization as long as $|X|$ is not too large. Currently, we do not have an approximate algorithm in this case for large $|X|$.

Using Lagrange multipliers, we may also obtain the following characterization of the minimizing distribution: $p(x|y) = p(x)^{1-\lambda} e^{\vec{\phi}(x)\cdot\vec{\psi}(y)+\gamma(y)}$. The role of the parameter λ can be seen to be an exponential weighting of the prior $p(x)$. This weighting is



(a) (b)

Figure 2.7: Application of MinMI to the first and second moment problem. Here $\vec{\phi}(x) = [x, x^2]$, $\vec{a}(y_1) = [-0.5, 1.75]$, $\vec{a}(y_2) = [0.5, 1.25]$ and $\bar{p}(y)$ is uniform. The x range was 500 equally spaced points between -5 and 5. **a.** The conditional distributions $p_{MI}(x|y_1)$ (red dashed line), $p_{MI}(x|y_2)$ (blue solid line) . **b.** The MaxEnt solution for the given problem.

similar to that seen in the Chernoff bound (see [27] page 312). Algorithmically, the regularized problem seems more complex than MinMI, and we do not currently have approximate algorithms for it. Since it is a convex problem one can always use standard convex optimization algorithms [18] when $O(|X|)$ resources are available. It will be interesting to explore this problem further, and when algorithms are available, compare its performance to that of MaxEnt and MinMI.

2.9 Applications

2.9.1 Moment Matching

To demonstrate some properties of the MinMI solution, we applied it to the well known problem of constraints on the first and second moments of a distribution. The MaxEnt solution to the above problem would be a Gaussian model of $p(x|y)$ with the appropriate mean and variance. The MinMI solution to this problem is shown in Figure 2.7, and is quite different from a Gaussian [6]. The two distributions $p_{MI}(x|y_1), p_{MI}(x|y_2)$ are structured to obey the moment constraints imposed by $\vec{a}(y)$ while keeping as little information as possible about the identity of Y. It can be seen that the solutions concentrate most of their joint mass around two points, while *trying* to maximize their overlap, thereby reducing information content.

[6] The exact solution may be closer to two delta functions, but due to numerical precision issues the algorithm converges to the distribution shown here.

2.9.2 Classification Experiments

We tested the MinMI classification scheme on 12 datasets from the UCI repository [94]. Only the discrete features in each database were considered. The algorithm of Section 2.7.3 was used to calculate the $f_{MI}(y|x)$ prediction function. We have found empirically that early stopping of the algorithm after 15 iterations improves classification results. A similar phenomenon is seen in MaxEnt implementations [98], where early stopping seems to prevent over fitting the training data. The features used as input to the MinMI algorithm were the singleton marginal distributions of each of the features, as described in Section 2.1.1. Classification performance was compared to that of Naive Bayes [7] and the corresponding first order conditional Log-Linear model [8]. The Naive Bayes model is obtained from the empirical singleton marginals $\bar{p}(x_i|y)$ simply by

$$p(y|x_1,\ldots,x_n) = \frac{\bar{p}(y)}{Z_x}\prod_i \bar{p}(x_i|y) \ . \tag{2.37}$$

The log linear model is given by

$$p(y|x_1,\ldots,x_n) = \frac{1}{Z_x}e^{\sum_i \psi(x_i,y)} \ , \tag{2.38}$$

where the function ψ is found by maximizing conditional likelihood. Recall also that the MinMI prediction function is given by (see Equation 2.11)

$$f_{MI}(y|x_1,\ldots,x_n) = \bar{p}(y)e^{\sum_i \psi(x_i,y)} \ . \tag{2.39}$$

All models thus have a similar parametric shape, but their parameters are obtained via different optimization schemes and principles. A comparison of Naive Bayes and Logistic Regression was also carried out in [97].

The results for all the datasets are shown in Figure 2.8. Both MinMI and Naive Bayes can be seen to outperform the Log-Linear model on small sample sizes as described previously in [97] for Naive Bayes. This result is intuitively plausible (and is also justified rigorously in [97]), since both MinMI and Naive Bayes use only first order statistics, which can be reliably estimated from small samples, whereas the Log-Linear model uses the raw data (see Section 2.6.3). MinMI outperforms Naive Bayes on three databases [9], and is outperformed by it on four [10]. On the other databases, performance is similar. In databases where MinMI does not perform as well as Naive Bayes, it is likely that the conditional independence assumption is valid, and therefore Naive Bayes

[7] Marginals used for Naive Bayes and MinMI were estimated using Laplace smoothing with a pseudo-count of 1, as in [97].

[8] In the linearly separable case the conditional model solution is not unique. As in [97] we randomly sample separating hyperplanes, by carrying out a random walk in version space. The reported performance is the average generalization error over the sampled hyperplanes.

[9] voting-records,credit and hypo

[10] heart-disease,lymphography, promoters and splice

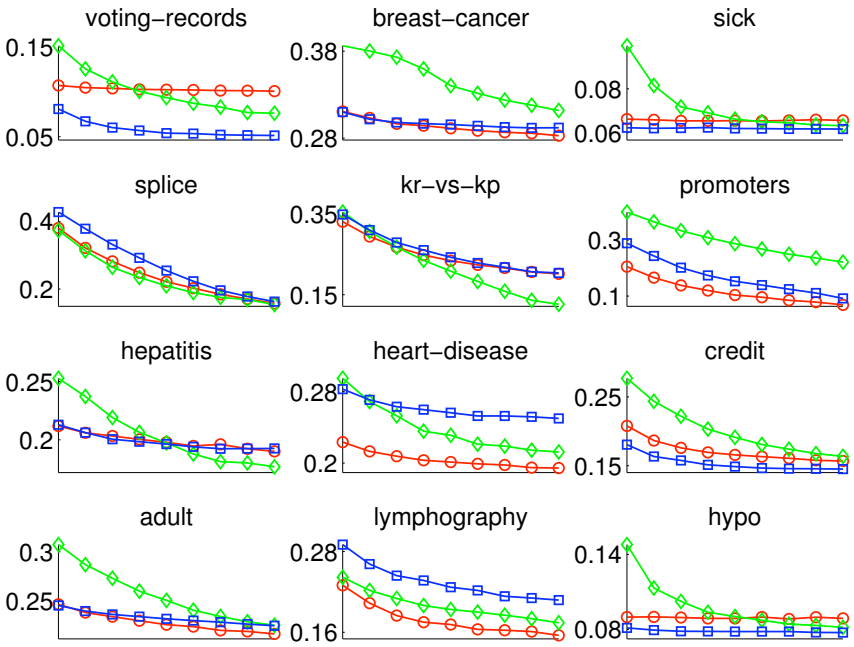


Figure 2.8: Classification error as a function of the number of samples used in training, for several UCI datasets. MinMI is blue line and squares, Naive Bayes is red line and circles, Log-Linear model is green line and diamonds. For each sample size, 1000 random splits of the data were performed. The samples sizes in the plots are $20, 30, \ldots, 100$. The marker sizes are larger than the standard errors.

may use it to average out noise in the features, whereas MinMI will not. On the other hand, MinMI is likely to outperform Naive Bayes in databases where there is a large set of correlated features, and the conditional independence assumption fails.

In the current experiments we used only singleton statistics. It would be interesting to explore the use of higher order statistics in classification. However, using all feature combinations is likely to result in over-fitting, so the method should be augmented with a feature selection mechanism, as in [43].

2.10 Discussion

We introduced the principle of minimum mutual information (MinMI) as a fundamental method for inferring a joint distribution in the presence of empirical conditional expectations. This principle replaces Maximum Entropy for such cases and in general is not equivalent to a maximum likelihood estimation of any parametric model.

It is interesting to note that the MinMI solution for a multivariate X does not satisfy the conditional independence properties which the corresponding graphical model possesses. This is clear already when singleton marginals are used as constraints. The resulting $p_{MI}(x|y)$ may in fact contain elaborate dependencies between the variables.

To see why this comes about consider the extreme case where all the conditional singleton marginals are constrained to be equal. It is easy to see that under $p_{MI}(x|y)$ the variables $X_1, \ldots, X_N$ will be completely dependent (i.e., $p_{MI}(x_1, \ldots, x_N|y) = p_{MI}(x_1|y)$).

It is important to stress that $p_{MI}(x)$ is not argued to be a *model* of the true underlying distribution. Rather, as the game theoretic analysis shows, it represents a worst case scenario with respect to prediction.

Although we did not address the case of continuous X domain directly, our formalism applies there as well. Consider a vector of continuous variables, with constraints on the means and covariances of subsets of its variables. The MinMI distribution in this case will be related to the corresponding Gaussian Markov field.

Another natural extension of the current work is feature induction [34]. As will be shown in the Chapter 4, one can look for features $\vec{\phi}(x)$ which maximize the minimum mutual information calculated in the current chapter, assuming both marginals are known. The extension to unknown marginals should provide a powerful tool for feature induction over variables sets, and will be an interesting subject for future research.

Chapter 3

Application of MinMI to Studying the Neural Code

The previous chapter presented the MinMI formalism as a method for calculating information given partial measurements of a system. In the current chapter we illustrate how this concept may be used to study questions related to neural coding under various scenarios. The structure of the neural code has been extensively studied over the last 50 years, with information theory playing an important methodological and conceptual role [116]. The first step in an information theoretic study is to calculate mutual information between a behavioral variable (e.g., movement [65], visual stimulus [14] etc) and some property of the neural response (e.g., spike counts of single neurons [49], joint statistics of pairs of neurons [113] and precise spike times [31] among others). Significant information values indicate that the given neural property may be used to predict the behavioral variable, and is thus likely to be physiologically involved in generating it (in the case of movement) or processing it (in case of external stimuli).

One caveat in the above paradigm is that even if a given neural response carries information about behavior, it is not clear exactly *which* property of the neural response is informative. To illustrate this, suppose one measures the information between 100 neurons $X_1, \ldots, X_{100}$ and a stimulus Y and finds it is relatively high [95]. It is still not clear which properties of the response convey the information: single neuron statistics, pairwise statistics, or precise firing patterns. To approach this difficulty, numerous information theoretic measures have been proposed to quantify information in higher order interactions [121, 100, 95, 23]. We discuss those in Section 3.1 and illustrate their relation to the MinMI principle.

The MinMI formalism allows several important extensions of information measurement in neural systems, which we demonstrate in this chapter. One advantage of our method is that it allows information estimation in scenarios where limited statistics are given. For example it allows one to ask questions about information in a large population of neurons even if no joint statistics can be estimated. Another advantage

is that because of its definition, the minimum information quantifies *only* the information available in the given statistics and does not assume any unnecessary coding mechanisms (e.g., independent coding by different neurons).

In what follows, we apply MinMI to several properties of neural codes, showing how it can be used to reveal properties which are not accessible by standard methods.

3.1 Synergy and Redundancy Measures

A central issue in neural coding is the importance of higher order statistics, and their contribution with respect to lower order statistics. A possible way of quantifying this contribution is to calculate the difference between higher order information and that in some model based on lower order statistics. A positive difference indicates *synergy*: information in higher order interactions, while a negative difference indicates *redundancy*. One so called synergy/redundancy measure has been suggested in [49, 121]

$$SynSum(X_1, \ldots, X_N, Y) = I(X_1, \ldots, X_N; Y) - \sum_i I(X_i; Y) \ , \tag{3.1}$$

which measures the difference between full information and the sum of individual informations. One shortcoming of the above measure is that the second term becomes dominant as N grows (the first is always bounded by $H(Y)$). Thus, large populations will always appear redundant.

Another possible measure of synergy compares the full information to the information in the case where neurons are conditionally independent (CI) given the stimulus

$$SynCI(X_1, \ldots, X_N, Y) = I(X_1, \ldots, X_N; Y) - I_{CI}(X_1, \ldots, X_N; Y) \ , \tag{3.2}$$

where I_{CI} is the information under the distribution $p_{CI}(x|y) \equiv \prod_{i=1}^{n} p(x_i|y)$ (this measure was denoted by ΔI_{noise} in [121]). Note that this measure does not necessarily grow with N and will equal zero when the neurons are CI. Another related measure based on the CI case, but not directly using information, was introduced in [100].

Both $SynSum$ and $SynCI$ compare the full information to that in first order statistics. Moreover, the typical implementation of these measures is for the two neuron case, where the only statistics less than full order are first order.

The generalization of synergy measures for higher order statistics, and $N > 2$ populations poses an important challenge. The $SynSum$ measure has been generalized to this scenario in [122], where it was decomposed into elements measuring synergy in k^{th} order correlations. The authors used the MaxEnt approach to estimate the expected entropy given only k^{th} order correlations.

MinMI offers an elegant approach for generalizing the $SynCI$ measure to higher orders. We first illustrate this for second order statistics in an $N > 2$ population. The $I^{(2)}$ measure quantifies the information available in a population given only its (first

and) second order statistics. To turn it into a synergy measure, we need to subtract the expected second order information in the CI model. If the neurons are CI, the pairwise statistics are expected to be $p(x_i, x_j|y) = p(x_i|y)p(x_j|y)$. We denote the minimum information in these pairwise statistics by $I^{(2)}_{CI}$. A natural measure of synergy is then the difference

$$SynI^{(2)}(X_1, \ldots, X_N, Y) = I^{(2)} - I^{(2)}_{CI} \ . \tag{3.3}$$

When the true population is conditionally independent, $SynI^{(2)} = 0$, as expected. Furthermore, when $N = 2$, we have that $SynI^{(2)} = SynCI$. Thus MinMI generalizes $SynCI$ to the study of pairwise interactions in large populations.

The $SynI^{(2)}$ measure may be generalized to studying k^{th} order correlations in large populations. The conditionally independent distribution in this case needs to be replaced with its k^{th} order equivalent. A possible candidate is the maximum entropy distribution for $(k-1)^{th}$ constraints [90, 122], which yields the CI distribution for $k = 2$.

3.2 Methods

The applications discussed in this chapter include both simulated and experimental data. In all cases, we calculate the $I^{(1)}, I^{(2)}$ information measures and show how they can be used to study coding mechanisms. The current section describes the experimental data briefly, and discusses some methodological issues in calculating the information quantities.

3.2.1 The Experimental Paradigm

The data presented here was obtained in experiments studying motor control in monkeys. The experiments were conducted by Ron Paz in the laboratory of Prof. Vaadia. Monkeys were trained to perform unimanual movements by operating two X-Y manipulanda. The movements were standard center-out reaching tasks with eight movement directions [51], and a delayed GO signal [1]. For comprehensive experimental details, see [107].

In each trial, the monkey was first presented with a signal indicating which hand will perform the movement (Laterality Cue). After a hold period ($1000 - 1500$ ms), a signal (Target Onset) was given indicating the direction to move to (one of eight). This was followed by an additional hold period ($1000 - 1500$ ms). During the hold periods, the monkey was instructed to hold the cursor in a circle located at the center of the screen. After the second hold period, the circle disappeared (Go Signal), and

[1] The complete behavioral paradigm included other components such as learning visuomotor transformations, but these are not relevant for our analysis

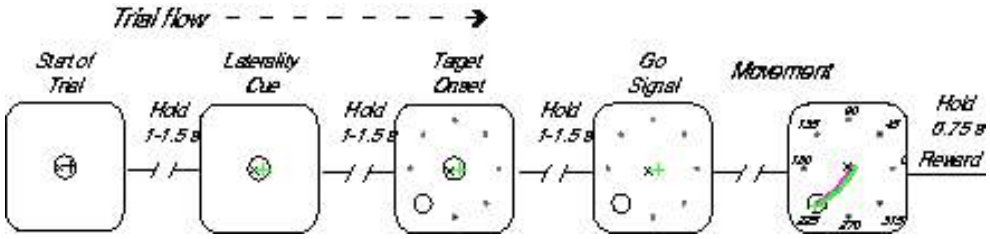


Figure 3.1: The center out movement task performed by behaving monkeys. The monkey is initially presented with a laterality cue, followed by the target location, and a go signal. These three signals are separated by hold periods of $1 - 1.5$ seconds.

the monkey was instructed to move to the specified target. The trial flow is illustrated in Figure 3.1.

During task performance, single-unit activity was recorded extra-cellularly from the primary and supplementary motor areas of the cortex. The data presented here is taken from 827 neurons in the primary motor cortex (MI). The number of successful repetitions per movement direction varied in the range $5 - 20$, resulting in $40 - 160$ for each laterality cue.

3.2.2 Quantization and Bias Correction

Spike trains were represented using the total number of spikes (i.e., spike count) in windows of different sizes (depending on the application), with a maximum size of $600ms$. Due to the low number of repetitions, we could not estimate the response distribution reliably from the raw spike count. To clarify why, consider for example a case where the spike counts vary in the range $0 - 20$ spikes, and the number of repetitions per y is 20. Then each spike count will be appear one time on average, precluding reliable estimate of its appearance probability.

To overcome this difficulty, we chose to quantize the spike counts into a small number of bins. There are several possible binning strategies, for example uniform (i.e., $1, 2, 3, 4, \ldots, 19, 20$), equal number of samples per bin, log scale etc. We chose a different, greedy scheme, where one searches for the binning which maximizes the mutual information. Because the number of possible binning schemes is exponential in the maximum spike count, we used a suboptimal search algorithm which unified bins that increased information. In calculating information we applied the Panzeri-Treves bias correction term [105] (see also [103] for a comprehensive survey of bias estimation.). This partly compensated for the increase of information with the number of bins. A similar quantization scheme was recently used in [96].

The above quantization algorithm was used to determine the bin sizes to be used in calculating the observed marginals, and the associated minimum information values. As in any mutual information estimation from finite samples, the minimum in-

formation statistic has a positive bias [2]. To overcome this bias, we used a bootstrap scheme: the entire information calculation (including quantization) was repeated N times, where each repetition was on a new sample which was generated by drawing trials uniformly with repetitions from the original set of trials. This resulted in N bootstrapped information values $I(1), \ldots, I(N)$. An estimate of the bias was then given by $I(0) - \frac{1}{N}\sum_{k=1}^{N} I(k)$ where $I(0)$ is the information calculated from the original sample. This resulted in an effective removal of the bias term as verified on simulated data.

All information values in this section are calculated in bits.

3.3 Results

Neural population codes may be studied at several levels, corresponding to different coding strategies. The basic level is the single neuron code. Next is the relation between the codes of different single neurons. Higher order interactions between neurons constitute yet another level. Finally, temporal structure may also be used to enhance coding efficiency. In the applications below, we show how the MinMI principle may be applied to the study of various neural coding schemes and quantify the level to which different populations use these schemes.

3.3.1 Two Binary Neurons and a Binary Stimulus

We begin with an illustration of MinMI calculation for the case of two artificial binary neurons X_1, X_2 so that each neuron has two possible responses. The stimulus Y is also taken to be binary. We assume that the two neurons were measured separately, so that only $p(x_1|y), p(x_2|y), p(y)$ are known and $p(x_1, x_2|y)$ is not known. We are interested in the minimum information available in a distribution $\hat{p}(x_1, x_2, y)$ satisfying the first order constraints $\hat{p}(x_i|y) = p(x_i|y)$, $i = 1, 2$. Note that any such distribution is completely defined by two numbers $\hat{p}(x_i = 1|y)$, since for each value of S, $\hat{p}(x_1, x_2|y)$ has four free parameters and has to satisfy three constraints (two first order constraints and one normalization constraint, $\sum_{x_1,x_2} \hat{p}(x_1, x_2|y) = 1$). In this specific case, the space of possible distributions $\hat{p}(x, y)$ can be visualized in two dimensions, as shown in Figure 3.2. The figure shows the value of the MI for each possible distribution in $\hat{p}(x, y)$ satisfying the constraints above. This is done for two different pairs of neurons, with different response distributions. The location of the MinMI distribution $p_{MI}(x_1, x_2|y)$ is also shown. A different distribution $\hat{p}(x, y)$, which yields more information, is the one in which the neurons are conditionally independent (CI) given the stimulus: $\hat{p}(x_1, x_2|y) = \hat{p}(x_1|y)\hat{p}(x_2|y)$. In such a case, averaging over the responses of the neurons will reduce

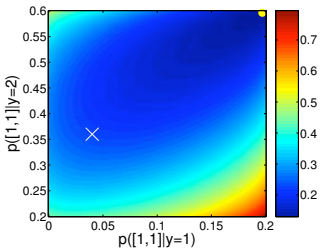
[2] To see why, consider a distribution with zero mutual information. Since information is non-negative, any estimation from finite samples will generate a non-negative value, and therefore the mean estimate will be strictly positive, yielding a biased estimate.

(a)

y	$p(x_1 = 1)$	$p(x_2 = 1)$
1	0.2	0.2
2	0.6	0.6

(b)

y	[0,0]	[0,1]	[1,0]	[1,1]
1	0.8	0	0	0.2
2	0.4	0	0	0.6



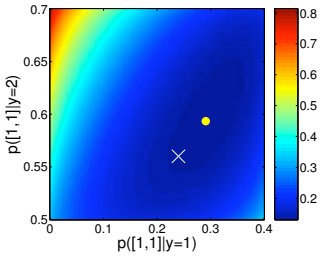
(c)

(d)

y	$p(x_1 = 1)$	$p(x_2 = 1)$
1	0.4	0.6
2	0.8	0.7

(e)

y	[0,0]	[0,1]	[1,0]	[1,1]
1	0.3	0.3	0.1	0.3
2	0.1	0.1	0.2	0.6



(f)

Figure 3.2: Illustration of $I^{(1)}$ for two binary neurons and a binary stimulus ($p(s) = 0.5$). Only the first order statistics of each neuron are assumed to be known. The results for two different neuronal responses are shown. Panel (a) gives the probability of each neuron firing for each of the two stimuli. Panel (b) shows the minimum information distribution $p_{MI}(x|y)$ for the statistics in (a). Panel (c) shows the information in all distributions satisfying with the given first order statistics. The yellow dot shows the location of the MinMI distribution in this information plane, and the white cross shows the CI distribution. The X and Y axes measure the probability of both neurons firing for stimuli $s = 1, 2$. Note that these two parameters specify the entire response distribution. Panels (d-f) follow the same conventions as panels (a-c).

noise, and thus yield more information about Y than in the MinMI distribution. The CI distribution is also shown (Figure 3.2c,f) and always yields more information than the MinMI one.

In the first example (Figure 3.2a-c) the two neurons have the same response distributions $p(x_1|y) = p(x_2|y)$. The MinMI distribution shown in the figure is the one in which the neurons are completely correlated, and thus lies on the boundary of the space of possible distributions. It is intuitively clear why this is the minimum: the two neurons are equivalent to a single neuron.

In contrast, when the two neurons differ in their response distributions (Figure 3.2d-f), they cannot be completely correlated. Thus, the information minimizing distribution will not lie on the boundary as in the previous example (compare Figure 3.2c with Figure 3.2f).

3.3.2 Coding Redundancy in Single Neurons

We next illustrate the use of MinMI in the study of single neuron codes and their combination in a population. An important question is this context is whether all neurons respond to similar stimulus properties, or do they provide complementary information. As an example, consider a population of neurons where each neuron is tuned to some preferred direction (PD) in the stimulus (i.e., movement direction in motor neurons, or orientation in visual neurons). The question now is how the PDs themselves are distributed among the neurons. In one extreme, all neurons have the same PD, while in the other extreme PDs are uniformly distributed among neurons. It is intuitively clear that the second scenario is advantageous in terms of stimulus coding. However, it is not clear how to quantify this intuition in terms of information, especially when the joint distribution of the population cannot be estimated.

The MinMI principle provides a natural framework for tackling the above problem. Ideally, in studying information in populations we are interested in the quantity $I(X_1, \ldots, X_N; Y)$. More specifically, we are interested in the contribution of single neuron codes to this information. Our $I^{(1)}$ measure provides precisely that. To illustrate how $I^{(1)}$ differentiates between different single neurons coding schemes, we simulate data from three hypothetical neuronal populations, with different degrees of overlap between single neuron codes [3]. Figure 3.3 shows the code structure for these populations and the respective $I^{(1)}$ values. The results correspond to the intuition mentioned above: low $I^{(1)}$ values correspond to populations with high overlap between single neuron codes, and high values correspond to low overlap. Note that the MinMI calculation is model-free, and thus does not use the concept of direction tuning or preferred direction. It can thus detect differences in population coding in considerably more complex scenarios, which could be very hard to visualize.

Dependence on Population Size

One of the overwhelming properties of the brain the enormous number of neurons comprising it (roughly 10^{10}). It is intuitively clear that this setup carries significant information processing potential. However, it is not always straightforward to analytically state how information depends on population size. Several works have studied this dependence, both from a theoretical [120], and practical [140] viewpoint. It is thus interesting to study how the MinMI measure grows with population size.

We applied $I^{(1)}$ calculation to the experimental data described in Section 3.2. Information about movement direction was calculated for two behavioral epochs: Target Onset and Go Signal. In both cases, we considered an epoch of 600 ms after the behavioral signal, and calculated $I^{(1)}$ for populations of increasing size [4]. Figure 3.4

[3] Parameters used in the tuning function are physiologically plausible, as verified on real data.

[4] In both cases we took only neurons that had significant mutual information when considered in

shows that $I^{(1)}$ grows monotonously with population size, with a steeper increase for smaller population sizes [5]. Moreover, it is clearly seen that movement related activity (after the Go Signal) is more informative than preparatory activity (after the Target Onset) for all population sizes, although this becomes more apparent for larger populations. This is not surprising since motor cortical neurons are typically less tuned during preparatory activity (see e.g. [107] for results on the current data).

The shape of the information curve here is similar to results in [140, 120]. Note, however that here, unlike in [120], no assumption is made about independence between neurons, and the only source of information is the individual neuronal responses. It is also interesting to note that under the conditional independence assumption, information will typically saturate its maximum value [66] as population size grows. This will occur since the system can eliminate all the noise in the responses by averaging. The MinMI measure, on the other hand, does not have this property. For example in the case where all neurons have identical responses, increasing population size will not increase information (see Section 3.3.1). This is clearly an advantage for discriminating between large populations. Considering the above observations, it is interesting to note that the information during movement (Go Signal) nearly saturates its full value of 3 bits [6]. This may be taken to indicate that the preferred directions of neurons are distributed in a manner which covers space efficiently, as in the rightmost panel in Figure 3.3.

3.3.3 Pairwise Coding in Populations

Second order statistics between neurons have been shown to play a part in neural coding, in the sense that their joint activity provides more information about a stimulus than their individual responses [65, 137]. Most research has been devoted to studying pairs of neurons, mostly due to sampling problems (but see [95]). It is clear, however, that if the second order statistics between all pairs in a population provide information about the *same* property of the stimulus, this should result in less information than if different pairs encode different stimulus properties. This situation is the pairwise equivalent of the single neuron coding issue discussed in the previous section.

The information available from the grouped pairwise responses in a population can be quantified using the $I^{(2)}$ measure. Figure 3.5 shows two toy populations, four neurons each, with identical pairwise statistics and therefore identical synergy values: the set of $SynSum(X_i, X_j, Y)$ $(i, j \in \{1, \ldots, 4\})$ values is identical in both populations. Furthermore, in this case $SynSum = SynCI$ since $I(X_i; Y) = 0$ for all neurons. In

isolation from the population. This resulted in 115 units for Target Onset, and 360 units for the Go Signal.

[5] The values for higher population sizes may be slightly positively biased since bootstrapped values sometimes reached the maximal values of 3 bits and thus the actual bias may be higher.

[6] For the Target Onset case, more neurons will be needed to determine the maximum information value.

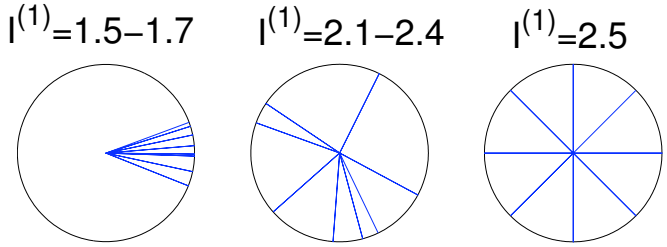


Figure 3.3: The information $I^{(1)}$ for different population coding schemes. We consider three populations of eight neurons responding to eight stimuli. The stimuli correspond to eight equally spaced directions on the circle ($s = \{0°, 45°, \ldots, 315°\}$). All neurons are cosine tuned with PDs given in the polar plots ($p(x_i|y) = Poiss\left(x_i|5 + 5\cos\left(y - \theta_i\right)\right)$, where $Poiss(r|\lambda)$ is probability of count r under a Poisson distribution with rate λ , and θ_i is the PD of neuron i. Responses above 15 spikes were clipped to a value of 15). Left panel shows *overlapping* tuning where all neurons have similar PDs (directions were drawn uniformly in the range $\pm 22.5°$). Middle panel shows tuning to *random* directions. In the right panel, neurons are tuned to *equally spaced* directions. $I^{(1)}$ values are given for each scenario (values for the *overlapping* and *random* tunings were obtained by drawing PDs 1000 times and calculating a 95% confidence interval).

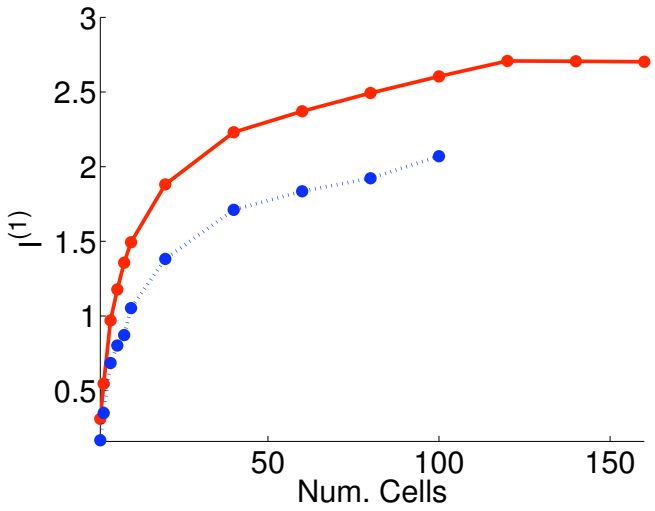


Figure 3.4: The information $I^{(1)}$ for increasing population sizes. The figure shows $I^{(1)}$ about movement direction as a function of population size for two behavioral epochs: Target Onset (blue dotted line) and Go Signal (red solid line). Values were generated by drawing random subsets of n neurons randomly, and calculating their $I^{(1)}$ values. Fifty subsets were generated and the average information values are shown.

	p_1	p_2
[0,0]	0.25	0.4
[0,1]	0.25	0.1
[1,0]	0.25	0.1
[1,1]	0.25	0.4

(a)

$I^{(2)} = 0.14$

y	x_1, x_2	x_3, x_4
1	p_1	p_1
2	p_2	p_1
3	p_1	p_2
4	p_2	p_2

(b)

$I^{(2)} = 0.7$

y	x_1, x_2	x_3, x_4
1	p_1	p_1
2	p_2	p_2
3	p_1	p_1
4	p_2	p_2

(c)

Figure 3.5: Information in populations from pairwise statistics. We consider the responses of four toy neurons $x_1, \ldots, x_4$ to four stimuli $y = \{1, \ldots, 4\}$ ($p(y) = 0.25$). Neurons x_1, x_2 are conditionally independent from x_3, x_4. Panel (a) defines the response distributions of two pairs of neurons. Note that the information in any single neuron is zero. Panels (b) and (c) give the response of the four neurons under two different scenarios by specifying the response of each pair. In both scenarios, pairwise synergy values ($SynSum$ and $SynCI$, which are equal in this case) are 0.7 for pairs (x_1, x_2) and (x_3, x_4) and zero for the other four pairs. However, the $SynI^{(2)}$ values for each distribution are different, as shown above panels (b) and (c).

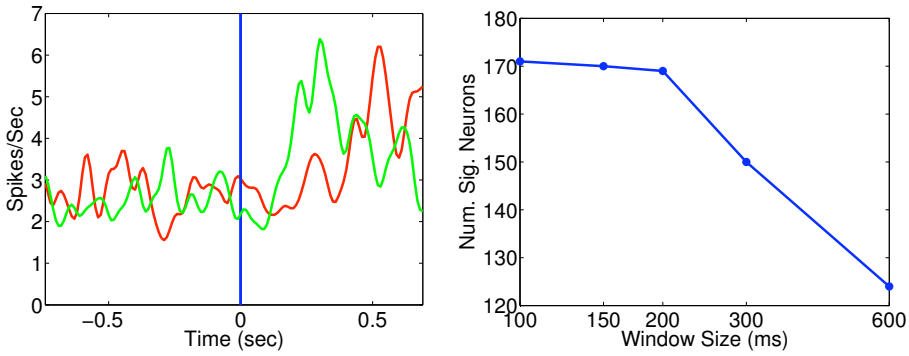
one population, all synergistic coding provides information about the same property of the stimulus, whereas in the other the pairwise codes are designed to provide disparate information. The difference between these two populations is clearly seen in their $I^{(2)}$ values. Thus the MinMI principle can be used to differentiate between populations with different pairwise code designs.

3.3.4 Temporal Coding

Temporal response profiles of single neurons may transmit information about behaviorally relevant variables [101, 102]. Intuitively, one could argue that if different behavioral parameters induce different response profiles, as measured by a Peri-Event Time Histogram (PETH, [108]), then the temporal response carries information about the variable. Our MinMI formalism allows us to make this statement explicit and to calculate the resulting information.

The response function of a neuron can be given by its response in a series of time bins $p(x_t|y)$, $t = 1 \ldots T$. A PETH is an example of such a profile where x_t is a binary variable, and one plots the rate function $p(x_t = 1|y)$ (usually scaled to spikes per second). The responses $p(x_t|y)$ are merely a set of first order statistics (disregarding correlations between bins) and thus we can calculate $I^{(1)}$ for these statistics, in order to obtain a measure of information in a PETH.

Figure 3.6 illustrates the application of MinMI to temporal coding in recordings from the primary motor cortex of behaving monkeys (see Section 3.2 for experimental



(a) (b)

Figure 3.6: Analysis of temporal coding using MinMI. Panel (a) shows the PETHs of the response to the laterality signal (left hand - red, right hand - green), for a neuron recorded in the primary motor cortex. The $I^{(1)}$ measure was significant for window size 200 ms and below but not for 300 ms or 600 ms. Panel (b) shows the number of neurons with significant $I^{(1)}$ ($p < 0.05$) as a function of the window size.

methods and data analysis). We consider the response to the binary laterality cue, which instructs the monkey which hand to move. Figure 3.6 shows a PETH of a neuron, where the spike count over a period of 600 ms is similar for both conditions. However, the temporal profiles differ between the two conditions. To analyze this coding using $I^{(1)}$ we partitioned the 600 ms period into $600, 300, 200, 150, 100$ ms windows, and calculated $p(x_t|y)$ and the corresponding $I^{(1)}$ for each partition. We then shuffled the trials between laterality signals and compared the shuffled values to the raw $I^{(1)}$ in order to test if the raw information was significantly different from zero. For the neuron in Figure 3.6a, we found that it was not significant for window sizes of 300 ms and above, but was significant for all lower sized windows. This indicates that MinMI may be used to detect information related to temporal structure.

We repeated the above procedure for the entire population of 827 neurons, and counted the number of significant neurons for each window size. Figure 3.6b shows this number as a function of window size. A large increase can be seen when moving from 600 ms to 200 ms, indicating relevant temporal structure at these time constants. The number then flattens for lower window sizes, suggesting no temporal information on these scales. Note that the number of significant neurons increases from 125 to 170, an increase of 35%.

3.4 Discussion

The current chapter illustrated the application of the MinMI principle to studying the neural code. In this context, the principle has three attractive properties. The

first is the ability to obtain a bound on information from a wide array of statistical observations and experimental scenarios. The second is that we can extend standard information measures (such as information in single neurons or pairs of neurons) to large populations, allowing the detection of complex coding schemes not evident when analyzing a small number of neurons (one or two) individually. The third is that our formalism allows us to measure information in specific aspects of the response such as first, second, and higher order responses. These advantages improve on current IT applications in neuroscience and provide a comprehensive framework for tackling fundamental issues in neural coding.

While the results presented here applied to neural coding, the MinMI principle is general and may be used in studying a wide array of complex systems. For instance, it may be used to estimate the information in a set of gene expression profiles about external conditions [47], and thus help in analyzing their functional role, and comparing different gene regulatory networks.

The MinMI measure is fundamentally different from maximum entropy (MaxEnt) based measures [90, 122]. This can be seen already at the single neuron level, where maximizing the entropy $H(X|Y)$ would yield a conditionally independent (CI) distribution. The MinMI distribution in this case is less informative (Figure 3.2), and illustrates that CI is an overly optimistic assumption with respect to information content.

Information in populations may be estimated without explicitly calculating the full joint statistics. A possible method for doing so, is via the *reconstruction method* which uses X to generate an approximation $\hat{Y}$ of Y [17]. A lower bound on the true information is then obtained from $I(Y;\hat{Y})$ [14, 95]. While these methods may be efficient in estimating the true value of the information, it is not clear what property (e.g. statistical interaction order) of the stimulus-response statistics generates this information. In contradistinction, the MinMI framework allows us to calculate information in a given statistical order *embedded* in a population of any size.

MinMI may also be used to study synergy in a given statistical order embedded in a population. For example, in Section 3.1 we extend pairwise synergy measures to populations with $N > 2$, using only pairwise statistics. This synergy measure is a single number which reflects the interaction of pairwise codes in the population. This approach differs from the common practice of analyzing pairs independently [31, 49, 65, 99], since the latter ignores population related effects (Section 3.3.3). Specifically interesting in this context is the question of interaction between pairwise codes (i.e., pairwise codes embedded in a population) is of considerable interest, and has received some attention in recent literature [113, 104] but is still largely unresolved. MinMI is an attractive tool in addressing this question, and could hopefully aid in understanding this aspect of the neural code.

Chapter 4

Sufficient Dimensionality Reduction

In the previous chapters we introduced the principle of Minimum Mutual Information as a method of characterizing the information available in statistical measurements. Throughout our discussion we assumed that the functions $\vec{\phi}(x)$, whose expected values we measured, were given in advance. Indeed, in many cases there are natural choices for such functions (e.g. responses of single cells or pairs etc.). However, in the general case it is interesting to ask what are the functions whose measurements provide the maximum information. The current chapter introduces a method for obtaining these functions. Due to its close link to the notion of statistical sufficiency we call our method Sufficient Dimensionality Reduction (SDR).

The starting point of our approach is the Minimum Mutual Information available in the measurement of a function $\vec{\phi}(x)$ under a distribution $\bar{p}(x, y)$. We denote this information by $I^{xy}_{min}[\vec{\phi}(x), \bar{p}]$ [1]. The quantity $I^{xy}_{min}[\vec{\phi}(x), \bar{p}]$ captures the amount of information available in measuring the expected value of $\vec{\phi}(x)$. We next look for a function $\vec{\phi}(x)$ which *maximizes* this information. This results in a Max-Min problem whose unknown is the function $\vec{\phi}(x)$.

This problem will be shown to be equivalent to finding a model of a special exponential form which is closest to the empirical distribution (contingency table) in the KL-divergence (or Maximum Likelihood) sense. We then present an iterative algorithm for solving these problems and prove its convergence to a solution. An interesting information geometric formulation of our algorithm is then suggested, which provides a covariant formulation of the extracted features. It also provides interesting sample complexity relations via the Cramer-Rao inequalities for the selected features. We conclude by demonstrating the performance of the algorithm, first on artificial data and then on a real-life document classification and retrieval problem.

[1] Although $I^{xy}_{min}[\vec{\phi}(x), \bar{p}]$ is very similar to I_{min} defined in the previous chapter, they are not equivalent due to the marginal constraints on $p(x)$ as will be seen later.

Most of the material in the current chapter was published in [56].

4.1 Problem Formulation

To illustrate our motivation, consider a die with an unknown outcome distribution. Suppose we are given the mean outcome of a die roll. What information does it provide about the probability to obtain 6 in rolling this die? One possible "answer" to this question was suggested by Jaynes [73] in his "Maximum Entropy (MaxEnt) principle".

Denote by $X = \{1, \ldots, 6\}$ the possible outcomes of the die roll, and by $\vec{\phi}(x)$ an observation, feature, or function of X. In the example of the expected outcome of the die, the observation is $\vec{\phi}(x) = x$, the specific outcome. Given the result of n rolls $x_1, \ldots x_n$, the empirical expected value is $\vec{a} = \frac{1}{n}\sum_{i=1}^{n}\vec{\phi}(x_i)$. MaxEnt argues that the "most probable" outcome distribution $\hat{p}(x)$ is the one with maximum entropy among all distributions satisfying the observation constraint, $\langle\vec{\phi}(x)\rangle_{\hat{p}(x)} = \vec{a}$. This distribution depends, of course, on the actual value of the observed expectation, $\vec{a}$.

The MaxEnt principle is considered problematic by many since it makes *implicit* assumptions about the distribution of the underlying "micro-states", that are not always justified. In this example there is in fact a uniform assumption on all the possible sequences of die outcomes that are consistent with the observations. MaxEnt also does not tell us *what* are the features whose expected values provide the *maximal* information about the desired unknown - in this case the probability to obtain 6. This question is meaningful only given an additional random variable Y which denotes (parameterizes) a set of possible distributions $p(x|y)$, and one measures feature quality with respect to this variable. In the die case we can consider the Y parameter as the probability to obtain 6. The optimal measurement, or observation, in this case is obviously the expected number of times 6 has occurred, namely, the expectation of the *single* feature $\phi(x) = \delta(6 - x)$. The interesting question is what is the general procedure for finding such features.

An important step towards formulating such a procedure is to quantify the information in a feature function $\vec{\phi}(x)$. This was exactly the goal of the previous chapter, which defined this information as the minimum information available in any distribution agreeing with empirically measured values of $\vec{\phi}(x)$. Here we use a nearly identical definition, but with the additional assumption of knowledge of the marginal of X. In the previous chapter we were interested in cases where $|X|$ may be extremely large, so that $p(x)$ may not be feasibly measured or even stored. Here we will limit the discussion to the setup where $p(x)$ can be measured. As we shall see, this will considerably simplify the analysis and algorithm,

Formally, we define $I^{xy}_{min}[\vec{\phi}(x), \bar{p}]$, the *information in the measurement* of $\vec{\phi}(x)$ on

$\bar{p}(x,y)$ as:

$$I^{xy}_{min}[\vec{\phi}(x),\bar{p}] \equiv \min_{\hat{p}(x,y)\in\mathcal{F}_{xy}(\vec{\phi}(x)),\bar{p})} I[\hat{p}(x,y)] \ . \tag{4.1}$$

The set $\mathcal{F}_{xy}(\vec{\phi}(x),\bar{p})$ is the set of distributions that satisfy the constraints, defined by

$$\mathcal{F}_{xy}(\vec{\phi}(x),\bar{p}) \equiv \left\{\hat{p}(x,y) : \begin{array}{c} \langle\vec{\phi}(x)\rangle_{\hat{p}(x|y)} = \langle\vec{\phi}(x)\rangle_{\bar{p}(x|y)} \\ \hat{p}(x) = \bar{p}(x) \\ \hat{p}(y) = \bar{p}(y) \end{array} \right\} .$$

The desired *most informative features* $\vec{\phi}^*(x)$ are precisely those whose measurements provide the *maximal* information about Y. Namely,

$$\vec{\phi}^*(x) = \arg\max_{\vec{\phi}(x)} I^{xy}_{min}[\vec{\phi}(x),\bar{p}] \ .$$

Plugging in the definition of I^{xy}_{min} we obtain the following Max-Min problem:

$$\vec{\phi}^*(x) = \arg\max_{\vec{\phi}(x)} \min_{\hat{p}(x,y)\in\mathcal{F}_{xy}(\vec{\phi}(x),\bar{p})} I[\hat{p}(x,y)] \ . \tag{4.2}$$

Notice that this variational principle *does not* define a generative statistical model and is in fact a model independent approach. As we show later, however, the resulting distribution $\hat{p}(x,y)$, is necessarily of a special exponential form and can be interpreted as a generative model in that class. There is no need, however, to make any assumption about the validity of such a model for the empirical data. The data distribution $\bar{p}(x,y)$ is in fact needed only in order to estimate the expectations $\langle\vec{\phi}(x)\rangle_{\bar{p}(x|y)}$ for every y (besides the marginals), given the candidate features $\vec{\phi}(x)$. In practice these expectations are estimated from finite samples and empirical distributions. From a machine learning view our method thus requires only "statistical queries" on the underlying joint distribution [78].

In the next section we show that the problem of finding the optimal functions $\vec{\phi}(x)$ is *dual* to the problem of extracting the optimal features for Y that capture information on the variable X, and in fact the two problems are solved simultaneously.

4.2 The Nature of the Solution

We first show that the problem as formulated in Equation 4.2 is equivalent to the problem of minimizing the KL divergence between the empirical distribution $\bar{p}(x,y)$ and a special family of distributions of an exponential form. To simplify notation, we sometimes omit the suffix of (x,y) from the distributions. Thus p_t stands for $p_t(x,y)$ and $\bar{p}$ for $\bar{p}(x,y)$

Minimizing the mutual information in Equation 4.1 under the linear constraints on the expectations $\langle\vec{\phi}(x)\rangle_{\bar{p}(x|y)}$ is equivalent to maximizing the joint entropy:

$$H[\hat{p}(x,y)] = -\sum_{x,y} \hat{p}(x,y)\log\hat{p}(x,y) \ ,$$

under these constraints, with the additional requirement that the marginals are not changed, $\hat{p}(x) = \bar{p}(x)$ and $\hat{p}(y) = \bar{p}(y)$. Due to the concavity of the entropy and the convexity of the linear constraints, there exists a unique maximum entropy distribution (for compact domains of x and y) which has the exponential form[2] [3][27]:

$$\hat{p}_\phi(x,y) = \frac{1}{Z} \exp\left(\vec{\phi}(x) \cdot \vec{\psi}(y) + A(x) + B(y)\right) \ , \tag{4.3}$$

where Z, the normalization (partition) function is given by:

$$Z \equiv \sum_{x,y} \exp\left(\vec{\phi}(x) \cdot \vec{\psi}(y) + A(x) + B(y)\right) \quad ,$$

and the functions $\vec{\psi}(y), A(x), B(y)$ are uniquely determined as Lagrange multipliers from the expectation values $\langle\vec{\phi}(x)\rangle_{\bar{p}(x|y)}$ and the marginal constraints. It is important to notice that while the distribution in Equation 4.3 which maximizes the entropy is unique, there is freedom in the choice of the functions in the exponent. This freedom is however restricted to linear transformations of the vector-functions $\vec{\psi}(y)$ and $\vec{\phi}(x)$ respectively, as long as the original variables X and Y remain the same, as we show later on.

The distributions of the form in Equation 4.3 can also be viewed as a distribution class parameterized by the continuous family of functions $\Theta = [\vec{\psi}(y), \vec{\phi}(x), A(x), B(y)]$ (note that we treat ψ and ϕ symmetrically). We henceforth denote this class by P_Θ.

The discussion above shows that for every candidate feature $\vec{\phi}(x)$, the minimum information in Equation 4.2 is the information in the distribution $\hat{p}_\phi$. As argued before, this is precisely the information in the measurement of $\vec{\phi}(x)$ about the variable Y. We now define the set of information minimizing distributions $\mathcal{P}_\Phi \subset P_\Theta$, as follows:

$$\mathcal{P}_\Phi \equiv \left\{\hat{p} \in P_\Theta : \exists\vec{\phi}(x) : \quad \hat{p} = \hat{p}_\phi\right\} .$$

It can be easily shown that $\hat{p}_\phi$ is the only distribution in P_Θ satisfying the constraints in $\mathcal{F}_{xy}(\vec{\phi}(x), \bar{p})$ (see [34]). Thus, $\mathcal{P}_\Phi$ can alternately be defined as:

$$\mathcal{P}_\Phi = \left\{\hat{p} \in P_\Theta : \exists\vec{\phi}(x) : \quad \hat{p} \in \mathcal{F}_{xy}(\vec{\phi}(x), \bar{p})\right\} . \tag{4.4}$$

Finding the optimal (most informative) features $\vec{\phi}^*(x)$ amounts now to finding the distribution in $\mathcal{P}_\Phi$ which maximizes the information:

$$\hat{p}^* = \arg\max_{\hat{p}\in\mathcal{P}_\Phi} I[\hat{p}] \ .$$

The optimal $\vec{\phi}^*(x)$ will then be the ϕ parameters of $\hat{p}^*$, as in Equation 4.2.

[2]Note that the unique distribution can actually be on the closure (boundary) of the set of such exponential forms. We do not address this point here in details.

[3]We explicitly state the dependence on $\vec{\phi}(x)$, since we will be varying it in the optimization.

For every $\hat{p} \in \mathcal{P}_\Phi$ one can easily show that

$$I[\hat{p}] = I[\bar{p}] - D_{KL}[\bar{p}|\hat{p}] \ . \tag{4.5}$$

Equation 4.5 has two important consequences. First, it shows that maximizing $I[\hat{p}]$ for $\hat{p} \in \mathcal{P}_\Phi$ is equivalent to minimizing $D_{KL}[\bar{p}|\hat{p}]$ for $\hat{p} \in \mathcal{P}_\Phi$:

$$\hat{p}^* = \arg\min_{\hat{p} \in \mathcal{P}_\Phi} D_{KL}[\bar{p}|\hat{p}] \ . \tag{4.6}$$

Second, Equation 4.5 shows that the information in $I[\hat{p}^*]$ cannot be larger than the information in the original data (the empirical distribution). This supports the intuition that the model $\hat{p}^*$ maintains only properties present in the original distribution that are captured by the selected features $\vec{\phi}^*(x)$.

A problem with Equation 4.6 is that it is a minimization over a subset of P_Θ, namely $\mathcal{P}_\Phi$. The following proposition shows that this is in fact equivalent to minimizing the same function over all of P_Θ. Namely, the closest distribution to the data in P_Θ satisfies the conditional expectation constraints, and is thus in $\mathcal{P}_\Phi$.

Proposition 1

$$\arg\min_{\hat{p} \in \mathcal{P}_\Phi} D_{KL}[\bar{p}|\hat{p}] = \arg\min_{\hat{p} \in P_\Theta} D_{KL}[\bar{p}|\hat{p}]$$

Proof: We need to show that the distribution which minimizes the right hand side is in $\mathcal{P}_\Phi$. Indeed, by taking the (generally functional) derivative of $D_{KL}[\bar{p}|\hat{p}]$ w.r.t. the parameters Θ in $\hat{p}$, one obtains the following conditions:

$$\begin{array}{lll} \forall y & \langle \vec{\phi}(x) \rangle_{\hat{p}(x|y)} = \langle \vec{\phi}(x) \rangle_{\bar{p}(x|y)} \\ \forall x & \langle \vec{\psi}(y) \rangle_{\hat{p}(y|x)} = \langle \vec{\psi}(y) \rangle_{\bar{p}(y|x)} \\ \forall x & \hat{p}(x) = \bar{p}(x) \\ \forall y & \hat{p}(y) = \bar{p}(y) \ . \end{array} \tag{4.7}$$

Clearly, this distribution satisfies the constraints in $\mathcal{F}_{xy}(\vec{\phi}(x), \bar{p})$ and from the definition in Equation 4.4 we conclude that it is in $\mathcal{P}_\Phi$. □

Our problem is thus equivalent to the minimization problem:

$$p^* = \arg\min_{\hat{p} \in P_\Theta} D_{KL}[\bar{p}|\hat{p}] \ . \tag{4.8}$$

Equation 4.8 is symmetric with respect to ϕ and ψ, thus removing the asymmetry between X and Y in the formulation of Equation 4.2.

Notice that this minimization problem can be viewed as a Maximum Likelihood fit to the given $\bar{p}(x, y)$ in the class P_Θ, known as an association model in the statistical literature (see [61]), though the context there is quite different. It is also interesting to write it as a matrix factorization problem (Matrix exponent here is element by element)

$$P = \frac{1}{Z} e^{\Phi\Psi} \ , \tag{4.9}$$

where P is the distribution $p(x, y)$, Φ is a $|X| \times (d+2)$ matrix whose $(d+1)^{th}$ column is ones, and Ψ is a $d+2 \times |Y|$ matrix whose $(d+2)^{th}$ row is ones (the vectors $A(x), B(y)$ are the $(d+2)^{th}$ column of Φ and $(d+1)^{th}$ row of Ψ). While the vectors $A(x), B(y)$ are clearly related to the marginal distributions $p(x)$ and $p(y)$, it is in general not possible to eliminate them and absorb them completely in the marginals. We therefore consider the more general form of P_Θ. In the maximum-likelihood formulation of the problem, however, nothing guarantees the quality of this fit, nor justifies the class P_Θ from first principles. We therefore prefer the information theoretic, model independent, interpretation of our approach. As will be shown, this interpretation provides us with powerful geometric structure and analogies as well.

4.3 Link to Statistical Sufficiency

A classic scenario in machine learning and statistics is that of identifying a source distribution from an IID sample generated by it. Given a sample $x^n = [x_1, \ldots, x_n]$ generated IID by an unknown distribution $p(x|y)$, a basic task in statistical estimation is to infer the value of y from the sample. It is commonly assumed that the source distribution belongs to some parametric family $p(x|y)$ where distributions' "indices", or "parameters" $y \in Y$ may take an infinite set of values [4].

A useful approach to this problem is to extract a small set of statistics or features, which are functions of the samples and use only their values for inferring the source parameters. Such statistics are said to be sufficient if they capture all the "relevant information" in the sample about the identity of $y \in Y$. As is well known, under certain regularity assumptions non-trivial sufficient statistics exist *if and only if* $p(x|y)$ belongs to an exponential family [110]. The structure of the exponential family is similar to that in Equation 4.3

$$p(x|y) = \frac{1}{Z_y} \exp\left(\vec{\phi}(x) \cdot \vec{\psi}(y) + A(x) + B(y)\right) \ , \tag{4.10}$$

where $\vec{\phi}(x)$ are the sufficient statistics. Since SDR approximates the empirical distribution via an exponential form, it can be understood as a method for finding the function $\vec{\phi}(x)$ which best approximates a sufficient statistic for the given data.

4.4 An Iterative Projection Algorithm

The previous section showed the equivalence of the Max-Min problem, Equation 4.2 to the KL divergence minimization problem in Equation 4.8. We now present an iterative algorithm which provably converges to a local minimum of Equation 4.8, and hence

[4] Our notation differs from that of the statistical literature, where y is commonly denoted by θ and is usually a continuous parameter.

solves the Max-Min problem as well. This minimization problem can be solved by a number of optimization tools, such as gradient descent or such iterative procedures as described by [60]. In what follows, we demonstrate some information geometric properties of this optimization procedure, thus constructing a general framework from which more general convergent algorithms can be generated.

The algorithm is described using the information geometric notion of *I-projections* described in Section 1.4 and used in the previous chapters. Recall that the *I-projection* of a distribution $q(x)$ on a set of distributions $\mathcal{F}$ is defined as the distribution in $\mathcal{F}$ which minimizes the KL-divergence $D_{KL}[p|q]$. We denote this distribution here by $IPR(q, \mathcal{F})$:

$$IPR(q, \mathcal{F}) \equiv \arg\min_{p \in \mathcal{F}} D_{KL}[p|q] \ .$$

We now focus on the case where the set $\mathcal{F}$ is determined by expectation values. Given a d dimensional feature function $\vec{\phi}(x)$ and a distribution $\bar{p}(x)$, we consider the set of distributions which agree with $\bar{p}(x)$ on the expectation values of $\vec{\phi}(x)$, and denote it by $\mathcal{F}_x(\vec{\phi}(x), \bar{p}(x))$. Namely,

$$\mathcal{F}_x(\vec{\phi}(x), \bar{p}(x)) \equiv \left\{ \hat{p}(x) : \langle \vec{\phi}(x) \rangle_{\hat{p}(x)} = \langle \vec{\phi}(x) \rangle_{\bar{p}(x)} \right\} \ ,$$

which is clearly convex due to the linearity of expectations. The *I-projection* in this case has the exponential form

$$IPR(q(x), \mathcal{F}_x(\vec{\phi}(x), \bar{p}(x))) = \frac{1}{Z^*} q(x) \exp\left(\vec{\lambda}^* \cdot \vec{\phi}(x)\right) \ ,$$

where $\vec{\lambda}^*$ is a vector of Lagrange multipliers corresponding to the expectation $\langle \vec{\phi}(x) \rangle_{\bar{p}(x)}$ constraints. In addition, for this special exponential form, the Pythagorean inequality is tight and becomes an equality [30]. This property is further linked to the notion of "geodesic lines" on the curved manifold of such distributions.

Before describing our algorithm, some additional notations are needed:

- $p_t(x, y)$ - the distribution after t iterations.
- $\vec{\psi}_t(y)$ - the $\vec{\psi}(y)$ functions for $p_t(x, y)$.
- $\vec{\phi}_t(x)$ - the $\vec{\phi}(x)$ functions for $p_t(x, y)$.
- Θ_t - the full parameter set for $p_t(x, y)$.

The iterative projection algorithm is outlined in Figure 4.1 and described graphically in Figure 4.2. *I-projections* of (exponential) distributions are applied iteratively: Once for fixed $\vec{\psi}(y)$ and their expectations, and then for fixed $\vec{\phi}(x)$ and their expectations. Interestingly, during the first projection, the functions $\vec{\phi}(x)$ are modified as Lagrange multipliers for $\langle \vec{\psi}(y) \rangle_{\hat{p}(y|x)}$, and vice-versa in the second projection. The iterations

Input: Joint (empirical) distribution $\bar{p}(x,y)$.

Output: $2d$ feature functions: $\vec{\psi}(y)$ and $\vec{\phi}(x)$, that result from $\hat{p}^*$, a solution of the variational problem Equation 4.2 and a (local) minimum of Equation 4.8 and the matrix factorization of Equation 4.9.

Initialization:

- Initialize $p_0(x,y) \in P_\Theta$ randomly, by choosing random $\vec{\phi}(x)$,$\vec{\psi}(y)$,$A(x)$ and $B(y)$.

Iterate:

- For all x, set:

$$\begin{aligned} p_{t+1}(y|x) &= IPR(p_t(y|x), \mathcal{F}_x(\vec{\psi}_t(y), \bar{p}(y|x))) \\ p_{t+1}(x,y) &= p_{t+1}(y|x)\bar{p}(x). \end{aligned}$$

- For all y, set:

$$\begin{aligned} p_{t+2}(x|y) &= IPR(p_{t+1}(x|y), \mathcal{F}_x(\vec{\phi}_{t+1}(x), \bar{p}(x|y))) \\ p_{t+2}(x,y) &= p_{t+2}(x|y)\bar{p}(y). \end{aligned}$$

- Halt on convergence (when small enough change in $p_t(x,y)$ is achieved).

Figure 4.1: The iterative projection algorithm.

can thus be viewed as alternating mappings between the two sets of d-dimensional functions, $\vec{\psi}(y)$ and $\vec{\phi}(x)$. This is also the direct goal of the variational problem.

We proceed to prove the convergence of the algorithm. We first show that every step reduces $D_{KL}[\bar{p}(x,y)|p_t(x,y)]$.

Proposition 2 $D_{KL}[\bar{p}|p_{t+1}] \leq D_{KL}[\bar{p}|p_t]$.

Proof: For every x, the following holds:

1. $p_{t+1}(y|x)$ is the *I-projection* of $p_t(y|x)$ on the set $\mathcal{F}_x(\vec{\psi}_t(y), \bar{p}(y|x))$.
2. $\bar{p}(y|x)$ is also in $\mathcal{F}_x(\vec{\psi}_t(y), \bar{p}(y|x))$.

Using the Pythagorean property, (which is an equality here) we have:

$$D_{KL}[\bar{p}(y|x)|p_t(y|x)] = D_{KL}[\bar{p}(y|x)|p_{t+1}(y|x)] + D_{KL}[p_{t+1}(y|x)|p_t(y|x)] \ .$$

Multiplying by $\bar{p}(x)$ and summing over all x values, we obtain:

$$D_{KL}[\bar{p}|p_t] - D_{KL}[\bar{p}(x)|p_t(x)] = D_{KL}[\bar{p}|p_{t+1}] + D_{KL}[p_{t+1}|p_t] - D_{KL}[\bar{p}(x)|p_t(x)] \ .$$

where we used $p_{t+1}(x) = \bar{p}(x)$. Elimination of $D_{KL}[\bar{p}(x)|p_t(x)]$ from both sides gives:

$$D_{KL}[\bar{p}|p_t] = D_{KL}[\bar{p}|p_{t+1}] + D_{KL}[p_{t+1}|p_t] \ . \tag{4.11}$$

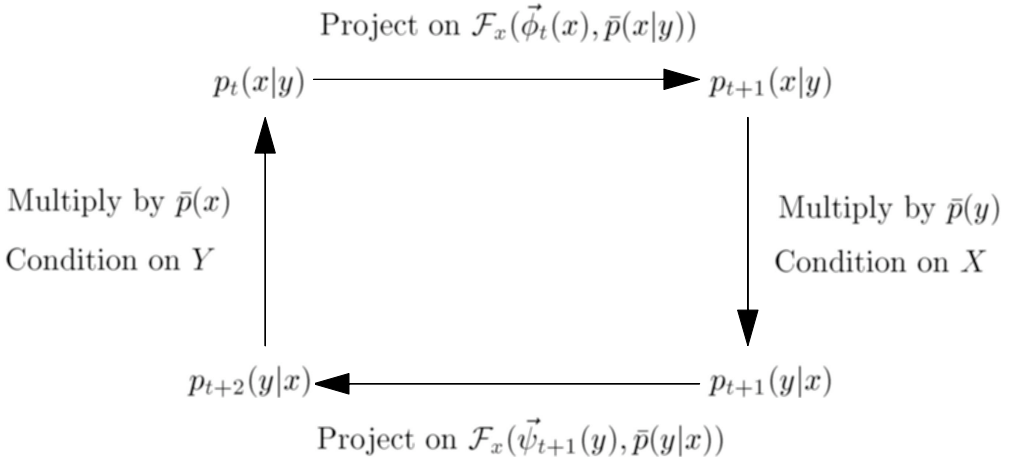


Figure 4.2: The iterative projection algorithm: The dual roles of the projections that determine $\vec{\psi}(y)$ and $\vec{\phi}(x)$ as dual Lagrange multipliers in each iteration. These iterations always converge and the diagram becomes commutative.

Using the non-negativity of $D_{KL}[p_{t+1}|p_t]$ the desired inequality is obtained:

$$D_{KL}[\bar{p}|p_t] \geq D_{KL}[\bar{p}|p_{t+1}] \ .$$

Note that equality is obtained if and only if $D_{KL}[p_{t+1}|p_t] = 0$. □

An analogous argument proves that:

$$D_{KL}[\bar{p}|p_{t+2}] \leq D_{KL}[\bar{p}|p_{t+1}] \ . \tag{4.12}$$

The following easily provable proposition states that the stationary points of the algorithm coincide with extremum points of the target function, $D_{KL}[\bar{p}|p_\Theta]$. Its proof uses the properties of the *I-projections* in the algorithm and the characterization of the extremum point in Equation 4.7.

Proposition 3 *If $p_t = p_{t+2}$ then the corresponding Θ_t satisfies $\frac{\partial D_{KL}[\bar{p}|p_\Theta]}{\partial \Theta}\Big|_{\Theta=\Theta_t} = 0$.*

In order to see that the algorithm indeed converges to a (generally local) minimum of $D_{KL}[\bar{p}|p_\Theta]$, note that $D_{KL}[\bar{p}|p_{2t}]$ is a monotonously decreasing bounded series, which therefore converges. Its difference series (see Equation 4.11,4.12),

$$D_{KL}[\bar{p}|p_t] - D_{KL}[\bar{p}|p_{t+2}] = D_{KL}[p_{t+1}|p_t] + D_{KL}[p_{t+2}|p_{t+1}] \ ,$$

therefore converges to zero. Taking t to infinity, we get $p_{t+2} = p_{t+1} = p_t$. Thus, the limit is a stationary point of the iterations, and from proposition 3 it follows that it is indeed a local minimum of $D_{KL}[\bar{p}|p_\Theta]$. [5]

[5] In order to take this limit, we use the fact that p_{t+1} and p_{t+2} are continuous functions of p_t because the *I-projection* is continuous in its parameters.

4.4.1 Implementation Issues

The description of the iterative algorithm assumes the existence of a module which calculates *I-projections* on linear (expectation) constraints. Because no general closed form solution for such a projection is available, it is found by successive iterations which asymptotically converge to the solution.

In this work, IPR was implemented using the *Generalized Iterative Scaling* (GIS) procedure [32], described in Figure 1.2.

We now briefly address the computational complexity of our algorithm. The overall time complexity naturally depends on the number of iterations needed till convergence. In the experiments described in this work we ran the algorithm for up to 100 iterations, which produced satisfactory results. The *I-projection* steps, which are performed using GIS, are iterative as well. We used 100-200 GIS iterations for each *I-projection* with a stopping condition when the ratio between empirical and model expectations was close enough to 1. Each step of the GIS has linear complexity in $|X|d$, where $|X|$ is the size of the X variable and d the number of features. Since in each SDR iteration we perform *I-projections* for all X's and Y's, each iteration performs $O(|X||Y||d|)$ operations.

The run-time of the algorithm is most influenced by the implementation of the *I-projection* algorithm. GIS is known to be slow to converge, but was used in this work for the simplicity of the presentation and implementation. Other methods that can speed this calculation by a factor of 20 have been suggested in the literature [88] and should be used to handle large datasets and many features. This is expected to make SDR comparable to SVD based algorithms in computational complexity.

The parameters were always initialized randomly, but different initial conditions did not affect the results noticeably. Initializing the parameters using the SVD of the log of $\bar{p}(x, y)$ generated a better initial distribution but did not improve the final results, nor convergence time.

4.5 Information Geometric Interpretation

The iterative algorithm and the exponential form provide us with an elegant information geometric insight and interpretation of our method. The values of the variable X are mapped into the d-dimensional differential manifold described by $\vec{\phi}(x)$, while values of the variable Y are mapped into a d-dimensional manifold described by $\vec{\psi}(y)$. Empirical samples of these variables are mapped into the same d-dimensional manifolds through the empirical expectations $\langle\vec{\phi}(x)\rangle_{\bar{p}(x|y)}$ and $\langle\vec{\psi}(y)\rangle_{\bar{p}(y|x)}$ respectively. These geometric embeddings, generated by the feature vectors Φ and Ψ, are in fact curved Riemannian differential manifolds with conjugate local metrics (see [2]).

The differential geometric structure of these manifolds is revealed through the normalization (partition) function of the exponential form, $Z(\phi; \psi)$. We note the following

relations:

$$\frac{\delta \log Z}{\delta \phi} = \langle \vec{\psi}(y) \rangle_{\bar{p}(y|x)} \tag{4.13}$$

$$\frac{\delta \log Z}{\delta \psi} = \langle \vec{\phi}(x) \rangle_{\bar{p}(x|y)} \ , \tag{4.14}$$

and the second derivatives,

$$\frac{\delta^2 \log Z}{\delta \phi_i \delta \phi_j} = \langle (\psi_i(y) - \langle \psi_i(y) \rangle)(\psi_j(y) - \langle \psi_j(y) \rangle) \rangle_{\bar{p}(y|x)} \tag{4.15}$$

$$\frac{\delta^2 \log Z}{\delta \psi_i \delta \psi_j} = \langle (\phi_i(x) - \langle \phi_i(x) \rangle)(\phi_j(x) - \langle \phi_j(x) \rangle) \rangle_{\bar{p}(x|y)} \ . \tag{4.16}$$

The last two matrices, which are positive definite, are also known as the Fisher information matrices for the two sets of parameters Φ and Ψ. Using the Information Geometry formalism of Amari, one can define the *natural* coordinates of those manifolds, as well as the *geodesic projections*, which are equivalent to the previously defined *I-projections*. The natural coordinates of the manifold are those that diagonalize the Fisher matrices locally, i.e. are locally uncorrelated. Moreover, the intrinsic geometric properties of these manifolds, such as their local curvature and geodesic distances are invariant with respect to transformations (including nonlinear) of the coordinates ϕ and ψ. Since our iterative algorithm can be formulated through covariant projections, its fixed (convergence) point is also invariant to local coordinate transformations, as long as the above coupling between the two manifolds is preserved.

This formulation suggests that the SDR reduced statistical description in terms of Φ and Ψ can be characterized in a way that is invariant to any $(1-1)$ transformation of X and Y. In particular it should be invariant to permutations of the rows and columns of the original co-occurrence matrix. This fact is illustrated in the next section. The information geometric formulation of the algorithm and its application to the study of geometric invariants requires further analysis.

4.5.1 Cramer-Rao Bounds and Uncertainty Relations

The special exponential form provides us with an interesting uncertainty relation between the conjugate manifolds Ψ and Φ and a way to deal with finite sample effects.

For a general parametric family, $p(x|\theta)$, with θ the parameter vector, the Fisher information matrix,

$$J_{i,j}(\theta) = \langle \left(\frac{\partial \log p(x|\theta)}{\partial \theta_i} \frac{\partial \log p(x|\theta)}{\partial \theta_j} \right) \rangle_x = \langle - \frac{\partial^2 \log p(x|\theta)}{\partial \theta_i \partial \theta_j} \rangle_x \ ,$$

provides bounds on the covariance matrix of *any* estimator of the parameter vector from a sample. Denoting such an estimator from an IID sample of size n, $x^{(n)}$, by $\hat{\theta}(x^n)$, the covariance matrix of the estimators, $Cov(\hat{\theta}_i, \hat{\theta}_j)$ is symmetric definite (if the

parameters are linearly independent) and can be diagonalized together with the Fisher information matrix. In particular, the diagonal elements of the two matrices satisfy the Cramer-Rao bound,

$$J_{i,i}(\theta)Var(\hat{\theta}_i(x^n)) \geq \frac{1}{n} \ .$$

For our special exponential form this inequality yields a particularly nice uncertainty relation between finite sample estimates of the feature vectors $\vec{\phi}(x)$ and $\vec{\psi}(y)$, since for the exponential form the Fisher information matrices are dual as well,

$$\begin{aligned} J_{i,j}(\psi) &= Cov(\hat{\phi}(x^n)) \\ J_{i,j}(\phi) &= Cov(\hat{\psi}(x^n)) \ , \end{aligned}$$

and the Cramer-Rao inequalities, for the diagonal terms, become

$$Var(\psi)Var(\hat{\phi}(x^n)) \geq \frac{1}{n} \tag{4.17}$$

$$Var(\phi)Var(\hat{\psi}(x^n)) \geq \frac{1}{n} \ , \tag{4.18}$$

as the Fisher information of the ϕ feature is just the variance of its adjoint ψ variable, and vice versa. In fact we know that this bound is tight precisely for exponential families, and Equations 4.17,4.18 are equalities.

These intriguing "uncertainty relations" between the conjugate features are strictly true only for the exponential form $\hat{p}^*$ and hold only approximately for the true variances. Yet they provide a way to analyze finite sample fluctuations in the estimates of the features.

The information-geometric structure of the problem allows us to interpret our algorithm as alternating (geodesic) projections between the two Riemannian manifolds of Ψ and Φ. These manifold allow an invariant formulation of the feature extraction problem through geometric quantities that do not depend on the choice of local coordinates, namely the specific choice of the functions $\phi(x)$ and $\psi(y)$. Among these invariants, the metric tensors of the manifolds provide us with the way to define and measure the geodesic projections. On the other hand, since these tensors are just the Fisher information matrices for the exponential forms, they provide us with *bounds on the finite sample fluctuations* of our feature functions.

4.6 Applications

The derivation of the SDR features suggests that they should be efficient in identifying a source distribution Y given a sample $X_1, \ldots, X_n$ by using just the empirical expectations $\frac{1}{n}\sum_{i=1}^{n}\vec{\phi}(x_i)$. We next show how the SDR algorithm can extract non-trivial structure from both artificial data, and real-life problems.

4.6.1 Illustrative Problems

In this section, some simple scenarios are presented where regression, either linear or non-linear, does not extract the information between the variables. SDR is then shown to find the appropriate regressors and uncover underlying structures in the data.

The construction of the SDR features is based on the assumption that only the knowledge of the function $\vec{\phi}(x)$ and its expected values $\langle\vec{\phi}(x)\rangle_{\bar{p}(x|y)}$ is required for estimating Y from X. Thus, assuming the SDR approximation is valid, we can replace $\bar{p}(x,y)$ with the above two functions. However, it is clear that one can use the Lagrange multipliers $\vec{\psi}(y)$ instead of the expected values, since there is a one to one correspondence between these two functions of Y. Finally, because the problem is symmetric w.r.t X and Y, we are also interested in the regressor averages $\langle\vec{\psi}(y)\rangle_{\bar{p}(y|x)}$, which together with $\vec{\psi}(y)$ should provide optimal information about X. Figure 4.3 depicts these plots for running SDR with $d = 1$ for the distribution:

$$p_1(y|x) \sim \mathcal{N}(0, 0.2 + 0.4|\sin 2x|) \ ,$$

where $\mathcal{N}(\mu,\sigma)$ is a normal distribution with mean μ and standard deviation σ. This distribution is of an exponential form and has a single sufficient statistic: y^2. The SDR single feature $\psi(y)$ turns out to be nearly identical to y^2, up to scaling and translation. The averages $\langle\psi(y)\rangle_{\bar{p}(y|x)}$ and the corresponding $\phi(x)$ indeed reveal the periodic structure in the data, and are thus appropriate regressors for this problem.

In Figure 4.3, the numerical value of x, y was used for plotting the SDR features. However, variables often cannot be assigned meaningful numerical values (e.g. terms in documents), and this approach cannot be used. One can still extract a description invariant representation in these cases by plotting the points $\vec{\phi}(x)$,$\vec{\psi}(y)$ in $\Re^d$. This allows the analysis of the functional relation between the two statistics, without assuming any order on the x or y domains. To illustrate this, consider a scrambled version of the distribution:

$$p_2(y|x) = \mathcal{N}(2\pi \sin x, 0.8 + 0.1x) \ ,$$

where both variables have undergone a random permutation. The scrambled distribution is shown in Figure 4.4, along with scatter plots of $\vec{\psi}(y)$ and $\vec{\phi}(x)$. These plots illustrate the structure of the differential manifolds described in the previous section. Since both scatter plots are clearly one dimensional curves, the *correct* order of x and y was recovered by traversing the curves, and the resulting "unscrambled distribution" is shown in Figure 4.4. It clearly demonstrates that the original continuous ordinal scale has been recovered.

Similar procedures can be used for recovering continuous structure from distributions with more than two statistics. Figure 4.5 shows the distribution:

$$p_3(y|x) = \frac{1}{Z_x} e^{-\frac{(y-\sin 2\pi x)^2}{2(0.8+0.1x)^2} - \frac{(y-\cos 2\pi x)^4}{(0.8+0.1x)^4}} \ .$$

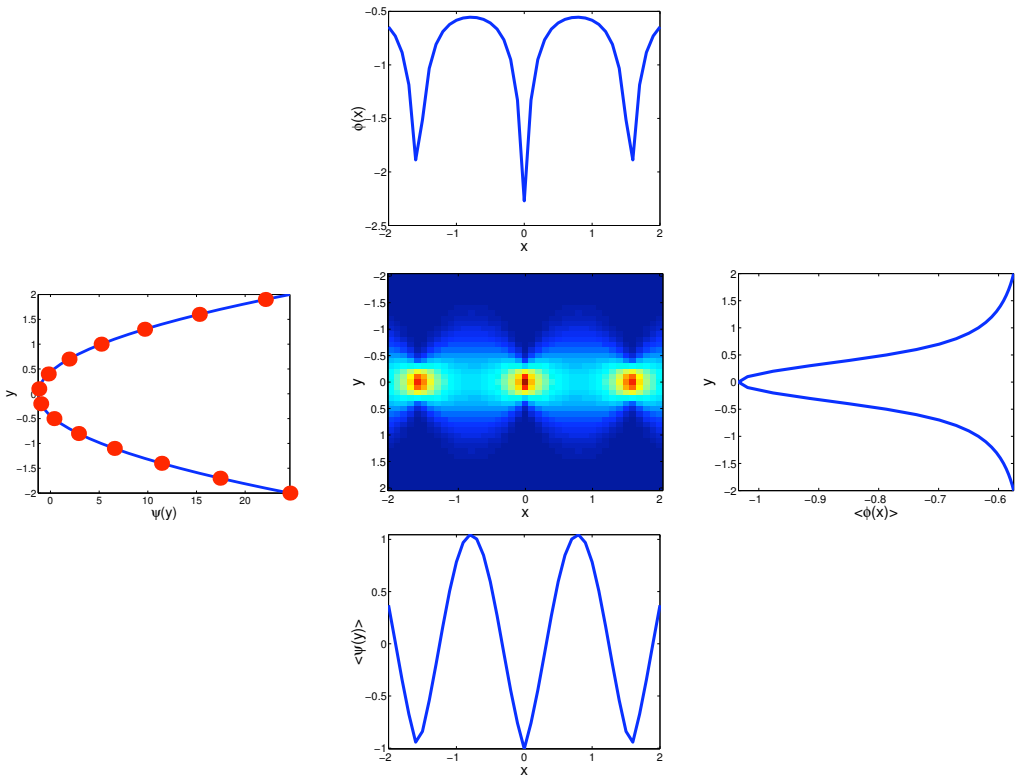


Figure 4.3: SDR feature extracted for the distribution $p_1(y|x) \sim \mathcal{N}(0, 0.2+0.4|\sin 2x|)$. **Middle**: The distribution p_1. **Left**: The SDR feature for y: $\vec{\psi}(y)$ in blue, and a scaled and translated y^2 in red dots. **Bottom**: Expected value $\langle\vec{\psi}(y)\rangle_{\bar{p}(y|x)}$ as a function of x. **Top**: The SDR feature for x: $\vec{\phi}(x)$. **Right**: Expected value $\langle\vec{\phi}(x)\rangle_{\bar{p}(x|y)}$ as a function of y.

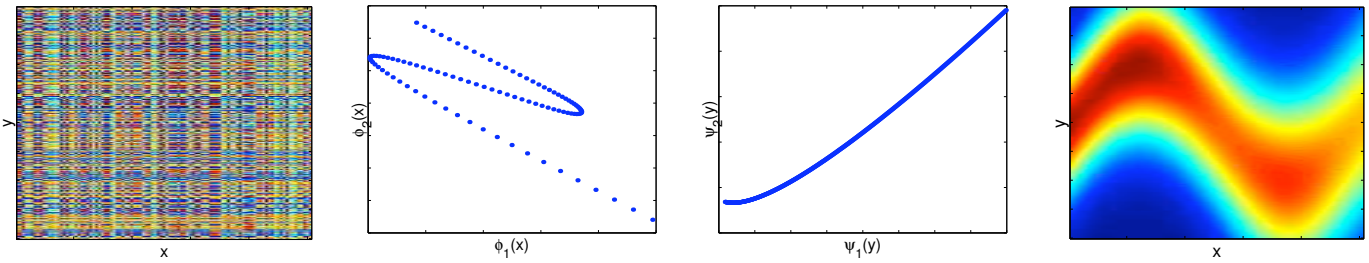


Figure 4.4: Reordering X and Y using SDR. From left to right: **1**: The scrambled version of $p_2(y|x) = \mathcal{N}(2\pi \sin x, 0.8+0.1x)$ **2**: Scatter plot of $\phi_2(x)$ vs. $\phi_1(x)$ - the $\vec{\phi}(x)$ curved manifold. **3**: Scatter plot of $\psi_2(y)$ vs. $\psi_1(y)$ - the $\vec{\psi}(y)$ manifold. **4**: Reordering of p_2 according to the $\vec{\phi}(x)$ and $\vec{\psi}(y)$ curves.

This distribution has four sufficient statistics, namely y, y^2, y^3, y^4. However, SDR can still be used with $d = 2$ to represent both X and Y as two dimensional curves, as shown in Figure 4.5. Although two statistics do not capture *all* the information in the distribution, the two curves still reveal the underlying parameterization. Specifically, the original order in the X and Y domains can be reconstructed.

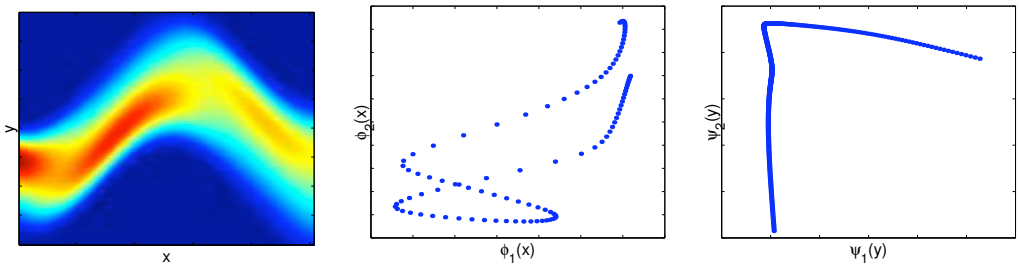


Figure 4.5: SDR analysis of $p_3(x, y)$. **Left** The distribution $p_3(x, y)$. **Middle**: Scatter plot for the manifold $\vec{\phi}(x)$. **Right**: Scatter plot for the manifold $\vec{\psi}(y)$.

4.6.2 Document Classification and Retrieval

The analysis of large databases of documents presents a challenge for machine learning and has been addressed by many different methods (for a recent review see [146]). Important applications in this domain include text categorization, information retrieval and document clustering.

Several works use a probabilistic and information theoretic framework for this analysis [68, 129], whereby documents and terms are considered stochastic variables. The probability of a term $w \in \{w_1, \ldots, w_{|W|}\}$ appearing in a document $doc \in \{doc_1, \ldots, doc_{|doc|}\}$ is denoted $p(w|doc)$ and is obtained from the normalized term count vector for this document:

$$p(w|doc) = \frac{n(doc, w)}{\sum_w n(doc, w)} \ ,$$

where $n(doc, w)$ is the number of occurrences of term w in document doc. The joint distribution $p(w, doc)$ is then calculated by multiplying by a document prior $p(doc)$ (e.g. uniform or proportional to document size). This stochastic relationship is then analyzed using an assumed underlying generative model, e.g. a linear mixture model as in [68], or a maximum entropy model as in [98]. An optimal model is found, and used for predicting properties of unseen documents.

In the current work, SDR is used for finding term features $\vec{\phi}(w)$ such that using their mean values alone we can infer information about document identity. This approach is nonlinear and thus significantly differs from linear based approaches such as LSI [33] or PLSI [68].

Our approach extends the works on maximum entropy in NLP [12, 34]. In these works, the set of features $\vec{\phi}(w)$ was predetermined, or was algorithmically chosen from

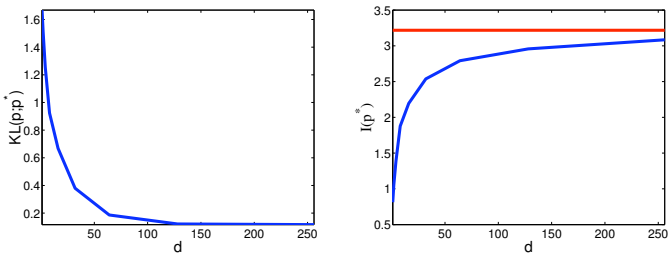


Figure 4.6: Left: KL between the original distribution and the model, for different d values. Right: Information in the model. The horizontal line marks the information in the original distribution.

a large set of predefined features. SDR does not make assumptions about the feature space (e.g. positivity, range etc.) and performs a completely unsupervised search for the optimal $\vec{\phi}(w)$.

Document Indexing Using SDR

The output of the SDR algorithm can be used to represent every document by an index in $\Re^d$. This transformation reduces the representation of a document from $|W|$, the number of terms in the corpus, to d dimensions. Such a dimensionality reduction serves two purposes: First, to extract relevant information and eliminate noise. Second, low dimensional vectors can allow faster, more efficient information retrieval, an important feature in real world applications. The resulting indices can be used for document retrieval or categorization.

A natural index is the expected value of $\vec{\phi}(w)$ for the given document: $\hat{\phi}(doc) \equiv \langle\vec{\phi}(w)\rangle_{p(w|doc)}$. Since the SDR formulation assumes only knowledge of expected values can be used, it is sensible to represent a document using this set of values. Alternatively, one can use the set of Lagrange multipliers $\vec{\psi}(doc)$ for each document in the training data. Given a new document, not in the training data, one can find $\vec{\psi}(doc)$ using a single *I-projection* of $p(w|doc)$ on the linear constraints imposed by $\vec{\phi}(w)$.

Document Classification Application

We used the 20Newsgroups database [81] to test how SDR can generate small features sets for document classification.

We start with an illustrative example, which shows how SDR features capture information about document content. We chose two different newsgroups with subjects: "alt.atheism" and "comp.graphics", and preselected 500 terms and 500 documents per subject, using the information gain criterion [146]. The projection algorithm was then run on the resulting $p(w, doc)$ matrix, for values of $d = 1, 2, 4, 8, \ldots, 256$. Figure 4.6 shows $D_{KL}[p|p^*]$, $I[p^*]$ as a function of d. It can be seen that as larger values of d

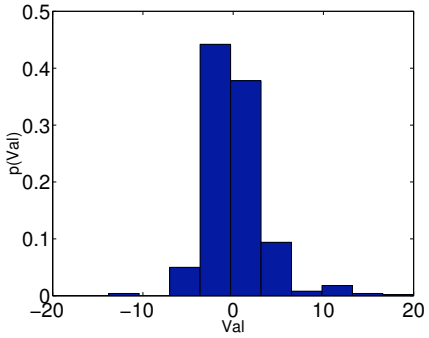


Figure 4.7: Histogram of $\vec{\phi}(w)$ values.

fatwa	correction
rushdie	gamma
islam	jfif
muslims	fullcolor
islamic	lossless

Figure 4.8: Terms with high (right) and low (left) $\phi_1(w)$ values

are used, the original distribution is approached in the KL sense and in the mutual information sense (as in Equation 4.5).

To gain insight into the nature of the SDR features, we look at $\vec{\phi}(w)$ obtained for the $d = 1$ case. A histogram of the values of $\vec{\phi}(w)$ is shown in Figure 4.7. It can be seen that the values are roughly symmetrical about 0. Figure 4.8 shows the 5 terms with the highest and lowest $\vec{\phi}(w)$ values. Clearly, the terms with high $\vec{\phi}(w)$ correspond to the "comp.graphics" subject, and the ones with low $\vec{\phi}(w)$ correspond to "alt.atheism". This single feature thus maps the terms into a continuous scale through which it assigns positive weights to one class, negative weights to the other, and negligible weights to terms which are possibly irrelevant to the classes.

We next performed classification on eight different pairs of newsgroups. The $\hat{\phi}(doc)$ index was used as input to a support vector machine SVM-Light [75], which was trained to classify documents according to their subjects. Baseline results were obtained by running the SVM classifier on the original $p(w|doc)$ vectors. The training and testing sets consisted of 500 documents per subject, and 500 terms. We experimented with values $d = 1, 2, 4, 8, \ldots, 256, 500$ to test how well we can classify with relatively few features. Figure 4.9 shows the fraction of the baseline performance that can be achieved using a given number of features. It can be seen that even when using only four features, 98% of the baseline performance can be achieved.

Thus, classification based on the SDR index achieves performance comparable to that of the baseline, even though it uses significantly less features (compared to the 500 features used by the baseline classifier). Importantly, the features were obtained

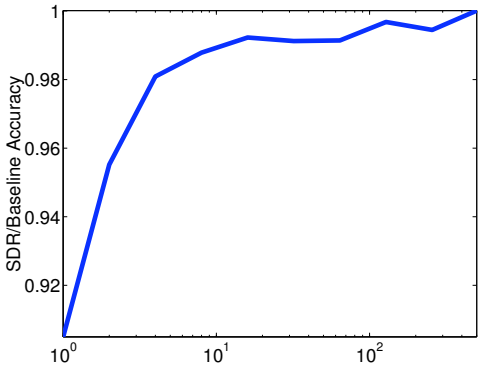


Figure 4.9: Fraction of baseline performance achieved by using a given number of SDR features. Results are averaged over eight newsgroup pairs.

in a wholly unsupervised manner, and have still discovered the document properties relevant for classification.

Information Retrieval Application

Automated information retrieval is commonly done by converting both documents and queries into a vector representation, and then measuring their similarity using the cosine of the angle between these vectors [33].

SDR offers a natural procedure for converting documents into d dimensional indices, namely $\hat{\phi}(doc)$. Retrieval can then be performed using the standard cosine similarity measure, although this is not necessarily the optimal procedure.

The following databases were used to test information retrieval: MED(1033 documents, 5381 terms), CRAN (1400 documents, 4612 terms) and CISI (1460 documents, 5609 terms).[6] For each database, precision was calculated as a function of recall at levels $0.1, 0.2, \ldots, 0.9$. The performance was then summarized as the mean precision over these 9 recall levels (see [68]).

We compared the performance of the SDR based indices with that of indices generated using the following algorithms:

- RAW-TF: Uses the original normalized term count vector $p(w|doc)$ as an index. The dimension of the index is the number of terms.

- Latent Semantic Indexing (LSI): LSI is, like SDR, a dimensionality reduction mechanism, which uses the Singular Value Decomposition (SVD) to obtain a low rank approximation of the term-frequency matrix: $p(doc, w) \approx USV$, where U is a $|doc| \times d$ matrix , S is $d \times d$ and V is $d \times |w|$. A new document vector $\vec{x}$ is then represented by the d dimensional vector $S^{-1}V\vec{x}$.

[6]The list of terms used can be obtained from *www.cs.utk.edu/∼lsi/corpora.html*

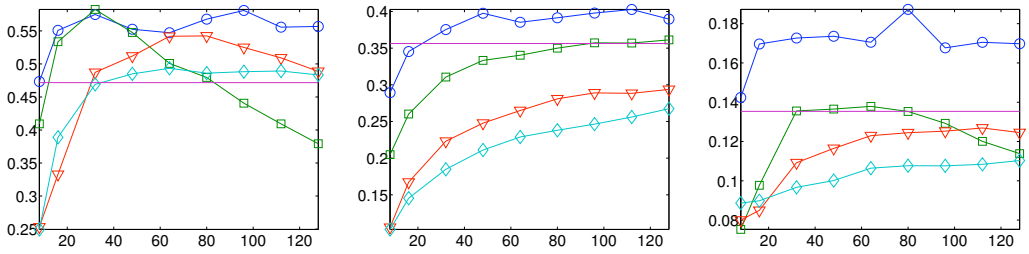


Figure 4.10: Mean precision as a function of index dimension for four indexing algorithms (SDR in circles, LSI in triangles, LogLSI in diamonds and LLE in squares), and three databases (from left to right: MED,CRAN,CISI). Horizontal line is the mean precision of the RAW-TF algorithm.

- Log LSI: Since SDR is related to performing a low rank approximation of the log of $p(w, doc)$, we also tried to perform LSI where the input matrix is $\log p(w, doc)$. In order to avoid taking a log of zero, we thresholded the matrix to 10^{-7}.

- Locally Linear Embedding (LLE) [118]: LLE performs a non linear approximation of the manifold on which the vectors $p(w|doc)$ lie. It maps these vectors into a d dimensional space, while preserving the neighborhood structure in the original, high-dimensional, space. LLE has been shown to perform well for image, as well as document data. For the experiments described here, we followed the procedure used by [118],[7] using the dot product between document vectors as the neighborhood metric, and taking 10 nearest neighbors (other values were experimented with, but 10 gave optimal performance). The version of LLE used here cannot index new vectors (although such an extension was given by [119]), so document and query data were indexed simultaneously. Note that this gives LLE an advantage over the other indexing algorithms, since the test data (i.e. the queries) is used in the training procedure.

For each of the databases, indexing dimensions of $d = 8, 16, 32, 48, \ldots, 128$ were used. Figure 4.10 shows the mean precision as a function of index dimension for the four indexing algorithms. Table 4.1 gives the peak performance on each of the databases for each of the algorithms (LogLSI is not included since it was inferior to LSI). It can be seen that SDR performs uniformly better than the other algorithms for most index sizes. Moreover, SDR achieves high performance for a small number of features, where other methods perform poorly. This suggests that SDR succeeds at capturing the low dimensional manifold on which the documents reside. The fact that LogLSI performs poorly shows that SDR is not equivalent to an SVD low rank approximation of $\log p(w, doc)$.

[7] Code available from http://www.cs.toronto.edu/~roweis/lle/code.html, with minor changes.

	Raw TF	LSI	LLE	SDR
MED	47.18 (5381)	54.25 (80)	58.25(32)	58.1 (96)
CRAN	35.63 (4612)	28.95 (128)	36.13(128)	40.28 (112)
CISI	13.54 (5609)	13.1 (128)	13.79(64)	18.72 (80)

Table 4.1: Mean precision results for information retrieval for several databases and different indexing algorithms. Numbers in brackets are the number of features for which the optimal performance is obtained

4.7 Discussion

The current work suggests an information theoretic feature extraction mechanism, where the features are used for calculating means over empirical samples, and these means are in turn used for inferring an unknown "variable" or other "relevant property" of the statistical data. The main goal of this procedure is to reveal possible *continuous* low-dimensional underlying structure that can "explain" the dependencies among the variables. Yet our applications suggest that the extracted continuous features can be used successfully for prediction of the relevant variable. An interesting question is in what sense is this resulting prediction close to optimal? In other words, how well can one infer using the extracted features compared the optimal inference procedure? The prediction error of an optimal procedure is the Bayes error [36]. When the true underlying distribution is of an exponential form with the correct dimensionality SDR is equivalent to maximum likelihood estimation of its parameters/features, and as such is consistent [36] and achieves asymptotically the optimal error in this case.

In the more interesting case, where the source distribution is not in the special exponential form, SDR should be interpreted as an induction principle, similar - but better motivated - than the maximum entropy principle. In fact it solves an *inverse* problem to that of MaxEnt, by finding the *optimal* set of constraints - or observables - that capture the mutual information in the empirical joint distribution. It is obvious that methods that exploit prior knowledge about the true distribution can do better in such cases. However, assuming that only empirical conditional expectations of given single variable functions are known about the source joint distribution (similar to "statistical queries" in machine learning, see [78]) our Max-Min mutual information principle is a most reasonable induction method.

We briefly address several other important issues that our procedure raises.

4.7.1 Information in Individual Features

The low rank decomposition calculated using SVD suggests a clear ordering on the features using their associated singular values. Moreover, the SVD solutions are nested (i.e. the optimal features for $d = 2$ are a subset of those for $d = 3$). Due to its non-linear nature, the SDR solution is not necessarily nested (this is also true for solutions of linear based methods like those in [68, 84]).

However, given a set of optimal SDR features $\vec{\phi}^*(x)$, the information in a single feature can be quantified in several ways. The simplest measure is $I^{xy}_{min}[\phi^*_i(x)]$ (as defined in Equation 4.1), which reflects the information obtained when measuring only $\phi^*_i(x)$. Another measure is $I^{xy}_{min}[\vec{\phi}^*(x)] - I^{xy}_{min}[\phi^*_1(x), \ldots, \phi^*_{i-1}(x), \phi^*_{i+1}(x), \phi^*_d(x)]$ which is the information lost as a result of not measuring $\phi^*_i(x)$. The two measures reflect different properties of the feature, and further research is required to test their usefulness, for example in assigning confidence to the measurement of different features.

4.7.2 Finite Samples

The basic assumption behind our problem formulation is that we have access to the true expectation values $\langle\vec{\psi}(y)\rangle$ and $\langle\vec{\phi}(x)\rangle$. These can be estimated *uniformly well* from a finite sample under the standard uniform convergence conditions. In other words, standard learning theoretical techniques can give us the sample complexity bounds, given the dimensions of X and Y and the reduced dimension d. For continuous X and Y further assumptions must be made, such as the fat-shattering dimension of the features.

When the source distribution is close to the exponential form - most of the mutual information is captured by the features - the Cramer-Rao bounds provide a much simpler method for analyzing the finite sample effects, as we discussed in section 5.1. We can then bound the prediction errors in terms of the empirical covariance matrices of the obtained features.

4.7.3 Diagonalization and Dimensionality Reduction

The optimal features are not unique in the following sense: since only the dot-product $\vec{\phi}(x) \cdot \vec{\psi}(y)$ appears in the distribution $\hat{p}(x, y)$, any invertible matrix R can be applied such that $\vec{\phi}(x)R^{-1}$ and $R\vec{\psi}(y)$ are equivalent to $\vec{\phi}(x)$ and $\vec{\psi}(y)$. Note, however, that although the resulting functions $\vec{\psi}(y)$ and $\vec{\phi}(x)$ may depend on the initial point of the iterations, the information extracted does not (for the same optimum).

One can remove this ambiguity by orthogonalization and scaling of the feature functions, for example by applying Singular Value Decomposition (SVD) to $\log \hat{p}(x, y)$ (for additional normalization schemes see [6]). Another option is to de-correlate $\vec{\phi}(x)$ using an appropriate linear transformation. By de-correlation, we mean a transformation on

$\vec{\phi}(x)$ such that the new functions satisfy

$$\sum_x p(x)\phi_i(x)\phi_j(x) = \delta_{ij} \ .$$

To calculate such a transformation, define $C \equiv \Phi * diag(p(x)) * \Phi^T$ (where Φ is a matrix whose rows are the original functions), calculate V, D the eigen-decomposition of C (i.e. $C = VDV^{-1}$), define the transformation $A \equiv D^{-0.5}V^{-1}$, and apply it to the original functions to obtain the transformed ones $\Phi_{Trans} \equiv A\Phi$.

Notice, however, that our procedure is very different from direct application of SVD to $\log \bar{p}(x, y)$. These two coincide *only* when the original joint distribution is already of the exponential form of Equation 4.3. In all other cases SVD based approximations (LSI included) will not preserve information as well as our features at the same dimension reduction.

4.7.4 Information Theoretic Interpretation

Our information MaxMin principle is close in its formal structure to the problem of channel capacity with some channel uncertainty (see e.g. [82]). This suggests the interesting interpretation for the features as channel characteristics. If the channel only enables the reliable transmission of d expected values, then our $\vec{\psi}(y)$ exploit this channel in an optimal way. The channel decoder of this case is provided by the vector $\vec{\phi}(x)$ and the decoding is performed through a dot-product of these two vectors. This intriguing interpretation of our algorithm obviously requires further analysis.

4.7.5 Relations to Other Methods

Dimension reduction and clustering algorithms have become a fundamental component in unsupervised large scale data analysis. SDR is a dimension reduction method in that it reduces the description of the distribution $p(x, y)$ from $|X||Y|$ components to $(d+1)(|X|+|Y|)$. There is a large family of algorithms with a similar purpose, which are based on a linear factorization of $p(x, y)$. For example [68] and [84] suggest finding positive matrices Q of size $|X| \times d$ and R of size $d \times |Y|$ such that $p = QR$. Their procedure relies on the fact that the rows (or columns) of p lie on a d dimensional plane. The SVD based LSI method [33] also performs a linear factorization of p, but it is not required to be positive.

Our approach is equivalent to approximating p (in the KL-divergence sense) by $\frac{1}{Z}e^{QR}$ where Q and R are matrices of size $|X| \times (d+2)$ and $(d+2) \times |Y|$ respectively. Since the exponent can be written as $e^{QR} = I + QR + \frac{(QR)^2}{2!} + \ldots$, (matrix powers are element by element) the above linear methods can be said to approximate its first two terms. However, it is important to note that the two methods (linear and exponential) assume different models of the data, and their success or failure is application dependent. The information retrieval experiments in the current work have shown that

even in document analysis, where linear methods have been successful, exponential factorization can improve performance.

An information theoretic approach to feature selection is also used by [148], in the context of texture modeling in vision. Their approach is similar to ours in that they arrive at a min-max entropy formulation. However, in contradistinction with the current work, they assume a specific underlying parametric model, and also define a finite feature set from which features are chosen using a greedy procedure. In this sense, their algorithm is more similar to the feature selection mechanism of [34]. There are also several works which search for approximate sufficient statistics (see [141, 50]) by directly calculating the information between the statistic and the parameter of interest. This results in a formalism different from ours, and usually necessitates some modeling assumptions to make computation feasible.

Since SDR finds a mapping from the X and Y variables into d dimensional space, it can be considered an embedding algorithm. As such, it is related to non-linear embedding algorithms, notably multi-dimensional scaling [29], Locally Linear Embedding [118] and IsoMAP [132]. Such algorithms try to preserve properties of points in high dimensional space (e.g. distance, neighborhood structure) in the embedded space. SDR is not formulated in such a way, but rather requires that the original points can be reconstructed optimally from the embedded points, where the quality of reconstruction is measured using the KL-divergence.

Relations to the Information Bottleneck

A closely related idea is the *Information Bottleneck (IB) Method* (originated in [135]) which aims at a clustering of the rows of $p(x, y)$ that preserves information. In a well defined sense, the IB method can be considered as a special case of SDR, when the features functions are restricted to a finite set of values that correspond to the clusters. However, clustering may not be the correct answer for many problems where the relationship between the variables comes from some hidden low dimensional continuous structures. In such cases clustering tends to quantize the data in a rather arbitrary way, while low dimensional features are simpler and easier for interpretation.

On the other hand, the motivation behind SDR is essentially the same as that of the IB, to find low-dimensional/complexity representations of one variable that preserve the mutual information about another variable. The algorithm in that case is however quite different and it solves in fact a more symmetric problem - find low dimensional representations of both variables (X and Y) such that the mutual information between them is captured as good as possible. This is closely related to the symmetric version of the information bottleneck that is discussed by [44]. As in the IB, the quality of the procedure can be described in terms of the "information curve" - the fraction of the mutual information that can be captured as a function of the reduced dimension.

An interesting open question, at this point, is if we can extend the SDR algorithm to deal with different dimensions simultaneously. One would like to move continuously through more and more complex representations, as done in the IB through a mutual information constraint on the complexity of the representation. There are good reasons to believe that such an extension, that may "soften" the notion of the dimension, is possible also with SDR.

4.8 A Euclidean Extension: Joint Embedding

The SDR algorithm results in two sets of *informative* features: $\vec{\phi}(x)$ and $\vec{\psi}(y)$. An interesting question in this context is what can be said about two values x, y such that $\vec{\phi}(x)$ and $\vec{\psi}(y)$ are close in the d dimensional space. Due to non-uniqueness of SDR solutions under affine transformations, distances between x and y features are apparently meaningless.

In order to obtain a meaningful distance measure between X and Y features we modified the SDR method to directly address Euclidean distances between feature functions [53]. We used a distribution model similar to that in Equation 4.3, but with Euclidean distance replacing the dot product [8]:

$$\hat{p}(y|x) = \frac{\bar{p}(y)}{Z(x)} e^{-d^2_{x,y}} \quad \forall x \in X, \forall y \in Y \ . \tag{4.19}$$

where $d^2_{x,y} \equiv |\vec{\phi}(x) - \vec{\psi}(y)|^2 = \sum_{k=1}^{d} (\phi_k(x) - \psi_k(y))^2$ is the Euclidean distance between $\vec{\phi}(x)$ and $\vec{\psi}(y)$ and $Z(x)$ is the partition function for each value of x.

The optimal features $\vec{\phi}(x), \vec{\psi}(y)$ are obtained by maximizing the likelihood of the empirical data with respect to the model above. Since the model explicitly uses the Euclidean distance between features, and is *not* invariant to affine transformation, we can expect x, y with high $\hat{p}(y|x)$ values to yield $\vec{\phi}(x), \vec{\psi}(y)$ values which are close in the Euclidean sense. We name the above method Co-Occurrence Data Embedding (CODE).

Figure 4.11 shows the application of the CODE algorithm to papers from the NIPS conference database [9]. We used the data to generate an authors-words matrix (as in the Roweis database). We could now embed authors and words into $\mathbb{R}^2$, by using CODE to model $p(word|author)$. The results are shown in Figure 4.11. It can be seen that authors are indeed mapped next to terms relevant to their work, and that authors dealing with similar domains are also mapped together. This illustrates how co-occurrence of words and authors may be used to induce a metric on authors alone.

[8] Here we use a conditional model, and also insert the empirical Y marginal explicitly. Various alternative models are possible, but this one was found to outperform the others).

[9] Data for NIPS 1-12 are from `http://www.cs.toronto.edu/~roweis/data.html` . These were augmented with data from NIPS volumes 13-17 Data available at `http://robotics.stanford.edu/~gal/` .

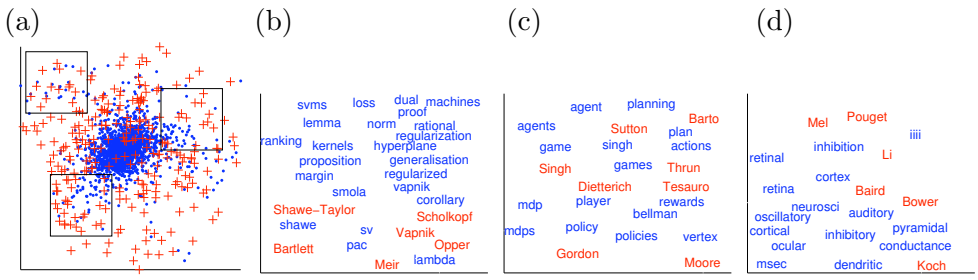


Figure 4.11: CODE Embedding of 2000 words and 250 authors from the NIPS database (the 250 authors with highest word counts were chosen; words selection criterion described in [53]). Left panel shows embeddings for authors (red crosses) and words (blue dots). Other panels show embedded authors (only first 100 shown) and words for the areas specified by rectangles. They can be seen to correspond to learning theory, control and neuroscience (from left to right).

The CODE algorithm was tested on several other text databases and provided features which were shown to outperform those obtained by other methods. We have also recently applied CODE to gene expression data, yielding interpretable results.

4.9 Conclusions and Further Research

We have presented a feature extraction method based on an information theoretic account of the notion of measurement. Our proposed method is a new dimensionality reduction technique. It is nonlinear, and aims directly at preserving mutual information in a given empirical co-occurrence matrix. This is achieved through an information variation principle that enables us to calculate simultaneously informative feature functions for *both* random variables. In addition we obtain an exponential model approximation to the given data which has precisely these features as *dual sets* of sufficient statistics. We described an alternating projection algorithm for finding these features and proved its convergence to a (local) optimum. This is in fact an algorithm for extracting optimal sets of constraints from statistical data.

Our experimental results show that our method performs well on language related tasks, and does better than other linear models, which have been used extensively in the past. Performance is enhanced both in using smaller feature sets, and in obtaining better accuracy. Initial experiments on image data also show promising results. Taken together, these results suggest that the underlying structure of non-negative data may often be captured better by SDR than using linear mixture models.

One immediate extension of our work is to the multivariate case. The natural measure of information in this case is the *multi-information* [44]. One can then generalize the bivariate SDR optimization principle, thus defining an optimal measurement on a large set of variables.

Chapter 5

Sufficient Dimensionality Reduction with Irrelevance Statistics

The previous chapter showed how the SDR method may be used to learn functions over a variable X which are useful for predicting a variable Y. However, it is often the case that Y has several different features, all of which may be relevant under different circumstances. Thus for instance, spoken utterances can be labeled by their contents or by the speaker's identity; face images can be categorized by either a person's identity or expression; protein molecules can be classified by their physical structure or biological function. All are valid alternatives for analyzing the data, but the question of the "correctness" or "relevance" depends on the task. The "noise" in one analysis is the "signal" for another.

The current chapter addresses this problem by utilizing additional *irrelevance* data as "side information". Such irrelevance information is very often available in terms of joint statistics of our variables in another context, but the irrelevant attributes are usually not explicit. A typical example is the analysis of gene expression data for some pathology, where the irrelevance information can be given in terms of the expression of control, healthy tissues. In this case it is essentially impossible to isolate the irrelevant variables, though they are implicitly expressed in the expression patterns and statistics. The goal of the new unsupervised learning algorithm is to identify structures which are characteristic to the relevant dataset, but do not describe well the irrelevance data. The idea of using such "side information" to enhance learning algorithms previously appeared in [25] and [145], which looked for relevant clusters in data.

The method presented in the current chapter, *SDR with Irrelevance Statistics* (SDR-IS), seeks features which are maximally informative about one, relevant, variable Y^+, while being minimally informative about another one Y^- provided as irrelevance information. Once the question is properly posed using information theoretic measures

the SDR formulation yields the solution to the new problem.

Features (i.e. statistics of empirical data) that carry no information about a parameter are known as ancillary statistics [42]. These are mainly used for estimating precision of standard estimators. SDR-IS extracts features that are approximately sufficient for the relevant variable Y^+, and at the same time approximately ancillary for Y^-. The quantitative nature of the approximation is determined by a trade-off between the information that the extracted features carry about Y^+ and the information they maintain about Y^-.

The next two sections formalize the problem of continuous feature extraction with irrelevance information using the previously introduced notion of "information in a measurement". We then derive its formal and algorithmic solutions, relate them to likelihood ratio maximization , and demonstrate their operation on synthetic and real world problems.

Most of the material in the current chapter was published in [54].

5.1 Problem Formulation

To formalize the above ideas, consider a scenario where two empirical joint distributions are given for three categorical random variables X, Y^+ and Y^-. The first is the main data, $\bar{p}^+ \equiv P(X, Y^+)$, which describes the joint distribution of Y^+ and X . The second is the irrelevance data, $\bar{p}^- \equiv P(X, Y^-)$, which is assumed to contain irrelevant structures in the main data. Our goal is to identify features of X that characterize its probabilistic relation to Y^+ but not its relation to Y^-. Note that Y^+ and Y^- need not come from the same space, or have the same size and dimension. Potentially, one may be continuous and the other discrete, although we do not treat the continuous case here.

We seek a d dimensional continuous feature of X which we denote $\vec{\phi}(x) : X \to \Re^d$, such that only its expected values $\langle \vec{\phi}(x) \rangle_{p(x|y^+)}$ characterize the stochastic dependence between X and Y^+, while the corresponding values for Y^-, namely $\langle \vec{\phi}(x) \rangle_{p(x|y^-)}$, do not characterize the dependence of Y^- on X. For example, the mean number of words in some semantic set may reveal a document's writing style, but tell us nothing of its content. Here, Y^+ would be a set of documents of different writing styles, and Y^- a set of documents with the same style but varying contents. X will represent the set of words.

The idea of using expected values of features to describe a distribution stands in the basis of the Maximum-Entropy (MaxEnt) approach [72, 34]. On one hand, these descriptions provide a natural way to efficiently estimate and represent distributions using parametric representations. Furthermore, the extracted parameters often provide compact description of the data in terms of interpretable features. While in standard MaxEnt the features are predetermined (or greedily optimized over a given set as in

[34]), the previous chapter presented the SDR method for finding features which are optimal for a given set of distributions over X, thus solving an *Inverse Maximum Entropy* problem. The continuous features $\vec{\phi}(x)$ can be any d dimensional function of a discrete variable X. To evaluate the "goodness" of $\vec{\phi}(x)$, we used the notion of *measurement information* $I^{xy}_{min}[\vec{\phi}(x), \bar{p}]$, defined in Section 4.1. Given this measure for the quality of $\vec{\phi}(x)$, the goal of relevant feature extraction is to identify features that are maximally informative about Y^+ while minimally informative about Y^-. This dual optimization task can be approached by minimizing the weighted difference

$$\mathcal{L}[\vec{\phi}(x)] = I^{xy}_{min}[\vec{\phi}(x), \bar{p}^+] - \lambda I^{xy}_{min}[\vec{\phi}(x), \bar{p}^-] \tag{5.1}$$

over $\vec{\phi}(x)$, where λ is a positive tradeoff parameter reflecting the weight to be assigned to the irrelevance data.

Using the definition of I^{xy}_{min}, the optimization problem in Equation 5.1 thus becomes:

$$\begin{aligned} \vec{\phi}^*(x) &= \arg\max \mathcal{L}[\vec{\phi}(x)] \\ &= \arg\max_{\vec{\phi}(x)} \min_{\hat{p}^+ \in \mathcal{P}(\vec{\phi}, \bar{p}^+)} I[\hat{p}^+] - \lambda \min_{\hat{p}^- \in \mathcal{P}(\vec{\phi}, \bar{p}^-)} I[\hat{p}^-] \end{aligned} \tag{5.2}$$

5.2 Solution Characterization

In order to characterize the solution of the variational problem in Equation 5.2, we now calculate its gradient and observe its vanishing points. We start by characterizing the form of the distribution $\hat{p}_\phi(X, Y)$ that achieves the minimum of $I^{xy}_{min}[\vec{\phi}(x), \bar{p}^\pm]$ (Equation 4.1). Since $I[\hat{p}(X,Y)] = H[\hat{p}(X)] + H[\hat{p}(Y)] - H[\hat{p}(X,Y)]$, and the marginals $\hat{p}(X)$, $\hat{p}(Y)$ are kept constant by the definition of $\mathcal{P}(\vec{\phi}(x), \bar{p})$, we have $I[\hat{p}(X,Y)] = const - H[\hat{p}(X,Y)]$. This turns Equation 4.1 into a problem of entropy maximization under linear constraints

$$\hat{p}_\phi(X,Y) = \max_{\hat{p}(X,Y) \in \mathcal{P}(\vec{\phi}(x)), \bar{p})} H[\hat{p}(X,Y)] \quad , \tag{5.3}$$

whose solutions are known to be of exponential form [34]

$$\hat{p}_\phi(x,y) = \frac{1}{Z} \exp\left(\vec{\phi}(x) \cdot \vec{\psi}_\phi(y) + A_\phi(x) + B_\phi(y)\right) \ . \tag{5.4}$$

The $\vec{\psi}_\phi(y)$,$A_\phi(x)$ and $B_\phi(y)$ are complex functions of $\vec{\phi}(x)$ that play the role of Lagrange multipliers in the maximum entropy problem derived from Equation 5.3. We explicitly note their dependence on $\vec{\phi}(x)$, to avoid confusion in describing the algorithm.

While $H[\hat{p}_\phi(X,Y)]$ is a complex function of $\vec{\phi}(x)$, its gradient can be derived analytically using the fact that $\hat{p}_\phi$ has the exponential form of Equation 5.4. In the appendix, section A.2.1, we show that this gradient is

$$\frac{\partial H[\hat{p}_\phi(X,Y)]}{\partial \vec{\phi}(x)} = \bar{p}(x) \left(\langle \vec{\psi}_\phi \rangle_{\hat{p}_\phi(y|x)} - \langle \vec{\psi}_\phi \rangle_{\bar{p}(y|x)} \right) \tag{5.5}$$

It is now straightforward to calculate the gradient of the functional in Equation 5.2. Denote by $\hat{p}_\phi^+$ and $\hat{p}_\phi^-$ the information minimizing distributions obtained in $I_{min}^{xy}[\phi, \bar{p}^+]$ and $I_{min}^{xy}[\phi, \bar{p}^-]$, and by $\vec{\psi}_\phi^+$ and $\vec{\psi}_\phi^-$ their corresponding Lagrange multipliers. The gradient is then

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \vec{\phi}(x)} &= \bar{p}^+(x) \left(\langle \vec{\psi}_\phi^+ \rangle_{\bar{p}^+(y^+|x)} - \langle \vec{\psi}_\phi^+ \rangle_{\hat{p}_\phi^+(y^+|x)} \right) \\ &\quad -\lambda \bar{p}^-(x) \left(\langle \vec{\psi}_\phi^- \rangle_{\bar{p}^-(y^-|x)} - \langle \vec{\psi}_\phi^- \rangle_{\hat{p}_\phi^-(y^-|x)} \right) \end{aligned} \tag{5.6}$$

Setting it to zero we obtain the characterization of the extremum point

$$\bar{p}^+(x) \Delta \langle \vec{\psi}_\phi^+ \rangle = \lambda \bar{p}^-(x) \Delta \langle \vec{\psi}_\phi^- \rangle \tag{5.7}$$

where $\Delta\langle\psi\rangle$ is the difference in the expectation of ψ taken according to the model and the true distribution.

To obtain some intuition into the last equation consider the following two observations. First, note that maximizing the information $I_{min}^{xy}[\phi, \bar{p}^+]$ requires to minimize the absolute difference between the expectancies of $\vec{\psi}_\phi^+$, as can be seen when taking $\lambda = 0$. Second, it can be shown that when minimizing $I_{min}^{xy}[\phi, \bar{p}^-]$ alone, some elements of $\vec{\phi}(x)$ must diverge. In these infimum points $\Delta\langle\vec{\psi}_\phi^-\rangle$ does not generally vanish. Taken together, these facts imply that for the $\lambda > 0$ case, the difference $\Delta\langle\vec{\psi}_\phi^+\rangle$ should generally be different from zero. This implies, as expected, that the resulting $\vec{\phi}(x)$ conveys less information than the $\lambda = 0$ solution. The optimal $\vec{\phi}(x)$ is thus bound to provide an inaccurate model for those aspects of $\bar{p}^+$ that also improve the model of $\bar{p}^-$.

An additional interesting interpretation of $\vec{\psi}_\phi^+$, $\vec{\psi}_\phi^-$ is that they reflect the relative importance of $\vec{\phi}(x)$ in $\hat{p}_\phi^+$, $\hat{p}_\phi^-$ for a given y. This view is prevalent in the boosting literature, where such coefficients function as the weights of the weak learners (see e.g. [83]). However, SDR-IS also optimizes the weak learners, searching for a small but optimal set of learners.

5.3 Algorithmic Considerations

Unlike the case of $\lambda = 0$ for which an iterative algorithm was described in the previous chapter, the $\lambda > 0$ case poses a special difficulty in developing such an algorithm. One could supposedly proceed by calculating $\vec{\psi}_\phi^+$, $\vec{\psi}_\phi^-$ assuming a constant value of $\vec{\phi}(x)$ and then calculate the resulting $\vec{\phi}(x)$ assuming $\vec{\psi}^+$ and $\vec{\psi}^-$ are constant. However, as was shown in [56], updating $\vec{\psi}_\phi^-$ will increase $I_{min}^{xy}[\vec{\phi}(x), \bar{p}^-]$ thereby decreasing the target function. Thus, such a procedure is not guaranteed to improve the target function. Possibly, an algorithm guaranteed to converge for a limited range of λ values can be devised, as done for IBSI [25], but this remains to be studied.

Fortunately, the analytic characterization of the gradient derived above allows one to use a gradient ascent algorithm for finding the optimal features $\vec{\phi}(x)$, for any given

value of λ. This requires to calculate a Maximum Entropy distribution on each of its iterations, namely, to calculate numerically the set of Lagrange multipliers $\vec{\psi}_\phi(y)$, $A_\phi(x)$ and $B_\phi(y)$ which appear in the gradient expression in Equation 5.5. This convex problem has a single maximum, and well studied algorithms exist for finding Maximum Entropy distributions under linear constraints [1]. These include GIS [32], IIS [34], or gradient based algorithms (see [88] for a review of different algorithms and their relative efficiency). In all the results described below we used the GIS algorithm.

5.4 Relation to Other Methods

5.4.1 Likelihood Ratio Maximization

Further intuition into the functional of Equation 5.2 can be obtained, using the result of [56] yielding that it equals up to a constant to

$$\mathcal{L}[\vec{\phi}(x)] \;=\; -D_{KL}[\bar{p}^+||\hat{p}^+_\phi] + \lambda D_{KL}[\bar{p}^-||\hat{p}^-_\phi] \;, \tag{5.8}$$

where $D_{KL}[p||q] \equiv \sum p_i \log(p_i/q_i)$ is the Kullback-Leibler divergence. When $\bar{p}^+$ and $\bar{p}^-$ share the same marginal distribution $\bar{p}(x)$, a joint distribution $\bar{p}(X, Y^+, Y^-)$ can be defined that coincides with the pairs-joint distributions $\bar{p}^+(X, Y^+)$ and $\bar{p}^-(X, Y^-)$,

$$\bar{p}(x, y^+, y^-) \equiv \bar{p}^+(y^+|x)\bar{p}^-(y^-|x)\bar{p}(x) \;. \tag{5.9}$$

The above distribution has the quality that Y^- and Y^+ are conditionally independent given X. In many settings, this is indeed a reasonable assumption. In this case

$$\begin{aligned}
\mathcal{L} \;&=\; -\sum_{x,y^+,y^-} \bar{p}(x, y^+, y^-) \log\left(\frac{\bar{p}^+(x, y^+)}{\hat{p}^+_\phi(x, y^+)}\right) \\
&\quad +\lambda \sum_{x,y^+,y^-} \bar{p}(x, y^+, y^-) \log\left(\frac{\bar{p}^-(x, y^-)}{\hat{p}^-_\phi(x, y^-)}\right) \\
&=\; -\left\langle \log\left(\frac{\bar{p}^+(x, y^+)}{\bar{p}^-(x, y^-)^\lambda}\frac{\hat{p}^-_\phi(x, y^-)^\lambda}{\hat{p}^+_\phi(x, y^+)}\right)\right\rangle_{\bar{p}(x,y^+,y^-)} \\
&=\; \left\langle \log\left(\frac{\hat{p}^+_\phi(x, y^+)}{\hat{p}^-_\phi(x, y^-)^\lambda}\right)\right\rangle_{\bar{p}(x,y^+,y^-)} + const
\end{aligned} \tag{5.10}$$

This suggests that in the special case of $\lambda = 1$, SDR-IS operates to maximize the expected log likelihood ratio, between the maximum entropy models $\hat{p}^+_\phi$ and $\hat{p}^-_\phi$. In the general case of $\lambda > 0$ a weighted log likelihood ratio is obtained. For vanishing λ, the irrelevance information is completely ignored and the problem reduces to unconstrained likelihood maximization of the maximum entropy model $\hat{p}^+_\phi$.

[1] Note that all the constraints in $\mathcal{P}(\vec{\phi}(x), \bar{p})$ are indeed linear.

5.4.2 Weighted vs. Constrained Optimization

The trade-off optimization problem of Equation 5.1, is related to the following constrained optimization problem

$$\vec{\phi}^*(x) = \arg \max_{\vec{\phi}(x): I^{xy}_{min}[\vec{\phi}(x),\bar{p}^-] \leq D} I^{xy}_{min}[\vec{\phi}(x), \bar{p}^+] \quad . \tag{5.11}$$

Although the Lagrangian for this problem is identical to the SDR-IS target functional of Equation (5.1), these two problems are not necessarily equivalent, since a constrained optimization problem like (5.11) may in principle be solved by the minimum point of Equation 5.1. However, under certain convexity conditions such problems can be shown to be equivalent. While we do not present a similar proof here, we found numerically in all the data described below, that the maximum points of Equation 5.11 were always achieved at the maximum of Equation (5.1) rather than at its minima.

5.4.3 Related Methods

Several methods previously appeared in the literature, which make use of auxiliary data or additional sources of information to enhance learning features of a main data set. The method of Oriented-PCA [37] uses a main data set with covariance S^+ and an irrelevance data set with covariance S^- to find features w that maximize the Signal to Noise Ratio $\frac{w^T S^+ w}{w^T S^- w}$. Constrained-PCA [37] finds principal components of the main data which are orthogonal to the irrelevance data. While these methods implicitly assume Gaussian distributions in input space, a kernelized version of OPCA was described in [92].

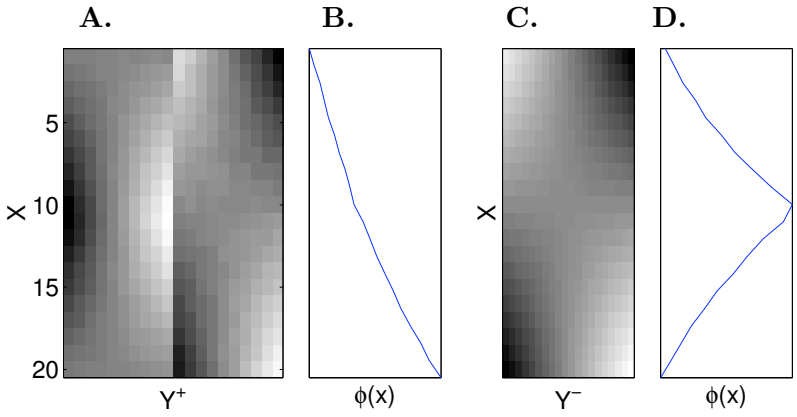


Figure 5.1: Demonstration of SDR-IS operation. **A.** A joint distribution $P(X, Y^+)$ that contains two distinct and conflicting structures (see text) **B.** Extracting a one-dimensional feature with $\lambda = 0$ identifies the top-to-bottom gradient. **C.** A joint distribution $P(X, Y^-)$ that contains a single structure similar to the right structure of $P(X, Y^+)$. **D.** Extracting a one-dimensional feature with $\lambda = 1$ successfully ignores the top-to-bottom gradient and extracts the weaker structure of $P(X, Y^+)$.

Another line of work uses auxiliary information in the form of equivalence constraints. The auxiliary data here is a set of relations that enforce similarity between the elements of the main data. These relations are used to improve dimensionality reduction [125], or to improve the distance metrics used for clustering [145].

Separating several conflicting structures in the data has also been addressed in [133] where a bilinear model was used to separate style from content. This model does not use auxiliary information, but rather assumes that the two structures can be represented by a linear model.

SDR-IS differs from the above methods in that it is a non-linear method for extracting continuous features, which are **least informative** about the irrelevance data. The relative importance of the irrelevance data is determined throught the tradeoff parameter λ.

5.5 Applications

We first illustrate the operation of SDR-IS on a synthetic example that demonstrates its main properties. Then, we describe its application to the problem of feature extraction for face recognition.

5.5.1 A Synthetic Example

To demonstrate the ability of our approach to uncover weak but interesting hidden structures in data, we designed a co-occurrence matrix that contains two competing sub-structures (see figure 5.1A). The right half of the matrix contains a top-to-bottom gradient, while its left half contains large variance at the middle values of X. The right structure was hand-crafted to be stronger in magnitude than the left one.

When SDR-IS is applied with no irrelevance information ($\lambda = 0$) and $d = 1$, it extracts the top-to-bottom gradient (Figure 5.1B). This $\phi(x)$ follows from the strong structure on the right part of 5.1A.

We now created a second co-occurrence matrix $P(X, Y^-)$ that contains a top-to-bottom structure similar to that of $P(X, Y^+)$ (Figure 5.1C). Applying SDR-IS with $\lambda = 1$ on both matrices now successfully ignores the strong top-to-bottom structure in $P(X, Y^+)$ and retrieves the weaker structure that emphasizes the mid values of X (Figure 5.1D). Importantly, this is done in an unsupervised manner, without explicitly pointing to the strong but irrelevant structure.

Further understanding of the operation of SDR-IS is gained by tracing its output as a function of the tradeoff parameter λ. Figure 5.2A plots the optimal features $\vec{\phi}(x)$ extracted for various λ values, revealing a phase transition around a critical value $\lambda = 0.26$. The reason for this behavior is that at the critical λ, the top-to-bottom feature $\vec{\phi}(x)$ bears larger loss (due to the information $I^{xy}_{min}[\vec{\phi}(x), \bar{p}^-]$ conveyed about Y^-) than gain. Figure 5.2B traces the values of $I^{xy}_{min}[\vec{\phi}(x), \bar{p}^+]$ and $I^{xy}_{min}[\vec{\phi}(x), \bar{p}^+]$,

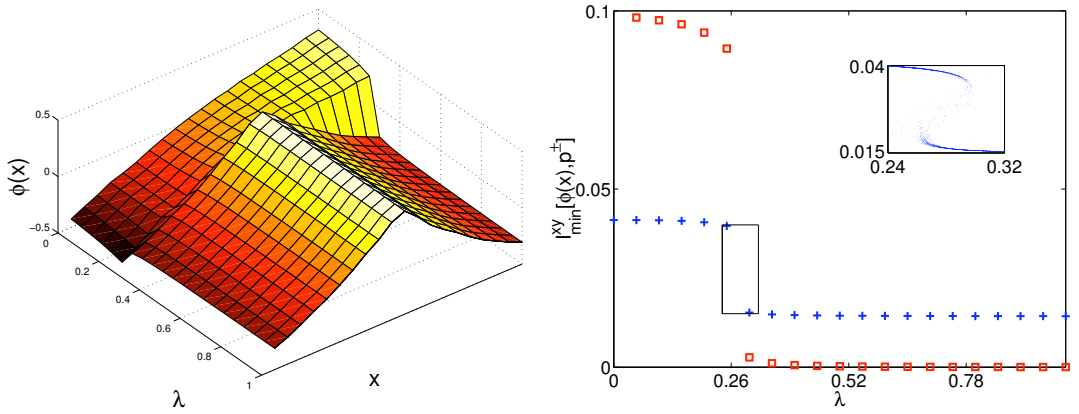


Figure 5.2: Operation of SDR-IS on the synthetic example of Figure 1 for various values of λ. **A.** The optimal $\vec{\phi}(x)$ extracted with SDR-IS. **B.** The information conveyed about Y^+ (crosses) and Y^- (squares), by the optimal $\vec{\phi}(x)$'s of the left panel. A phase transition around 0.26 is observed both in the information values and the $\vec{\phi}(x)$'s. The inset shows the spinodal metastable points of $I^{xy}_{min}[\vec{\phi}(x), \bar{p}^+]$ around the phase transition point (black box).

again revealing a pronounced phase transition, and an S shaped (spinodal) curve of $I^{xy}_{min}[\vec{\phi}(x), \bar{p}^+]$, indicating the co-existence of three local maxima and the metastable region (inset of Figure 5.2B). Such spinodal curves are typical to phase transition phenomena obeserved in numerous physical systems. Plotting the SDR-IS functional of Equation 5.1 as a function of λ (not shown) also reveals a discontinuity in its first derivative, indicating a first order phase transition. These discontinuities reflect the removal of "irrelevant" features from $\vec{\phi}(x)$, and can thus be used to select interesting values of λ.

The irrelevant structures in the above example were hand crafted to be strongly and cleanly manifested in $p(x, y^-)$. The next section studies the application of SDR-IS to real data, in which structures are much more covert.

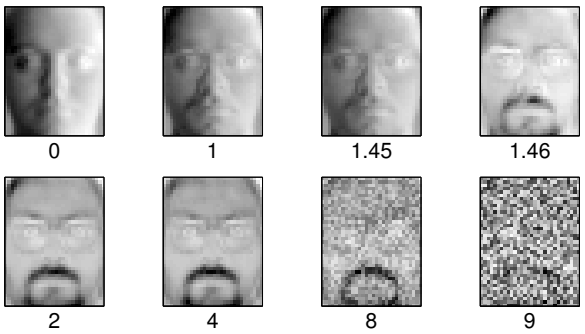


Figure 5.3: Extracting a single feature using SDR-IS, for various λ values. An apparent phase transition is observed around $\lambda = 1.45$. $\bar{p}^+$ was created by taking images of all men in the AR database with neutral face expressions and light either from the right or the left (a total of 100 images). $\bar{p}^-$ was similarly created with 100 female images. Positive λ values reveal features that differentiate between men but not between women.

5.5.2 Face Images

Face recognition poses a challenge to relevant features extraction since these must be invariant to various interfering structures, such as face expression and light conditions. Such nuisance structures are often more pronounced in the data than the subtle features required to recognize a person.

We tested SDR-IS on this task using the AR database [91], a collection of faces with various face expressions, light conditions and occlusions. Each image was translated into a joint probability matrix, by considering the normalized grey levels of the pixel x in the image y as the probability $p(x|y)$, and setting $p(y)$ uniform. This normalization scheme views every image y as a distribution $p(x|y)$ which stands for the probability of observing a photon at a given pixel x. To demonstrate the operation of SDR-IS on this data we first trained it to extract a single feature, for various λ values. The experiment details and resulting $\vec{\phi}(x)$ are given in Figure 5.3. When λ is low (small weight for irrelevance information) the main structure captured is the direction of light source (right vs. left). As λ increases the optimal $\vec{\phi}(x)$ first changes only slightly, but then a phase transition occurs around $\lambda = 1.45$, and a second structure emerges. This phase transition can be well observed when tracing the values of $I^{xy}_{min}[\vec{\phi}(x), \bar{p}^+]$ and $I^{xy}_{min}[\vec{\phi}(x), \bar{p}^-]$ as a function of λ (Figure 5.4), and results from the same reasons discussed in the synthetic example described above. This result suggests that such information curves can be used to identify "interesting" values of λ and their corresponding features even for high dimensional and complex data. As λ further increases, the algorithm focuses on minimizing information about the irrelevance information, disregarding information about the main data. This results in the noisy features seen in Figure 5.3 for high λ values.

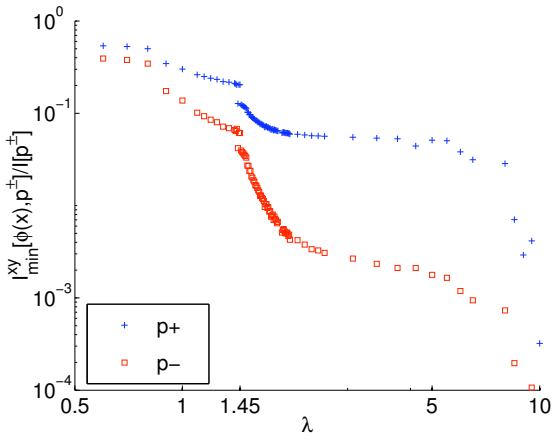


Figure 5.4: Normalized information about the main data $I^{xy}_{min}[\vec{\phi}(x), \bar{p}^+]$ and the irrelevance data $I^{xy}_{min}[\vec{\phi}(x), \bar{p}^-]$, as a function of λ, for the data of Figure 5.3. Note the phase transition in both information levels for $\lambda = 1.45$.

To quantify the performance of SDR-IS in a comparative manner, we used it in a

difficult task of unsupervised feature extraction for face recognition, and compared its performance with three methods: PCA - the most widely used dimensionality reduction method; Constrained PCA (CPCA); and oriented PCA (OPCA) - two methods that utilize the same irrelevance data as SDR-IS [37]. We created $p(X, Y^+)$ with images of five different men, under all the different conditions of face expression and light conditions (a total of 26 images per person). As irrelevance data we used all 26 images of another randomly chosen man. The task of clustering these images into the five correct sets is hard since the nuisance structures are far more dominant than the relevant structure of inter subject variability, in face of light and face expression invariances.

All methods were used to reduce the dimensionality of the images. PCA representations were obtained by projecting on the principal components. The SDR-IS representation was obtained by replacing each image y with its expected SDR-IS feature values $\langle\vec{\phi}(x)\rangle_{p(x|y)}$. This follows our motivation of using expected values alone to represent y.

To quantify the effectiveness of the reduced representations in preserving person identity, we calculated the number of same-class (same-person) neighbors out of the k nearest neighbors of each image[2]. This was averaged over all images and k's and normalized, yielding the precision index [3].

Optimal parameters (dimensionality and λ) for all methods, were chosen to maximize the precision index for a training set. Reported results were obtained on a separate testing set. This entire procedure was repeated 10 times on randomly chosen subsets of the database. Figure 4 compares the effectiveness of SDR-IS with the one obtained with PCA based methods. SDR-IS was found to achieve more than 30 percent improvement over the second best method.

We further compared the performance of the four methods for each predefined dimensionality d. Figure 5.6 shows that SDR-IS dominates the other methods over all d values. This is more pronounced for low values of d, which agrees with the intuition that the irrelevance data allows SDR-IS to focus on the more relevant features.

5.6 Discussion

The method introduced in this chapter addresses the fundamental problem of extracting relevant structure in an unsupervised manner, a problem for which only few principled approaches were suggested. We focused on continuous features of categorical variables

[2] As a metric for measuring distances between images, we tested both the L2 norm and the Mahalanobis distance in the reduced representation. We report the Mahalanobis results only, since L2 results were considerably worse for PCA.

[3] We also evaluated the methods by clustering the low dimensional vectors into five groups and comparing the resulting clusters with the true ones. This resulted in qualitatively similar result, albeit noisier. We prefer the method presented here since it does not depend on a noisy second phase of clustering.

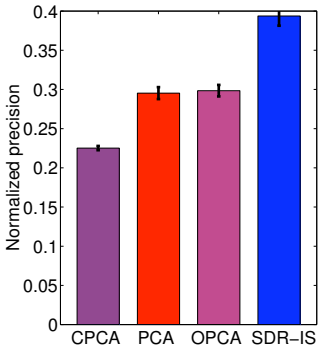


Figure 5.5: Performance of SDR-IS compared with other methods. Performance is normalized between 0 (obtained with random neighboring) and 1 (all nearest neighbors are of the same class). The average over ten cross validation sets is shown. SDR-IS achieves 30 percent improvement over the second best method (OPCA).

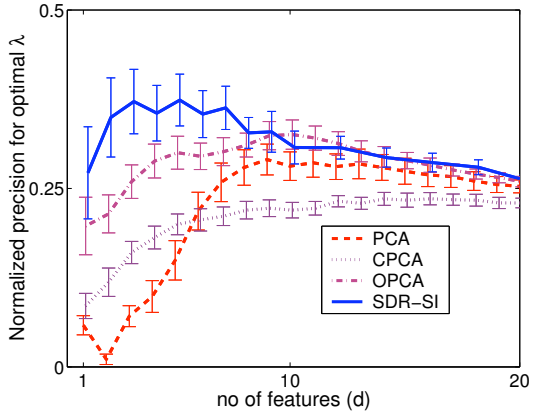


Figure 5.6: Performance of SDR-IS compared with other methods as a function of dimensionality d for the AR data. The mean performance over 10 testing sets is reported, and bars denote standard error of the mean over these sets. In SDR-IS, a value of λ was chosen for each d, to maximize performance over the training set

and used the information theoretic notion of *information in the expectation value of a measurement*, to derive algorithms that extract the most informative features, by utilizing information about irrelevant properties. Such an information theoretic approach makes no assumptions about the origin of the empirical data and is thus different from the more common generative modeling methodology.

Our formalism can be extended to the case of multiple relevance and irrelevance variables $(Y_1^+, ..., Y_{n^+}^+)$ and $(Y_1^-, ..., Y_{n^-}^-)$, with joint distributions $\bar{p}_i^+ \equiv \bar{p}(x, y_i^+)$ and $\bar{p}_i^- \equiv \bar{p}(x, y_i^-)$. Following a similar weighted optimization problem we write the Lagrange form of the functional $\mathcal{L} = \sum_{i=1}^{n^+} \lambda_i I_{min}^{xy}[\vec{\phi}(x), \bar{p}_i^+] - \sum_{i=1}^{n^-} \lambda_i I_{min}^{xy}[\vec{\phi}(x), \bar{p}_i^-]$, which can be maximized as in the two variables case.

An interesting property that is revealed when applying SDR-IS both to synthetic

and real life data, is the emergence of phase-transitions. These are discontinuous changes in the information values, that occur at specific values of λ. They are paralleled by abrupt changes in the features $\vec{\phi}(x)$ and thus provide a natural way to focus on important values of λ that characterize the inherent features of the data. As demonstrated in Figure (5.4), we were able to follow the metastable region of the phase transition, which appears to behave like a first-order transition in thermodynamics.

An interesting algorithmic problem, not fully answered at this point, is to design an iterative-projection algorithm, similar to the SDR, for solving the implicit equations for the optima. This can improve time complexity and convergence of the algorithm, making it even more practical.

Chapter 6

Information Minimization and Dimension - The Gaussian Information Bottleneck

The previous chapters addressed the problem of finding features whose measurements provide useful information. However, they did not address at depth the question of the dimensionality of the feature function $\vec{\phi}(x)$ [1].

The current chapter addresses the question of the interaction between dimensionality reduction and preservation of information. This will be done by analyzing dimensionality reduction of Gaussian variables via the Information Bottleneck approach [135].

Extracting relevant aspects of complex data is a fundamental task in machine learning and statistics. The problem is often that the data contains many structures, which make it difficult to define which of them are relevant and which are not in an unsupervised manner. For example, speech signals may be characterized by their volume level, pitch, or content; pictures can be ranked by their luminosity level, color saturation or importance with regard to some task.

This problem was addressed in a principled manner by the information bottleneck (IB) approach [135]. Given the joint distribution of a "source" variable X and another "relevance" variable Y, IB operates to compress X, while preserving information about Y. The variable Y thus implicitly defines what is relevant in X and what is not. Formally, this is cast as the following variational problem

$$\min_{p(t|x)} \mathcal{L} : \mathcal{L} \equiv I(X;T) - \beta I(T;Y) \ , \tag{6.1}$$

where T represents the compressed representation of X via the conditional distributions $p(t|x)$, while the information that T maintains on Y is captured by the distribution $p(y|t)$. This formulation is general and does not depend on the type of the X, Y

[1] Although suggestions for choice of dimension were given in Section 4.7.3.

distribution. The positive parameter β determines the tradeoff between compression and preserved relevant information, as the Lagrange multiplier for the constrained optimization problem $\min_{p(t|x)} I(X;T) - \beta\,(I(T;Y) - const)$. Since T is a function of X it is independent of Y given X, thus the three variables can be written as the Markov chain $Y - X - T$. From the information inequality we thus have $I(X;T) - \beta I(T;Y) \geq (1-\beta)I(T;Y)$, and therefore for all values of $\beta \leq 1$, the optimal solution of the minimization problem is degenerated $I(T;X) = I(T;Y) = 0$. As we will show below, the range of degenerated solutions is even larger for Gaussian variables and depends on the eigen spectrum of the variables covariance matrices.

The rationale behind the IB principle can be viewed as model-free "looking inside the black-box" system analysis approach. Given the input-output (X,Y) "black-box" statistics, IB aims to construct efficient representations of X, denoted by the variable T, that can account for the observed statistics of Y. IB achieves this using a single tradeoff parameter to represent the tradeoff between the complexity of the representation of X, measured by $I(X;T)$, and the accuracy of this representation, measured by $I(T;Y)$. The choice of mutual information for the characterization of complexity and accuracy stems from Shannon's theory, where information minimization corresponds to optimal compression in Rate Distortion Theory, and its maximization corresponds to optimal information transmission in Noisy Channel Coding.

From a machine learning perspective, IB may be interpreted as regularized generative modeling. Under certain conditions $I(T;Y)$ can be interpreted as an empirical likelihood of a special mixture model, and $I(T;X)$ as penalizing complex models [130]. While this interpretation can lead to interesting analogies, it is important to emphasize the differences. First, IB views $I(X;T)$ not as a regularization term, but rather corresponds to the distortion constraint in the original system. As a result, this constraint is useful even when the joint distribution is known exactly, because the goal of IB is to obtain compact representations rather than to estimate density. Interestingly, $I(T;X)$ also characterizes the complexity of the representation T as the expected number of bits needed to specify the t for a given x. In that role it can be viewed as an expected "cost" of the internal representation, as in MDL. As is well acknowledged now source coding with distortion and channel coding with cost are dual problems [see for example 124, 112]. In that information theoretic sense, IB is *self dual*, where the resulting source and channel are perfectly matched [as in 48].

The information bottleneck approach has been applied so far mainly to categorical variables, with a discrete T that represents (soft) clusters of X. It has been proved useful for a range of applications from documents clustering [128] through neural code analysis [38] to gene expression analysis [44, 126] (for a more detailed review of IB clustering algorithms see [127]). However, its general information theoretic formulation is not restricted, both in terms of the nature of the variables X and Y, as well as of the compression variable T. It can be naturally extended to nominal, categorical, and con-

tinuous variables, as well as to dimension reduction rather than clustering techniques. The goal of this chapter is to apply the IB for the special, but very important, case of Gaussian processes which has become one of the most important generative classes in machine learning. In addition, this is the first concrete application of IB to dimension reduction with continuous compressed representation, and as such exhibits interesting dimension related phase transitions.

The general solution of IB for continuous T yields the same set of self-consistent equations obtained already in [135], but solving these equations for the distributions $p(t|x)$, $p(t)$ and $p(y|t)$ without any further assumptions is a difficult challenge, as it yields non-linear coupled eigenvalue problems. As in many other cases, however, we show here that the problem turns out to be analytically tractable when X and Y are joint multivariate Gaussian variables. In this case, rather than using the fixed point equations and the generalized Blahut-Arimoto algorithm as proposed in [135], one can explicitly optimize the target function with respect to the mapping $p(t|x)$ and obtain a closed form solution of the optimal dimensionality reduction.

The optimal compression in the Gaussian Information Bottleneck (GIB) is defined in terms of the compression-relevance tradeoff (also known as the "Information Curve", or "Accuracy-Complexity" tradeoff), determined by varying the parameter β. The optimal solution turns out to be a noisy linear projection to a subspace whose dimensionality is determined by the parameter β. The subspaces are spanned by the basis vectors obtained as in the well known *Canonical Correlation Analysis* (CCA) [69], but the exact nature of the projection is determined in a unique way via the parameter β. Specifically, as β increases, additional dimensions are added to the projection variable T, through a series of critical points (structural phase transitions), while at the same time the relative magnitude of each basis vector is rescaled. This process continues until all the relevant information about Y is captured in T. This demonstrates how the IB principle can provide a continuous measure of model complexity in information theoretic terms.

The idea of maximization of relevant information was also taken in the *Imax* framework of Becker and Hinton [8, 7], which followed Linsker's idea of information maximization [85, 86]. In the Imax setting, there are two one-layer feed forward networks with inputs X_a, X_b and outputs neurons Y_a, Y_b; the output neuron Y_a serves to define relevance to the output of the neighboring network Y_b. Formally, the goal is to tune the incoming weights of the output neurons, such that their mutual information $I(Y_a; Y_b)$ is maximized. An important difference between *Imax* and the IB setting, is that in the *Imax* setting, $I(Y_a; Y_b)$ is invariant to scaling and translation of the Y's since the compression achieved in the mapping $X_a \rightarrow Y_a$ is not modeled explicitly. In contrast, the IB framework aims to characterize the dependence of the solution on the explicit compression term $I(T; X)$, which is a *scale sensitive* measure when the transformation is noisy. This view of compressed representation T of the inputs X is useful when

dealing with neural systems that are stochastic in nature and limited in their responses amplitudes and are thus constrained to finite $I(T;X)$.

The current chapter starts by defining the problem of relevant information extraction for Gaussian variables. Section 3 gives the main result of the chapter: an analytical characterization of the optimal projections, which is then developed in Section 4. Section 5 develops an analytical expression for the GIB compression-relevance tradeoff - the information curve. Section 6.5 shows how the general IB algorithm can be adapted to the Gaussian case, yielding an iterative algorithm for finding the optimal projections. The relations to canonical correlation analysis and coding with side-information are discussed in Section 6.8.

Most of the material in the current chapter was published in [24].

6.1 Gaussian Information Bottleneck

We now formalize the problem of Information Bottleneck for Gaussian variables. Let (X,Y) be two jointly multivariate Gaussian variables of dimensions n_x, n_y and denote by Σ_x,Σ_y the covariance matrices of X,Y and by Σ_{xy} their cross-covariance matrix [2]. The goal of GIB is to compress the variable X via a stochastic transformation into another variable $T \in R^{n_x}$, while preserving information about Y. The dimension of T is not explicitly limited in our formalism, since we will show that the effective dimension is determined by the value of β.

It is shown in [57] that the optimum for this problem is obtained by a variable T which is also jointly Gaussian with X. The formal proof uses the entropy power inequality as in [13], and is rather technical, but an intuitive explanation is that since X and Y are Gaussians, the only statistical dependencies that connect them are bilinear. Therefore, a linear projection of X is sufficient to capture all the information that X has on Y. The Entropy-power inequality is used to show that a linear projection of X, which is also Gaussian in this case, indeed attains this maximum information.

Since every two centered random variables X and T with jointly Gaussian distribution can be presented through the linear transformation $T = AX + \xi$, where $\xi \sim N(0, \Sigma_\xi)$ is another Gaussian that is independent of X, we formalize the problem using this representation of T, as the following minimization,

$$\min_{A,\Sigma_\xi} \mathcal{L} \equiv I(X;T) - \beta I(T;Y) \tag{6.2}$$

over the noisy linear transformations of A, Σ_ξ

$$T = AX + \xi; \quad \xi \sim N(0, \Sigma_\xi) \ . \tag{6.3}$$

Thus T is normally distributed $T \sim N(0, \Sigma_t)$ with $\Sigma_t = A\Sigma_x A^T + \Sigma_\xi$.

[2] For simplicity we assume that X and Y have zero means and Σ_x, Σ_y are full rank. Otherwise X and Y can be centered and reduced to the proper dimensionality.

Interestingly, the term ξ can also be viewed as an additive noise term, as commonly done in models of learning in neural networks. Under this view, ξ serves as a regularization term whose covariance determines the scales of the problem. While the goal of GIB is to find the optimal projection parameters A, Σ_ξ jointly, we show below that the problem factorizes such that the optimal projection A does not depend on the noise, which does not carry any information about Y.

6.2 The Optimal Projection

The first main result of this chapter is the characterization of the optimal A,Σ_ξ as a function of β

Theorem 6.2.1 *The optimal projection $T = AX + \xi$ for a given tradeoff parameter β is given by $\Sigma_\xi = I_x$ and*

$$A = \left\{ \begin{array}{cc} \left[\mathbf{0}^T; \ldots; \mathbf{0}^T\right] & 0 \leq \beta \leq {\beta^c}_1 \\ \left[\alpha_1 \mathbf{v}_1^T, \mathbf{0}^T; \ldots; \mathbf{0}^T\right] & \beta_1^c \leq \beta \leq {\beta^c}_2 \\ \left[\alpha_1 \mathbf{v}_1^T; \alpha_2 \mathbf{v}_2^T; \mathbf{0}^T; \ldots; \mathbf{0}^T\right] & {\beta^c}_2 \leq \beta \leq {\beta^c}_3 \\ \vdots & \end{array} \right\} \tag{6.4}$$

where $\{\mathbf{v}_1^T, \mathbf{v}_2^T, \ldots, \mathbf{v}_{n_x}^T\}$ are left eigenvectors of $\Sigma_{x|y}\Sigma_x^{-1}$ sorted by their corresponding ascending eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_{n_x}$, ${\beta^c}_i = \frac{1}{1-\lambda_i}$ are critical β values, α_i are coefficients defined by $\alpha_i \equiv \sqrt{\frac{\beta(1-\lambda_i)-1}{\lambda_i r_i}}$, $r_i \equiv \mathbf{v}_i^T \Sigma_x \mathbf{v}_i$, $\mathbf{0}^T$ is an n_x dimensional row vector of zeros, and semicolons separate rows in the matrix A.

This theorem asserts that the optimal projection consists of eigenvectors of $\Sigma_{x|y}\Sigma_x^{-1}$, combined in an interesting manner: For β values that are smaller than the smallest critical point ${\beta^c}_1$, compression is more important than any information preservation and the optimal solution is the degenerated one $A \equiv 0$. As β is increased, it goes through a series of critical points ${\beta^c}_i$, at each of which another eigenvector of $\Sigma_{x|y}\Sigma_x^{-1}$ is added to A. Even though the rank of A increases at each of these transition points, A changes continuously as a function of β since at each critical point ${\beta^c}_i$ the coefficient α_i vanishes. Thus β parameterizes a sort of "continuous rank" of the projection.

To illustrate the form of the solution, we plot the landscape of the target function $\mathcal{L}$ together with the solution in a simple problem where $X \in R^2$ and $Y \in R$. In this case A has a single non-zero row, thus A can be thought of as a row vector of length 2, that projects X to a scalar $A : X \to R$, $T \in R$. Figure 6.1 shows the target function $\mathcal{L}$ as a function of the (vector of length 2) projection A. In this example, the largest eigenvalue is $\lambda_1 = 0.95$, yielding ${\beta^c}_1 = 20$. Therefore, for $\beta = 15$ (Figure 6.1A) the zero solution is optimal, but for $\beta = 100 > \beta^c$ (Figure 6.1B) the corresponding eigenvector is a feasible solution, and the target function manifold contains two mirror minima. As

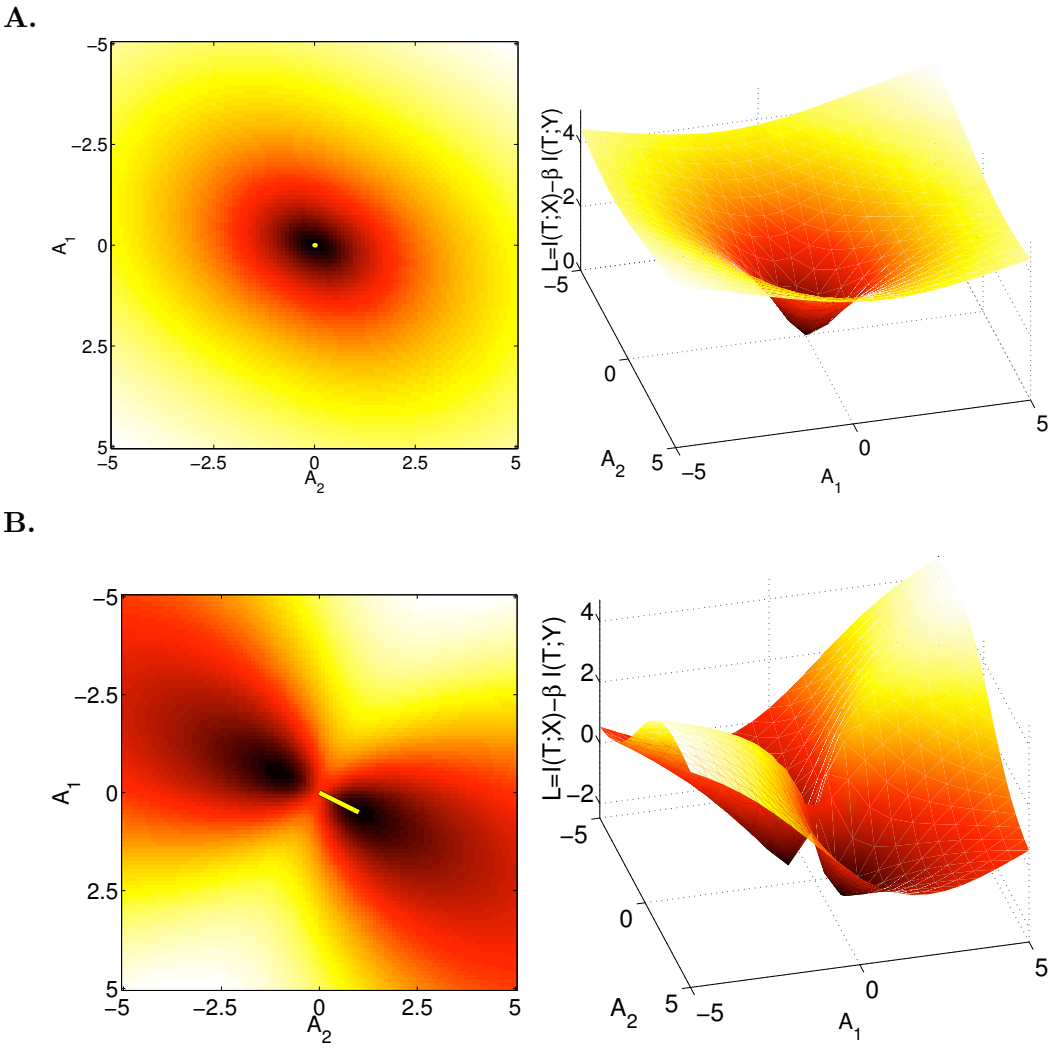


Figure 6.1: The surface of the target function $\mathcal{L}$ calculated numerically as a function of the optimization parameters in two illustrative examples with a scalar projection $A : R^2 \rightarrow R$. Each row plots the target surface $\mathcal{L}$ both in 2D (left) and 3D (right) as a function of the (two dimensional) projections A. **A.** For $\beta = 15$, the optimal solution is the degenerated solution $A \equiv 0$. **B.** For $\beta = 100$, a non degenerate solution is optimal, together with its mirror solution. The $\Sigma_{x|y}\Sigma_x^{-1}$- eigenvector of smallest eigenvalue, with a norm computed according to theorem 6.2.1 is superimposed, showing that it obtains the global minimum of $\mathcal{L}$. Parameters' values $\Sigma_{xy} = [0.1\ 0.2]$, $\Sigma_x = I_2$, $\Sigma_\xi = 0.3I_{2\times 2}$.

β increases from 1 to ∞, these two minima, starting as a single unified minimum at zero, split at β^c, and then diverge apart to ∞.

We now turn to prove theorem 6.2.1.

6.3 Deriving the Optimal Projection

We first rewrite the target function as

$$\mathcal{L} = I(X;T) - \beta I(T;Y) = h(T) - h(T|X) - \beta h(T) + \beta h(T|Y) \tag{6.5}$$

where h is the (differential) entropy of a continuous variable

$$h(X) \equiv -\int_X f(x) \log f(x)\, dx \quad .$$

Recall that the entropy of a d dimensional Gaussian variable is

$$h(X) = \frac{1}{2} \log \left((2\pi e)^d |\Sigma_x| \right)$$

where $|x|$ denotes the determinant of x, and Σ_x is the covariance of X. We therefore turn to calculate the relevant covariance matrices. From the definition of T we have $\Sigma_{tx} = A\Sigma_x$, $\Sigma_{ty} = A\Sigma_{xy}$ and $\Sigma_t = A\Sigma_x A^T + \Sigma_\xi$. Now, the conditional covariance matrix $\Sigma_{x|y}$ can be used to calculate the covariance of the conditional variable $T|Y$, using the Schur complement formula [see e.g., 87]

$$\Sigma_{t|y} = \Sigma_t - \Sigma_{ty}\Sigma_y^{-1}\Sigma_{yt} = A\Sigma_{x|y}A^T + \Sigma_\xi$$

The target function can now be rewritten as

$$\begin{aligned} \mathcal{L} &= \log(|\Sigma_t|) - \log(|\Sigma_{t|x}|) - \beta \log(|\Sigma_t|) + \beta \log(|\Sigma_{t|y}|) \\ &= (1-\beta)\log(|A\Sigma_x A^T + \Sigma_\xi|) - \log(|\Sigma_\xi|) + \beta \log(|A\Sigma_{x|y}A^T + \Sigma_\xi|) \end{aligned} \tag{6.6}$$

Although $\mathcal{L}$ is a function of both the noise Σ_ξ *and* the projection A, Lemma A.3.1 in Appendix A shows that for every pair (A, Σ_ξ), there is another projection $\tilde{A}$ such that the pair $(\tilde{A}, I)$ obtains the same value of $\mathcal{L}$. This is obtained by setting $\tilde{A} = \sqrt{D^{-1}}VA$ where $\Sigma_\xi = VDV^T$, which yields $\mathcal{L}(\tilde{A}, I) = \mathcal{L}(A, \Sigma_\xi)$[3]. This allows us to simplify the calculations by replacing the noise covariance matrix Σ_ξ with the identity matrix I_d.

To identify the minimum of $\mathcal{L}$ we differentiate $\mathcal{L}$ w.r.t. to the projection A using the algebraic identity $\frac{\delta}{\delta A}\log(|ACA^T|) = (ACA^T)^{-1}2AC$ which holds for any symmetric matrix C.

$$\frac{\delta \mathcal{L}}{\delta A} = (1-\beta)(A\Sigma_x A^T + I_d)^{-1}2A\Sigma_x + \beta(A\Sigma_{x|y}A^T + I_d)^{-1}2A\Sigma_{x|y} \tag{6.7}$$

Equating this derivative to zero and rearranging, we obtain necessary conditions for an internal minimum of $\mathcal{L}$, which we explore in the next two sections.

[3] Although this holds only for full rank Σ_ξ, it does not limit the generality of the discussion since low rank matrices yield infinite values of $\mathcal{L}$ and are therefore suboptimal.

6.3.1 Scalar Projections

For clearer presentation of the general derivation, we begin with a sketch of the proof by focusing on the case where T is a scalar, that is, the optimal projection matrix A is a now a single row vector. In this case, both $A\Sigma_x A^T$ and $A\Sigma_{x|y}A^T$ are scalars, and we can write

$$\left(\frac{\beta-1}{\beta}\right)\left(\frac{A\Sigma_{x|y}A^T+1}{A\Sigma_x A^T+1}\right)A = A\left[\Sigma_{x|y}\Sigma_x^{-1}\right] \quad . \tag{6.8}$$

This equation is therefore an eigenvalue problem in which the eigenvalues depend on A. It has two types of solutions depending on the value of β. First, A may be identically zero. Otherwise, A must be the eigenvector of $\Sigma_{x|y}\Sigma_x^{-1}$, with an eigenvalue $\lambda = \frac{\beta-1}{\beta}\,\frac{A\Sigma_{x|y}A^T+1}{A\Sigma_x A^T+1}$

To characterize the values of β for which the optimal solution does not degenerate, we find when the eigenvector solution is optimal. Denote the norm of Σ_x w.r.t. A by $r = \frac{A\Sigma_x A^T}{||A||^2}$. When A is an eigenvector of $\Sigma_{x|y}\Sigma_x^{-1}$, Lemma A.3.2 shows that r is positive and that $A\Sigma_{x|y}\Sigma_x^{-1}\Sigma_x A^T = \lambda r||A||^2$. Rewriting the eigenvalue and isolating $||A||^2$, we have

$$0 < ||A||^2 = \frac{\beta(1-\lambda)-1}{r\lambda} \; . \tag{6.9}$$

This inequality provides a constraint on β and λ that is required for a non-degenerated type of solution

$$\lambda \le \frac{\beta-1}{\beta} \quad \text{or} \quad \beta \ge (1-\lambda)^{-1} \quad , \tag{6.10}$$

thus defining a critical value $\beta^c(\lambda) = (1-\lambda)^{-1}$. For $\beta \le \beta^c(\lambda)$, the weight of compression is so strong that the solution degenerates to zero and no information is carried about X or Y. For $\beta \ge \beta^c(\lambda)$ the weight of information preservation is large enough, and the optimal solution for A is an eigenvector of $\Sigma_{x|y}\Sigma_x^{-1}$. The feasible regions for non degenerated solutions and the norm $||A||^2$ as a function of β and λ are depicted in Figure 6.2.

For some β values, several eigenvectors can satisfy the condition for non degenerated solutions of equation (6.10). Appendix A.3.3 shows that the optimum is achieved by the eigenvector of $\Sigma_{x|y}\Sigma_x^{-1}$ with the smallest eigenvalue. Note that this is also the eigenvector of $\Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx}\Sigma_x^{-1}$ with the largest eigenvalue. We conclude that for scalar projections

$$A(\beta) = \begin{cases} \sqrt{\frac{\beta(1-\lambda)-1}{r\lambda}}v_1 & 0 < \lambda \le \frac{\beta-1}{\beta} \\ 0 & \frac{\beta-1}{\beta} \le \lambda \le 1 \end{cases} \tag{6.11}$$

where v_1 is the eigenvector of $\Sigma_{x|y}\Sigma_x^{-1}$ with the smallest eigenvalue.

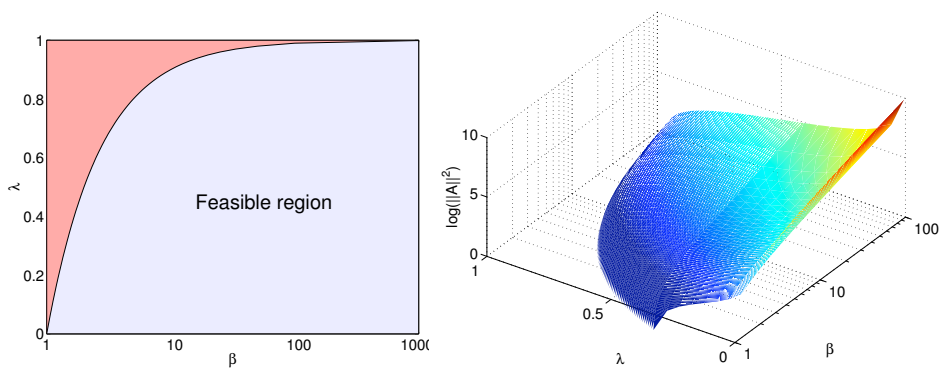


Figure 6.2: **A.** The regions of (β,λ) pairs that lead to the zero (red) and eigenvector (blue) solutions. **B.** The norm $||A||^2$ as a function of β and λ over the feasible region.

6.3.2 The High-Dimensional Case

We now return to the proof of the general, high dimensional case, which follows the same lines as the scalar projection case. Setting the gradient in equation (6.7) to zero and reordering we obtain

$$\frac{\beta-1}{\beta}\left[(A\Sigma_{x|y}A^T+I_d)(A\Sigma_x A^T+I_d)^{-1}\right]A = A\left[\Sigma_{x|y}\Sigma_x^{-1}\right] \ . \tag{6.12}$$

Equation (6.12) shows that the multiplication of $\Sigma_{x|y}\Sigma_x^{-1}$ by A must reside in the span of the rows of A. This means that A should be spanned by up to n_t eigenvectors of $\Sigma_{x|y}\Sigma_x^{-1}$. We can therefore represent the projection A as a mixture $A = WV$ where the rows of V are left normalized eigenvectors of $\Sigma_{x|y}\Sigma_x^{-1}$ and W is a mixing matrix that weights these eigenvectors. The form of the mixing matrix W, that characterizes the norms of these eigenvectors, is described in the following lemma, which is proved in Appendix A.3.4.

Lemma 6.3.1 *The optimum of the cost function is obtained with a diagonal mixing matrix W of the form*

$$W = diag\left[\sqrt{\frac{\beta(1-\lambda_1)-1}{\lambda_1 r_1}};\ldots;\sqrt{\frac{\beta(1-\lambda_k)-1}{\lambda_k r_k}};0;\ldots;0\right] \tag{6.13}$$

where $\{\lambda_1,\ldots,\lambda_k\}$ are $k \leq n_x$ eigenvalues of $\Sigma_{x|y}\Sigma_x^{-1}$ with critical β values ${\beta^c}_1,\ldots,{\beta^c}_k \leq \beta$. $r_i \equiv \mathbf{v}_i^T\Sigma_x\mathbf{v}_i$ as in theorem 6.2.1.

The proof is presented in appendix A.3.4.

We have thus characterized the set of all minima of $\mathcal{L}$, and turn to identify which of them achieve the global minima.

Corollary 6.3.2
The global minimum of $\mathcal{L}$ is obtained with all λ_i that satisfy $\lambda_i < \frac{\beta-1}{\beta}$

The proof is presented in appendix A.3.4.

Taken together, these observations prove that for a given value of β, the optimal projection is obtained by taking all the eigenvectors whose eigenvalues λ_i satisfy $\beta \geq \frac{1}{1-\lambda_i}$, and setting their norm according to $A = WV$ with W determined as in Lemma 6.3.1. This completes the proof of Theorem 6.2.1.

6.4 The GIB Information Curve

The information bottleneck is targeted at characterizing the tradeoff between information preservation (accuracy of relevant predictions) and compression. Interestingly, much of the structure of the problem is reflected in the *information curve*, namely, the maximal value of relevant preserved information (accuracy), $I(T;Y)$, as function of the complexity of the representation of X, measured by $I(T;X)$. This curve is related to the rate-distortion function in lossy source coding, as well as to the achievability limit in source coding with side-information [143, 27]. It was shown to be concave under general conditions [52], but its precise functional form depends on the joint distribution and can reveal properties of the hidden structure of the variables. Analytic forms for the information curve are known only for very special cases, such as Bernoulli variables and some intriguing self-similar distributions. The analytic characterization of the Gaussian IB problem allows us to obtain a closed form expression for the information curve in terms of the relevant eigenvalues.

To this end, we substitute the optimal projection $A(\beta)$ into $I(T;X)$ and $I(T;Y)$ and rewrite them as a function of β

$$\begin{aligned} I_\beta(T;X) &= \frac{1}{2}\log\left(|A\Sigma_x A^T + I_d|\right) \\ &= \frac{1}{2}\log\left(|(\beta(I-D)-I)D^{-1}|\right) \\ &= \frac{1}{2}\sum_{i=1}^{n(\beta)}\log\left((\beta-1)\frac{1-\lambda_i}{\lambda_i}\right) \\ I_\beta(T;Y) &= I(T;X) - \frac{1}{2}\sum_{i=1}^{n(\beta)}\log\beta(1-\lambda_i) \quad , \end{aligned} \tag{6.14}$$

where D is a diagonal matrix whose entries are the eigenvalues of $\Sigma_{x|y}\Sigma_x^{-1}$ as in appendix A.3.4, and $n(\beta)$ is the maximal index i such that $\beta \geq \frac{1}{1-\lambda_i}$. Isolating β as a function of $I_\beta(T;X)$ in the correct range of n_β and then $I_\beta(T;Y)$ as a function of $I_\beta(T;X)$ we have

$$I(T;Y) = I(T;X) - \frac{n_I}{2}\log\left(\prod_{i=1}^{n_I}(1-\lambda_i)^{\frac{1}{n_I}} + e^{\frac{2I(T;X)}{n_I}}\prod_{i=1}^{n_I}\lambda_i{}^{\frac{1}{n_I}}\right) \tag{6.15}$$

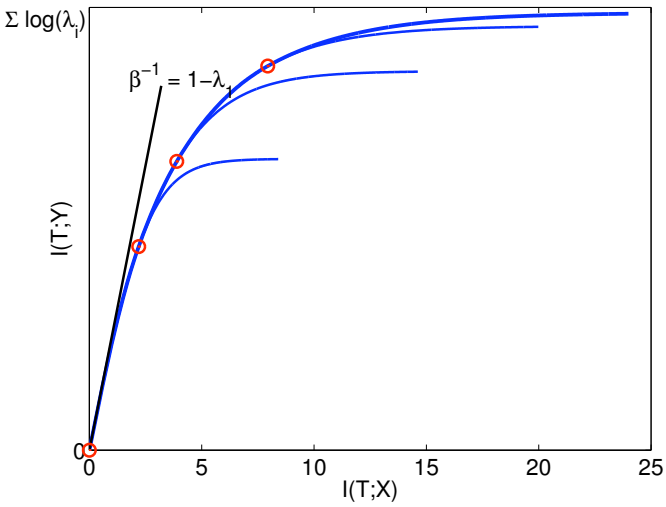


Figure 6.3: GIB information curve obtained with four eigenvalues $\lambda_i = 0.1,0.5,0.7,0.9$. The information at the critical points are designated by circles. For infinite β, curve is saturated at the log of the determinant $\sum \log \lambda_i$. For comparison, information curves calculated with smaller number of eigenvectors are also depicted (all curves calculated for $\beta < 1000$). The slope of the un-normalized curve at each point is the corresponding β^{-1}. The tangent at zero, with slope $\beta^{-1} = 1-\lambda_1$, is super imposed on the information curve.

where the products are over the *first* $n_I = n_{\beta(I(T;X))}$ eigenvalues, since these obey the critical β condition, with $c_{n_I} \leq I(T;X) \leq c_{n_I+1}$ and $c_{n_I} = \sum_{i=1}^{n_I-1} \log \frac{\lambda_{n_I}}{\lambda_i} \frac{1-\lambda_i}{1-\lambda_{n_I}}$.

The GIB curve, illustrated in Figure A.1, is continuous and smooth, but is built of several segments: as $I(T;X)$ increases additional eigenvectors are used in the projection. The derivative of the curve, which is equal to β^{-1}, can be easily shown to be continuous and decreasing, therefore the information curve is concave everywhere, in agreement with the general concavity of information curve in the discrete case [143, 52]. Unlike the discrete case where concavity proofs rely on the ability to use a large number of clusters, concavity is guaranteed here also for segments of the curve, where the number of eigenvectors are limited a-priori.

At each value of $I(T;X)$ the curve is bounded by a tangent with a slope $\beta^{-1}(I(T;X))$. Generally in IB, the data processing inequality yields an upper bound on the slope at the origin, $\beta^{-1}(0) < 1$, in GIB we obtain a tighter bound: $\beta^{-1}(0) < 1 - \lambda_1$. The asymptotic slope of the curve is always zero, as $\beta \to \infty$, reflecting the law of diminishing return: adding more bits to the description of X does not provide higher accuracy about T. This relation between the spectral properties of the covariance matrices raises interesting questions for special cases where the spectrum can be better characterized, such as random-walks and self-similar processes.

6.5 An Iterative Algorithm

The GIB solution is a set of scaled eigenvectors, and as such can be calculated using standard techniques. For example gradient ascent methods were suggested for learning CCA [7, 16]. An alternative approach is to use the general iterative algorithm for IB problems [135]. This algorithm that can be extended to continuous variables and representations, but its practical application for arbitrary distributions leads to a non-linear generalized eigenvalue problem whose general solution can be difficult. It is therefore interesting to explore the form that the iterative algorithm assumes once it is applied to Gaussian variables. Moreover, it may be possible to later extend this approach to more general parametric distributions, such as general exponential forms, for which linear eigenvector methods may no longer be adequate.

The general conditions for the IB stationary points were presented by [135] and can be written for a continuous variable x by the following self consistent equations for the unknown distributions $p(t|x)$, $p(y|t)$ and $p(t)$:

$$\begin{aligned} p(t) &= \int_X dx\, p(x)p(t|x) \\ p(y|t) &= \frac{1}{p(t)} \int_X dx\, p(x,y)p(t|x) \\ p(t|x) &= \frac{p(t)}{Z(\beta)} e^{-\beta D_{KL}[p(y|x)|p(y|t)]} \end{aligned} \tag{6.16}$$

where $Z(\beta)$ is a normalization factor (partition function) and is independent of x. It is important to realize that those conditions assume nothing about the representation variable T and should be satisfied by *any* fixed point of the IB Lagrangian. When X, Y and T have finite cardinality, those equations can be iterated directly in a Blahut-Arimoto like algorithm,

$$\begin{aligned} p(t_{k+1}|x) &= \frac{p(t_k)}{Z_{k+1}(x,\beta)} e^{-\beta D_{KL}[p(y|x)|p(y|t_k)]} \\ p(t_{k+1}) &= \int_X dx\, p(x)p(t_{k+1}|x) \\ p(y|t_{k+1}) &= \frac{1}{p(t_{k+1})} \int_X dx\, p(x,y)p(t_{k+1}|x)\ . \end{aligned} \tag{6.17}$$

where each iteration results in a distribution over the variables T_k, X and Y. The second and third equations calculate $p(t_{k+1})$ and $p(y|t_{k+1})$ using standard marginalization, and the Markov property $Y - X - T_k$. These iterations were shown to converge to the optimal T by [135].

For the general continuous T such an iterative algorithm is clearly not feasible. We show here, how the fact that we are confined to Gaussian distributions, can be used to turn those equations into an efficient parameter updating algorithm. We conjecture that algorithms for parameters optimizations can be defined also for parametric

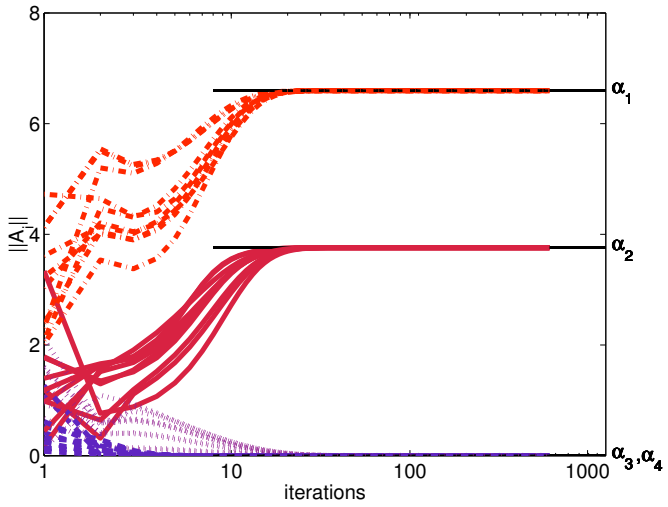


Figure 6.4: The norm of projection on the four eigenvectors of $\Sigma_{x|y}\Sigma_x^{-1}$, as evolves along the operation of the iterative algorithm. Each line corresponds to the length of the projection of one row of A on the closest eigenvector. The projection on the other eigenvectors also vanishes (not shown). β was set to a value that leads to two non vanishing eigenvectors. The algorithm was repeated 10 times with different random initialization points, showing that it converges within 20 steps to the correct values α_i.

distribution other than Gaussians, such as other exponential distributions that can be efficiently represented with a small number of parameters.

In the case of Gaussian $p(x, y)$, when $p(t_k|x)$ is Gaussian for some k, so are $p(t_k)$, $p(y|t_k)$ and $p(t_{k+1}|x)$. In other words, the set of Gaussians $p(t|x)$ is invariant under the above iterations. To see why this is true, notice that $p(y|t_k)$ is Gaussian since T_k is jointly Gaussian with X. Also, $p(t_{k+1}|x)$ is Gaussian since $D_{KL}[p(y|x)|p(y|t_k)]$ between two Gaussians contains only second order moments in y and t and thus its exponential is Gaussian. This is in agreement with the general fact that the optima (which are fixed points of 6.17) are Gaussian [57]. This invariance allows us to turn the IB algorithm that iterates over distributions, into an algorithm that iterates over the parameters of the distributions, being the relevant degrees of freedom in the problem.

Denote the variable T at time k by $T_k = A_k X + \xi_k$, where $\xi_k \sim \mathcal{N}(0, \Sigma_{\xi_k})$. The parameters A and Σ at time $k+1$ can be obtained by substituting T_k in the iterative IB equations. As shown in Appendix A.3.5, this yields the following update equations

$$\begin{aligned} \Sigma_{\xi_{k+1}} &= \left(\beta\Sigma_{t_k|y}^{-1} - (\beta-1)\Sigma_{t_k}^{-1}\right)^{-1} \\ A_{k+1} &= \beta\Sigma_{\xi_{k+1}}\Sigma_{t_k|y}^{-1}A_k\left(I - \Sigma_{y|x}\Sigma_x^{-1}\right) \end{aligned} \tag{6.18}$$

where $\Sigma_{t_k|y}, \Sigma_{t_k}$ are the covariance matrices calculated for the variable T_k.

This algorithm can be interpreted as repeated projection of A_k on the matrix $I - \Sigma_{y|x}\Sigma_x^{-1}$ (whose eigenvectors we seek) followed by scaling with $\beta\Sigma_{\xi_{k+1}}\Sigma_{t_k|y}^{-1}$. It

thus has similar form to the power method for calculating the dominant eigenvectors of the matrix $\Sigma_{y|x}\Sigma_x^{-1}$ [35, 58]. However, unlike the naive power method, where only the single dominant eigenvector is preserved, the GIB iterative algorithm maintains several different eigenvectors, and their number is determined by the continuous parameter β and emerges from the iterations: All eigenvectors whose eigenvalues are smaller than the critical β vanish to zero, while the rest are properly scaled. This is similar to an extension of the naive power method known as *Orthogonal Iteration*, in which the projected vectors are renormalized to maintain several non vanishing vectors [74].

Figure 6.4 demonstrates the operation of the iterative algorithm for a four dimensional X and Y. The tradeoff parameter β was set to a value that leads to two vanishing eigenvectors. The norm of the other two eigenvectors converges to the correct values, which are given in Theorem 6.2.1.

The iterative algorithm can also be interpreted as a regression of X on T via Y. This can be seen by writing the update equation for A_{k+1} as

$$A_{k+1} = \Sigma_{\xi_{k+1}}\Sigma_{t_k|y}^{-1}\left(\Sigma_{yt_k}\Sigma_y^{-1}\right)\left(\Sigma_{yx}\Sigma_x^{-1}\right). \tag{6.19}$$

Since $\Sigma_{yx}\Sigma_x^{-1}$ describes the optimal linear regressor of X on Y, the operation of A_{k+1} on X can be described by the following diagram

$$X \xrightarrow{\Sigma_{yx}\Sigma_x^{-1}} \mu_{y|x} \xrightarrow{\Sigma_{yt_k}\Sigma_y^{-1}} \mu_{t_k|\mu_{y|x}} \xrightarrow{\Sigma_{\xi_{k+1}}\Sigma_{t_k|y}^{-1}} T_{k+1} \tag{6.20}$$

where the last step scales and normalizes T.

6.6 Relation To Other Works

6.6.1 Canonical Correlation Analysis and Imax

The GIB projection derived above uses weighted eigenvectors of the matrix $\Sigma_{x|y}\Sigma_x^{-1} = I - \Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx}\Sigma_x^{-1}$. Such eigenvectors are also used in *Canonical correlations Analysis* (CCA) [69, 134, 15], a method of descriptive statistics that finds linear relations between two variables. Given two variables X, Y, CCA finds a set of basis vectors for each variable, such that the correlation coefficient between the projection of the variables on the basis vectors is maximized. In other words, it finds the bases in which the correlation matrix is diagonal and the correlations on the diagonal are maximized. The bases are the eigenvectors of the matrices $\Sigma_y^{-1}\Sigma_{yx}\Sigma_x^{-1}\Sigma_{xy}$ and $\Sigma_x^{-1}\Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx}$, and the square roots of their corresponding eigenvalues are the *canonical correlation coefficients*. CCA was also shown to be a special case of continuous Imax [8, 7], when the Imax networks are limited to linear projections.

Although GIB and CCA involve the spectral analysis of the same matrices, they have some inherent differences. First of all, GIB characterizes not only the eigenvectors but also their norm, in a way that that depends on the trade-off parameter β. Since

CCA depends on the correlation coefficient between the compressed (projected) versions of X and Y, which is a *normalized* measure of correlation, it is invariant to a rescaling of the projection vectors. In contrast, for any value of β, GIB will choose one particular rescaling given by theorem 6.2.1.

While CCA is symmetric (in the sense that both X and Y are projected), IB is non symmetric and only the X variable is compressed. It is therefore interesting that both GIB and CCA use the same eigenvectors for the projection of X.

6.6.2 Multiterminal Information Theory

The Information Bottleneck formalism was recently shown [52] to be closely related to the problem of source coding with side information [143]. In the latter, two *discrete* variables X, Y are encoded separately at rates R_x, R_y, and the aim is to use them to perfectly reconstruct Y. The bounds on the achievable rates in this case were found in [143] and can be obtained from the IB information curve.

When considering continuous variables, lossless compression at finite rates is no longer possible. Thus, mutual information for continuous variables is no longer interpretable in terms of the actual number of encoding bits, but rather serves as an optimal measure of information between variables. The IB formalism, although coinciding with coding theorems in the discrete case, is more general in the sense that it reflects the tradeoff between compression and information preservation, and is not concerned with exact reconstruction.

Lossy reconstruction can be considered by introducing distortion measures as done for source coding of Gaussians with side information by [144] and by [13] [see also 111], but these focus on the region of achievable rates under constrained distortion and are not relevant for the question of finding the representations which capture the information between the variables. Among these, the formalism closest to ours is that of [13] where the distortion in reconstructing X is assumed to be small (high-resolution scenario). However, their results refer to encoding rates and as such go to infinity as the distortion goes to zero. They also analyze the problem for scalar Gaussian variables, but the one-dimensional setting does not reveal the interesting spectral properties and phase transitions which appear only in the multidimensional case discussed here.

6.6.3 Gaussian IB with Side Information

When handling real world data, the relevance variable Y often contains multiple structures that are correlated to X, although many of them are actually irrelevant. The information bottleneck with side information ($IBSI$) [25] alleviates this problem using side information in the form of an *irrelevance* variable Y^- about which information is removed. $IBSI$ thus aims to minimize

$$\mathcal{L} = I(X;T) - \beta \left(I(T;Y^+) - \gamma I(T;Y^-) \right) \tag{6.21}$$

This formulation can also be extended to the Gaussian case, in a manner similar to the original GIB functional. Looking at its derivative w.r.t. to the projection A yields

$$\begin{aligned}\frac{\delta\mathcal{L}}{\delta A} = (\quad & 1-\beta+\beta\gamma \quad)(A\Sigma_x A^T + I_d)^{-1} 2A\Sigma_x \\ + \quad & \beta \qquad\quad (A\Sigma_{x|y^+} A^T + I_d)^{-1} 2A\Sigma_{x|y^+} \\ - \quad & \beta\gamma \qquad\quad (A\Sigma_{x|y^-} A^T + I_d)^{-1} 2A\Sigma_{x|y^-} \ .\end{aligned}$$

While GIB relates to an eigenvalue problem of the form $\lambda A = A\Sigma_{x|y}\Sigma_x^{-1}$, GIB with side information ($GIBSI$) requires to solve of a matrix equation of the form $\lambda' A + \lambda^+ A\Sigma_{x|y^+}\Sigma_x^{-1} = \lambda^- A\Sigma_{x|y^-}\Sigma_x^{-1}$, which is similar in form to a generalized eigenvalue problem. However, unlike standard generalized eigenvalue problems, but as in the GIB case analyzed in this chapter, the eigenvalues themselves depend on the projection A.

6.7 Practical Implications

The GIB approach can be viewed as a method for finding the best linear projection of X, under a constraint on $I(T;X)$. Another straightforward way to limit the complexity of the projection is to specify its dimension in advance. Such an approach leaves open the question of the relative weighting of the resulting eigenvectors. This is the approach taken in classical CCA, where the number of eigenvectors is determined according to a statistical significance test, and their weights are then set to $\sqrt{1-\lambda_i}$. This expression is the correlation coefficient between the i^{th} CCA projections on X and Y, and reflects the amount of correlation captured by the i^{th} projection. The GIB weighting scheme is different, since it is derived to preserve maximum information under the compression constraint. To illustrate the difference, consider the case where $\beta = \frac{1}{1-\lambda_3}$, so that only two eigenvectors are used by GIB. The CCA scaling in this case is $\sqrt{1-\lambda_1}$, and $\sqrt{1-\lambda_2}$. The GIB weights are (up to a constant) $\alpha_1 = \sqrt{\frac{\lambda_3-\lambda_1}{\lambda_1 r_1}}, \alpha_2 = \sqrt{\frac{\lambda_3-\lambda_2}{\lambda_2 r_2}}$,, which emphasizes large gaps in the eigenspectrum, and can be very different from the CCA scaling.

This difference between CCA scaling and GIB scaling may have implications on two aspects of learning in practical applications. First, in applications involving compression of Gaussian signals due to limitation on available band-width. This is the case in the growing field of sensor networks in which sensors are often very limited in their communication bandwidth due to energy constraints. In these networks, sensors communicate with other sensors and transmit information about their local measurements. For example, sensors can be used to monitor chemicals' concentrations, temperature or light conditions. Since only few bits can be transmitted, the information has to be compressed in a relevant way, and the relative scaling of the different eigenvectors becomes important [as in transform coding 62]. As shown above, GIB describes the optimal transformation of the raw data into information conserving representation.

The second aspect where GIB becomes useful is in interpretation of data. Today, canonical correlation analysis is widely used for finding relations between multi-variate continuous variables, in particular in domains which are inherently high dimensional such as meteorology [139] chemometrics [3] and functional MRI of brains [45]. Since GIB weights the eigenvectors of the normalized cross correlation matrix in a different way than CCA, it may lead to very different interpretation of the relative importance of factors in these studies.

6.8 Discussion

We applied the information bottleneck method to continuous jointly Gaussian variables X and Y, with a continuous representation of the compressed variable T. We derived an analytic optimal solution as well as a new general algorithm for this problem (GIB) which is based solely on the spectral properties of the covariance matrices in the problem. The solutions for GIB are characterized in terms of the trade-off parameter β between compression and preserved relevant information, and consist of eigenvectors of the matrix $\Sigma_{x|y}\Sigma_x^{-1}$, continuously adding up vectors as more complex models are allowed. We provide an analytic characterization of the optimal tradeoff between the representation complexity and accuracy - the "information curve" - which relates the spectrum to relevant information in an intriguing manner. Besides its clean analytic structure, GIB offers a way for analyzing empirical multivariate data when only its correlation matrices can be estimated. In that case it extends and provides new information theoretic insight to the classical Canonical Correlation Analysis.

The most intriguing aspect of GIB is in the way the dimensionality of the representation changes with increasing complexity and accuracy, through the continuous value of the trade-off parameter β. While both mutual information values vary continuously on the smooth information curve, the dimensionality of the optimal projection T increases discontinuously through a cascade of structural (second order) phase transitions, and the optimal curve moves from one analytic segment to another. While this transition cascade is similar to the bifurcations observed in the application of IB to clustering through deterministic annealing, this is the first time such dimensional transitions are shown to exist in this context. The ability to deal with all possible dimensions in a single algorithm is a novel advantage of this approach compared to similar linear statistical techniques as CCA and other regression and association methods.

Interestingly, we show how the general IB algorithm which iterates over distributions, can be transformed to an algorithm that performs iterations over the distributions' *parameters*. This algorithm, similar to multi-eigenvector power methods, converges to a solution in which the number of eigenvectors is determined by the parameter β, in a way that emerges from the iterations rather than defined a-priori.

For multinomial variables, the IB framework can be shown to be related in some

limiting cases to maximum-likelihood estimation in a latent variable model [130]. It would be interesting to see whether the GIB-CCA equivalence can be extended and give a more general understanding of the relation between IB and statistical latent variable models.

While the restriction to a Gaussian joint distribution deviates from the more general distribution independent approach of IB, it provides a precise example to the way representations with different dimensions can appear in the more general case. We believe that this type of dimensionality-transitions appears for more general distributions, as can be revealed in some cases by applying the Laplace method of integration (a Gaussian approximation) to the integrals in the general IB algorithm for continuous T.

The more general exponential forms, can be considered as a kernelized version of IB [see 92] and appear in other minimum-information methods [such as SDR, 56]. these are of particular interest here, as they behave like Gaussian distributions in the joint kernel space. The Kernel Fisher-matrix in this case will take the role of the original cross covariance matrix of the variables in GIB.

Another interesting extension of our work is to networks of Gaussian processes. A general framework for that problem was developed in [44] and applied for discrete variables. In this framework the mutual information is replaced by multi-information, and the dependencies of the compressed and relevance variables are specified through two Graphical models. It is interesting to explore the effects of dimensionality changes in this more general framework, to study how they induce topological transitions in the related graphical models, as some edges of the graphs become important only beyond corresponding critical values of the tradeoff parameter β.

Chapter 7

Discussion and Concluding Remarks

Information theory offers an elegant and systematic approach to analyzing the information one variable conveys about the other. As such, it is not surprising that it has made a significant impact on fields like machine learning and neural coding. However, due to its very general scope, it has not been straightforward to turn it into a practical tool in these fields. The current dissertation advances the use of information theory in these fields by combining its use with the notion of partial measurements. In the previous chapters we presented several methods which show how information may be efficiently measured in various scenarios that have not been addressed by existing methods. This idea was shown to result in a novel classification algorithm, and to aid in the analysis of the neural code. The SDR method, which combines maximization and minimization of information, was shown to yield excellent practical results in extracting useful features from large databases. Finally, by analyzing the information theoretic tradeoff between complexity and accuracy via the Gaussian Information Bottleneck, we have obtained insight into the emergence of dimensionality in the representation of data.

It will be interesting to see if the insight obtained in the GIB analysis may be transferred to SDR-like methods. Since Gaussian distributions have an exponential form, as do the SDR models, it seems like there could be a formalism which incorporates these two approaches. However, since SDR is not a clustering method like the Information Bottleneck, this would require a novel information theoretic notion of compression in the SDR case, which is yet to be found. One option could involve the number of bits needed to specify the SDR function, or a related measure of complexity.

The MinMI method is specifically designed to handle multivariate scenarios (e.g., a population of neurons), and approximate algorithms were presented for such cases. The SDR method, on the other hand, assumes that the size of X is small enough, since $O(|X|)$ computational resources are needed to run it. This results from focusing on features $\vec{\phi}(x)$ which depend on *all* of X. SDR can be modified to look for features

of subsets of X, i.e. $\phi(x_i)$ or $\phi(x_i, x_j)$. This should allow practical algorithms for the multivariate case, and will make the method applicable to tasks such as language modeling [89].

The methods presented here have used the mutual information between two variables X and Y[1]. While this partition into two variables is appropriate in many cases, one may sometimes be interested in the dependence between larger sets of variables. The measure which generalizes mutual information to such cases is known as the multi-information [131]. It will be interesting to consider generalizations of the methods presented here to this more general case. These could potentially be used to study information flow in networks, and to find features of the network that govern this flow. A relevant recent work in that respect is the extension of the Information Bottleneck method to multi-information [44]. We expect some of the concepts introduced there to be applicable to the methods in this thesis.

MinMI was shown to apply to a wide range of issues in neural coding, from single and pairwise coding, to coding in the temporal domain. The long term goal of such a method is of course to aid in discovering new neural coding mechanisms. We briefly mention a few areas of research where the application of MinMI may be particularly appropriate.

- Redundancy reduction - The concept of redundancy reduction in the brain was introduced by Barlow [4] as a possible cortical design principle, and has been an influential paradigm in brain research. A recent work [23] quantified redundancy along the auditory pathway, and has found that it is indeed reduced. As illustrated in Chapter 3, MinMI can be used to quantify the difference in redundancy between populations, and is thus an attractive tool in studying such problems. The measure used in [23] assumed conditional independence between neurons. Since MinMI does not make any such assumptions, it may result in a more sensitive measure. This should especially be true for large populations where the conditional independence assumption results in a saturation of the information values.

- Pairwise coding in populations - The study of coding via correlations in cells has been a very active field of research in neuroscience [100, 65, 137]. However, most studies focused on analysis of isolated pairs, without taking into account the relations between pairwise codes in the population (see [113, 104] for approaches which do address this case). MinMI appears to be an appropriate tool for measuring information in this scenario, and should be a valuable instrument for addressing this problem in analyzing experimental data.

- Temporal Coding - As shown in Chapter 3, MinMI can find a significant number

[1] Note that this does not limit X itself from being multivariate.

of cells which use response profiles to encode stimulus values. This result has already been demonstrated on neuro-physiological data from different domains and cortical areas (ongoing work, not shown here). One immediate implication of this analysis is that it increases the size of the population of neurons which participate in coding, and could thus help in uncovering phenomena in which a relatively small number of neurons participate. Furthermore, we expect it will be helpful in understanding what time constants in the neuronal responses are relevant for decoding stimulus values.

- Precise firing events - There is an ongoing debate in neuroscience regarding the importance of precise firing times, and correlation between neurons in small time windows (see e.g., [1, 31]). While we did not address this issue in the neural coding applications, MinMI could be adapted to this case by considering partial measurements related to precise timing (such as the mean number of coincident spikes in a short time window). This, as in the cases mentioned above, will help in studying this property when embedded in a population of cells.

As noted earlier, the MinMI method is closely related to the classical Rate Distortion Theory. However, MinMI minimizes information subject to a set of expectation constraints on a distribution, as opposed to the distortion constraints in rate distortion theory. Thus, although some results may carry over from the information theoretic literature, several new and interesting problems arise. One example is the algorithmic challenges which one faces due to the possibly exponentials size of the X variable. It will be interesting to study the information theoretic implications of MinMI. In other words, find a communication problem to which MinMI is the answer.

The methods throughout this thesis made frequent use of concepts from both the machine learning literature (such as classification error and feature extraction) and from information theory. While this is not the first use of information theory in machine learning, it does provide several novel connections between these two fields, among which are an information theoretic interpretation of dimension in feature extraction (as in GIB), and of matrix factorization (as in SDR). We would like to hope this will stimulate research in the interface between these two important fields, in a mutually beneficial manner.

Appendix A

Proofs

In the current appendix, we provide proofs of various propositions given in the text.

A.1 MinMI Results

A.1.1 Convergence Proof for the MinMI Algorithm

Since $p_{t+1}(x|y)$ is the projection of $p_t(x)$ on $\mathcal{F}(\vec{\phi}(x), \vec{a}(y))$ and $p_t(x|y)$ is also in $\mathcal{F}(\vec{\phi}(x), \vec{a}(y))$, by the definition of the previous iteration, we have by the Pythagorean equality for *I-projections* (see Section 1.4 and [30]) that

$$\begin{aligned} D_{KL}[p_t(x|y)|p_t(x)] &= D_{KL}[p_t(x|y)|p_{t+1}(x|y)] \\ &\quad + D_{KL}[p_{t+1}(x|y)|p_t(x)] \ . \end{aligned}$$

Averaging the above over $\bar{p}(y)$ and rearranging

$$\begin{aligned} I[p_t(x,y)] &= \langle D_{KL}[p_t(x|y)|p_{t+1}(x|y)]\rangle_{\bar{p}(y)} + \\ &\quad + I[p_{t+1}(x,y)] + D_{KL}[p_{t+1}(x)|p_t(x)] \ . \end{aligned}$$

The right hand side has $I[p_{t+1}(x,y)]$ plus some positive quantity. We can thus conclude that

$$I[p_{t+1}(x,y)] \leq I[p_t(x,y)] \ , \tag{A.1}$$

and therefore the algorithm reduces the mutual information in each iteration.

To see that the algorithm indeed converges to the minimum, note that since $I[p_t(x,y)]$ is a monotonous, lower bounded, decreasing series, its difference series converges to zero as t goes to infinity

$$\langle D_{KL}[p_t(x|y)|p_{t+1}(x|y)]\rangle_{\bar{p}(y)} + D_{KL}[p_{t+1}(x)|p_t(x)] \to 0 \ .$$

The above expression is zero if and only if $p_t(x|y) = p_{t+1}(x|y)$ and $p_t(x) = p_{t+1}(x)$. This implies that at the limit $p_\infty(x)$ and $p_\infty(x|y)$ are invariant to the MinMI iterations, and thus the fixed point equations characterizing the minimum of the information in Equation 2.6 are satisfied by $p_\infty(x,y)$ so that $p_\infty(x,y) = p_{MI}(x,y)$

A.1.2 Convex Duality

Convex optimization problems have a special role in the optimization literature, due to the effectiveness with which they can be solved (see [26] for details). Additionally, convex duality theorems state that every convex problem has a dual problem with an identical optimum. We shall use this property in what follows. Below we give a very brief introduction to convex duality. We begin with some definitions

Definition 1 *A function $f(x)$ is convex if for every x_1, x_2 in its domain and $0 \leq \lambda \leq 1$*

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2) \ . \tag{A.2}$$

A convex optimization problem is a constrained optimization problem where the optimized function is convex, and constrained functions are also convex [1]. Formally, let x be a variable in $\Re^n$, $f_0(x), \ldots, f_m(x)$ a set of convex functions, A a $p \times n$ matrix, and b a vector of size p. We consider the following constrained minimization problem

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0 \quad i = 1 \ldots m \\ & Ax = b \end{array} \ , \tag{A.3}$$

A common method for approaching constrained optimization is by constructing the Lagrangian function, defined by

$$\mathcal{L}(x, \lambda, \nu) = f_0(x) + \sum_i \lambda_i f_i(x) + \nu^T (Ax - b) \ , \tag{A.4}$$

where $\lambda_i \geq 0, \nu_i$ are parameters referred to as *Lagrange multipliers*. The vectors λ, ν are referred to as the *dual variables* associated with problem A.3. It can be shown that the set of values (x^*, λ^*, ν^*) minimizing $\mathcal{L}(x, \lambda, \nu)$ yield the optimal x^*.

Following [26] (page 216) we define the Lagrange dual function $g : \Re^m, \Re^p \rightarrow \Re$ as the minimum value of the Lagrangian over x for a given value of the multipliers λ, ν

$$g(\lambda, \nu) = \inf_x \mathcal{L}(x, \lambda, \nu) \ . \tag{A.5}$$

It can be shown that $g(\lambda, \nu)$ lower bounds the value of $f_0(x)$ for any feasible x, and $\lambda \geq 0$. Thus maximization of $g(\lambda, \nu)$ is guaranteed to give a lower bound on the optimum of problem A.3. In fact, in most cases the maximum of $g(\lambda, \nu)$ achieves the minimum of problem A.3. This is known as *strong duality*. A simple condition which guarantees strong duality is *Slater's condition*. It states that if there exists a *strictly feasible* value of x, i.e. $f_i(x) < 0 \quad i = 1 \ldots m$ (note the strict inequality) then strong duality holds. Formally, define c^* as the minimum value of problem A.3, then under Slater's condition the following holds:

$$\max_{\lambda \geq 0, \nu} g(\lambda, \nu) = c^* \ . \tag{A.6}$$

[1] More generally, these are problems where the constraints are given by a convex set, but here, as in [26], we focus on this more restricted setting

The problem $\max_{\lambda \geq 0, \nu} g(\lambda, \nu)$ is known as the *dual problem* of problem A.3. This is a very useful result since often the dual problem is more easily solved than the primal one. Furthermore, the dual problem may yield an insight into the solution of the primal problem, as will be shown later.

Although the dual problem seems to be unconstrained (aside from positivity of λ), it is often written as a constrained problem, as we now explain. In many cases, the function $g(\lambda, \nu)$ takes the value $-\infty$ for a set of dual parameter values. Thus it can be written as

$$g(\lambda, \nu) = \begin{cases} \bar{g}(\lambda, \nu) & (\lambda, \nu) \in \mathcal{F}_{dual} \\ -\infty & (\lambda, \nu) \notin \mathcal{F}_{dual} \end{cases} , \tag{A.7}$$

where $\tilde{g}(\lambda, \nu)$ are the non infinite minima, and $\mathcal{F}_{dual}$ is a set in the domain of (λ, ν). The dual problem can then be written as

$$\begin{array}{ll} \text{maximize} & \tilde{g}(\lambda, \nu) \\ \text{subject to} & (\lambda, \nu) \in \mathcal{F}_{dual} \end{array} . \tag{A.8}$$

The set $\mathcal{F}_{dual}$ is often convex, and thus the dual problem is also a constrained convex optimization problem.

Finally, we mention another useful result, which relates the primal and dual optimal assignments, rather than their optimal values. The dual parameters λ^*, ν^* which maximize the dual problem can be shown to be those which minimize the Lagrangian in Equation A.4 (see [26] page 248). Thus, if the primal optimum x^* is characterized in terms of its Lagrange multipliers (as in the Maximum Entropy case), then the dual solution can be used to obtain it [2].

A.1.3 Convex Duality for MinMI

Our convex duality proof is different from standard duality derivations, due to some non-trivial properties of the current problem. We borrow on ideas used in Gallager's (see [46], page 462), and Chiang and Boyd [26].

The Lagrangian for the primal problem is

$$\begin{aligned} \mathcal{L}(p, \vec{\psi}(y), \gamma(y), \lambda_{xy}) &= I[\hat{p}(x,y)] - \sum_y \bar{p}(y)\vec{\psi}(y) \cdot \left(\sum_x \hat{p}(x|y)\vec{\phi}(x) - \vec{a}(y) \right) \\ &\quad - \sum_y \bar{p}(y)\gamma(y) \left(\sum_x \hat{p}(x|y) - 1 \right) - \sum_{x,y} \bar{p}(y)\lambda_{xy}\hat{p}(x|y) , \end{aligned}$$

where we multiplied all dual parameters (or Lagrange multipliers) by $\bar{p}(y)$ for convenience. The dual function is defined as:

$$g(\vec{\psi}(y), \gamma(y), \lambda_{xy}) = \min_{\hat{p}(x|y)} \mathcal{L}(p, \vec{\psi}(y), \gamma(y), \lambda_{xy}) . \tag{A.9}$$

[2] This theorem requires some additional technical details such as strict convexity and strong duality, which hold in the cases we consider.

Deriving w.r.t $\hat{p}(x|y)$ and equating to zero we obtain

$$\log \hat{p}(x|y) - \log \hat{p}(x) - \vec{\phi}(x) \cdot \vec{\psi}(y) - \gamma(y) - \lambda_{xy} = 0 \ . \tag{A.10}$$

Thus the minimizing $\hat{p}(x|y)$ depends on the dual parameters in the following way

$$\hat{p}(x|y) = \hat{p}(x) e^{\vec{\phi}(x) \cdot \vec{\psi}(y) + \gamma(y) + \lambda_{xy}} \ . \tag{A.11}$$

Plugging this back in to get $g(\lambda)$ we have

$$\begin{aligned} g(\vec{\psi}(y), \gamma(y), \lambda_{xy}) &= \sum_y \bar{p}(y) \sum_x \hat{p}(x|y) \left(\vec{\phi}(x) \cdot \vec{\psi}(y) + \gamma(y) + \lambda_{xy} \right) \\ &\quad - \sum_y \bar{p}(y) \vec{\psi}(y) \cdot \left(\sum_x \hat{p}(x|y) \vec{\phi}(x) - \vec{a}(y) \right) \\ &\quad - \sum_y \bar{p}(y) \gamma(y) \left(\sum_x \hat{p}(x,y) - 1 \right) - \sum_{x,y} \bar{p}(y) \lambda_{xy} \hat{p}(x,y) \\ &= \sum_y \bar{p}(y) \left(\vec{\psi}(y) \cdot \vec{a}(y) + \gamma(y) \right) \ . \end{aligned}$$

Although it seems as if we are done, this is not the case. Note that Equation A.11 is not a closed form solution for $\hat{p}(x|y)$ in the sense that its right hand side contains $\hat{p}(x) = \sum_y \hat{p}(x|y)\bar{p}(y)$ which depends on $\hat{p}(x|y)$. Thus, given the values of the dual parameters, we cannot calculate $\hat{p}(x|y)$. This should lead us to suspect that the value we obtained for $g(\vec{\psi}(y), \gamma(y), \lambda_{xy})$ may not be valid for all values of dual parameters, because in some cases the minimizing $\hat{p}(x|y)$ will be infinite and will not be revealed when setting the derivative to zero. While these infinite minimizers cannot be the solution to the primal problem (since they do not satisfy the normalization condition), we need to make sure we do not consider their corresponding dual parameters when maximizing $g(\vec{\psi}(y), \gamma(y), \lambda_{xy})$. We call these parameters *bad*, and those yielding a finite minimizer *good*. It is easy to see that such bad parameters exist: multiply Equation A.11 by $\bar{p}(y)$ and sum over all values of y. This yields

$$\hat{p}(x) = \hat{p}(x) \sum_y \bar{p}(y) e^{\vec{\psi}(y)\vec{\phi}(x) + \gamma(y) + \lambda_{xy}} \ . \tag{A.12}$$

Now, *assuming* $\hat{p}(x) \neq 0$ (this is not a valid assumption but we use it to illustrate the point) we may divide by it to obtain

$$\sum_y \bar{p}(y) e^{\vec{\psi}(y)\vec{\phi}(x) + \gamma(y) + \lambda_{xy}} = 1 \ , \tag{A.13}$$

which, by the non-negativity of λ_{xy} becomes

$$\sum_y \bar{p}(y) e^{\vec{\psi}(y)\vec{\phi}(x) + \gamma(y)} \leq 1 \ . \tag{A.14}$$

Thus we have a necessary condition characterizing the good dual parameters. This statement however is not yet precise, since we have ignored the $\hat{p}(x) = 0$ case. Our strategy in what follows is: find a set which *contains* all good dual parameters (Proposition 1), although it may also contain some bad ones. Then show that the value of $g(\vec{\psi}(y), \gamma(y), \lambda_{xy})$ for dual parameters in this set still lower bounds the primal problem (Proposition 2) , and therefore we lose nothing by considering the extra bad values.

We begin with a necessary condition for a set dual parameters to be good.

Proposition 1 *If a dual parameter set is good, then* $\sum_y \bar{p}(y) e^{\vec{\psi}(y)\vec{\phi}(x)+\gamma(y)} \leq 1$ *for all* x *values.*

Proof: *The proof follows the arguments used in Gallager's proof (see [46], page 462) in analyzing the rate distortion function. Denote by* $p(x|y)$ *the minimizer of Equation A.9 for a given set of dual parameters. If there is no* x *such that* $p(x) = 0$*, the argument in the previous paragraph holds, and Equation A.14 then shows that the required condition holds. Now, assume that there exist an* $\hat{x}$ *such that* $p(\hat{x}) = 0$*. We will show that the Equation A.14 holds for this* $\hat{x}$*. For every* y *there must exist* $\hat{x}(y)$ *such that* $p(\hat{x}(y)|y)$ *is finite, and thus achieves a zero gradient.*

We now define a new distribution $p\prime(x|y)$ *using the minimizer* $p(x|y)$*. For a* small $\epsilon > 0$ *define*

$$\begin{aligned} p\prime(\hat{x}|y) &= \epsilon \cdot e^{\vec{\psi}(y)\vec{\phi}(x)+\gamma(y)+\lambda_{\hat{x}(y)y}} \quad \forall y \\ p\prime(\hat{x}(y)|y) &= p(\hat{x}(y)|y) - p\prime(\hat{x}|y) \quad \forall y \\ p\prime(x|y) &= p(x|y) \quad \forall y, x \neq \hat{x}(y) \ . \end{aligned}$$

If ϵ *is small enough then the above will be a valid conditional distribution for all* y*. Define* $\mathcal{L}_x(\vec{\psi}(y), \gamma(y), \lambda_{xy})$ *as the part of the Lagrangian that depends on* x*, namely (we ignore factors which do not depend on* $p(x|y)$ *and suppress the dependency on the dual parameters for brevity)*

$$\mathcal{L}_x(p) = \sum_y \bar{p}(y) p(x|y) \left[\log \frac{p(x|y)}{p(x)} - \vec{\phi}(x) \cdot \vec{\psi}(y) - \gamma(y) - \lambda_{xy} \right] \ . \tag{A.15}$$

Note that $\mathcal{L}(p) = \sum_x \mathcal{L}_x(p)$ *(again, up to constants which do not depend on* $p(x|y)$*).*

Then

$$\mathcal{L}(p\prime) - \mathcal{L}(p) = \mathcal{L}_{\hat{x}}(p\prime) - \mathcal{L}_{\hat{x}}(p) + \sum_{x \neq \hat{x}} [\mathcal{L}_x(p\prime) - \mathcal{L}_x(p)] = \mathcal{L}_{\hat{x}}(p\prime) + \sum_{x \neq \hat{x}} [\mathcal{L}_x(p\prime) - \mathcal{L}_x(p)] \ ,$$

where we have used the fact that $p(\hat{x}|y)$ *is zero for all* y *since* $p(\hat{x}) = 0$*. Since* $\frac{\partial \mathcal{L}_x(p)}{\partial p(\hat{x}(y)|y)} = 0$*, the sum over* $x \neq \hat{x}$ *in Equation A.16 has no first-order variation in* ϵ*, so that to first order in* ϵ

$$\begin{aligned}
\mathcal{L}(p\prime) - \mathcal{L}(p) &= \sum_y \bar{p}(y) p(\hat{x}|y) \log \frac{\epsilon}{\sum_y \bar{p}(y) p\prime(\hat{x}|y)} \\
&= \sum_y \bar{p}(y) p(\hat{x}|y) \log \frac{1}{\sum_{\hat{y}} \bar{p}(\hat{y}) e^{\vec{\psi}(\hat{y})\vec{\phi}(x)+\gamma(\hat{y})}} \\
&= -c \cdot \log \sum_{\hat{y}} \bar{p}(\hat{y}) e^{\vec{\psi}(\hat{y})\vec{\phi}(x)+\gamma(\hat{y})} \ ,
\end{aligned}$$

where $c \equiv \sum_y \bar{p}(y) p(\hat{x}|y)$ *is some positive constant. Since* $p(x|y)$ *minimizes* $\mathcal{L}(p)$*, the above expression must be negative, implying that*

$$\sum_{\hat{y}} \bar{p}(\hat{y}) e^{\vec{\psi}(\hat{y})\vec{\phi}(x)+\gamma(\hat{y})} \leq 1 \ . \tag{A.16}$$

This is the condition we set out to prove.

The above proposition implies that we lose nothing by restricting the set of dual parameters to those satisfying the given constraints. That is, the optimal set of dual parameters is contained in this constrained set. However, it may still be that a set of parameters satisfies the above constraint, but is not a valid minimizer of Equation A.9. Although this may be possible, the following proposition implies that the value of g in that case is always below the optimal value of the primal problem, and thus duality is conserved.

Proposition 2 *For every primal feasible distribution* $\hat{p}(x|y) \in \mathcal{F}(\vec{\phi}(x), \vec{a}(y))$ *and set of dual parameters which satisfy the conditions in Proposition 1, the following holds:* $I[p] \geq g(\vec{\psi}(y), \gamma(y), \lambda_{xy})$

Proof: *We look at the difference* $I[p] - g(\vec{\psi}(y), \gamma(y), \lambda_{xy})$

$$\begin{aligned}
I[p] - g(\vec{\psi}(y), \gamma(y), \lambda_{xy}) &= \sum_{x,y} \bar{p}(y)\hat{p}(x|y) \log \frac{\hat{p}(x|y)}{\hat{p}(x)} - \sum_y (\vec{\psi}(y)\vec{a}(y) + \gamma(y)) \\
&= \sum_{x,y} \bar{p}(y)\hat{p}(x|y) \log \frac{\hat{p}(x|y)}{\hat{p}(x)} - \sum_y (\vec{\psi}(y) \sum_x \vec{\phi}(x)\hat{p}(x|y) + \gamma(y)) \\
&= \sum_{x,y} \bar{p}(y)\hat{p}(x|y) \log \frac{\hat{p}(x|y)}{\hat{p}(x)} - \sum_{x,y} \bar{p}(y)\hat{p}(x|y)(\vec{\psi}(y)\vec{\phi}(x) + \gamma(y)) \\
&= \sum_{x,y} p(y)p(x|y) \log \frac{p(x|y)}{p(x)e^{\vec{\psi}(y)\vec{\phi}(x)+\gamma(y)}} \ .
\end{aligned}$$

Now use the inequality $\log x \leq x - 1$*.*

$$\begin{aligned}
g(\vec{\psi}(y), \gamma(y), \lambda_{xy}) - I[p] &= \sum_{x,y} \bar{p}(y)\hat{p}(x|y) \log \frac{\hat{p}(x)e^{\vec{\psi}(y)\vec{\phi}(x)+\gamma(y)}}{\hat{p}(x|y)} \\
&\leq \sum_{x,y} \bar{p}(y)(\hat{p}(x)e^{\vec{\psi}(y)\vec{\phi}(x)+\gamma(y)} - 1) \\
&= 0 \ .
\end{aligned}$$

Putting it all together, we have the following: Any dual parameter set which satisfies the constraints in Proposition 1 lower bounds $I[p]$ for any feasible primal parameters. Furthermore the optimal dual parameters satisfy the constraints of Proposition 1 (since they are minimizers of Equation A.9). Since convex duality guarantees that the maximum of $g(\vec{\psi}(y), \gamma(y), \lambda_{xy})$ coincides with the minimum of the primal problem, we have the desired result [3].

MinMI with Inequality Expectation Constraints

Here we briefly cover some results for the MinMI problem where the expectations may lie in the range $\vec{a}(y) \pm \vec{\beta}(y)$. The Lagrangian in this case needs to include two multipliers: $\vec{\psi}^+(y)$ for the inequality $\langle \vec{\phi}(x) \rangle_{\hat{p}(x|y)} \leq \vec{a}(y) + \vec{\beta}(y)$ and $\vec{\psi}^-(y)$ for the inequality $\langle \vec{\phi}(x) \rangle_{\hat{p}(x|y)} \geq \vec{a}(y) - \vec{\beta}(y)$.

The Lagrangian is then (after some algebra)

$$\begin{aligned}
\mathcal{L}(p, \vec{\psi}(y), \gamma(y), \lambda_{xy}) &= I[\hat{p}(x,y)] + \sum_y \bar{p}(y)(\vec{\psi}^+(y) - \vec{\psi}^-(y)) \cdot \left(\sum_x \hat{p}(x|y)\vec{\phi}(x) - \vec{a}(y) \right) \\
&\quad - \sum_y \vec{\beta}(y) \cdot (\vec{\psi}^+(y) + \vec{\psi}^-(y)) - \sum_y \bar{p}(y)\gamma(y) \left(\sum_x \hat{p}(x|y) - 1 \right) \\
&\quad - \sum_{x,y} \bar{p}(y)\lambda_{xy}\hat{p}(x|y) \ ,
\end{aligned}$$

Deriving this w.r.t. $\hat{p}(x|y)$ and defining $\vec{\psi}(y) \equiv \vec{\psi}^+(y) - \vec{\psi}^-(y)$ yields the same form of solution as in the standard MinMI case (although with different constraints).

To derive the dual, we substitute the minimizing $\hat{p}(x|y)$ into the Lagrangian, yielding

$$g(\vec{\psi}^+(y), \vec{\psi}^-(y), \gamma(y), \lambda_{xy}) = \sum_y \bar{p}(y) \left((\vec{\psi}^+(y) - \vec{\psi}^-(y)) \cdot \vec{a}(y) + \gamma(y) - \vec{\beta}(y) \cdot (\vec{\psi}^+(y) + \vec{\psi}^-(y)) \right)$$

To simplify this function, we use a trick from [41]. We claim that the maximum of g is obtained with at most one of the multipliers $\vec{\psi}^+(y), \vec{\psi}^-(y)$ being non zero for all values of y. To see this, assume we have a y where both are non zero. Then we can decrease them both by the same amount, until the lower one becomes zero. We have not changed the term depending on $\vec{\psi}^+(y) - \vec{\psi}^-(y)$, but have decreased the one depending on $\vec{\psi}^+(y) + \vec{\psi}^-(y)$ thus yielding a higher value of g [4].

[3] This of course is true only when Slater's condition holds. As long as the expected values are obtained from some empirical data, Slater's condition will hold, as explained in the footnote in Section 1.4.1 regarding duality in MaxEnt

[4] To make this argument exact, one needs to show that this decrease in values does not change the dual feasibility of the parameters. This is in fact true since the dual constraints can be shown only to depend on $\vec{\psi}^+(y) - \vec{\psi}^-(y)$

Using the above insight and defining $\vec{\psi}(y) \equiv \vec{\psi}^{+}(y) - \vec{\psi}^{-}(y)$, we obtain the following expression for the dual function:

$$g(\vec{\psi}(y), \vec{\psi}(y), \gamma(y), \lambda_{xy}) \quad = \quad \sum_{y} \bar{p}(y) \left(\vec{\psi}(y) \cdot \vec{a}(y) + \gamma(y) - \vec{\beta}(y) |\vec{\psi}(y)| \right) \ .$$

The rest of the duality proof is very similar to the one for the standard case. The main difference is in Proposition 2, where the difference $g(\vec{\psi}(y), \gamma(y), \lambda_{xy}) - I[p]$ turns out to be $\sum_y \bar{p}(y) |\vec{\psi}(y)| (-|\beta(y)| + \beta(y))$ which again is zero since $\vec{\beta}(y)$ is positive.

A.1.4 Minimax Theorem for MinMI

We first bound the max expression from below. For any $f(y|x)$ satisfying the *sub normalization* constraint $\sum_y f(y|x) \leq 1$ the following holds

$$\max_{\hat{p}(x,y) \in \mathcal{P}(\vec{a})} -\langle \log f(y|x) \rangle_{\hat{p}(x,y)} \geq -\langle \log f(y|x) \rangle_{p_{MI}(x,y)} \geq -\langle \log f_{MI}(y|x) \rangle_{p_{MI}(x,y)} = H[p_{MI}(y|x)] \ . \tag{A.17}$$

The first inequality follows since $p_{MI}(x,y) \in \mathcal{P}_{x|y}\big(\vec{\phi}(x), \vec{a}(y), \bar{p}(y)\big)$. The second is an information inequality for unnormalized distribution proved in Proposition 3 below. The equality is since $f_{MI}(y|x) = p_{MI}(y|x)$ when $p_{MI}(x) > 0$.

We now have a lower bound on the result of the min max. To see that $f_{MI}(y|x)$ achieves it, note that for every $\hat{p}(x,y) \in \mathcal{P}_{x|y}\big(\vec{\phi}(x), \vec{a}(y), \bar{p}(y)\big)$

$$-\langle \log f_{MI}(y|x) \rangle_{\hat{p}(x,y)} = H[p_{MI}(y|x)] \tag{A.18}$$

The function $f_{MI}(y|x)$ thus achieves a value for the expression in the max which is the minimum possible for this maximization, by the previous bound. Thus $f_{MI}(y|x)$ is the solution to the given min max problem.

Proposition 3 *Let $f(y|x)$ be a subnormalized function. Then*

$$\langle \log f_{MI}(y|x) \rangle_{p_{MI}(x,y)} \leq \langle \log f(y|x) \rangle_{p_{MI}(x,y)} \tag{A.19}$$

Proof:

$$\langle \log f_{MI}(y|x) \rangle_{p_{MI}(x,y)} - \langle \log f(y|x) \rangle_{p_{MI}(x,y)} = \langle \log \frac{f_{MI}(y|x)}{f(y|x)} \rangle_{p_{MI}(x,y)} \tag{A.20}$$

We can limit the above expectation to values of x where $p_{MI}(x,y) \neq 0$. In these points $f_{MI}(y|x)$ is normalized to one, and we can write $p_{MI}(y|x)$ instead. We thus have

$$\begin{aligned} \langle \log \frac{p_{MI}(y|x)}{f(y|x)} \rangle_{p_{MI}(x,y)} \quad &= \quad \langle \log \frac{p_{MI}(y|x)}{f_N(y|x)} \rangle_{p_{MI}(x,y)} - \langle \log a(x) \rangle_{p_{MI}(x)} \\ &= \quad \langle D_{KL}[p_{MI}(y|x) | f_N(y|x)] \rangle_{p_{MI}(x)} - \langle \log a(x) \rangle_{p_{MI}(x)} \end{aligned}$$

where $f_N(y|x)$ is the normalized version of $f(y|x)$ and $a(x)$ is the normalization factor. Note that since $f(y|x)$ normalized to less than 1, we have $a(x) \leq 1$, and thus the expression is positive.

A.2 SDR-SI Results

A.2.1 Deriving the Gradient of the Joint Entropy

To calculate the gradient of the entropy $H[\hat{p}_\phi(x,y)]$, we first prove three useful properties of the distribution $\hat{p}_\phi$. Since $\hat{p}_\phi$ is in $\mathcal{P}(\vec{\phi}(x), \bar{p})$, it satisfies the marginal constraints: $\hat{p}_\phi(x) = \sum_{y'} \hat{p}_\phi(x,y') = \bar{p}(x)$, $\hat{p}_\phi(y) = \sum_{x'} \hat{p}_\phi(x',y) = \bar{p}(y)$, as well as the expectation constraints $\sum_{x'} \vec{\phi}(x')(\hat{p}_\phi(x',y) - \bar{p}(x',y)) = 0$. Deriving the three constraints equations w.r.t. $\vec{\phi}(x)$ yields

$$\sum_{y'} \frac{\partial \hat{p}_\phi(x,y')}{\partial \phi(x)} = 0; \quad \sum_{x'} \frac{\partial \hat{p}_\phi(x',y)}{\partial \phi(x)} = 0 \tag{A.21}$$

for the marginal constrains, and

$$\hat{p}_\phi(x,y) - \bar{p}(x,y) + \sum_{x'} \phi(x') \frac{\partial \hat{p}_\phi(x',y)}{\partial \phi(x)} = 0 \tag{A.22}$$

for the expectation constraints.

The derivative of the entropy can now be written as

$$\begin{aligned} \frac{\partial H[\hat{p}_\phi]}{\partial \vec{\phi}(x)} &= -\sum_{x',y'} \frac{\partial \hat{p}_\phi(x',y')}{\partial \vec{\phi}(x)} \\ &\quad -\sum_{x',y'} \frac{\partial \hat{p}_\phi(x',y')}{\partial \vec{\phi}(x)} \log \hat{p}_\phi(x',y') \\ &= -\sum_{x',y'} \frac{\partial \hat{p}_\phi(x',y')}{\partial \vec{\phi}(x)} \log \hat{p}_\phi(x',y') \ , \end{aligned} \tag{A.23}$$

where the last equality stems from the vanishing derivative of the marginal constraints in Equation A.21. Plugging in the exponential form of $\hat{p}_\phi$ from Equation 5.4, and using Equation A.21 again, we have

$$\frac{\partial H[\hat{p}_\phi]}{\partial \vec{\phi}(x)} = -\sum_{x',y'} \frac{\partial \hat{p}_\phi(x',y')}{\partial \vec{\phi}(x)} \vec{\phi}(x') \cdot \vec{\psi}_\phi(y') \ .$$

Now using Equation A.22 for the derivative of the expectation constraints, we finally obtain

$$\frac{\partial H[\hat{p}_\phi]}{\partial \vec{\phi}(x)} = p(x) \left(\langle \vec{\psi}_\phi \rangle_{\hat{p}_\phi(y|x)} - \langle \vec{\psi}_\phi \rangle_{\bar{p}(y|x)} \right) \ . \tag{A.24}$$

A.3 GIB Results

A.3.1 Invariance to the Noise Covariance Matrix

Lemma A.3.1 *For every pair (A, Σ_ξ) of a projection A and a full rank covariance matrix Σ_ξ, there exists a matrix $\tilde{A}$ such that $\mathcal{L}(\tilde{A}, I_d) = \mathcal{L}(A, \Sigma_\xi)$, where I_d is the $n_t \times n_t$ identity matrix.*

Proof: Denote by V the matrix which diagonalizes Σ_ξ, namely $\Sigma_\xi = VDV^T$, and by c the determinant $c \equiv |\sqrt{D^{-1}}V| = |\sqrt{D^{-1}}V^T|$. Setting $\tilde{A} \equiv \sqrt{D^{-1}}VA$, we have

$$\begin{aligned}
\mathcal{L}(\tilde{A}, I) &= (1-\beta)\log(|\tilde{A}\Sigma_x\tilde{A}^T + I_d|) - \log(|I_d|) + \beta\log(|\tilde{A}\Sigma_{x|y}\tilde{A}^T + I_d|) \quad \text{(A.25)} \\
&= (1-\beta)\log(c|A\Sigma_x A^T + \Sigma_\xi|c) - \log(c|\Sigma_\xi|c) + \beta\log(c|A\Sigma_{x|y}A^T + \Sigma_\xi|c) \\
&= (1-\beta)\log(|A\Sigma_x A^T + \Sigma_\xi|) - \log(|\Sigma_\xi|) + \beta\log(|A\Sigma_{x|y}A^T + \Sigma_\xi|) \\
&= \mathcal{L}(A, \Sigma_\xi) \ ,
\end{aligned}$$

where the first equality stems from the fact that the determinant of a matrix product is the product of the determinants. □

A.3.2 Properties of Eigenvalues of $\Sigma_{x|y}\Sigma_x^{-1}$ and Σ_x

Lemma A.3.2 *Denote the set of left normalized eigenvectors of* $\Sigma_{x|y}\Sigma_x^{-1}$ *by* $\mathbf{v}_i$ *(*$||\mathbf{v}_i|| = 1$*) and their corresponding eigenvalues by* λ_i*. Then*

1. *All the eigenvalues are real and satisfy* $0 \leq \lambda_i \leq 1$
2. $\exists r_i > 0$ *s.t.* $\mathbf{v}_i^T\Sigma_x\mathbf{v}_j = \delta_{ij}r_i$.
3. $\mathbf{v}_i^T\Sigma_{x|y}\mathbf{v}_j = \delta_{ij}\lambda_i r_i$.

The proof is standard [see e.g 58] and is brought here for completeness.
Proof:

1. The matrices $\Sigma_{x|y}\Sigma_x^{-1}$ and $\Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx}\Sigma_x^{-1}$ are positive semi definite (PSD), and their eigenvalues are therefore positive [5].Since $\Sigma_{x|y}\Sigma_x^{-1} = I - \Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx}\Sigma_x^{-1}$, the eigenvalues of $\Sigma_{x|y}\Sigma_x^{-1}$ are bounded between 0 and 1.

2. Denote by V the matrix whose rows are $\mathbf{v}_i^T$. The matrix $V\Sigma_x^{\frac{1}{2}}$ is the eigenvector matrix of $\Sigma_x^{-\frac{1}{2}}\Sigma_{x|y}\Sigma_x^{-\frac{1}{2}}$ since $\left(V\Sigma_x^{\frac{1}{2}}\right)\Sigma_x^{-\frac{1}{2}}\Sigma_{x|y}\Sigma_x^{-\frac{1}{2}} = V\Sigma_{x|y}\Sigma_x^{-\frac{1}{2}} = \left(V\Sigma_{x|y}\Sigma_x^{-1}\right)\Sigma_x^{\frac{1}{2}} = DV\Sigma_x^{\frac{1}{2}}$. From the fact that $\Sigma_x^{-\frac{1}{2}}\Sigma_{x|y}\Sigma_x^{-\frac{1}{2}}$ is symmetric, $V\Sigma_x^{\frac{1}{2}}$ is orthogonal, and thus $V\Sigma_x V^T$ is diagonal.

3. Follows from 2: $\mathbf{v}_i^T\Sigma_{x|y}\Sigma_x^{-1}\Sigma_x\mathbf{v}_j = \lambda_i\mathbf{v}_i^T\Sigma_x\mathbf{v}_j = \lambda_i\delta_{ij}r_i$.

□

A.3.3 Optimal Eigenvector

For some β values, several eigenvectors can satisfy the conditions for non degenerated solutions (equation 6.10). To identify the optimal eigenvector, we substitute the value

[5]To see why $\Sigma_{x|y}\Sigma_x^{-1}$ is PSD, note that it has the same eigenvalues as $\Sigma_x^{-\frac{1}{2}}\Sigma_{x|y}\Sigma_x^{-\frac{1}{2}}$ (Section A.4.5), where the latter has a square root $\Sigma_x^{-\frac{1}{2}}\Sigma_{x|y}^{\frac{1}{2}}$ and is therefore PSD.

of $||A||^2$ from equation (6.9) $A\Sigma_{x|y}A^T = r\lambda||A||^2$ and $A\Sigma_x A^T = r||A||^2$ into the target function $\mathcal{L}$ of equation (6.6), and obtain

$$\mathcal{L} = (1-\beta)\log\left(\frac{(1-\lambda)(\beta-1)}{\lambda}\right) + \beta\log\left(\beta(1-\lambda)\right) \ . \tag{A.26}$$

Since $\beta \geq 1$, this is monotonically increasing in λ and is minimized by the eigenvector of $\Sigma_{x|y}\Sigma_x^{-1}$ with the smallest eigenvalue. Note that this is also the eigenvector of $\Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx}\Sigma_x^{-1}$ with the largest eigenvalue.

A.3.4 Optimal Mixing Matrix

Lemma A.3.3 *The optimum of the cost function is obtained with a diagonal mixing matrix W of the form*

$$W = diag\left[\sqrt{\frac{\beta(1-\lambda_1)-1}{\lambda_1 r_1}}; \ldots; \sqrt{\frac{\beta(1-\lambda_k)-1}{\lambda_k r_k}}; 0; \ldots; 0\right] \ , \tag{A.27}$$

where $\{\lambda_1, \ldots, \lambda_k\}$ are $k \leq n_x$ eigenvalues of $\Sigma_{x|y}\Sigma_x^{-1}$ with critical β values ${\beta^c}_1, \ldots, {\beta^c}_k \leq \beta$. $r_i \equiv \mathbf{v}_i^T\Sigma_x\mathbf{v}_i$ as in theorem 6.2.1.

Proof: We write $V\Sigma_{x|y}\Sigma_x^{-1} = DV$ where D is a diagonal matrix whose elements are the corresponding eigenvalues, and denote by R the diagonal matrix whose i^{th} element is r_i. When $k = n_x$, we substitute $A = WV$ into equation (6.12), and eliminate V from both sides to obtain

$$\frac{\beta-1}{\beta}\left[(WDRW^T + I_d)(WRW^T + I_d)^{-1}\right]W = WD \ . \tag{A.28}$$

Use the fact that W is full rank to multiply by W^{-1} from the left and by $W^{-1}(WRW^T + I_d)W$ from the right

$$\frac{\beta-1}{\beta}(DRW^TW + I_d) = D(RW^TW + I_d) \ . \tag{A.29}$$

Rearranging, we have,

$$W^TW = [\beta(I-D) - I](DR)^{-1} \ , \tag{A.30}$$

which is a diagonal matrix.

While this does not uniquely characterize W, we note that using properties of the eigenvalues from lemma A.3.2, we obtain

$$|A\Sigma_x A^T + I_d| \ = \ |WV\Sigma_x V^TW^T + I_d| = |WRW^T + I_d| \ .$$

Note that WRW^T has left eigenvectors W^T with corresponding eigenvalues obtained from the diagonal matrix W^TWR. Thus if we substitute A into the target function in equation (6.6), a similar calculation yields

$$\mathcal{L} = (1-\beta)\sum_{i=1}^{n}\log\left(||\mathbf{w}_i^T||^2 r_i + 1\right) + \beta\sum_{i=1}^{n}\log\left(||\mathbf{w}_i^T||^2 r_i\lambda_i + 1\right) \ . \tag{A.31}$$

where $||\mathbf{w}_i^T||^2$ is the i^{th} element of the diagonal of W^TW. This shows that $\mathcal{L}$ depends only on the norm of the columns of W, and all matrices W that satisfy (A.30) yield the same target function. We can therefore choose to take W to be the diagonal matrix which is the (matrix) square root of (A.30)

$$W = \sqrt{[\beta(I-D)-I](DR)^{-1}} \ , \tag{A.32}$$

which completes the proof of the full rank ($k = n_x$) case.

In the low rank ($k < n_x$) case W does not mix all the eigenvectors, but only k of them. To prove the lemma for this case, we first show that any such low rank matrix is equivalent (in terms of the target function value) to a low rank matrix that has only k non zero rows. We then conclude that the non zero rows should follow the form described in the above lemma.

Consider a $n_x \times n_x$ matrix W of rank $k < n_x$, but without any zero rows. Let U be the set of left eigenvectors of WW^T (that is, $WW^T = U\Lambda U^T$). Then, since WW^T is Hermitian, its eigenvectors are orthonormal, thus $(UW)(WU)^T = \Lambda$ and $W' = UW$ is a matrix with k non zero rows and $n_x - k$ zero lines. Furthermore, W' obtains the same value of the target function, since

$$\begin{aligned}
\mathcal{L} &= (1\text{–}\beta)\log(|W'RW'^T + \Sigma_\xi^2|) + \beta\log(|W'DRW'^T + \Sigma_\xi^2|) \qquad \text{(A.33)}\\
&= (1\text{–}\beta)\log(|UWRW^TU^T + UU^T\Sigma_\xi^2|) + \beta\log(|UWDRW^TU^T + UU^T\Sigma_\xi^2|)\\
&= (1\text{–}\beta)\log(|U||WRW^T + \Sigma_\xi^2||U^T|) + \beta\log(|U||UWDRW^TU^T + \Sigma_\xi^2||U^T|)\\
&= (1\text{–}\beta)\log(|WRW^T + \Sigma_\xi^2|) + \beta\log(|WDRW^TT + \Sigma_\xi^2|) \ ,
\end{aligned}$$

where we have used the fact that U is orthonormal and hence $|U| = 1$. To complete the proof note that the non zero rows of W' also have $n_x - k$ zero columns and thus define a square matrix of rank k, for which the proof of the full rank case apply, but this time by projecting to a dimension k instead of n_x. □

This provides a characterization of all local minima. To find which is the global minimum, we prove the following corollary.

Corollary A.3.4
The global minimum of $\mathcal{L}$ is obtained with all λ_i that satisfy $\lambda_i < \frac{\beta-1}{\beta}$

Proof: Substituting the optimal W of equation (A.32) into equation (A.31) yields

$$\mathcal{L} = \sum_{i=1}^{k}(\beta-1)\log\lambda_i + \log(1-\lambda_i) + f(\beta) \ . \tag{A.34}$$

Since $0 \le \lambda \le 1$ and $\beta \ge \frac{1}{1-\lambda}$, $\mathcal{L}$ is minimized by taking all the eigenvalues that satisfy $\beta > \frac{1}{(1-\lambda_i)}$. □

A.3.5 Deriving the Iterative Algorithm

To derive the iterative algorithm in section 6.5, we assume that the distribution $p(t_k|x)$ corresponds to the Gaussian variable $T_k = A_k X + \xi_k$. We show below that $p(t_{k+1}|x)$ corresponds to $T_{k+1} = A_{k+1}X + \xi_{k+1}$ with $\xi_{k+1} \sim N(0, \Sigma_{\xi_{k+1}})$ and

$$\begin{aligned} \Sigma_{\xi_{k+1}} &= \left(\beta\Sigma^{-1}_{t_k|y} - (\beta-1)\Sigma^{-1}_{t_k}\right)^{-1} \\ A_{k+1} &= \beta\Sigma_{\xi_{k+1}}\Sigma^{-1}_{t_k|y}A_k\left(I - \Sigma_{y|x}\Sigma^{-1}_x\right) \ . \end{aligned} \tag{A.35}$$

We first substitute the Gaussian $p(t_k|x) \sim N(A_k x, \Sigma_{\xi_k})$ into the equations of (6.17), and treat the second and third equations. The second equation $p(t_k) = \int_x p(x)p(t_k|x)dx$, is a marginal of the Gaussian $T_k = A_k X + \xi_k$, and yields a Gaussian $p(t_k)$ with zero mean and covariance

$$\Sigma_{t_k} = A_k\Sigma_x A_k^T + \Sigma_{\xi_k} \ . \tag{A.36}$$

The third equation, $p(y|t_k) = \frac{1}{p(t_k)}\int_x p(x,y)p(t_k|x)dx$ defines a Gaussian with mean and covariance matrix given by:

$$\begin{aligned} \mu_{y|t_k} &= \mu_y + \Sigma_{yt_k}\Sigma^{-1}_{t_k}(t_k - \mu_{t_k}) = \Sigma_{yt_k}\Sigma^{-1}_{t_k}t_k \equiv B_k t_k \\ \Sigma_{y|t_k} &= \Sigma_y - \Sigma_{yt_k}\Sigma^{-1}_{t_k}\Sigma_{t_k y} = \Sigma_y - A_k\Sigma_{xy}\Sigma^{-1}_{t_k}\Sigma_{yx}A_k^T \ , \end{aligned} \tag{A.37}$$

where we have used the fact that $\mu_y = \mu_{t_k} = 0$, and define the matrix $B_k \equiv \Sigma_{yt_k}\Sigma^{-1}_{t_k}$ as the regressor of t_k on y. Finally, we return to the first equation of (6.17), that defines $p(t_{k+1}|x)$ as

$$p(t_{k+1}|x) = \frac{p(t_k)}{Z(x,\beta)}e^{-\beta D_{KL}[p(y|x)|p(y|t_k)]} \ . \tag{A.38}$$

We now show that $p(t_{k+1}|x)$ is Gaussian and compute its mean and covariance matrix.

The KL divergence between the two Gaussian distributions, in the exponent of equation (A.38) is known to be

$$\begin{aligned} 2D_{KL}[p(y|x)|p(y|t_k)] &= \log\frac{|\Sigma_{y|t_k}|}{|\Sigma_{y|x}|} + Tr(\Sigma^{-1}_{y|t_k}\Sigma_{y|x}) \\ &+ (\mu_{y|x} - \mu_{y|t_k})^T\Sigma^{-1}_{y|t_k}(\mu_{y|x} - \mu_{y|t_k}) \ . \end{aligned} \tag{A.39}$$

The only factor which explicitly depends on the value of t in the above expression is $\mu_{y|t_k}$ derived in equation (A.37), is linear in t. The KL divergence can thus be rewritten as

$$D_{KL}[p(y|x)|p(y|t_k)] = c(x) + \frac{1}{2}(\mu_{y|x} - B_k t_k)^T\Sigma^{-1}_{y|t_k}(\mu_{y|x} - B_k t_k)$$

Adding the fact that $p(t_k)$ is Gaussian we can write the log of equation (A.38) as a quadratic form in t

$$\log p(t_{k+1}|x) = Z(x) + (t_{k+1} - \mu_{t_{k+1}|x})^T\Sigma_{\xi_{k+1}}(t_{k+1} - \mu_{t_{k+1}|x})$$

where

$$\begin{aligned}
\Sigma_{\xi_{k+1}} &= \left(\beta B_k^T \Sigma_{y|t_k}^{-1} B_k + \Sigma_{t_k}^{-1}\right)^{-1} \\
\mu_{t_{k+1}|x} &= A_{k+1} x \\
A_{k+1} &= \beta \Sigma_{\xi_{k+1}} B_k^T \Sigma_{y|t_k}^{-1} \Sigma_{yx} \Sigma_x^{-1} x \ .
\end{aligned} \tag{A.40}$$

This shows that $p(t_{k+1}|x)$ is a Gaussian $T_{k+1} = A_{k+1}x + \xi_{k+1}$, with $\xi \sim N(0, \Sigma_{\xi_{k+1}})$.

To simplify the form of $A_{k+1}, \Sigma_{\xi_{k+1}}$, we use the two following matrix inversion lemmas [6], which hold for any matrices E, F, G, H of appropriate sizes when E, H are invertible.

$$\begin{aligned}
(E - FH^{-1}G)^{-1} &= E^{-1} + E^{-1}F(H - GE^{-1}F)^{-1}GE^{-1} \\
(E - FH^{-1}G)^{-1}FH^{-1} &= E^{-1}F(H - GE^{-1}F)^{-1} \ .
\end{aligned} \tag{A.41}$$

Using $E \equiv \Sigma_{t_k}$, $F \equiv \Sigma_{yt_k}$, $H \equiv \Sigma_y$, $G \equiv \Sigma_{yt_k}$, $B_k = \Sigma_{yt_k}\Sigma_{t_k}^{-1}$ in the first lemma we obtain

$$\Sigma_{t_k|y}^{-1} = \Sigma_{t_k}^{-1} + B_k^T \Sigma_{y|t_k}^{-1} B_k \quad .$$

Replacing this into the expression for $\Sigma_{\xi_{k+1}}$ in equation (A.40) we obtain

$$\Sigma_{\xi_{k+1}} = \left(\beta \Sigma_{t_k|y}^{-1} - (\beta - 1)\Sigma_{t_k}^{-1}\right)^{-1} \ . \tag{A.42}$$

Finally, using again $E \equiv \Sigma_{t_k}$, $F \equiv \Sigma_{t_k y}$, $H \equiv \Sigma_y$, $G \equiv \Sigma_{yt_k}$ in the second matrix lemma, we have $\Sigma_{t_k|y}^{-1}\Sigma_{t_k y}\Sigma_y^{-1} = \Sigma_{t_k}^{-1}\Sigma_{t_k y}\Sigma_{y|t_k}^{-1}$, which turns the expression for A_{k+1} in equation (A.40) into

$$\begin{aligned}
A_{k+1} &= \beta \Sigma_{\xi_{k+1}} \Sigma_{t_k|y}^{-1} \Sigma_{t_k y} \Sigma_y^{-1} \Sigma_{yx} \Sigma_x^{-1} \\
&= \beta \Sigma_{\xi_{k+1}} \Sigma_{t_k|y}^{-1} A_k \Sigma_{xy} \Sigma_y^{-1} \Sigma_{yx} \Sigma_x^{-1} \\
&= \beta \Sigma_{\xi_{k+1}} \Sigma_{t_k|y}^{-1} A_k (I - \Sigma_{x|y} \Sigma_x^{-1}) \ ,
\end{aligned} \tag{A.43}$$

which completes the derivation of the algorithm as described in (6.17).

A.4 Optimality of the Gaussian Solution for GIB

This section is somewhat longer and more technical than the previous ones. In Chapter 6 we derived the Gaussian Information Bottleneck solution under the assumption that the optimal distribution is itself Gaussian. While this seems like a reasonable assumption, its proof, given here, is somewhat involved.

In order to prove the validity of the assumption we show that the information curve achieved with the Gaussian solution (see Section 6.4) is the best possible, i.e. it achieves the highest information $I(T;Y)$ under a given level of compression $I(T;X)$.

[6] The first equation is the standard inversion lemma [see e.g., 77, page 571]. The second can be easily verified from the first.

This immediately implies that it yields the optimal value for the GIB minimization problem in Equation 6.1.

Formally, define the *true* information curve as [7]

$$R(I_x) = \max_{p(t|x):T\to X\to Y, I(T;X)=I_x} I(T;Y) \ . \tag{A.44}$$

The function $R(I_x)$ is defined as the maximum information one can keep about Y when forced to compress X by I_x.

In Chapter 6 we carry out this optimization under the limitation that $p(t|x)$ is Gaussian [8]. Denote by $G(T,X)$ the set of conditional Gaussian $p(t|x)$ distributions, then the Gaussian information curve is defined as

$$R^G(I_x) = \max_{p(t|x):T\to X\to Y, I(T;X)=I_x, p(t|x)\in G(T,X)} I(T;Y) \ . \tag{A.45}$$

This function is calculated explicitly in section 6.4 and will be described in what follows, since we will use a notation more convenient for the current proof.

The main result of the current section is that the optimization problems in Equations A.44 and A.45 are equivalent. This is stated below

Theorem A.4.1 *$R(I_x) = R^G(I_x)$ and thus the maximum in Equation A.44 can be achieved with a Gaussian $p(t|x)$*

In [13] a similar problem is solved for one dimensional Gaussian variables, in the context of lossy coding with side information. Here we treat the multi-dimensional case, which has several interesting properties, and technical difficulties which are not present in the scalar case.

To prove the theorem, one needs to show that $R^G(I_x) \geq R(I_x)$ for all values of I_x. This is proven in the remainder of the text. The section is organized as follows: Section A.4.1 defines notations and characterizes certain covariance matrices. Section A.4.2 describes the curve $R^G(I_x)$ which was calculated in Chapter 6. Section A.4.3 transforms the variables X, Y into a representation which is easier to work with. Finally, Section A.4.4 gives the proof of Theorem A.4.1.

A.4.1 Notations and Matrix Properties

In what follows, we use various forms of covariance matrices, which we outline in this section.

Notation 1 *Denote by X_i the univariate Gaussian variable which is the i-th coordinate of the multivariate random variable X. We use Y_i with a similar notation.*

[7] We explicitly state the Markov chain condition here. Of course it is also assumed in Chapter 6

[8] Which implies X, T are jointly Gaussian, since X itself is Gaussian

Notation 2 *Denote by $X^{(m,d)}$ the multivariate Gaussian of dimension d, $X^{(m,d)} \equiv (X_m, ..., X_{m+d-1})$. Also, denote $X^{(1,d)} \equiv X^{(d)}$.*

Notation 3 *Denote by Σ_{xy} the covariance matrix of X and Y. Thus the $(i,j)^{th}$ element of Σ_{xy} is $E(X_i Y_j) - E(X_i)E(Y_j)$. The dimensions of Σ_{xy} are (n_x, n_y).*

Notation 4 *Denote by Σ_x the covariance matrix of X. Thus the $(i,j)^{th}$ element of Σ_x is $E(X_i X_j) - E(X_i)E(X_j)$. The dimensions of Σ_x are (n_x, n_x).*

Consider the variable $W = [X, Y]$ produced by concatenating X, Y. Its covariance matrix is thus given by

$$\Sigma_w = \begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_y \end{bmatrix}$$

The distribution $p(x|y)$ is known to be Gaussian as well. The covariance of this Gaussian does not depend on the value of y conditioned on. We denote the conditional covariance by $\Sigma_{x|y}$. Note that it is a square matrix of size n_x. The following identity expresses $\Sigma_{x|y}$ as a function of the unconditional covariances [87]

$$\Sigma_{x|y} = \Sigma_x - \Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx} \ . \tag{A.46}$$

In what follows, we make frequent use of the matrices $\Sigma_{x|y}\Sigma_x^{-1}$ and $\Sigma_{y|x}\Sigma_y^{-1}$. These are normalized conditional covariance matrices, which reflect by how much the knowledge of Y reduces the *spread* of X. For one dimensional variables the scalar $\Sigma_{x|y}\Sigma_x^{-1}$ will be zero if X, Y are deterministically related, and one if X, Y are independent. In the multidimensional case, the values of interest are the eigenvalues of the matrices. As in the scalar case, eigenvalues close to zero reflect a deterministic relation between X and Y, and eigenvalues close to one reflect independence.

Notation 5 *Denote the eigenvalues of $\Sigma_{x|y}\Sigma_x^{-1}$ by $\lambda_1, \ldots, \lambda_{n_x}$, and assume they are sorted in an ascending order.*

The following Lemma states the eigenvalues of $\Sigma_{x|y}\Sigma_x^{-1}, \Sigma_{y|x}\Sigma_y^{-1}$ are essentially identical.

Lemma A.4.2 *Assume $n_x \geq n_y$. Denote by $\gamma_1, \ldots, \gamma_{n_y}$ the sorted eigenvalues of $\Sigma_{y|x}\Sigma_y^{-1}$. Then $\gamma_i = \lambda_i$ for $i = 1, \ldots, n_y$. Furthermore $\lambda_i = 1$ for $n_y < i \leq n_x$*

Proof: Assume v is a left eigenvector of $\Sigma_{x|y}\Sigma_x^{-1}$ with eigenvalue λ, then

$$\begin{aligned}
v(I - \Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx}\Sigma_x^{-1}) &= \lambda v \\
v(\Sigma_{xy}\Sigma_y^{-1} - \Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx}\Sigma_x^{-1}\Sigma_{xy}\Sigma_y^{-1}) &= \lambda v \Sigma_{xy}\Sigma_y^{-1} \\
v\Sigma_{xy}\Sigma_y^{-1}(I - \Sigma_{yx}\Sigma_x^{-1}\Sigma_{xy}\Sigma_y^{-1}) &= \lambda v \Sigma_{xy}\Sigma_y^{-1} \\
\left(v\Sigma_{xy}\Sigma_y^{-1}\right)\Sigma_{y|x}\Sigma_y^{-1} &= \lambda \left(v\Sigma_{xy}\Sigma_y^{-1}\right) \ .
\end{aligned}$$

Thus $v\Sigma_{xy}\Sigma_y^{-1}$ is an eigenvector of $\Sigma_{y|x}\Sigma_y^{-1}$ with eigenvalue λ, and all the eigenvalues of $\Sigma_{x|y}\Sigma_x^{-1}$ are eigenvalues of $\Sigma_{y|x}\Sigma_y^{-1}$, which shows that there is a correspondence between the eigenvalue sets.

To see that $\lambda_i = 1$ for $n_y < i \leq n_x$, note that $\Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx}\Sigma_x^{-1}$ has rank n_y at most, so it has an eigenvalue zero with multiplicity at least $n_x - n_y$. Thus the last $n_x - n_y$ eigenvalues of $\Sigma_{x|y}\Sigma_x^{-1}$ are equal to 1. □

A.4.2 The Structure of $R^G(I_x)$

In Section 6.4 we derived the analytic form of $R^G(I_x)$. The matrix $\Sigma_{x|y}\Sigma_x^{-1}$ discussed in the previous section plays a central role in the derivation. We show in 6.4 that the curve $R^G(I_x)$ is made up of a set of segments, which form a continuous curve. In order to understand the functional form of the segments, we need the following definitions. Define $\hat{c}(k), \hat{n}(I)$ as the functions

$$\begin{aligned} \hat{c}(k) &= \begin{cases} 0 & k = 1 \\ \sum_{i=1}^{k-1} \log \frac{\lambda_k}{\lambda_i} \frac{1-\lambda_i}{1-\lambda_k} & k > 1 \end{cases} \\ \hat{n}(I) &= \arg \max_{1 \leq k \leq n_x} \hat{c}(k) \leq I \ . \end{aligned}$$

Note that $\hat{n}(I)$ partitions the values of I into n_x segments. In Section 6.4 we show that $R^G(I_x)$ has a slightly different form for each such segment, and is given by

$$R^G(I_x) = I_x - \frac{\hat{n}(I_x)}{2} \log \left(\prod_{i=1}^{\hat{n}(I_x)} (1-\lambda_i)^{\frac{1}{\hat{n}(I_x)}} + e^{\frac{2I_x}{\hat{n}(I_x)}} \prod_{i=1}^{\hat{n}(I_x)} \lambda_i^{\frac{1}{\hat{n}(I_x)}} \right) \quad . \tag{A.47}$$

The function $R^G(I_x)$ can be shown to be continuous and concave in I. An example of $R^G(I_x)$ is given in Figure A.1, showing the points $I_x = \hat{c}(k)$ where $\hat{n}(I_x)$ switches values.

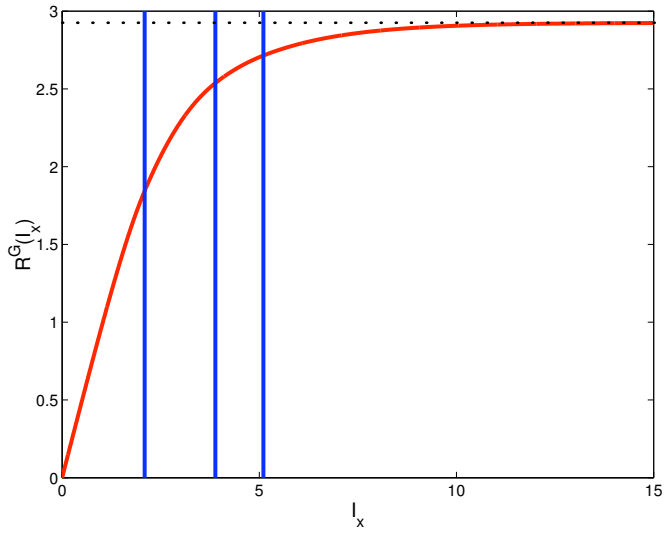


Figure A.1: The curve $R^G(I_x)$ obtained with four eigenvalues λ_i=0.01,0.4,0.8,0.9. Values of I_x corresponding to $\hat{c}(2), \hat{c}(3), \hat{c}(4)$ (i.e where $\hat{n}(I_x)$ switches values) are marked by vertical lines. The asymptotic value of $I(X;Y)$ is marked by the horizontal line.

A.4.3 Canonical Representation

Before proving the main theorem, we will transform the variables X and Y to a canonical representation, which is easier to manipulate. This is done via the eigenvectors of the matrices $\Sigma_{x|y}\Sigma_x^{-1}$ and $\Sigma_{y|x}\Sigma_y^{-1}$

Notation 6 *Denote A_x the eigenvector matrix of $\Sigma_{x|y}\Sigma_x^{-1}$ and B_y that of $\Sigma_{y|x}\Sigma_y^{-1}$.*

Define $\tilde{X} \equiv A_x X$ and $\tilde{Y} \equiv B_y Y$. The following Lemma gives some useful properties of the new variables (see proof in section A.4.5)

Lemma A.4.3 *The matrices A_x, B_y can be scaled so that the following properties hold*

1. *A_x, B_y are full rank*
2. *$\tilde{X}$ and $\tilde{Y}$ are spherized: $\Sigma_{\tilde{x}} = I_X$ and $\Sigma_{\tilde{y}} = I_Y$*
3. *The $(i,j)^{th}$ element of the matrix $\Sigma_{\tilde{x}\tilde{y}}$ is $\sqrt{1-\lambda_i}\delta_{ij}$. Thus X_i, Y_j have covariance $\sqrt{1-\lambda_i}$ if $i = j$, and are uncorrelated otherwise.*

To illustrate the above properties, we give below the full covariance matrix between variables with dimensions $n_x = 3, n_y = 2$

$$\Sigma_{\tilde{x}\tilde{y},\tilde{x}\tilde{y}} = \begin{pmatrix} 1 & 0 & 0 & \sqrt{1-\lambda_1} & 0 \\ 0 & 1 & 0 & 0 & \sqrt{1-\lambda_2} \\ 0 & 0 & 1 & 0 & 0 \\ \sqrt{1-\lambda_1} & 0 & 0 & 1 & 0 \\ 0 & \sqrt{1-\lambda_2} & 0 & 0 & 1 \end{pmatrix}$$

As a consequence of the above properties, we have the following corollary

Corollary A.4.4 *For all* $1 \leq i \leq \min(n_x, n_y)$

$$\tilde{Y}_i = \sqrt{1-\lambda_i}\tilde{X}_i + N_i \quad with \quad N_i \sim \mathcal{N}(0, \sqrt{\lambda_i}). \tag{A.48}$$

Thus $Y^{(d)} = DX^{(d)} + \xi$ *where* $\xi \sim \mathcal{N}(0, N)$ *and* D, N *are diagonal with* $\sqrt{1-\lambda_i}, \sqrt{\lambda_i}$ *on their diagonal.*

This follows from the fact that two Gaussians can be presented as a noisy linear transformation, and since $E(\tilde{X}_i\tilde{Y}_i) = \sqrt{1-\lambda_i}$ and $V(\tilde{X}_i) = V(\tilde{Y}_i) = 1$ the transformation between $\tilde{X}_i, \tilde{Y}_i$ must be the one given above.

Also, since $\Sigma_{\tilde{x}|\tilde{y}} = \Sigma_{\tilde{x}} - \Sigma_{\tilde{x}\tilde{y}}\Sigma_{\tilde{y}}^{-1}\Sigma_{\tilde{y}\tilde{x}}$ and the matrices $\Sigma_{\tilde{x}|\tilde{y}}, \Sigma_{\tilde{x}}$ have the structure given above, we have that

Corollary A.4.5 $\Sigma_{\tilde{x}|\tilde{y}}$ *is diagonal. Thus* $\tilde{x}$ *are conditionally independent given* $\tilde{y}$

Since A_x, B_y are non singular, they do not affect the mutual information between X, Y and T

$$I(\tilde{X};T) = I(X;T) \qquad I(\tilde{Y};T) = I(Y;T) \ .$$

Our optimization problem depends on X, Y only through the information values $I(X;T), I(Y;T)$, and thus we can use $\tilde{X}, \tilde{Y}$ instead of X, Y. Furthermore, the variables X cannot provide any information about $\tilde{Y}^{(n_x+1,n_y)}$ (when $n_y > n_x$), and thus we can assume that $n_y \leq n_x$ by discarding the excess Y variables.

From this point on, we will refer to the variables X and Y as if they were given in their canonical representation in the first place.

A.4.4 Proof of Theorem A.4.1

We start by formulating $R(I_x)$ as a minimization problem. Using the Markov relation $T \to X \to Y$, we have $I(T;Y) = I(T;X) - I(T;X|Y)$ and thus

$$R(I_x) = \max_{p(t|x):T\to X\to Y, I(T;X)=I_x} I(T;Y) = I_x - \min_{p(t|x):T\to X\to Y, I(T;X)=I_x} I(T;X|Y) \quad . \tag{A.49}$$

This, together with the form of $R^G(I_x)$ given in Equation A.47 implies that Theorem 1 is equivalent to showing that

Theorem A.4.6 *For every Markov chain* $T \to X \to Y$ *, where* $I(T;X) = I_x$ *it holds that*

$$I(T;X|Y) \geq \frac{\hat{n}(I_x)}{2} \log \left(\prod_{i=1}^{\hat{n}(I_x)} (1-\lambda_i)^{\frac{1}{\hat{n}(I_x)}} + e^{\frac{2I_x}{\hat{n}(I_x)}} \prod_{i=1}^{\hat{n}(I_x)} \lambda_i^{\frac{1}{\hat{n}(I_x)}} \right) \tag{A.50}$$

The following lemma, which will be proven in section A.4.5, bounds $I(T;X|Y)$ for subsets of the variable X.

Lemma A.4.7 *For all d such that $d \leq n_x$, define $k = \min(n_y, d)$. Then*

$$I(T;X^{(d)}|Y) \geq \frac{k}{2}\log\left(\prod_{i=1}^{k}(1-\lambda_i)^{\frac{1}{k}} + e^{\frac{2I(T;X^{(d)})}{k}}\prod_{i=1}^{k}\lambda_i^{\frac{1}{k}}\right) \tag{A.51}$$

This of course provides a bound for $I(T;X|Y)$ by selecting $d = n_x$. However this bound is not tight, whereas that of Theorem A.4.6 is.

We prove Theorem A.4.6 by showing by induction that it holds for all values of $\hat{n}(I_x)$. The induction is over $k = n_y - \hat{n}(I_x)$ which is the number of segments between segment number $\hat{n}(I_x)$ and the last segment (note that we have at most n_y segments).

When $k = 0$, i.e. $\hat{n}(I_x) = n_y$ the claim follows from application of Lemma A.4.7 with $d = n_x$.

In the induction step, we assume the theorem holds for I_x such that $n_y - \hat{n}(I_x) < k$ and show it holds for all I_x such that $n_y - \hat{n}(I_x) = k$. Thus from now on, we limit ourselves to I_x such that $\hat{n}(I_x) = n_y - k$. We separate X into two sets of variables. One contains its first $n_y - k$ elements $X^1 \equiv X^{(1,n_y-k)}$, and the other, $X^2 \equiv X^{(n_y-k+1,n_x)}$ contains the rest.

Note that

$$\begin{aligned}
I(T;X|Y) &= I(T;X^1,X^2|Y) \\
&= h(X^1,X^2|Y) - h(X^1,X^2|T,Y) \\
&= h(X^1|Y) + h(X^2|Y) - h(X^1,X^2|T,Y) \\
&= h(X^1|Y) + h(X^2|Y) - h(X^1|T,Y) - h(X^2|X^1,T,Y) \\
&\geq h(X^1|Y) + h(X^2|Y) - h(X^1|T,Y) - h(X^2|T,Y) \\
&= I(T;X^1|Y) + I(T;X^2|Y)
\end{aligned} \tag{A.52}$$

where the first equality is the definition of mutual information, the second equality is a result of X^1, X^2 being independent conditioned on Y, the third is the entropy chain rule, and the inequality is since conditioning reduces entropy.

Now, consider two variables T^1, T^2 such that the following holds:

$$\begin{aligned}
I(T^1;X^1|Y) &= I(T;X^1|Y) \\
I(T^2;X^2|Y) &= I(T;X^2|Y) \\
I(T^1,T^2;X^1,X^2) &= I(T;X)
\end{aligned} \tag{A.53}$$

and the Markov chains $T_1 \to X \to Y$, and $T_2 \to X \to Y$ are satisfied. Denote the set of variables which satisfy the above by $\mathcal{T}$ (note we are not placing any limitation on

(T_1, T_2)). Then it immediately follows that $(T^1, T^2) = (T, T)$ is in $\mathcal{T}$. Therefore

$$\begin{aligned} I(T;X|Y) &\geq I(T;X^1|Y) + I(T;X^2|Y) \\ &\geq \min_{(T^1,T^2)\in\mathcal{T}} I(T^1;X^1|Y) + I(T^2;X^2|Y) \ , \end{aligned} \tag{A.54}$$

where the first inequality follows from Equation A.52 and the second from the fact that $\mathcal{T}$ contains (T, T).

Denote the eigenvalues corresponding to the variables in X^2 by $\lambda'_1, \ldots, \lambda'_k$.

To use the induction step, note that X^2 is $n_{X^2} = k$ dimensional and therefore $n_{X^2} - \hat{n}(I(T^2;X^2)) < k$. Thus we can bound $I(T^2;X^2|Y)$ using the induction step. Denote $n'_I = \hat{n}(I(T^2;X^2))$, then

$$I(T^2;X^2|Y) \geq \frac{n'_I}{2} \log \left(\prod_{i=1}^{n'_I} (1-\lambda'_i)^{\frac{1}{n'_I}} + e^{\frac{2I(T^2,X^2)}{n'_I}} \prod_{i=1}^{n'_I} {\lambda'_i}^{\frac{1}{n'_I}} \right)$$

We denote the function on the right hand side by $f_2(I(T^2;X^2))$. We can also use Lemma A.4.7 to bound $I(T^1;X^1|Y)$ so that

$$I(T^1;X^1|Y) \geq \frac{n_y - k}{2} \log \left(\prod_{i=1}^{n_y-k} (1-\lambda_i)^{\frac{1}{n_y-k}} + e^{\frac{2I(T^1,X^1)}{n_y-k}} \prod_{i=1}^{n_y-k} \lambda_i^{\frac{1}{n_y-k}} \right)$$

We denote the function on the right hand side by $f_1(I(T^1;X^1))$. Thus we have

$$I(T^1;X^1|Y) + I(T^2;X^2|Y) \geq f_1(I(T^1;X^1)) + f_2(I(T^2;X^2)) \tag{A.55}$$

The right hand side of the above expression, which we want to minimize over $\mathcal{T}$ involves only $I(T^1;X^1), I(T^2;X^2)$. To minimize it over $\mathcal{T}$, we scan all values of $I(T^1;X^1), I(T^2;X^2)$ which are obtained in $\mathcal{T}$. To see what these values are, note that since X^1 and X^2 are independent, the following holds

$$\begin{aligned} I(T;X) = I(T^1,T^2;X^1,X^2) &= h(X^1,X^2) - h(X^1,X^2|T^1,T^2) \\ &= h(X^1) + h(X^2) - h(X^1|T^1,T^2) - h(X^2|X^1,T^1,T^2) \\ &\geq h(X^1) + h(X^2) - h(X^1|T^1,T^2) - h(X^2|T^1,T^2) \\ &\geq h(X^1) + h(X^2) - h(X^1|T^1) - h(X^2|T^2) \\ &= I(T^1;X^1) + I(T^2;X^2) \quad . \end{aligned}$$

The values of $I(T^1;X^1), I(T^2;X^2)$ which are obtained in $\mathcal{T}$ thus satisfy $I(T^1;X^1) + I(T^2;X^2) \leq I(T;X)$. This region is depicted in Figure A.2.

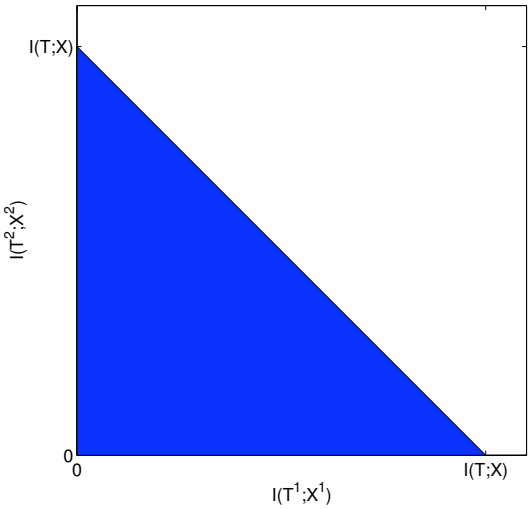


Figure A.2: The shaded region shows the values of $I(T^1;X^1), I(T^2;X^2)$ which can be achieved in $\mathcal{T}$. Note that there may be points in the region which are not achievable, but all achievable points are included in it.

Thus the minimization in Equation A.54 satisfies

$$\min_{(T^1,T^2)\in\mathcal{T}} f_1(I(T^1;X^1)) + f_2(I(T^2;X^2)) \geq \min_{\substack{0\leq\tilde{c}\leq I(T;X)\\ 0\leq a\leq\tilde{c}}} f_1(\tilde{c}-a) + f_2(a) \quad \text{(A.56)}$$

Where we have inequality since a given pair of values

$$\begin{aligned} I(T^1;X^1) &= \tilde{c}-a \\ I(T^2;X^2) &= a \end{aligned}$$

may not be achieved by an element in $\mathcal{T}$.

The minimum of Equation A.56 is given in the following lemma, which is proven in section A.4.5

Lemma A.4.8 *If $\hat{n}(I(T;X)) = n_y - k$ then*

$$\min_{\substack{0\leq\tilde{c}\leq I(T;X)\\ 0\leq a\leq\tilde{c}}} f_1(\tilde{c}-a) + f_2(a) = f_1(I(T;X)) \quad \text{(A.57)}$$

Thus we have that for $I(T;X)$ such that $\hat{n}(I(T;X)) = n_y - k$ the following holds (by combining Equations A.54,A.56 and the definition of f_1)

$$I(T;X|Y) \geq \frac{n_y-k}{2}\log\left(\prod_{i=1}^{n_y-k}(1-\lambda_i)^{\frac{1}{n_y-k}} + e^{\frac{2I(T;X)}{n_y-k}}\prod_{i=1}^{n_y-k}\lambda_i^{\frac{1}{n_y-k}}\right) \quad \text{(A.58)}$$

which is what we wanted to prove in the induction. We therefore have proved Theorem A.4.6, and its equivalent Theorem A.4.1, which is the main result of this section.

A.4.5 Proof of Lemmas

In this section we give proofs for the Lemmas stated in the text above.

Proof of Lemma A.4.7

In what follows we use $Y^{(d)} \equiv Y$ for $d > n_y$. First, we bound $I(T; X^{(d)}|Y)$ by an expression depending only on $Y^{(d)}$. We have that

$$I(T; X^{(d)}|Y) \geq I(T; X^{(d)}|Y^{(d)}) \ , \tag{A.59}$$

which is a result of the following use of the information chain rule

$$\begin{aligned} I(X^{(d)}; T, Y^{(d+1,n_y)}|Y^{(d)}) &= I(X^{(d)}; Y^{(d+1,n_y)}|Y^{(d)}) + I(X^{(d)}; T|Y) \\ &= I(X^{(d)}; T|Y^{(d)}) + I(X^{(d)}; Y^{(d+1,n_y)}|T, Y^{(d)}) \ . \end{aligned}$$

Using the fact that $I(X^{(d)}; Y^{(d+1,n_y)}|Y^{(d)}) = 0$, we get the desired inequality.

To simplify Equation A.59 we use the following lemma:

Lemma A.4.9 *The chain $T \to X^{(d)} \to Y^{(d)}$ is Markov*

Proof: By the Markovity of $T \to X \to Y$ and using the information chain rule, we have

$$0 = I(Y; T|X) = I(Y^{(d)}, Y^{(d+1,n_y)}; T|X) \quad = \quad I(Y^{(d)}; T|X) + I(Y^{(d+1,n_y)}; T|X, Y^{(d)}) \ ,$$

which implies

$$I(Y^{(d)}; T|X) = 0 \ . \tag{A.60}$$

Using the information chain rule

$$\begin{aligned} I(Y^{(d)}; T, X^{(d+1,n_x)}|X^{(d)}) &= I(Y^{(d)}; T|X^{(d)}) + I(Y^{(d)}; X^{(d+1,n_x)}|T, X^{(d)}) \\ &= I(Y^{(d)}; X^{(d+1,n_x)}|X^{(d)}) + I(Y^{(d)}; T|X^{(d+1,n_x)}, X^{(d)}) = 0 \ , \end{aligned}$$

where we have used $I(Y^{(d)}; X^{(d+1,n_x)}|X^{(d)}) = 0$ and Equation A.60. We have that a sum of two informations is zero, and therefore

$$I(Y^{(d)}; T|X^{(d)}) = 0 \ , \tag{A.61}$$

which implies the desired Markov chain. □

We can now write Equation A.59 using the above Markov chain

$$\begin{aligned} I(T; X^{(d)}|Y) \geq I(T; X^{(d)}|Y^{(d)}) &= I(T; X^{(d)}) - I(T; Y^d) \\ &= I(T; X^{(d)}) - h(Y^{(d)}) + h(Y^{(d)}|T) \end{aligned} \tag{A.62}$$

If $d > n_y$ we have

$$I(T; X^{(d)}|Y) \geq I(T; X^{(d)}) - h(Y) + h(Y|T) \ . \tag{A.63}$$

Equations A.62,A.63 involve $h(Y^{(k)}|T)$ where $k \equiv \min(n_y, d)$. We now bound this conditional entropy. Using Corollary A.4.4, write $h(Y^{(k)}|T) = h(DX^{(k)} + N|T)$. To bound this entropy, we use the conditional form of the entropy power inequality (see e.g. [13]), which states that if U, V, W, T are continuous random variables, and $U = V + W$ are n dimensional, then

$$e^{\frac{2}{n}h(U|T)} \geq e^{\frac{2}{n}h(V|T)} + e^{\frac{2}{n}h(W|T)} \ . \tag{A.64}$$

Corollary A.4.4 states that $Y^{(k)} = DX^{(k)} + N$ and therefore

$$\begin{aligned} e^{\frac{2}{k}h(Y^{(k)}|T)} &\geq e^{\frac{2}{k}h(DX^{(k)}|T)} + e^{\frac{2}{k}h(N|T)} \\ &= e^{\frac{2}{k}h(X^{(k)}|T)+\frac{2}{k}\log|D|} + e^{\frac{2}{k}h(N)} \\ &= e^{\frac{2}{k}h(X^{(k)}|T)+\frac{2}{k}\log|D|} + e^{\frac{1}{k}\log(2\pi e)^k|N|} \\ &= e^{-\frac{2}{k}I(T;X^{(k)})+\frac{2}{k}h(X^{(k)})}|D|^{\frac{2}{k}} + (2\pi e)|N|^{\frac{2}{k}} \\ &= (2\pi e)e^{-\frac{2}{k}I(T;X^{(k)})}|D|^{\frac{2}{k}} + (2\pi e)|N|^{\frac{2}{k}} \\ &= (2\pi e)e^{-\frac{2}{k}I(T;X^{(k)})}\prod_{i=1}^{k}(1-\lambda_i)^{\frac{1}{k}} + (2\pi e)\prod_{i=1}^{k}\lambda_i^{\frac{1}{k}} \quad . \end{aligned} \tag{A.65}$$

Taking the log, $\frac{2}{k}h(Y^{(k)}|T) \geq \log 2\pi e + \log\left(e^{-\frac{2}{k}I(X^{(k)};T)}\prod(1-\lambda_i)^{\frac{1}{d}} + \prod\lambda_i^{\frac{1}{d}}\right)$ and subtracting $\frac{2}{k}h(Y^{(k)}) = \log(2\pi e)$, we obtain

$$-\frac{2}{k}I(Y^{(k)};T) \geq \log\left(e^{-\frac{2}{k}I(T;X^{(k)})}\prod_{i=1}^{k}(1-\lambda_i)^{\frac{1}{k}} + \prod_{i=1}^{k}\lambda_i^{\frac{1}{k}}\right) \ . \tag{A.66}$$

Substituting Equation A.66 into A.62, we finally obtain

$$\begin{aligned} I(T; X^{(d)}|Y) &= \frac{k}{2}\left(\frac{2}{k}I(T;X^{(k)}) - \frac{2}{k}I(T;Y^{(k)})\right) \\ &\geq \frac{k}{2}\log\left(e^{-\frac{2}{k}I(T;X^{(k)})}\prod_{i=1}^{k}(1-\lambda_i)^{\frac{1}{k}} + \prod_{i=1}^{k}\lambda_i^{\frac{1}{k}}\right) + \frac{k}{2}e^{\log\frac{2}{k}I(T;X^{(k)})} \\ &= \frac{k}{2}\log\left(\prod_{i=1}^{k}(1-\lambda_i)^{\frac{1}{k}} + e^{\frac{2}{k}I(T;X^{(k)})}\prod_{i=1}^{k}\lambda_i^{\frac{1}{k}}\right). \end{aligned} \tag{A.67}$$

Proof of Lemma A.4.8

We first wish to minimize the function

$$F(a) = f_1(\tilde{c} - a) + f_2(a) \tag{A.68}$$

w.r.t the variable a.

In what follows, we show that $F(a)$ is monotonously increasing and therefore has its minimum at $a = 0$. The derivative of $F(a)$ w.r.t. a is given by

$$\begin{aligned}
\frac{d}{da}F(a) &= \frac{e^{\frac{2a}{n'_I}}\prod_{i=1}^{n'_I}\lambda'^{\frac{1}{k}}_i}{\prod_{i=1}^{n'_I}(1-\lambda'_i)^{\frac{1}{n'_I}} + e^{\frac{2a}{n'_I}}\prod_{i=1}^{n'_I}\lambda'^{\frac{1}{k}}_i} - \frac{e^{\frac{2(\tilde{c}-a)}{n_y-k}}\prod_{i=1}^{n_y-k}\lambda_i^{\frac{1}{n_y-k}}}{\prod_{i=1}^{n_y-k}(1-\lambda_i)^{\frac{1}{n_y-k}} + e^{\frac{2(\tilde{c}-a)}{n_y-k}}\prod_{i=1}^{n_y-k}\lambda_i^{\frac{1}{n_y-k}}} \\
&= \frac{1}{e^{-\frac{2a}{n'_I}}\prod_{i=1}^{n'_I}\left(\frac{1-\lambda'_i}{\lambda'_i}\right)^{\frac{1}{n'_I}} + 1} - \frac{1}{e^{-\frac{2(\tilde{c}-a)}{n_y-k}}\prod_{i=1}^{n_y-k}\left(\frac{1-\lambda_i}{\lambda_i}\right)^{\frac{1}{n_y-k}} + 1} \\
&= e^{-\frac{2(\tilde{c}-a)}{n_y-k}}\prod_{i=1}^{n_y-k}\left(\frac{1-\lambda_i}{\lambda_i}\right)^{\frac{1}{n_y-k}} - e^{-\frac{2a}{n'_I}}\prod_{i=1}^{n'_I}\left(\frac{1-\lambda'_i}{\lambda'_i}\right)^{\frac{1}{n'_I}}
\end{aligned}$$

where the last equality is up to a multiplicative positive constant.

Showing that $\frac{d}{da}F(a) > 0$ is equivalent to showing

$$\frac{e^{-\frac{2(\tilde{c}-a)}{n_y-k}}\prod_{i=1}^{n_y-k}\left(\frac{1-\lambda_i}{\lambda_i}\right)^{\frac{1}{n_y-k}}}{e^{-\frac{2a}{n'_I}}\prod_{i=1}^{n'_I}\left(\frac{1-\lambda'_i}{\lambda'_i}\right)^{\frac{1}{n'_I}}} \geq 1 \tag{A.69}$$

We develop this inequality and show that it must be true. Isolating a:

$$e^{2a(\frac{1}{n_y-k}+\frac{1}{n'_I})} \geq e^{\frac{2\tilde{c}}{n_y-k}}\prod_{i=1}^{n'_I}\left(\frac{1-\lambda'_i}{\lambda'_i}\right)^{\frac{1}{n'_I}}\prod_{i=1}^{n_y-k}\left(\frac{\lambda_i}{1-\lambda_i}\right)^{\frac{1}{n_y-k}} \tag{A.70}$$

$$\begin{aligned}
2a\frac{n_y-k+n'_I}{n'_I(n_y-k)} &\geq \frac{2\tilde{c}}{n_y-k} + \frac{1}{n'_I}\sum_{i=1}^{n'_I}\log\left(\frac{1-\lambda'_i}{\lambda'_i}\right) + \frac{1}{n_y-k}\sum_{i=1}^{n_y-k}\log\left(\frac{\lambda_i}{1-\lambda_i}\right) \\
a &\geq \frac{\tilde{c}n'_I}{n_y-k+n'_I} + \frac{n_y-k}{2(n_y-k+n'_I)}\sum_{i=1}^{n'_I}\log\left(\frac{1-\lambda'_i}{\lambda'_i}\right) \\
&+ \frac{n'_I}{2(n_y-k+n'_I)}\sum_{i=1}^{n_y-k}\log\left(\frac{\lambda_i}{1-\lambda_i}\right) \\
&= \frac{\tilde{c}n'_I}{n_y-k+n'_I} + \frac{n_y-k}{n_y-k+n'_I}\frac{1}{2}\sum_{i=1}^{n'_I}\log\left(\frac{1-\lambda'_i}{\lambda'_i}\right)\left(\frac{\lambda_{n_y-k+1}}{1-\lambda_{n_y-k+1}}\right) \\
&+ \frac{n'_I}{n_y-k+n'_I}\frac{1}{2}\sum_{i=1}^{n_y-k}\log\left(\frac{\lambda_i}{1-\lambda_i}\right)\left(\frac{1-\lambda_{n_y-k+1}}{\lambda_{n_y-k+1}}\right) \\
&= \frac{\tilde{c}n'_I}{n_y-k+n'_I} + \frac{n_y-k}{n_y-k+n'_I}\frac{1}{2}\sum_{i=1}^{n'_I}\log\left(\frac{1-\lambda'_i}{\lambda'_i}\right)\left(\frac{\lambda_{n_y-k+1}}{1-\lambda_{n_y-k+1}}\right) \\
&- \frac{n'_I}{n_y-k+n'_I}\hat{c}(n_y-k+1)
\end{aligned} \tag{A.71}$$

Note that the second term above is negative since λ_{n_y-k+1} is smaller than all the other λ in the sum, and $x/(1-x)$ is increasing. So every item in the log is smaller than 1. Thus, if we satisfy the inequality without the second term, we will surely satisfy it with the second term. The inequality thus becomes

$$\begin{aligned} a &\geq \frac{\tilde{c}n'_I}{n_y-k+n'_I} - \frac{n'_I}{n_y-k+n'_I}\hat{c}(n_y-k+1) \qquad (A.72) \\ &= \frac{n'_I}{n_y-k+n'_I}\left(\tilde{c}-\hat{c}(n_y-k+1)\right) \ . \end{aligned}$$

The last expression is negative, since we assumed that $\hat{n}(I(T;X)) = n_y - k$ and thus $\tilde{c} \leq I(T;X) < \hat{c}(n_y-k+1)$. Since $a \geq 0$ by definition, the last inequality is always satisfied, which proves that $F(a)$ is increasing.

The minimum of $F(a)$ is thus achieved at $a=0$, and its value is $F(0) = f_1(\tilde{c})$ since $f_2(0) = 0$. The overall minimization thus becomes

$$\min_{\substack{0 \leq \tilde{c} \leq I(T;X) \\ 0 \leq a \leq \tilde{c}}} f_1(\tilde{c}-a) + f_2(a) = \min_{0 \leq \tilde{c} \leq I(T;X)} f_1(\tilde{c}) = f_1(I(T;X)) \ ,$$

where the last transition follows from $f_1(\tilde{c})$ being a decreasing function.

Proof of Lemma A.4.3

We first observe that if v is a left eigenvector of $\Sigma_x^{-\frac{1}{2}}\Sigma_{x|y}\Sigma_x^{-\frac{1}{2}}$ with eigenvalue λ then $v\Sigma_x^{-\frac{1}{2}}$ is a left eigenvector of $\Sigma_{x|y}\Sigma_x^{-1}$ with the same eigenvalue. This can be seen from

$$\begin{aligned} v\Sigma_x^{-\frac{1}{2}}\Sigma_{x|y}\Sigma_x^{-\frac{1}{2}} &= \lambda v \\ \left(v\Sigma_x^{-\frac{1}{2}}\right)\Sigma_{x|y}\Sigma_x^{-1} &= \lambda v\Sigma_x^{-\frac{1}{2}} \end{aligned}$$

Since $\Sigma_x^{-\frac{1}{2}}\Sigma_{x|y}\Sigma_x^{-\frac{1}{2}}$ is a symmetric matrix, it has an orthonormal (and therefore independent) set of eigenvectors, which we denote by C_x. We have just shown that $A_x = C_x\Sigma_x^{-\frac{1}{2}}$ which implies that A_x is full rank, thus proving Section 1 of the lemma. We similarly denote D_y the eigenvector matrix of $\Sigma_y^{-\frac{1}{2}}\Sigma_{y|x}\Sigma_y^{-\frac{1}{2}}$ and obtain a similar result for B_y.

To show that the covariance of $\tilde{X}$ is sphered (Section 2 of the Lemma)

$$\Sigma_{\tilde{x}} = A_x\Sigma_x A_x^T = C_x\Sigma_x^{-\frac{1}{2}}\Sigma_x\Sigma_x^{-\frac{1}{2}}C_x^T = C_xC_x^T = I \ ,$$

and one can obtain $\Sigma_{\tilde{y}} = B_y\Sigma_y B_y^T = I$ in a similar fashion.

To study the covariance between $\tilde{X}$ and $\tilde{Y}$ recall from Lemma A.4.2 that B_y can be obtained from A_x via a transformation $B_y = RA_x\Sigma_{xy}\Sigma_y^{-1}$ where R is a scaling matrix,

of size (n_y, n_x) with non-zero elements only on the diagonal. To find R we use the fact that $B_y \Sigma_y B_y^T = I$.

$$\begin{aligned} I = B_y \Sigma_y B_y^T &= R A_x \Sigma_{xy} \Sigma_y^{-1} \Sigma_y \Sigma_y^{-1} \Sigma_{yx} A_x^T R^T \\ &= R A_x \left(\Sigma_{xy} \Sigma_y^{-1} \Sigma_{yx} \Sigma_x^{-1} \right) \Sigma_x A_x^T R^T \\ &= R(I - D) A_x \Sigma_x A_x^T R \\ &= R(I - D) R^T \ . \end{aligned}$$

Thus $R_{ii} = (1 - \lambda_i)^{-\frac{1}{2}}$. We can now find the covariance $\Sigma_{\tilde{x}\tilde{y}}$

$$\begin{aligned} \Sigma_{\tilde{x}\tilde{y}} &= A_x \Sigma_{xy} B_y^T = A_x \Sigma_{xy} \Sigma_y^{-1} \Sigma_{yx} A_x^T R^T \\ &= A_x \left(\Sigma_{xy} \Sigma_y^{-1} \Sigma_{yx} \Sigma_x^{-1} \right) \Sigma_x A_x^T R^T \\ &= (I - D) A_x \Sigma_x A_x^T R^T \\ &= (I - D) A_x \Sigma_x A_x^T R^T \\ &= (I - D) R^T \ . \end{aligned}$$

which implies that the $(i, j)^t h$ element of $\Sigma_{\tilde{x}\tilde{y}}$ is $\sqrt{1 - \lambda_i} \delta_{ij}$, as in Lemma A.4.3.

Appendix B

Table of Symbols

MI	Mutual Information
MinMI	Minimum Mutual Information
MaxEnt	Maximum Entropy
CI	Conditionally Independent (i.e., $p(x_1, \ldots, x_n\|y) = \prod_i p(x_i\|y)$)
SDR	Sufficient Dimensionality Reduction
SDR-IS	SDR with Irrelevance Statistics
IB	Information Bottleneck
GIB	Gaussian Information Bottleneck
CCA	Canonical Correlation Analysis
X, Y	Random variables indicating respect.: Features,Classes (in machine learning) or Response,Stimulus (in neural coding)
$\|X\|$	The number of different values the random variable X may take
$p(x, y)$	Joint distribution of the true system
$\bar{p}, \hat{p}$	Empirical and model distributions
$H(X)$	The entropy of a discrete variable X
$I(X;Y)$	The MI between two variables X and Y
$I[p(x, y)]$	The MI between variables with a joint distribution $p(x, y)$
$D_{KL}[p(x)\|q(x)]$	The KL divergence between the distributions $p(x)$ and $q(x)$
$\vec{\phi}(x)$	A vector of features of X
$\langle f(x) \rangle_{p(x)}$	The expected value of the function $f(x)$ w.r.t the distribution $p(x)$
$\mathcal{P}_x(\vec{\phi}(x), \vec{a})$	The set of distributions over X such that the expected value of $\vec{\phi}(x)$ is $\vec{a}$
$\mathcal{P}_x(\vec{\phi}(x), \vec{a}, \vec{\beta})$	Same as $\mathcal{P}_x(\vec{\phi}(x), \vec{a})$, but with the expected value in the range $\vec{a} \pm \vec{\beta}$
$\mathcal{P}_{x\|y}\big(\vec{\phi}(x), \vec{a}(y), \bar{p}(y)\big)$	The set of distributions $p(x, y)$ such that the conditional expected value of $\vec{\phi}(x)$ is $\vec{a}(y)$, and whose marginal is $\bar{p}(y)$
$\mathcal{F}_{xy}(\vec{\phi}(x), \bar{p})$	The set of distributions $p(x, y)$ that agree with $\bar{p}(x, y)$ on the expected values of $\vec{\phi}(x)$ and on its marginals
$\mathcal{F}_x(\vec{\phi}(x), \bar{p})$	The set of distributions $p(x)$ that agree with $\bar{p}(x)$ on the expected values of $\vec{\phi}(x)$
$I_{min}\left[\vec{\phi}(x), \vec{a}(y), \bar{p}(y)\right]$	Minimum information in the set $\mathcal{P}_{x\|y}\big(\vec{\phi}(x), \vec{a}(y), \bar{p}(y)\big)$
$I^{xy}_{min}[\vec{\phi}(x), \bar{p}]$	Minimum information in the set $\mathcal{P}(\vec{\phi}(x), \bar{p})$
$I^{(k)}$	Minimum information in the set of k^{th} order marginals
$IPR(q, \mathcal{F})$	The *I-projection* of the distribution $q(x)$ on the set of distributions $\mathcal{F}$

Bibliography

[1] M. Abeles. Time is precious. *Science*, 304(5670):523–524, 2004.

[2] S. Amari and N. Nagaoka. *Methods of Information Geometry*. Oxford University Press, 2000.

[3] H. Antti, E. Holmes, and J. Nicholson. Multivariate solutions to metabonomic profiling and functional genomics. part 2 - chemometric analysis, 2002. http://www.acc.umu.se/ tnkjtg/chemometrics/editorial/oct2002.

[4] H.B. Barlow. Sensory mechanisms, the reduction of redundancy, and intelligence. In *Mechanisation of thought processes*, pages 535–539. Her Majesty's stationary office, London, 1959.

[5] A. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Trans. Inform. Theory*, IT-44(6):2743–2760, 1998.

[6] M.P. Becker and C.C. Clogg. Analysis of sets of two-way contingency tables using association models. *J. Amer. Statist. Assoc.*, 84(405):142–151, 1989.

[7] S. Becker. Mutual information maximization: Models of cortical self-organization. *Network: Computation in Neural Systems*, 7:7–31, 1996.

[8] S. Becker and G.E. Hinton. A self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355(6356):161–163, 1992.

[9] P. N. Belhumeur, J. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Trans. Pattern Anal. Machine Intell.*, 19(7):711–720, 1997.

[10] A.J. Bell and T.J. Sejnowski. An information maximization approach to blind seperation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995.

[11] A.J. Bell and T.J. Sejnowski. The independent components of natural scenes are edge filters. *Vision Res.*, 37(23):3327–3338, 1997.

[12] A.L. Berger, S.A. Della-Pietra, and V.J. Della-Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71, 1996.

[13] T. Berger and R. Zamir. A semi-continuous version of the berger-yeung problem. *IEEE Trans. Inform. Theory*, 45(5):1520–1526, 1999.

[14] W. Bialek, F. Rieke, and R.R de Ruyter van Steveninck. Reading a neural code. *Science*, 252(5014):1854–1857, 1991.

[15] M. Borga. Canonical correlation: a tutorial. http://people.imt.liu.se/magnus/cca, January 2001.

[16] M. Borga, H. Knutsson, and T. Landelius. Learning canonical correlations. In *Proceedings of the 10th Scandinavian Conference on Image Analysis*, Lappeenranta, Finland, June 1997.

[17] A. Borst and F.E. Theunissen. Information theory and neural coding. *Nat. Neurosci.*, 2(11):947–957, 1999.

[18] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge Univ. Press, 2004.

[19] J.P. Boyle and R.L. Dijkstra. A method for finding projections onto the intersections of convex sets in hilbert spaces. In *Lecture Notes in Statistics 37*, pages 28–47. Springer, 1986.

[20] L. M. Bregman. The method of successive projection for finding a common point of convex sets. *Sov. Math. Dokl.*, 6:688–692, 1965.

[21] N. Brenner, S.P. Strong, R. Koberle, R. de Ruyter van Steveninck, and W. Bialek. Synergy in a neural code. *Neural Computation*, 13(7):1531–1552, 2000.

[22] J. K. Chapin, K. A. Moxon, R. S. Markowitz, and M. A. L. Nicolelis. Real-time control of a robot arm using simultaneously recorded neurons in the motor cortex. *Nat. Neurosci.*, 2(7):664–670, 1999.

[23] G. Chechik, A. Globerson, M.J. Anderson, E.D. Young, I. Nelken, and N. Tishby. Group redundancy measures reveal redundancy reduction in the auditory pathway. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, Cambridge, MA, 2002. MIT Press.

[24] G. Chechik, A. Globerson, N. Tishby, and Y. Weiss. Gaussian information bottleneck. *Journal of Machine Learning Research*, 6:165–188, 2005.

[25] G. Chechik and N. Tishby. Extracting relevant structures with side information. In S. Thrun S. Becker and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 857–864, Cambridge, MA, 2003. MIT Press.

[26] M. Chiang and S. Boyd. Geometric programming duals of channel capacity and rate distortion. *IEEE Trans. Inform. Theory*, IT-50(2):245–258, 2004.

[27] T.M. Cover and J.A. Thomas. *The elements of information theory*. Wiley, New York, 1991.

[28] T.M. Cover and J.A Thomas. *Elements of information theory*. Wiley, 1991.

[29] T. Cox and M. Cox. *Multidimensional Scaling*. Chapman and Hall, 1994.

[30] I. Csiszar. I-divergence geometry of probability distributions and minimization problems. *Ann. Probab.*, 3(1):146–158, 1975.

[31] Y. Dan, J.M. Alonso, W.M. Usrey, and R.C. Reid. Coding of visual information by precisely correlated spikes in the lateral geniculate nucleus. *Nat. Neurosci.*, 1(6):501–507, 1998.

[32] J.N. Darroch and D. Ratcliff. Generalized iterative scaling for log-linear models. *Ann. Math. Statist.*, 43(5):1470–1480, 1972.

[33] S.C. Deerwester, S.T. Dumais, T.K. Landauer, G.W. Furnas, and R.A. Harshman. Indexing by latent semantic analysis. *J. Amer. Soc. Inform. Sci.*, 41(6):391–407, 1990.

[34] S.A. Della-Pietra, V.J. Della-Pietra, and J.D. Lafferty. Inducing features of random fields. *IEEE Trans. Pattern Anal. Machine Intell.*, 19(4):380–393, 1997.

[35] J.W. Demmel. *Applied numerical linear algebra*. Society for Industrial and Applied Mathematics, Philadelphia, 1997.

[36] L. Devroye, L. Gyorfi, and G. Lugosi. *A Probabilistic Thoery of Pattern Recognition*. Springer, 1996.

[37] K.I. Diamantaras and S.Y. Kung. *Principal Component Neural Networks: Theory and Applications*. John Wiley, 1996.

[38] A.G. Dimitrov and J.P. Miller. Neural coding and decoding: Communication channels and quantization. *Network: Computation in Neural Systems*, 12(4):441–472, 2001.

[39] J. P. Donoghue. Connecting cortex to machines: recent advances in brain interfaces. *Nat. Neurosci.*, 5:1085–1088, 2002.

[40] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience Publication, 2000.

[41] M. Dudik, S.J. Phillips, and R. E. Schapire. Performance guarantees for regularized maximum entropy density estimation. In J.Shawe-Taylor and Y. Singer, editors, *Proceedings of the 17th Annual Conference on Computational Learning Theory*, pages 472–486. Springer, 2004.

[42] R.A. Fisher. *Statistical Methods and Scientific Inference*. Edinburgh: Oliver and Boyd, 1956.

[43] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29(2-3):131–163, 1997.

[44] N. Friedman, O. Mosenzon, N. Slonim, and N. Tishby. Multivariate information bottleneck. In J.S. Breese and D. Koller, editors, *Proceedings of the 17th Annual Conference on Uncertainty in Artificial Intelligence (UAI-01)*, pages 152–161, San Francisco, CA, 2001. Morgan Kaufmann.

[45] O. Friman, M. Borga, P. Lundberg, and H. Knutsson. Adaptive analysis of fMRI data. *NeuroImage*, 19(3):837–845, 2003.

[46] R.G. Gallager. *Information Theory and Reliable Communication*. John Wiley and Sons, 1968.

[47] A.P. Gasch, P.T. Spellman, C.M. Kao, O. Carmel-Harel, M.B. Eisen, G. Storz, D. Botstein D, and P.O. Brown. Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell.*, 11(12):4241–4257, 2000.

[48] M. Gastpar, B. Rimoldi, and M. Vetterli. To code, or not to code: lossy source-channel communication revisited. *IEEE Trans. Inform. Theory*, IT-49(5):1147–1158, 2003.

[49] T.J. Gawne and B.J. Richmond. How independent are the messages carried by adjacent inferior temporal cortical neurons? *J Neurosci.*, 13(7):2758–2771, 1993.

[50] D. Geiger, A. Rudra, and L.T. Maloney. Features as sufficient statistics. In M.I. Jordan, M.J. Kearns, and S.A. Solla, editors, *Advances in Neural Information Processing Systems 10*, pages 794–800, Cambridge, MA, 1998. MIT Press.

[51] A.P. Georgopoulos, R.E. Kettner, and A.B. Schwartz. Primate motor cortex and free arm movements to visual targets in three dimensional space. i. relations between single cell discharge and direction of movement. *J. Neurosci.*, 8(8):2913–2927, 1988.

[52] R. Gilad-Bachrach, A. Navot, and N. Tishby. An information theoretic tradeoff between complexity and accuracy. In B. Schölkopf and M. K. Warmuth, editors, *Proceedings of the 16th Annual Conference on Computational Learning Theory*, pages 595–609. Springer, 2003.

[53] A. Globerson, G. Chechik, F. Pereira, and N.Tishby. Euclidean embedding of co-occurrence data. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 497–504, Cambridge, MA, 2005. MIT Press.

[54] A. Globerson, G. Chechik, and N. Tishby. Sufficient dimensionality redcution with side information. In C. Meek and U. Kjaerulff, editors, *Proceedings of the 19th Annual Conference on Uncertainty in Artificial Intelligence (UAI-03)*, pages 281–288. Morgan Kaufmann, San Francisco, CA, 2003.

[55] A. Globerson and N.Tishby. The minimum information principle in discriminative learning. In M. Chickering and J. Halpern, editors, *Proceedings of the 20th Annual Conference on Uncertainty in Artificial Intelligence (UAI-04)*, pages 193–200, Arlington, Virginia, 2004. AUAI Press.

[56] A. Globerson and N. Tishby. Sufficient dimensionality reduction. *Journal of Machine Learning Research*, 3:1307–1331, 2003.

[57] A. Globerson and N. Tishby. On the optimality of the gaussian information bottleneck curve. Technical report, Hebrew University, January 2004.

[58] G.H. Golub and C.F.V. Loan. *Matrix Computations*. The Johns Hopkins University Press, Baltimore, 1989.

[59] J. Gondzio, O. du Merle, R. Sarkissian, and J.P. Vial. Accpm - a library for convex optimization based on an analytic center cutting plane method. *European Journal of Operational Research*, 94:206–211, 1996.

[60] L.A. Goodman. Simple models for the analysis of association in cross-classifications having ordered categories. *J. Amer. Statist. Assoc.*, 74(367):537–552, 1979.

[61] L.A. Goodman. The analysis of cross-classified data having ordered and/or unordered categories: Association models, correlation models, and asymmetry models for contingency tables with or without missing entries. *Ann. Statist.*, 13(1):10–69, 1985.

[62] V.K. Goyal. Theoretical foundations of transform coding. *Signal Processing Magazine, IEEE*, 18(5):9–21, 2001.

[63] D. Grünwald and A.P. Dawid. Game theory, maximum entropy, minimum discrepancy, and robust bayesian decision theory. *Ann. Statist.*, 32:1367–1433, 2004.

[64] T. Hastie, R. Tibshirani, and J.H. Friedman. *The elements of statistical learning: data mining, inference, and pr ediction*. New York: Springer-Verlag, 2001.

[65] N.G. Hatsopoulos, C.L. Ojakangas, L. Paninski, and J.P. Donoghue. Information about movement direction obtained from synchronous activity of motor cortical neurons. *Proc. Natl. Acad. Sci. USA*, 95:15706–15711, 1998.

[66] D. Haussler and M. Opper. General bounds on the mutual information between a parameter and n conditionally independent observations. In *Proceedings of the 8th Annual Conference on Computational Learning Theory*, pages 402–411. 1995.

[67] M.E. Hellman and J. Raviv. Probability of error, equivocation, and the chernoff bound. *IEEE Trans. Inform. Theory*, IT-16(4):368–372, 1970.

[68] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1):177–196, 2001.

[69] H. Hotelling. The most predictable criterion. *Journal of Educational Psychology,*, 26:139–142, 1935.

[70] K. Huang, H. Yang, I. King, M.R. Lyu, and L. Chan. The minimum error minimax probability machine. *Journal of Machine Learning Research*, 5:1253–1286, 2004.

[71] T Jaakkola, M. Meila, and T. Jebara. Maximum entropy discrimination. In S. A. Solla, T. K. Leen, and K.-R. Muller, editors, *Advances in Neural Information Processing Systems 12*, pages 470–477, Cambridge, MA, 2000. MIT Press.

[72] E. T. Jaynes. On the rational of maximum entropy methods. *Proceedings of the IEEE*, 70:939–952, 1982.

[73] E.T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106:620–630, 1957.

[74] A. Jennings and G.W. Stewart. Simultaneous iteration for partial eigensolution of real matrices. *J. Inst. Math Appl*, 15:351–361, 1975.

[75] T. Joachims. Text categorization with support vector machines: learning with many relevant features. In C. Nedellec and C. Rouveirol, editors, *Proceedings of the 10th European Conference on Machine Learning*, pages 137–142. Springer-Verlag, 1998.

[76] M.I. Jordan, editor. *Learning in graphical models*. MIT press, Cambridge, MA, 1998.

[77] S.M. Kay. *Fundamentals of Statistical Signal Processing. Volume I, Estimation Theory*. Prentice-Hall, 1993.

[78] M. Kearns. Efficient noise-tolerant learning from statistical queries. In *Proceedings of the twenty-fifth annual ACM symposium on Theory of computing*, pages 392–401, 1993.

[79] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In A.P. Danyluk C.E. Brodley, editor, *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289. Morgan Kaufmann, 2001.

[80] G. Lanckriet, L. Ghaoui, C. Bhattacharyya, and M.I. Jordan. A robust minimax approach to classification. *Journal of Machine Learning Research*, 3:555–582, 2002.

[81] K. Lang. Newsweeder: Learning to filter news. In S.J. Russell A. Prieditis, editor, *Proceedings of the 12th International Conference on Machine Learning*, pages 331–339. Morgan Kaufmann, 1995.

[82] A. Lapidoth and P. Narayan. Reliable communication under channel uncertainty. *IEEE Trans. Inform. Theory*, IT-44(6):2148–2177, 1998.

[83] G. Lebanon and J. Lafferty. Boosting and maximum likelihood for exponential models. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 447–454, Cambridge, MA, 2002. MIT Press.

[84] D. Lee and H. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[85] R. Linsker. Self-organization in a perceptual network. *IEEE Computer*, 21(3):105–117, 1988.

[86] R. Linsker. Local synaptic learning rules suffice to maximize mutual information in a linear network. *Neural Computation*, 4(5):691–702, 1992.

[87] J.R. Magnus and H. Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. John Wiley and Sons, 1988.

[88] R. Malouf. A comparison of algorithms for maximum entropy parameter estimation. In *Proceedings of the 6th Conference on Natural Language Learning*, pages 49–55, 2002.

[89] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts, 1999.

[90] L. Martignon, G. Deco, K.B. Laskey, M.E. Diamond, W. Freiwald, and E. Vaadia. Neural coding: Higher-order temporal patterns in the neurostatistics of cell assemblies. *Neural Computation*, 12(11):2621–2653.

[91] A.M. Martinez and R. Benavente. The AR face database. Technical Report 24, CVC, June 1998.

[92] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, A. Smola, and K. Müller. Invariant feature extraction and classification in kernel spaces. In S.A. Solla, T.K. Leen, and K.R. Muller, editors, *Advances in Neural Information Processing Systems 12*, pages 526–532, Cambridge, MA, 2000. MIT Press.

[93] G. A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, 63(2):81–97, 1956.

[94] P. Murphy and D. Aha. Uci repository of machine learning databases (machine readable data repository. *UCI Dept. of Info. and Comp. Sci.*, 1998.

[95] N.S. Narayanan, E.Y. Kimchi, and M. Laubach. Redundancy and synergy of neuronal ensembles in motor cortex. *J. Neurosci.*, 25(17):4207–4216, 2005.

[96] I. Nelken, G. Chechik, T.D. MrsicFlogel, A.J. King, and J.W.H. Schupp. Encoding stimulus information by spike numbers and mean response time in primary auditory cortex. *J. Computational Neurosci., In Press*, 2005.

[97] A.Y. Ng and M.I. Jordan. On discriminative vs. generative classifiers. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 605–610, Cambridge, MA, 2002. MIT Press.

[98] K. Nigam, J. Lafferty, and A. McCallum. Using maximum entropy for text classification. In *IJCAI-99 Workshop on Machine Learning for Information Filtering*, pages 61–67, 1999.

[99] S. Nirenberg, S.M. Carcieri, A.L. Jacobs, and P.E. Latham. Retinal ganglion cells act largely as independent encoders. *Nature*, 411(6838):698–701, 2001.

[100] S. Nirenberg and P.E. Latham. Decoding neuronal spike trains: how important are correlations? *Proc. Natl. Acad. Sci.*, 100:7348–7353, 2003.

[101] L.M. Optican and B.J. Richmond. Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex III. Information theoretic analysis. *J. Neurophysiol.*, 57(1):162–177, 1987.

[102] L.C. Osborne, W. Bialek, and S.G. Lisberger. Time course of information about motion direction in visual area MT of macaque monkeys. *J. Neurosci.*, 24(13):3210–3222, 2004.

[103] L. Paninski. Estimation of entropy and mutual information. *Neural Computation.*, 15:1191–1254, 2003.

[104] S. Panzeri, , S. R. Schultz, A. Treves, and E.T. Rolls. Correlations and the encoding of information in the nervous system. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 266:1001–1012., 1999.

[105] S. Panzeri and A. Treves. Analytical estimates of limited sampling biases in different information measures. *Network: Computation in Neural Systems*, 7:87–107, 1996.

[106] J. Park and A. Darwiche. Approximating map using stochastic local search. In J. Breese and D. Koller, editors, *Proceedings of the 17th Annual Conference on Uncertainty in Artificial Intelligence (UAI-01)*, pages 403–410, San Francisco, CA, 2001. Morgan Kaufmann.

[107] R. Paz, T. Boraud, C. Natan, H. Bergman, and E. Vaadia. Preparatory activity in motor cortex reflects learning of local visuomotor skills. *Nat. Neurosci.*, 6(8):882–890, 2003.

[108] D.H. Perkel, G.L. Gerstein, and G.P. Moore. Neuronal spike trains and stochastic point processes. I. The single spike train. *Biophys. J.*, 7(4):391–418, 1967.

[109] S.J. Phillips, M. Dudik, and R. E. Schapire. A maximum entropy approach to species distribution modeling. In C.E. Brodley, editor, *Proceedings of the 21st International Conference on Machine Learning*, pages 655–662. Morgan Kaufmann, 2004.

[110] E. Pitman. Sufficient statistics and intrinsic accuracy. *Proc. of the Cambridge Phil. Soc.*, 32:567–579, 1936.

[111] S.S. Pradhan. On rate-distortion function of gaussian sources with memory with side information at the decoder. Technical report, Berkeley, 1998.

[112] S.S. Pradhan, J. Chou, and K. Ramchandran. Duality between source coding and channel coding and its extension to the side information case. *IEEE Trans. Inform. Theory*, 49(5):1181–1203, 2003.

[113] J. Puchalla, E. Schneidman, R.A. Harris, and M.J. Berry. Redundancy in the population code of the retina. *Neuron*, 46:493–504, 2005.

[114] L. Rabiner and B.H. Juang. *Fundumentals of speech recognition*. Prentice Hall, 1993.

[115] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C.H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J.P. Mesirov, T. Poggio, W. Gerald, M. Loda, E.S. Lander, and T.R. Golub. Multiclass cancer diagnosis using tumor gene expression signatures. *PNAS*, 98(26), 2001.

[116] F. Rieke, D. Warland, R.R de Ruyter van Steveninck, and W. Bialek. *Spikes*. MIT Press, 1997.

[117] M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nat. Neurosci.*, 2(11):1019–1025, 1999.

[118] S.T. Roweis and L.K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–6, 2000.

[119] S.T. Roweis, L.K. Saul, and G. Hinton. Global coordination of local linear models. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 889–896, Cambridge, MA, 2002. MIT Press.

[120] I. Samengo. Independent neurons representing a finite set of stimuli: dependence of the mutual information on the number of units sampled. *Network: Computation in Neural Systems*, 12(1):21–31, 2001.

[121] E. Schneidman, W. Bialek, and M.J. Berry. Synergy, redundancy, and independence in population codes. *J. Neurosci.*, 23(37):11539–11553, 2003.

[122] E. Schneidman, S. Still, M.J. Berry, and W. Bialek. Network information and connected correlations. *Physical Review Letters*, 91:238701, 2003.

[123] C.E. Shannon. A mathematical theory of communication. *The Bell systems technical journal*, 27:379–423,623–656, 1948.

[124] C.E. Shannon. Coding theorems for a discrete source with a fidelity criterion. In *Institute for Radio Engineers, International Convention Record*, volume 7, part 4, pages 142–163, New York, NY, USA, March 1959.

[125] N. Shental, T. Hertz, D. Weinshall, and M. Pavel. Adjustment learning and relevant component analysis. In A. Heyden, G. Sparr, M. Nielsen, and P. Johansen, editors, *Proceedings of the 7th European Conference of Computer Vision*, pages 776–792. Springer, 2002.

[126] J. Sinkkonen and S. Kaski. Clustering based on conditional distribution in an auxiliary space. *Neural Computation*, 14(1):217–239, 2001.

[127] N. Slonim. *Information Bottlneck theory and applications*. PhD thesis, Hebrew University of Jerusalem, 2003.

[128] N. Slonim and N. Tishby. Agglomerative information bottleneck. In S. A. Solla, T. K. Leen, and K.-R. Muller, editors, *Advances in Neural Information Processing Systems 12*, pages 612–623. MIT Press, Cambridge, MA, 2000.

[129] N. Slonim and N. Tishby. Document clustering using word clusters via the information bottleneck method. In P. Ingwersen N.J. Belkin and M-K. Leong, editors, *Research and Development in Information Retrieval (SIGIR)*, pages 208–215. ACM press, New York, 2000.

[130] N. Slonim and Y. Weiss. Maximum likelihood and the information bottleneck. In S. Thrun S. Becker and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 335–342, Cambridge, MA, 2003. MIT Press.

[131] M. Studenty and J. Vejnarova. The multiinformation function as a tool for measuring stochastic dependence. In M.I. Jordan, editor, *Learning in Graphical Models*, pages 261–297. MIT press, Cambridge, MA, 1998.

[132] J.B. Tenenbaum, V. de Silva, and J.C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[133] J.B. Tenenbaum and W.T. Freeman. Separating style and content with bilinear models. *Neural Computation*, 12(6):1247–1283, 2000.

[134] B. Thompson. *Canonical correlation analysis: Uses and interpretation.*, volume 47. Thousands Oak, CA Sage publications, 1984.

[135] N. Tishby, F.C. Pereira, and W. Bialek. The information bottleneck method. In B. Hajek and R.S. Sreenivas, editors, *Proc. of 37th Allerton Conference on communication and computation*, pages 368–377, 1999.

[136] K. Torkkola. Feature extraction by non-parametric mutual information maximization. *Journal of Machine Learning Research*, 3:1415–1438, 2003.

[137] E. Vaadia, I. Haalman, M. Abeles, H. Bergman, Y. Prut, H. Slovin, and A. Aertsen. Dynamics of neuronal interactions in monkey cortex in relation to behavioral events. *Nature*, 373(6514):515–518, 1995.

[138] V. Vapnik. *Statistical Learning Theory*. John Wiley and Sons, 1998.

[139] H. von Storch and F.W. Zwiers. *Statistical Analysis in Climate Research*. Cambridge University Press, 1999.

[140] J. Wessberg, C.R. Stambaugh, J.D. Kralik, P.D. Beck, M. Laubach, J.K. Chapin, J. Kim, S.J. Biggs, M.A. Srinivasan, and M.A. Nicolelis. Real-time prediction of hand trajectory by ensembles of cortical neurons in primates. *Nature*, 408(6810):305–306.

[141] D.R. Wolf and E.I. George. Maximally informative statistics. *Revista de la Real Academia de Ciencias, Special edition on Bayesian Statistics*, 93(3):381–386, 1999.

[142] D.H. Wolpert. On the connection between in-sample testing and generalization error. *Complex Systems*, 6:47–94, 1992.

[143] A.D. Wyner. On source coding with side information at the decoder. *IEEE Trans. Inform. Theory*, IT-21(3):294–300, 1975.

[144] A.D. Wyner. The rate distortion function for source coding with side information at the decoder II: General sources. *Information and Control*, 38:60–80, 1978.

[145] E.P. Xing, A.Y. Ng, M.I. Jordan, and S. Russell. Distance metric learning with application to clustering with side-information. In S. Thrun S. Becker and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 505–512, Cambridge, MA, 2003. MIT Press.

[146] Y. Yang and X. Liu. A re-examination of text categorization methods. In *Research and Development in Information Retrieval (SIGIR)*, pages 42–49. ACM press, New York, 1999.

[147] J.S. Yedidia, W.T. Freeman, and Y. Weiss. Understanding belief propagation and its generalizations. In *IJCAI 2001 Distinguished Lecture track*.

[148] S.C. Zhu, Z.N. Wu, and D. Mumford. Minimax entropy principle and its application to texture modeling. *Neural Computation*, 9(8):1627–1660, 1997.