Variational Information Maximization in Stochastic Environments

Felix Agakov

Doctor of Philosophy
Institute for Adaptive and Neural Computation
School of Informatics
University of Edinburgh
2005

Abstract

Information maximization is a common framework of unsupervised learning, which may be used for extracting informative representations y of the observed patterns x. The key idea there is to maximize mutual information (MI), which is a formal measure of coding efficiency. Unfortunately, exact maximization of MI is computationally tractable only in a few special cases; more generally, approximations need to be considered. Here we describe a family of variational lower bounds on mutual information which gives rise to a formal and theoretically rigorous approach to information maximization in large-scale stochastic channels. We hope that the results presented in this work are potentially interesting for maximizing mutual information from several perspectives. First of all, our method optimizes a proper lower bound, rather than a surrogate objective criterion or an approximation of MI (which may only be accurate under specific asymptotic assumptions, and weak or even undefined when the assumptions are violated). Secondly, the flexibility of the choice of the variational distribution makes it possible to generalize and improve simple bounds on MI. For example, we may introduce tractable *auxiliary variational* bounds on MI, which may be used to improve on any simple generic approach without altering properties of the original channel. Thirdly, the suggested variational framework is typically simpler than standard variational approaches to maximizing the conditional likelihood in stochastic autoencoder models, while it leads to the same fixed points in its simplest formulation; this gives rise to more efficient optimization procedures. Finally, in some cases the variational framework results in optimization procedures which only require local computations, which may be particularly attractive from the neuro-biological perspective. Possibly the most important contribution of this work is a rigorous and general framework for maximizing the mutual information in intrinsically intractable channels. We show that it gives rise to simple, stable, and easily generalizable optimization procedures, which outperform and supersede many of the common approximate information-maximizing techniques. We demonstrate our results by considering clustering, dimensionality reduction, and binary stochastic coding problems, and discuss a link to approximate statistical inference.

Acknowledgements

I thank my teachers for their support and patience. My family for their support and love. And Mikhail Khodorkovsky for his honor and courage.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Felix Agakov*)

Table of Contents

List of Figures

Chapter 1

Introduction

Possibly one of the most important tasks faced by an intelligent organism may be formulated as learning about the environment, and using the knowledge to address challenges offered by the external world. Knowledge about the world is represented internally by means of synaptic structures or neural activations. Generally, it is believed that internal representations are formed in a way which allows them to capture useful information about the external environment. The goal of learning may in this case be formulated as finding informative representations about the regularities occurring in the external world.

Many of machine learning methods address fundamentally similar tasks, where one of the key goals is to find informative representations of the observations about the environment. One obvious application related to modeling of biological systems is learning neural population codes for a set of input stimuli. Some other applications include speech analysis and object tracking, face recognition and automated medical diagnostics, where the extracted representations can be used for making important practical decisions. Unfortunately, the problem of finding internal representations of the environment may become notoriously difficult in the presence of noise. In this thesis we consider a class of machine learning methods making a recourse to information theory, and address the fundamental computational problems of applying the methods for finding informative regularities in stochastic environments.

1.1 Background

In recent years we witnessed a significant transformation of methods aimed at designing intelligent systems. Historically, such methods were based mainly on modeling heuristic expert knowledge rather than the problem domain. While being sometimes efficient for addressing relatively simple problems of local inference, the approaches lacked formal mechanisms for resolving global queries not representable in the basis of specified heuristics and axioms. Moreover, they could not be easily applied in incomplete and uncertain domains, i.e. in the majority of practically interesting and challenging situations (e.g. Jaynes (1996), Jensen (1996)). Late 80s and early 90s were marked by an increased popularity of non-symbolic methods and development of the learning paradigm, where one of the

principal objectives was to define a utility function and optimize its expectation with respect to a (generally, parametric) model of the environment. While definitions of the objective criteria used by the earlier models were not always properly understood, it was soon realized that many of the sensible optimization criteria for *supervised*[1] training of early connectionist models had strong probabilistic foundations (e.g. Neal (1992), Bishop (1995)). The probabilistic view was important, as it helped to shift the heuristics from the level of function definition to the level of model specification, which was arguably easier to understand by the majority of experts. Moreover, it suggested a unifying framework for interpretation and analysis of the regression and classification algorithms.

At the same time, a particular level of attention was paid to the development of *unsupervised*[2] learning techniques (see e.g. Hinton and Sejnowski (1999)), where the idea was to learn statistical models, or extract statistical regularities of the observations (see e.g. Dayan (2001) and references therein for a high-level introduction). Possibly the largest class of the unsupervised methods is based on estimation of probability densities of the observed data, where the fundamental idea is to estimate parameters of a model which would give rise to the observations. Specifically, *maximum likelihood* methods aim at fitting a constrained distribution $p(\mathsf{x}|\mathbf{\Theta})$ to a set of the empirical observations $\{\mathsf{x}\}$ by estimating the deterministic parameters $\mathbf{\Theta}$. One way of introducing constraints on $p(\mathsf{x})$ is to consider a generative *latent variable* model $\mathcal{M}_L = p(\mathsf{x}|\mathsf{y}, \mathbf{\Theta}_{x|y})p(\mathsf{y}|\mathbf{\Theta}_y)$, where y defines a vector of latent variables, $p(\mathsf{y}|\mathbf{\Theta}_y)$ is the marginal distribution of the latent variables parameterized by $\mathbf{\Theta}_y$, and $p(\mathsf{x}|\mathsf{y}, \mathbf{\Theta}_{x|y})$ is the conditional distribution parameterized by $\mathbf{\Theta}_{x|y}$. The goal of learning by maximizing the likelihood would in this case involve fitting $p(\mathsf{x}|\mathbf{\Theta}_y, \mathbf{\Theta}_{x|y}) = \int_{\mathsf{y}} p(\mathsf{y}|\mathbf{\Theta}_y)p(\mathsf{x}|\mathsf{y}, \mathbf{\Theta}_{x|y})\mathrm{d}\mathsf{y}$ to the observed data, which would involve maximization of the likelihood for $\mathbf{\Theta}_y$ and $\mathbf{\Theta}_{x|y}$ (see e.g. Everitt (1984), Hertz et al. (1991), Bishop (1995), Hinton and Sejnowski (1999) for references, comprehensive descriptions of common unsupervised learning techniques, and general discussions of density estimation methods).

Another class of unsupervised learning methods is based on the idea of finding *most informative* descriptions y of the training patterns x, for example by learning the optimal mapping $p(\mathsf{y}|\mathsf{x}, \mathbf{\Theta}_{y|x})$. The coding efficiency is typically quantified by *mutual information* between x and y, which is defined as $I(\mathsf{x}, \mathsf{y}) \stackrel{\text{def}}{=} H(\mathsf{y}) - H(\mathsf{y}|\mathsf{x})$. Here $H(\mathsf{y}|\mathsf{x}) \stackrel{\text{def}}{=} -\int_{\mathsf{x}}\int_{\mathsf{y}} p(\mathsf{y}|\mathsf{x}, \mathbf{\Theta}_{\mathsf{y}|\mathsf{x}})p(\mathsf{x}) \log p(\mathsf{y}|\mathsf{x}, \mathbf{\Theta}_{\mathsf{y}|\mathsf{x}})\mathrm{d}\mathsf{y}\mathrm{d}\mathsf{x}$ is the *conditional entropy* and $H(\mathsf{y}) \stackrel{\text{def}}{=} -\int_{\mathsf{y}} p(\mathsf{y}) \log p(\mathsf{y})\mathrm{d}\mathsf{y}$ is the *marginal entropy*, which for discrete variables may be thought of as measures of uncertainty (see a more detailed discussion in Section 1.3). Additionally, $p(\mathsf{x})$ defines the distribution of the source patterns, and $p(\mathsf{y}) = \int_{\mathsf{x}} p(\mathsf{y}|\mathsf{x}, \mathbf{\Theta}_{\mathsf{y}|\mathsf{x}})p(\mathsf{x})\mathrm{d}\mathsf{x}$ is the marginal distribution of their descriptions. The mutual information may therefore be interpreted as the decrease in uncertainty of the training patterns x given their representations y, which is

[1] In many cases supervised learning may be thought of as an optimization framework for evaluating parameters of the mapping from the inputs to the explicitly defined target outputs.

[2] Unsupervised learning defines an optimization framework for the cases when the observations cannot be represented as associative pairs of inputs and the corresponding target outputs (or environmental values, *cf* reinforcement learning).

one of the fundamental measures of information theory (see e.g. McEliece (1977), Cover and Thomas (1991)). While the idea of maximizing mutual information has been long known in communication and information theory (see e.g. Blahut (1972), Arimoto (1972), McEliece (1977)), it was introduced to the machine learning community relatively recently (Linsker (1988)). Since then it has been used in a variety of contexts, including clustering (e.g Dhillon and Guan (2003), Jenssen et al. (2003)), population coding (e.g. Brunel and Nadal (1998), Stocks and Mannella (2001)), and feature extraction (Principe et al. (2000), Torkkola and Campbell (2000), Gokcay and Principe (2002)). A variety of other unsupervised learning techniques making a recourse to information theory are conceptually different from Linsker's *infomax* (Linsker (1988), Linsker (1989a), Linsker (1989b)), and utilize the mutual information in different contexts. For example, the family of "information bottleneck" methods (Tishby et al. (1999), Slonim et al. (2001), Chechik and Tishby (2002), Dhillon et al. (2002), Still and Bialek (2004)) proposes to maximize the amount of information which the compressed codes of the original data contain about some variables of relevance, while minimizing the amount of information between the data and the codes. The semi-supervised learning methods of Szummer and Jaakkola (2002) and Corduneanu and Jaakkola (2003) suggest to use the local mutual information (computed for each small region of the data space) as a local regularizer of a conditionally trained model. A fundamentally different *Imax* objective criterion (see e.g. Becker (1992), Becker and Hinton (1992)) suggests to optimize the information content between the code variables, rather than between the sources and the codes. The *redundancy reduction* methods (Barlow (1989), Atick (1992), Field (1994)) may also be viewed as specific approximations of the exact mutual information between the source patterns and their encoded representations (Nadal and Parga (1994), Nadal et al. (1998)). These are just a few of the machine learning methods to mention which require computations of the mutual information under various modeling assumptions.

We may generally view these learning methods as optimization procedures, which optimize different objective criteria defined for specific probabilistic models. The models may be conveniently described within the probabilistic *graphical modeling* paradigm, which offers a convenient framework for graphical representations of joint probability distributions via local constraints (Pearl (1988), Lauritzen and Spiegelhalter (1988), Cowell et al. (1999)). The objective functions, including the likelihood and the mutual information, may then be expressed by computing specific expectations for the considered probabilistic models (generally, the objectives may be interpreted as functionals of the joint probability distribution). The probabilistic view offers many advantages, including a formal generalization of inference (performed by the sound rules of probability theory), and theoretically motivated definitions of objective criteria for optimization. Moreover, reliance on the formal calculus of uncertainty makes the graphical models suitable for reasoning in uncertain (e.g. noisy or partially hidden) domains.

Unfortunately, this theoretical rigor comes at the expense of computational complexity of learning, which may be prohibitively high for many problems of interest. This motivates necessity of accurate and computationally tractable

approximations of the optimized criteria. The fundamental cause of computational intractability of rigorous learning methods in most of the high-dimensional densely-connected graphical models is the high complexity of computing the expectations over the hidden variables (typically, this complexity would be low only for a limited number of special cases). For example, fitting the model $p(\mathsf{x}, \mathsf{y})$ to a set of training patterns $\{\mathsf{x}\}$ by maximizing the exact likelihood would require computations of the marginal $p(\mathsf{x})$, which involves averaging $p(\mathsf{x}|\mathsf{y})$ over the marginal distribution of the hidden variables $p(\mathsf{y})$. Analogously, the exact evaluation of the mutual information between the sources x and the codes y requires computations of the entropy of the hidden variables, which is a negated expectation of $\log p(\mathsf{y})$ over $p(\mathsf{y})$. Clearly, as long as the integrals cannot be computed analytically (which happens, for example, when $p(\mathsf{y})$ is not in the tractable family and the integrands do not decouple in y), the complexity of computing the objectives is generally exponential in $|\mathsf{y}|$.

In this work we will focus particularly on a discussion of variational lower bounds on the generally intractable mutual information for intrinsically intractable stochastic channels. While variational bounds on the generally intractable likelihoods are well-known and widely applied in a variety of approximate learning problems (see e.g. Jaakkola (1997), Saul and Jordan (1998), Barber and Wiegerinck (1998), Neal and Hinton (1998) and many others (see Jordan et al. (1998) for references and overview)), surprisingly little work in machine learning appears to have been done on developing a general framework for optimizing the mutual information for the cases when it cannot be computed exactly. In fact, most of the current methods consider maximizing the mutual information for low-scale channels (when the computations may be tractable), or consider various numerical approximations which are generally accurate only in the asymptotic limits (see e.g. Brunel and Nadal (1998), Kang and Sompolinsky (2001), Hoch et al. (2003)). Other somewhat heuristic objective criteria (e.g. *"information potentials"*) are sometimes used as surrogate objectives (Torkkola and Campbell (2000), Gokcay and Principe (2002)), but generally little is known about the conditions when such approximations may be accurate for maximizing the exact mutual information (Torkkola (2000)). Other methods suggest to optimize proper bounds, but may only be applied for a specific choice of the variational distribution (e.g. Linsker (1992), Jaakkola and Jordan (1998), Lawrence et al. (1998)), which complicates their extensions to richer families of methods.

In this work we will describe a family of variational *lower bounds* on mutual information $I(\mathsf{x}, \mathsf{y})$, which gives rise to a formal and theoretically justified approach to information maximization in noisy channels. We will outline a simple variational Information Maximization (IM) algorithm, reminiscent of the generalized variational EM algorithm for likelihood training (Neal and Hinton (1998)), and demonstrate that it provides a simple and effective tool for learning encoders and variational decoders in a principled manner. We will show that in the simplest case when the variational distribution is unconstrained, the proposed variational optimization procedure reduces to a generally intractable form of the Arimoto-Blahut (Arimoto (1972), Blahut (1972)) algorithm for maximization of channel capacity. In the case when the variational decoder is constrained to be a linear

Gaussian, a specific application of the bound reduces to Linsker's *as-if* Gaussian criterion (Linsker (1992)). Then we will discuss a richer family of the *auxiliary variational* lower bounds on the mutual information, which formally generalizes on the simple variational bounds.

In the later chapters, we will investigate the relation of our approach to the variational likelihood training in generative models and conditional likelihood training in stochastic autoencoders. Specifically, we will show that conventional methods of maximizing the variational Jensen's bounds on the conditional likelihood in stochastic autoencoders may be interpreted as a complicated way of optimizing a simple bound on the mutual information. Additionally, standard approaches to training (noiseless) autoencoders may be shown to optimize the simplest form of the variational lower bound on the mutual information between the sources and the codes for a noiseless encoding mapping. Furthermore, we will discuss applications of the considered framework to dimensionality reduction, clustering, and population coding. We will show that for a specific parameterization of the encoder distribution, our method favorably compares with the common approximation of Brunel and Nadal (1998) and Kang and Sompolinsky (2001). We will also compare the variational method with an alternative approximation of the mutual information inspired by the recent work of Szummer and Jaakkola (2002) and Corduneanu and Jaakkola (2003), and empirically demonstrate that in the considered cases our variational technique is more preferable.

Probably the most important contribution of this work is a simple and general framework for maximizing the mutual information in high-dimensional stochastic channels. We show that it formally generalizes and outperforms some of the common information-maximizing techniques. Curious side-products of the framework are the general relations between the conditional and the information-theoretic learning methods (Chapter 3), a compression method for continuous data (Chapter 4), an information-theoretic clustering method applicable for learning kernel functions (Chapter 5), and a local learning rule for population coding in a set of point-neuron models (Chapter 6).

In the rest of this chapter we will briefly discuss general differences of Maximum Likelihood and Maximum Mutual Information methods, and introduce the basic concepts of information theory. The presentation in Section 1.2 is high-level; we will return to it in more details in Chapter 3.

1.2 Maximum Likelihood and Maximum Mutual Information

One of the principal goals of generative graphical modeling is finding a model which describes how the observed data is generated from hidden variable representations. A principled objective function which is typically optimized during learning in this case is the likelihood of the model's parameters. The probabilistic graphical formulation (see e.g. Pearl (1988), Jordan (1998), Jordan (2005)) may be used to simplify computations of the likelihood in a principled way, where the optimized objective may be formally obtained from the model's structure and

parameterization. At the stage of problem specification, this helps to shift the focus of applying expert heuristics from developing *ad hoc* algorithms and *ad hoc* optimization criteria to designing graphical *models* (which may have a semantic meaning and may arguably be easier to specify, analyze, and apply for answering inferential queries).

The choice of the likelihood as an objective function for optimization is attractive for a number of reasons. In particular, many practically useful objective functions used rather heuristically in a variety of applied domains (from engineering to neuroscience) may be shown to correspond to likelihood maximization in specific graphical models (e.g. Bishop (1995)). Moreover, if the joint distribution defined by a chosen model has the same parametric form as the true distribution, then likelihood-based learning may be viewed as obtaining statistical estimators of the unknown parameters from a set of samples given by the training set. Subject to very general regulatory conditions, the resulting maximum-likelihood estimators have attractive statistical properties, such as asymptotic consistency, efficiency, and normality (Fisher (1922), Fisher (1925), Cramer (1946), Fisher (1950)).

In many cases, the distribution of the data expressed from a specified model does not perfectly match the underlying process (for example, due to the restrictiveness of the inductive bias). However, optimization of the likelihood is still justifiable in this case. Specifically, for i.i.d. observations, maximization of the likelihood with respect to the deterministic parameters of the model may be viewed as minimization of the Kullback-Leibler divergence (Kullback and Leibler (1951), Kullback (1959)) between the empirical distribution and the model. Since the KL divergence is theoretically minimal if and only if both distributions are identical, we may expect that a trained model will approximate the empirical distribution, subject to the modeling constraints. As the size of a training dataset increases, a properly chosen model trained by maximizing the likelihood will tend to approximate the underlying data-generating process. An optimized model may then be used for generating new data, inferring latent variable representations of the observations, etc.

It is intuitive that for a latent variable model to provide a good fit to the observations, the latent variables should capture useful statistical properties of the visible variables. By approximating the empirical distribution, generative models trained by maximizing the likelihood tend to find latent variable representations which are useful for generating the training data. This, however, does not necessarily quantify predictability of the hidden variables from the observations. In many practical situations it may be useful not only to be able to generate the data and answer inferential queries about the hidden states, but also to quantify how much information about the hidden variables is contained in a given set of observations. For example, one may be interested in obtaining a model where the observed symptoms are most informative about the hidden diseases, or where the training data points are most predictive of their possible cluster labels. It is known from information theory that a principal measure of informativeness is given by the mutual information (e.g. McEliece (1977), Cover and Thomas (1991)). Therefore, if the goal of modeling is re-defined as finding the most in-

formative representations of the training data (in the entropy loss sense) then maximization of the mutual information (Linsker (1988)) between the hidden and the visible domains is, intuitively, a reasonable learning strategy to consider.

An alternative justification for maximization of the mutual information comes from the field of communication and information theory, which is closely related to graphical modeling and machine learning (e.g. MacKay (2003)). In this context we can construct a probabilistic graphical model for a noisy communication problem where the visible data vectors $\{\mathsf{x}\}$, passed through a noisy channel, give rise to perturbed received representations $\{\mathsf{y}\}$. A classic result in information and communication theory is that the mutual information $I(\mathsf{x}, \mathsf{y})$ between the transmitted and the received vectors gives a lower bound on the channel capacity, which defines the maximum source rate for which reliable communication may theoretically be achieved (Shannon (1948), McEliece (1977)). Then it is clear that maximization of mutual information between the transmitted and the received vectors may be interpreted as maximizing the lower bound on the information transmission over the noisy channel (Arimoto (1972), Blahut (1972), Cover and Thomas (1991)), which is a formally justifiable, but generally computationally intractable procedure.

1.3 Information-Maximization in Noisy Channels

The reliable communication of information over noisy channels is a fundamental problem, ranging from from the construction of good error-correcting codes (see e.g. Pretzel (1996), MacKay (2003)) to neural sensory processing (Barlow (1989), Nadal et al. (1998)). A principal goal of information transmission is to communicate the source data over a channel without error. The problem is complicated in the presence of a channel noise, which transforms the data sent through the channels. In many practical cases we have a limited control over the communication system and cannot eliminate the noise completely. We would therefore like to find an efficient way to minimize possible effects of the inherently irreducible channel noise.

One of the key ideas of information theoretic approaches to communication in noisy channels is to reduce possible effects of the noise by encoding the source vectors $\{\mathsf{x}\}$ into codewords $\{\mathsf{t}\}$, which is typically addressed by introducing a known form of a systematic redundancy. At the communication stage, the codewords $\{\mathsf{t}\}$ are transmitted through the noisy channel, which leads to the perturbed representations $\{\mathsf{y}\}$ at the receiver's end. Then the receiver applies a known decoding rule to obtain reconstructions $\{\tilde{\mathsf{x}}\}$ of the original sources (see figure 1.1). Intuitively, we would like to construct the codes in such a way that despite the noise, the received vectors preserve much of the useful information about the original sources; in other words, $\{\mathsf{y}\}$ should be strongly predictive of $\{\mathsf{x}\}$. The focus of coding theory and information theory is mainly on finding optimal encoder and decoder systems and analyzing their performance (McEliece (1977), Cover and Thomas (1991), MacKay (2003)).

Traditionally, information theoretic approaches were applied to data compression and communication problems (Shannon (1948), McEliece (1977)). However,

Figure 1.1: A schematic representation of a noisy channel. The source vector x is transformed into a codeword t (for example, by introducing a known systematic redundancy). The codeword is transmitted over the noisy channel, leading the perturbed representation y. The receiver decodes y to produce an estimate $\tilde{\mathsf{x}}$ of the original source vector.

relatively recently they have been used for a wider range of machine learning tasks, from feature extraction (Becker (1992), Becker and Hinton (1992), Fisher and Principe (1998), Torkkola and Campbell (2000)) to clustering (Dhillon et al. (2002), Dhillon and Guan (2003), Jenssen et al. (2003)), neural population coding (Linsker (1989b), Linsker (1992), Brunel and Nadal (1998), Pouget et al. (1998), Zhang and Sejnowski (1999)), audio-video fusion (Fisher et al. (2000)), and independent components analysis (Bell and Sejnowski (1995), Shriki et al. (2002)). The family of *information bottleneck* (Tishby et al. (1999), Friedman et al. (2001)) approaches used a specific information-theoretic definition of the rate distortion function, with a number of applications to feature extraction (Slonim et al. (2001), Slonim (2002), Chechik and Tishby (2002)) and dimensionality reduction (Chechik et al. (2003), Still and Bialek (2004)). Moreover, several techniques were developed for combining likelihood and information-theoretic training, with applications to semi-supervised classification (Szummer and Jaakkola (2002), Corduneanu and Jaakkola (2003)). The successful applications of information theoretic techniques for a wide range of machine learning problems advocates their use as a general method for learning in noisy environments, which may go beyond the conventional applications to coding and communication.

In this section we briefly review maximization of information transfer as a general strategy for training probabilistic graphical models, and point out computational difficulties of the exact formulation, as well as a few common ways to address the issue of intractability. Initially, however, we will introduce the fundamental definitions of information theory, which we will use throughout the subsequent discussions.

1.3.1 Information Content, Entropy, and Mutual Information

Before discussing the information maximization principle in the context of learning in probabilistic graphical models, it is instructive to define the *information content* of a random event, as well as the marginal and conditional *entropy* associated with its distributions. First, we will assume that the random variable X takes discrete values x in the range $\mathcal{R}_X$ with probability $p(\mathsf{X} = \mathsf{x})$ (to simplify the notation we will write $p(\mathsf{x})$ to denote both the probability of the discrete event x or, for continuous variables, the probability density function; also, when it is clear from the context, we will use lower-case letters to indicate both random variables

and their current states). Then we will discuss the corresponding definitions for the case of continuous random variables.

1.3.1.1 Information content

We define the information content (or *degree of surprise*) of an event as a continuous monotonically decreasing functional $I(\mathsf{x}) \stackrel{\text{def}}{=} f\{p(\mathsf{x})\} \in [0, \infty), \quad \mathsf{x} \in \mathcal{R}_X,\ f \in \mathcal{F}$ of the distribution $P(\mathsf{x})$, such that a linear decrease in certainty leads to a generally super-linear increase in informativeness. Another requirement which we impose on I is that the combined information content of a set of independent events $\mathsf{x}^{(1)}, \ldots, \mathsf{x}^{(M)}$ is additive, i.e.

$$I(\{\mathsf{x}^{(1)}, \ldots, \mathsf{x}^{(M)}\}) \stackrel{\text{def}}{=} f\{p(\mathsf{x}^{(1)}, \ldots, \mathsf{x}^{(M)})\} \equiv \sum_{m=1}^{M} f\{p(\mathsf{x}^{(m)})\}, \qquad (1.1)$$
$$\forall i = 1, \ldots, M . \mathsf{x}^{(i)} \in \mathcal{R}_X,\ f \in \mathcal{F}.$$

Here $\mathcal{F}$ indicates the space of the functionals which satisfy the specified constraints. From (1.1) it is clear that

$$f\{p(\mathsf{x})^M\} = M f\{p(\mathsf{x})\}. \qquad (1.2)$$

By the constraints on f and the continuity principle it is easy to show that (1.2) holds $\forall M \in (-\infty, \infty)$ (see e.g. Bishop (1995)). It is also clear that a family of functionals satisfying (1.2) includes $f\{p(\mathsf{x})\} = -\log_b p(\mathsf{x})$, where the base $b > 1$, thus leading to[3]

$$I(\mathsf{x}) = -\log_b p(\mathsf{x}), \quad b > 1. \qquad (1.3)$$

By convention, the information content is expressed in bits ($b = 2$) or in *nats* ($b = e$); unless noted otherwise, we will usually assume that $b = e$. From (1.3) it is clear that the information content of the unlikely events ($p(\mathsf{x}) \to 0$) is exponentially large, while the information content of events occurring deterministically ($p(\mathsf{x}) = 1$) equals to zero.

1.3.1.2 Marginal entropy of discrete events

The absolute entropy, which is a fundamental measure of uncertainty of a discrete random variable $\mathsf{x} \sim p(\mathsf{x})$, is defined as the average information content

$$H(\mathsf{x}) \stackrel{\text{def}}{=} -\langle \log p(\mathsf{x}) \rangle_{p(\mathsf{x})}, \qquad (1.4)$$

where $\langle f(\mathsf{x}) \rangle_{p(\mathsf{x})} \stackrel{\text{def}}{=} \sum_{\mathsf{x}} f(\mathsf{x}) p(\mathsf{x})$ denotes the expectation of $f(\mathsf{x})$ over $p(\mathsf{x})$ (assuming that x is discrete). If $p(\mathsf{x})$ is a marginal distribution of a more complex model, (1.4) is commonly referred to as the *marginal* entropy. Originally developed in statistical physics for describing macroscopic states of closed systems (see e.g. Landau and Lifshitz (1996) for the statistical mechanics formulation), it has

[3] For $\mathcal{R}_x \equiv \mathbb{N}$ and the normalized distribution function $p(x) = (1/2)^x$, the information content is expressed in bits as $I(x) = f(p(x)) = x f(1/2) = -\log_2 p(x)$ with $f(p(x)) \equiv -\log_2 p(x)$.

been widely used in information theory (e.g. Shannon (1948), McEliece (1977)) and other applied domains. In this context, it may be formally thought of as a quantification of the average amount of information provided by the observation of a random variable x, distributed according to $p(\mathsf{x})$.

The entropy defined by (1.4) satisfies a number of useful properties. For example, it has been shown that $H(\mathsf{x})$ lower-bounds an average length of a binary message needed to encode x (Shannon (1948)). Additionally, it is easy to see that for discrete variables the entropy is non-negative, reaching zero only in the deterministic case. For this case it is also easy to show that $H(\mathsf{x})$ is maximized when $p(\mathsf{x}) = 1/|\mathsf{x}|$ is uniform, leading to $H(\mathsf{x}) \leq \log |\mathsf{x}|$, where $|\mathsf{x}|$ is the number of states. Therefore, informally we may think of $H(\mathsf{x})$ as a measure of "sharpness" of $p(\mathsf{x})$, with the maximum obtained for the flattest (uniform) distribution. A number of other useful properties and interpretations are described in a variety of texts on information theory, e.g. Pinsker (1964), McEliece (1977), Cover and Thomas (1991).

1.3.1.3 Conditional entropy of discrete events

For a pair of random vectors x, y, we may analogously define the conditional entropy $H(\mathsf{x}|\mathsf{y})$ (*equivocation* of y about x). It is given as a functional of the conditional distribution, averaged over the joint probability of the combined events

$$H(\mathsf{x}|\mathsf{y}) \stackrel{\text{def}}{=} -\langle \log p(\mathsf{x}|\mathsf{y}) \rangle_{p(\mathsf{x},\mathsf{y})} = \langle H(p(\mathsf{x}|\mathsf{y})) \rangle_{p(\mathsf{y})}. \tag{1.5}$$

Here we defined $H(p(\mathsf{x}|\mathsf{y})) \stackrel{\text{def}}{=} -\langle \log p(\mathsf{x}|\mathsf{y}) \rangle_{p(\mathsf{x}|\mathsf{y})}$ to be the entropy of the conditional distribution $p(\mathsf{x}|\mathsf{y})$. For the communication channel shown on Figure 1.1, the conditional entropy (1.5) may be interpreted as the uncertainty of the receiver about the source x for the fixed received vector y, averaged over the distribution of the received vectors $p(\mathsf{y})$.

As before, it is possible to interpret the conditional entropy as a quantification of an average sharpness. Note that a large value of $H(p(\mathsf{x}|\mathsf{y}))$, achievable for flat posteriors $p(\mathsf{x}|\mathsf{y})$, would indicate a large uncertainty in decoding a specific codeword y by using $p(\mathsf{x}|\mathsf{y})$. Analogously, large values of the conditional entropy $H(\mathsf{x}|\mathsf{y})$ would indicate a large uncertainty of decoding with the exact posterior *on average*. Intuitively, the receiver may be interested in channels leading to small values of the conditional entropy $H(\mathsf{x}|\mathsf{y})$, though generally the reconstruction error will depend on the receiver's decoding strategy.

1.3.1.4 Mutual Information

Finally, the *mutual information* is defined as a change in the average information content of x due to the knowledge of y:

$$I(\mathsf{x},\mathsf{y}) \stackrel{\text{def}}{=} H(\mathsf{x}) - H(\mathsf{x}|\mathsf{y}) = H(\mathsf{y}) - H(\mathsf{y}|\mathsf{x}). \tag{1.6}$$

Fundamentally, it may be interpreted as a measure of predictability of y from x and, due to the symmetry, of x from y. From (1.6) it is easy to obtain the

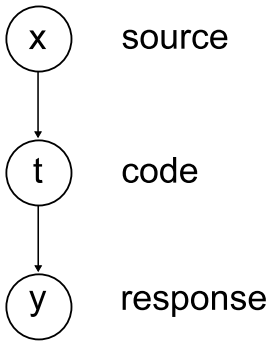


Figure 1.2: Graphical model of a noisy communication channel. Generally, we assume that we cannot eliminate the channel noise, which is implied in the parameterization of $p(\mathsf{y}|\mathsf{t})$.

equivalent formulation

$$I(\mathsf{x},\mathsf{y}) = KL(p(\mathsf{x},\mathsf{y})\|p(\mathsf{x})p(\mathsf{y})). \tag{1.7}$$

Clearly, the mutual information is minimized when $p(\mathsf{x},\mathsf{y}) = p(\mathsf{x})p(\mathsf{y})$, i.e. when the received vectors y are independent of the transmitted sources x. It is easy to see that if both x and y are discrete, the mutual information is bounded as $I(\mathsf{x},\mathsf{y}) \in [0, \log\min\{|\mathsf{x}|,|\mathsf{y}|\}]$, where $|\mathsf{x}|$ and $|\mathsf{y}|$ correspond to the number of states of the input and output variables.

1.3.1.5 Generalization to the continuous case

It may be shown that definitions similar to (1.4) – (1.7) are also applicable for non-discrete random vectors (e.g. Cover and Thomas (1991)). In this case $H(\mathsf{x})$ and $H(\mathsf{x}|\mathsf{y})$ are referred to as *differential entropies*, and the averages in (1.4) and (1.5) are computed by integrating over $\mathcal{R}_x$ and $\mathcal{R}_y$. It may be shown that in this case the differential entropies diverge to $-\infty$, which makes it difficult to interpret them as a measure of certainty in x. In fact, the differential entropies are not even invariant under deterministic invertible transformations (McEliece (1977), Bell and Sejnowski (1995)). However, by computing discrete quantizations of the random variables, it may be shown that the mutual information $I(\mathsf{x},\mathsf{y})$ computed over the continuous range satisfies properties of the mutual information for the discrete case, and therefore has similar properties and a similar fundamentally useful interpretation. Specifically, from definition (1.7) and the property of non-negativity of the Kullback-Leibler divergence, we may see that $I(\mathsf{x},\mathsf{y}) \geq 0$ for both discrete and continuous quantities.

1.3.1.6 Information maximization in graphical models

Figure 1.2 shows a simple graphical model of a noisy communication channel[4]. In a stochastic context, the quantitative part of the model is specified by the source distribution $p(\mathsf{x})$, the encoding transformation $p(\mathsf{t}|\mathsf{x})$, and a generally noisy non-invertible channel mapping $p(\mathsf{y}|\mathsf{t})$. The key idea of information maximization is to choose a mapping from source variables (inputs) x to response variables (outputs) y such that the outputs contain as much information as possible about which of the inputs was transmitted. The principal measure of information transfer in this context is the mutual information $I(\mathsf{x},\mathsf{y})$, defined by expression (1.6). The general aim will be to set any adjustable parameters of $p(\mathsf{y}|\mathsf{x}) = \langle p(\mathsf{y}|\mathsf{t})\rangle_{p(\mathsf{t}|\mathsf{x})}$ in order to maximize $I(\mathsf{x},\mathsf{y})$. Generally, we will presume that we are limited in the way we can adjust $p(\mathsf{y}|\mathsf{t})$; specifically, we may assume that we cannot explicitly modify the channel noise. For example, for deterministic mappings $p(\mathsf{t}|\mathsf{x}) \sim \delta(\mathsf{t}-\mathsf{t}(\mathsf{x}))$ and Gaussian channels $p(\mathsf{y}|\mathsf{t}) \sim \mathcal{N}_{\mathsf{y}}(\mathsf{t},\sigma_e^2\mathsf{I})$, we easily get $p(\mathsf{y}|\mathsf{x}) \sim \mathcal{N}_{\mathsf{y}}(\mathsf{t}(\mathsf{x}),\sigma_e^2\mathsf{I})$. For the fixed noise variance σ_e^2, our purpose in this specific case would be to modify the encoder mapping $\mathsf{t}(\mathsf{x})$. In a more general context, we will be trying to learn the modifiable parameters of $p(\mathsf{y}|\mathsf{x})$ to maximize the mutual information. With a slight abuse of terminology, we will refer to $p(\mathsf{y}|\mathsf{x})$ as the encoder distribution.

One specific case of interest we consider in this chapter is when the source is given by the empirical distribution

$$\tilde{p}(\mathsf{x}) = \frac{1}{M}\sum_{m=1}^{M}\delta\left(\mathsf{x}-\mathsf{x}^{(m)}\right), \tag{1.8}$$

for a finite set of unique training patterns $\{\mathsf{x}^{(m)}|m=1,\ldots,M\}$ (generalization to the repeating patterns is straight-forward; we omit it here for clarity). Unless stated otherwise, we will also assume that the channel is *memoryless*, i.e. each perturbed pattern $\mathsf{y}^{(i)}$ depends only on the corresponding source vector $\mathsf{x}^{(i)}$. In this case, the theoretically optimal decoder which maximizes the mutual information is given by Bayes rule (e.g. Papoulis (1984)) as

$$p(\mathsf{x}|\mathsf{y}) = \sum_{k=1}^{M}\delta\left(\mathsf{x}-\mathsf{x}^{(k)}\right)\omega_k(\mathsf{y}), \quad \omega_k(\mathsf{y}) \stackrel{\text{def}}{=} \frac{p(\mathsf{y}|\mathsf{x}^{(k)})}{\sum_{m=1}^{M}p(\mathsf{y}|\mathsf{x}^{(m)})}, \tag{1.9}$$

where we assumed that the adjustable parameters of the encoder $p(\mathsf{y}|\mathsf{x})$ in the definition of $\omega(\mathsf{y})$ are fixed at their optimal values. Clearly, (1.9) is a mixture of Dirac delta functions, which reduces to a single delta peak only for the case when stochastic images $\mathcal{I}$ of distinct source vectors are non-intersecting under the encoder $p(\mathsf{y}|\mathsf{x})$, i.e. $\forall i,j \in \{1,\ldots,M\}.i \neq j.\mathcal{I}(\mathsf{x}_i) \cap \mathcal{I}(\mathsf{x}_j) = \emptyset$. Formally, this is never the case if $\forall\mathsf{y}.\forall\mathsf{x}.p(\mathsf{y}|\mathsf{x}) \neq 0$, which leads to a mixture form of the Bayesian decoder $p(\mathsf{x}|\mathsf{y})$.

[4]Several equivalent graphical representations are possible. The model shown on Figure 1.2 is chosen in order to illustrate the similarity with the noisy communication channel (Figure 1.1). Here we presumed that $p(\mathsf{y}|\mathsf{t}) = \langle p(\mathsf{y}|\mathsf{t},\mathsf{e})\rangle_{p(\mathsf{e})}$, where $\mathsf{e} \sim p(\mathsf{e})$ is the implied random noise.

1.4 Computational Problems of the Exact Formulation

Unfortunately, the choice of the theoretically optimal decoder (1.9) may complicate evaluation of the true mutual information $I(\mathsf{x}, \mathsf{y})$, as well as learning of the optimal encoder $p(\mathsf{y}|\mathsf{x})$. The reason for it is that in the expression of the mutual information (1.6), the complexity of computing the conditional entropy $H(\mathsf{x}|\mathsf{y})$ for the mixture distribution $p(\mathsf{x}|\mathsf{y})$ is in general exponential in $|\mathsf{y}|$. Indeed, from (1.6) and (1.9) we get

$$I(\mathsf{x}, \mathsf{y}) = \left\langle \log \sum_{m=1}^{M} p\left(\mathsf{y}|\mathsf{x}^{(m)}\right) \right\rangle_{p(\mathsf{y})} - \sum_{m=1}^{M} \left\langle \log p\left(\mathsf{y}|\mathsf{x}^{(m)}\right) \right\rangle_{p\left(\mathsf{y}|\mathsf{x}^{(m)}\right)}, \tag{1.10}$$

which involves integration over the y variables. For a model shown on Figure 1.2, evaluation of the second term in (1.10) is computationally tractable as long as $p(\mathsf{y}|\mathsf{x})$ has a simple structure (for example, if it is factorized in y). However, evaluation of the first term in (1.10) (which is, in fact, the marginal entropy of the outputs $H(\mathsf{y})$) is in general problematic, since the marginal $p(\mathsf{y}) \propto \sum_{m=1}^{M} p(\mathsf{y}|\mathsf{x}^{(m)})$, being a mixture, is coupled in y. Therefore, computation of the mutual information generally involves integration over an exponential number of states of the output vectors. Moreover, it is important to note that the representational complexity of storing the optimal decoders is also in general exponential in $|\mathsf{y}|$, which motivates a choice of parametric or structural constraints on decoder's specification.

1.4.1 Tractable Special Cases

The problem of optimizing (1.6) has been well explored for a number of tractable special cases. Specifically, it has been looked at in the context of maximizing the channel capacity for systems with very low-dimensional output spaces (Arimoto (1972), Linsker (1989b)). Indeed, for low-dimensional discrete outputs the effective number of states of y may be reasonably small, so that the summations in (1.10) may be computed exactly. For the continuous case (e.g. for $\mathsf{y} \in \mathbb{R}^{|\mathsf{y}|}$), it may be impossible to express the mutual information analytically. However, due to the fact that y is low-dimensional, the integrals in (1.10) may be efficiently approximated numerically (for example, by applying the mean-value theorem or using any of the known methods of numerical integration, see e.g. Korn and Korn (1968), Press et al. (1992), Dahlquist (2003)).

Additionally, computations may be tractable in the case when the encoder $p(\mathsf{y}|\mathsf{x}) \sim \delta(\mathsf{y} - \mathsf{g}(\mathsf{x}; \boldsymbol{\Theta}))$ is deterministic and invertible (Bell and Sejnowski (1994), Nadal and Parga (1994)), so that the mutual information is easily expressed as

$$I(\mathsf{y}, \mathsf{x}) = c - \left\langle \log \frac{p(\mathsf{x})}{|\mathsf{J}_\mathsf{x}|} \right\rangle_{p(\mathsf{x})} = c + H(\mathsf{x}) + \langle \log |\mathsf{J}_\mathsf{x}| \rangle_{p(\mathsf{x})}. \tag{1.11}$$

Here c is a constant which does not affect the optimization surface for $p(\mathsf{y}|\mathsf{x})$, $\mathsf{J}_\mathsf{x} = \{\partial g_i(\mathsf{x})/\partial x_j\}$ is the Jacobian of the invertible mapping, and the average is

computed over the source vectors x. It is easy to see that if $p(\mathsf{x})$ is the empirical distribution (1.8), the objective (1.11) is tractable. Moreover, for specific choices of the invertible mapping $g(\mathsf{x};\boldsymbol{\Theta}) : \mathsf{x} \rightarrow \mathsf{y}$, optimization of (1.11) with respect to the encoder parameters $\boldsymbol{\Theta}$ leads to the *infomax* (Linsker, 1989a) formulation of ICA (Bell and Sejnowski, 1995).

Other tractable cases include linear Gaussian channels with the Gaussian distribution of the sources (Linsker, 1989a), or in fact any model where $p(\mathsf{x},\mathsf{y})$ is jointly Gaussian. For this case, maximization of the mutual information reduces to learning the structured covariance matrix of the Gaussian decoder $p(\mathsf{x}|\mathsf{y})$. Generally, for all of the discussed special cases, it is tractable to compute $I(\mathsf{x},\mathsf{y})$ and its gradients exactly. In order to find the optimal encoder, one may use iterative algorithms for maximizing the channel capacity (Arimoto (1972), Blahut (1972)), apply numerical optimization procedures (e.g. Bishop (1995), Luenberger (1998)), or in some cases find the optimal solutions analytically (Linsker (1989a), Linsker (1989b)).

1.5 Common Approximations of Mutual Information

We will now briefly describe several common methods which may be used to address the computational intractability of the exact formulation of the information-maximization problem. We focus specifically on Linsker's *as-if* Gaussian approximation, as it will prove to be relatable to the variational approach to information maximization which we will introduce in Chapter 2. See Section 6.3.1 for a detailed discussion of Brunel and Nadal's Fisher approximation criterion (Brunel and Nadal (1998)).

1.5.1 As-if Gaussian Approximation of $I(\mathsf{x},\mathsf{y})$

In Section 1.4.1 we mentioned several choices of the encoding distribution leading to the exact computations of the mutual information. For more general encoding distributions, it may be necessary to consider approximations of $I(\mathsf{x},\mathsf{y})$. Here we will discuss the method introduced by (Linsker (1992)), which utilizes the simple simple *as-if Gaussian* approximation of the mixture distribution. The key idea here is to approximate $I(\mathsf{x},\mathsf{y})$ by assuming that the joint distribution $p(\mathsf{x})p(\mathsf{y}|\mathsf{x}) \approx p_G(\mathsf{x},\mathsf{y}) \sim \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ is a Gaussian, independently of the exact form of $p(\mathsf{x},\mathsf{y})$. Without loss of generality, we will presume that the data is centered, i.e. $\boldsymbol{\mu} = \mathbf{0}$.

Note that the conditional entropy in (1.6) may in this case be approximated by

$$H(\mathsf{x}|\mathsf{y}) \approx H_G(\mathsf{x}|\mathsf{y}) \stackrel{\text{def}}{=} -\langle \log p_G(\mathsf{x}|\mathsf{y}) \rangle_{p_G(\mathsf{x},\mathsf{y})} = (1/2)\log(2\pi e)^{|\mathsf{x}|}|\boldsymbol{\Sigma}_{\mathsf{x}|\mathsf{y}}|, \qquad (1.12)$$

where $\mathbf{\Sigma}_{\mathsf{x}|\mathsf{y}}$ is the covariance of the decoder $p_G(\mathsf{x}|\mathsf{y})$ expressed from the joint Gaussian $p_G(\mathsf{x},\mathsf{y})$. If the joint covariance is partitioned as

$$\mathbf{\Sigma} \stackrel{\text{def}}{=} \langle [\mathsf{x}\ \mathsf{y}][\mathsf{x}\ \mathsf{y}]^T \rangle_{p(\mathsf{x},\mathsf{y})} - \langle [\mathsf{x}\ \mathsf{y}] \rangle_{p(\mathsf{x},\mathsf{y})} \langle [\mathsf{x}\ \mathsf{y}]^T \rangle_{p(\mathsf{x},\mathsf{y})} \tag{1.13}$$

$$\stackrel{\text{def}}{=} \begin{pmatrix} \mathbf{\Sigma}_{xx} & \mathbf{\Sigma}_{xy} \\ \mathbf{\Sigma}_{yx} & \mathbf{\Sigma}_{yy} \end{pmatrix} \in \mathbb{R}^{(|\mathsf{x}|+|\mathsf{y}|)\times(|\mathsf{x}|+|\mathsf{y}|)}, \tag{1.14}$$

we may express the conditional covariance as

$$\mathbf{\Sigma}_{\mathsf{x}|\mathsf{y}} = \mathbf{\Sigma}_{xx} - \mathbf{\Sigma}_{xy}\mathbf{\Sigma}_{yy}^{-1}\mathbf{\Sigma}_{yx} \tag{1.15}$$

(see e.g. von Mises (1964)). Then a substitution into the expression for mutual information (1.6) leads to the objective function

$$2I_G(\mathsf{x},\mathsf{y}) \quad = \quad \log|\mathbf{\Sigma}_{xx}| - \log|\mathbf{\Sigma}_{xx} - \mathbf{\Sigma}_{xy}\mathbf{\Sigma}_{yy}^{-1}\mathbf{\Sigma}_{yx}|. \tag{1.16}$$

For the assumed case of $\boldsymbol{\mu} = 0$, the objective (1.16) reduces to

$$2I_G(\mathsf{x},\mathsf{y}) \quad = \quad const - \log|\langle \mathsf{x}\mathsf{x}^T \rangle - \langle \mathsf{x}\mathsf{y}^T \rangle \langle \mathsf{y}\mathsf{y}^T \rangle^{-1} \langle \mathsf{y}\mathsf{x}^T \rangle|, \tag{1.17}$$

where the averages are computed over the source and the channel distributions $p(\mathsf{x})$, $p(\mathsf{y}|\mathsf{x})$. Note that if $p(\mathsf{x})$ is the empirical distribution, the as-if Gaussian approximation (1.17) of the mutual information $I(\mathsf{x},\mathsf{y})$ is a function of the encoder $p(\mathsf{y}|\mathsf{x})$ alone; optimization of (1.17) corresponds to learning the encoder's parameters.

It turns out that the as-if Gaussian objective $I_G(\mathsf{x},\mathsf{y})$ corresponds to a proper variational lower bound on the true mutual information $I(\mathsf{x},\mathsf{y})$ under the assumption of a Gaussian decoder (see Section 2.1). In other words, Linsker's approximation (Linsker (1992)) may be seen as a special case of a much more general variational procedure.

We will also show that in some simple cases, the encoder parameters obtained by maximizing (1.16) may be expressed analytically. Specifically, for linear Gaussian channels with the isotropic noise $p(\mathsf{y}|\mathsf{x}) \sim \mathcal{N}_y(\mathsf{W}\mathsf{x}, \sigma^2\mathsf{I})$, where $\mathsf{W} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{x}|}$ and $|\mathsf{y}| < |\mathsf{x}|$, the right singular vectors of the optimal weights correspond to rotations of the principle eigenvectors of the sample covariance $\mathsf{S} \stackrel{\text{def}}{=} \langle \mathsf{x}\mathsf{x}^T \rangle$ (see the discussion in Section 4.1.1). Equivalently, the result may be obtained in the noiseless limit of undercomplete linear projections $\mathsf{y} = \mathsf{W}\mathsf{x}$.

1.5.2 Other Approximations of $I(\mathsf{x},\mathsf{y})$

Note that a general way to evaluate intractable averages over the hidden variables is by drawing independent samples from appropriate distributions, and using Monte-Carlo estimations of the objective criteria (see e.g. Neal (1993), Gamerman (1997)). The method is conceptually attractive, as conceptually similar techniques can be used to approximate arbitrary averages over the latent variables. However, in practice sampling techniques have several important drawbacks. First of all, sampling independent points from the equilibrium distribution may be extremely time-consuming, since the system may get trapped in a local

mode of the distribution (e.g. Hertz et al. (1991), Neal (1993)). This problem is partially addressed by a number of auxiliary sampling techniques, such as hybrid Monte Carlo for continuous spaces (e.g. Neal (1993)), or the Swendsen-Wang (Swendsen and Wang, 1987) and partial decoupling (Higdon (1998), Morris (1999)) algorithms for discrete variable spaces, which may help to improve convergence time and reduce the time gap between independent samples (see e.g. MacKay (1998) for an introductory discussion). A somewhat more fundamental problem is assessing convergence to the equilibrium (and therefore the quality of the resulting approximations). In the future, we may potentially consider applications of sampling methods in the information-theoretic context; however, due to the general difficulties of assessing the quality of the resulting estimates, we do not consider them in the suggested work. In what follows, we will only consider analytical approximations of $I(\mathsf{x}, \mathsf{y})$, and compare our results against other analytical approximations.

One of the relatively recent analytical approximations of mutual information is based on the idea of interpreting different dimensions $\{y_i\}$ of the codewords $\mathsf{y} \in \mathbb{R}^{|\mathsf{y}|}$ as independent samples from a distribution parameterized by the source vectors x (see Brunel and Nadal (1998)). Specifically, if the encoding distribution $p(\mathsf{y}|\mathsf{x}) = \prod_{i=1}^{|\mathsf{y}|} p_i(y_i|\mathsf{x})$ and $p_i(y_i = s|\mathsf{x}) \equiv p_j(y_j = s|\mathsf{x})$ for all $i, j = 1, \ldots, |\mathsf{y}|$, then one may make a recourse to the results of statistical parameter estimation theory (e.g. Cramer (1946)) and approximate a lower bound on mutual information $I(\mathsf{x}, \mathsf{y})$ by applying the data-processing and the Cramer-Rao (Cover and Thomas (1991)) inequalities. While a direct application of the inequalities leads to a computationally intractable bound on $I(\mathsf{x}, \mathsf{y})$, the bound may in some cases be efficiently approximated by considering further numerical relaxations (Brunel and Nadal (1998)). Generally, the suggested approximations of the lower bound on $I(\mathsf{x}, \mathsf{y})$ are accurate under the asymptotic assumption of infinitely large code spaces ($|\mathsf{y}| \to \infty$), but they may lead to unstable behavior of the resulting learning algorithms for $|\mathsf{x}| < |\mathsf{y}|$. The identical results were also obtained by Kang and Sompolinsky (2001), who considered different numerical approximations applied in the same asymptotic limit $|\mathsf{y}| \to \infty$.

Another approximation of $I(\mathsf{x}, \mathsf{y})$ may be obtained by a simple re-formulation of the local approximations of Szummer and Jaakkola (2002) and Corduneanu and Jaakkola (2003), who propose to approximate the mutual information numerically for each infinitely small symmetric region in the source space. By applying the chain rule on mutual information, it is straight-forward to show that their criterion may be used to approximate a lower bound on $I(\mathsf{x}, \mathsf{y})$ (see discussion in Section 6.3.2). A number of other somewhat heuristic objective criteria is often used to approximate $I(\mathsf{x}, \mathsf{y})$ (see e.g. Principe et al. (2000), Torkkola and Campbell (2000), Gokcay and Principe (2002)), though the resulting approximations may be rather difficult to justify for maximizing the mutual information (Torkkola (2000)). Other work focuses specifically on noiseless overcomplete communication channels (where $|\mathsf{y}| > |\mathsf{x}|$ and $p(\mathsf{y}|\mathsf{x}) \sim \delta(\mathsf{y} - \mathsf{f}(\mathsf{x}))$ (see Shriki et al. (2002))) and cannot be easily applied in the case of stochastic encoding mappings.

1.6 Thesis Overview

In the following chapters we will discuss a principled and computationally tractable approach to maximization of the information transfer in large-scale stochastic models. Particularly, we will be interested in the family of variational approaches, where the goal may informally be formulated as transforming optimization of a generally intractable functional to optimization of its tractable bound. The tractability of the variational methods is typically achieved by decoupling the degrees of freedom of the original functional at the cost of introducing additional variational parameters. Variational methods have been long described in extensive mathematics and physics literature (see e.g. Gelfand and Fomin (1963), Ewig (1985), Fox (1987), Riley et al. (2002)), and applied to graphical modeling in the context of bounding the partition functions (see e.g. Jaakkola (1997), Jordan et al. (1998), Wainwright et al. (2002), Wainwright (2002)). We will be mainly interested in deriving tractable bounds on the mutual information.

In Chapter 2 we will discuss a simple and general variational approach to maximizing a proper (generic) lower bound on $I(\mathsf{x},\mathsf{y})$ in a noisy channel, and show that in the special case when the variational distribution is unconstrained, the method gives rise to the family of Blahut-Arimoto type algorithms (Arimoto (1972), Blahut (1972)). Generally, however, constraining the variational distributions is important in order to ensure that the method remains computationally tractable. We will also show that optimization of Linsker's *as-if* Gaussian objective criterion (Linsker (1992)) corresponds to a specific way of optimizing the variational lower bound on $I(\mathsf{x},\mathsf{y})$ for a specific choice of the variational decoder distribution. Finally, we will introduce an *auxiliary variational* approach to approximate information maximization, which formally generalizes on the simple generic bound without altering properties of the original channel.

In Chapter 3 we will explore general relations of the variational information-maximizing algorithm to maximum likelihood learning in generative models and conditional likelihood learning in chains. We will outline general differences between encoder models of communication channels and generative models. We will also show that the likelihood of a generative model may be viewed as a lower bound on $I(\mathsf{x},\mathsf{y})$ for the corresponding model of the stochastic channel, where the encoding distribution is the exact posterior of the generative model. A practical side-effect of this study is an information-theoretic objective for training generative models. We will also demonstrate a close relation between optimizing the generic variational bound on $I(\mathsf{x},\mathsf{y})$ and conditional likelihood training in stochastic autoencoders. Additionally, we will show that common approaches to training noiseless autoencoders maximize proper lower bounds on $I(\mathsf{x},\mathsf{y})$ (under the assumption of noiseless encoding mappings).

In Chapter 4 we will consider an application of the variational information maximizing approach to noisy constrained dimensionality reduction. We will prove a simple result that for constrained linear Gaussian channels, optimization of Linsker's bound cannot improve on the PCA projections. On the other hand, the richer family of the auxiliary variational lower bounds on $I(\mathsf{x},\mathsf{y})$ leads to significant improvements over Linsker's (PCA) bounds. This result is potentially

interesting from the communication-theoretic perspective, as it demonstrates a simple and computationally efficient way to produce tighter bounds on the capacity of a communication channel without altering its properties (e.g. without communicating more data across the channels). Additionally, we will discuss a simple information-theoretic approach to constrained dimensionality reduction for hybrid channels $\mathsf{x} \rightarrow \{\mathsf{y}, z\}$ (where $\mathsf{x} \in \mathbb{R}^{|\mathsf{x}|}$, $\mathsf{y} \in \mathbb{R}^{|\mathsf{y}|}$, and $z \in \{1, \ldots, |z|\}$), which may significantly improve reconstructions of the sources $\{\mathsf{x}\}$ from their lower-dimensional representations $\{\mathsf{y}\}$ at a small increase in the transmission cost (given by $|z|$). We will also point out a curious link between maximizing $I(\mathsf{x}, \{\mathsf{y}, z\})$ in a hybrid channel and maximizing the likelihood for a mixture of constrained Factor Analysis-type models with the uniform (rather than the spherical) distribution of the factors.

Chapter 5 demonstrates applications of the information-maximizing framework for the case of nonlinear encoder models. Specifically, we will discuss several ways of applying the framework to information-theoretic clustering. We will empirically demonstrate that the resulting information-theoretic clustering approaches favorably compare with the conventional clustering techniques. Moreover, we will show that the information-maximizing framework may be used to learn kernel functions, which may indeed be of a practical benefit for visualizing the underlying structure of the data. Moreover, we will review some of the theoretical properties of the variational information-maximizing framework for nonlinear Gaussian encoding distributions; for example, we will show that the Gaussian Process Latent Variable Model (Lawrence (2003)) arises as a special case of our information-theoretic formulation.

Chapter 6 explores applicability of the variational information-maximizing framework in the context of learning high-dimensional binary representations of continuous source patterns. We will consider the situation when the binary encodings are conditionally independent, and compare our variational approach with Brunel and Nadal's Fisher approximation of mutual information (Brunel and Nadal (1998)). Moreover, we will use the results of Szummer and Jaakkola (2002) and Corduneanu and Jaakkola (2003) to derive another straight-forward approximation of $I(\mathsf{x}, \mathsf{y})$. Our empirical results indicate that for the considered case, the variational approach is most preferable, while both the *local* approximation of $I(\mathsf{x}, \mathsf{y})$ (based on the work of Corduneanu and Jaakkola (2003)) and the generic variational approach significantly outperform the common Fisher approximation for undercomplete projections (the compression paradigm). Additionally, we will demonstrate that for a considered encoding distribution it is possible to derive a *local* learning rule, which may potentially be attractive from the neuro-biological perspective. This extends the results of Linsker (1997), who has previously showed the existence of local approximations of nonlinear optimization procedures on $I(\mathsf{x}, \mathsf{y})$ for invertible encoding mappings $\mathsf{x} \mapsto \mathsf{y}$ (*cf* Nadal and Parga (1994), Bell and Sejnowski (1995)).

In Chapter 7 we will change the perspective and consider a seemingly unrelated problem of lower-bounding the normalizing constant (log partition function) of a probability distribution. Specifically, we will introduce an auxiliary variable extension of any structured mean field theory, which can be useful in the context

of approximate probabilistic inference. While the method described there is of a potential interest as a general approach to approximate inference, it demonstrates a curious link to our variational information-maximizing framework. In particular, we will show that the improvement of the proposed bound on $\log Z$ over a convex combination of simpler bounds given by the standard theories is defined by a specific form of the generic lower bound on mutual information. The variational information-maximizing framework may therefore be seen as addressing an integral subgoal of variational inference. We will also show that the existing variational mixture methods (see Jaakkola and Jordan (1998), Lawrence et al. (1998)) may be viewed as specific numerically difficult instances of the auxiliary variational approach.

Finally, Chapter 8 summarizes the results and outlines directions for extending the presented work.

Chapter 2

Variational Information Maximization

Here we describe a family of variational lower bounds on mutual information $I(\mathsf{x}, \mathsf{y})$, which gives rise to a formal and theoretically justified approach to information maximization in noisy channels. In Section 2.1 we describe the variational Information Maximization (IM) algorithm, reminiscent of the generalized variational EM algorithm for likelihood training, and demonstrate that it provides a simple and effective tool for learning encoders and variational decoders in a principled manner. We show that in the simplest case when the simplest form of the bound is used and the decoder is unconstrained, the proposed variational optimization procedure reduces to a generally intractable form of the Arimoto-Blahut (Arimoto (1972), Blahut (1972)) algorithm for maximization of channel capacity. However, by choosing appropriate parametric constraints on the encoder-decoder pair, we may avoid intractabilities in a principled manner. The resulting algorithm would be optimizing a simple *generic* lower bound on mutual information, subject to the specific constraints on the variational decoder.

Then we outline general properties of the bound. In Section 2.2.1 we describe effects of choosing specific sparse decoder structures on the generic lower bound on $I(\mathsf{x}, \mathsf{y})$. Not surprisingly, we show that by considering richer decoder structures, we may indeed obtain tighter bounds on the intractable measure of information content. We also discuss a simple relation of the variational bound to the common *as-if Gaussian* approximation of the mutual information (Linsker, 1992). As we show in Section 2.2.2, independently of the choice of the encoder distribution, Linsker's approach corresponds to a special case of our formulation, where the variational decoder is constrained to be a linear Gaussian.

Finally, in Section 2.3 we introduce a richer family of *auxiliary variational* lower bounds, which generalize on the simpler generic bounds on $I(\mathsf{x}, \mathsf{y})$. The key idea there is to introduce additional variables, which may be used for capturing useful features of the source patters, and for introducing global dependencies to the decoded sources. Importantly, we show that the projections to the auxiliary space may be defined in a way which does not alter the original channel. By constraining the mappings to the auxiliary space to be in a tractable family and by imposing appropriate constraints on the variational decoders, the resulting bound

is a formal tractable generalization of the simpler generic approach described in Section 2.1.

2.1 Generic Lower Bound on Mutual Information

Since the exact evaluation and optimizaton of the information transfer is in general computationally intractable, our central aim here will be to maximize a lower bound on the mutual information in a tractable way. Using the formulation $I(\mathsf{x},\mathsf{y}) = H(\mathsf{x}) - H(\mathsf{x}|\mathsf{y})$, we are interested in optimising $I(\mathsf{x},\mathsf{y})$ with respect to the encoder $p(\mathsf{y}|\mathsf{x})$. In many cases which we consider $p(\mathsf{x})$ is simply the empirical distribution (1.8). In principle, this may be generalized to any case where averaging over the sources $p(\mathsf{x})$ is tractable. Nevertheless, since the distribution of the source patterns is not a function of the channel parameters, it has no effect on the optimization surface for $p(\mathsf{y}|\mathsf{x})$. Thus, to express the objective function we need to bound the intractable entropic term $H(\mathsf{x}|\mathsf{y})$ suitably.

2.1.1 Definition of a Simple Lower Bound on Mutual Information

In order to derive a simple lower bound on mutual information, we consider the Kullback-Leibler divergence $KL(p(\mathsf{x}|\mathsf{y})||q(\mathsf{x}|\mathsf{y}))$ between the posterior $p(\mathsf{x}|\mathsf{y})$ and its variational approximation $q(\mathsf{x}|\mathsf{y})$. Non-negativity of the divergence (e.g. Cover and Thomas (1991)) implies

$$\langle \log p(\mathsf{x}|\mathsf{y}) \rangle_{p(\mathsf{x}|\mathsf{y})} - \langle \log q(\mathsf{x}|\mathsf{y}) \rangle_{p(\mathsf{x}|\mathsf{y})} \geq 0 \;\Rightarrow\; \underbrace{\langle \log p(\mathsf{x}|\mathsf{y}) \rangle_{p(\mathsf{x}|\mathsf{y})p(\mathsf{y})}}_{-H(\mathsf{x}|\mathsf{y})} \geq \langle \log q(\mathsf{x}|\mathsf{y}) \rangle_{p(\mathsf{x}|\mathsf{y})p(\mathsf{y})} . \tag{2.1}$$

This leads to

$$I(\mathsf{x},\mathsf{y}) \geq H(\mathsf{x}) + \langle \log q(\mathsf{x}|\mathsf{y}) \rangle_{p(\mathsf{x},\mathsf{y})} \stackrel{\text{def}}{=} \tilde{I}(\mathsf{x},\mathsf{y}), \tag{2.2}$$

where $q(\mathsf{x}|\mathsf{y})$ is an arbitrary distribution, which saturates the bound for $q(\mathsf{x}|\mathsf{y}) \equiv p(\mathsf{x}|\mathsf{y})$. This agrees with the intuition that for a given channel $p(\mathsf{y}|\mathsf{x})$, the optimal decoder should correspond to the Bayesian posterior $p(\mathsf{x}|\mathsf{y})$, though for this case the computation of the mutual information (and its derivatives) may be intractable.

The bound (2.2) explicitly includes both the encoder $p(\mathsf{y}|\mathsf{x})$ and decoder $q(\mathsf{x}|\mathsf{y})$ and has a form similar to the criterion optimized by the Blahut-Arimoto algorithms for channel capacity (Blahut (1972), Arimoto (1972)). However, by analogy with a related variational extension (Neal and Hinton (1998)) of the classic expectation maximization algorithm (Dempster et al. (1977)), it is practical to extend the conventional formulation and *constrain* the decoder $q(\mathsf{x}|\mathsf{y})$ to lie in a tractable family (Barber and Agakov (2003)). These constraints are important, as they help to avoid computational intractability of the theoretically optimal, but practically infeasible decoders derived in the unconstrained formulation (see e.g. Cover and Thomas (1991)). The bound $\tilde{I}(\mathsf{x},\mathsf{y})$ is then optimized for both the

encoder $p(\mathsf{y}|\mathsf{x})$ and the variational distribution $q(\mathsf{x}|\mathsf{y})$ (which is generally different from the posterior $p(\mathsf{x}|\mathsf{y})$), subject to the imposed distribution and tractability constraints. The variational distribution $q(\mathsf{x}|\mathsf{y})$, which we use in the specification of the objective function, may be simply thought of as an auxiliary entity to facilitate computation and optimization of the otherwise intractable mutual information $I(\mathsf{x},\mathsf{y})$. Importantly, our principal goal of maximizing the bound $\tilde{I}(\mathsf{x},\mathsf{y})$ is, as in the exact formulation, to learn the optimal adjustable parameters of the *encoder* $p(\mathsf{y}|\mathsf{x})$.

Other (for example, mean-field type) relaxations of the mutual information may potentially be considered (see e.g. Jaakkola and Jordan (1998)). However, our current experience suggests that for certain choices of the decoder $q(\mathsf{x}|\mathsf{y})$, the variational bound (2.2) considered above is particularly computationally convenient. We will now outline a procedure for optimizing the bound and discuss some of its fundamental properties.

2.1.2 The Variational IM Algorithm

Let $\mathcal{P}$ and $\mathcal{Q}$ denote families of the encoders $p(\mathsf{y}|\mathsf{x})$ and variational decoders $q(\mathsf{x}|\mathsf{y})$ respectively. To assure applicability of the learning for the case of large scale models, the distribution families must be chosen in such a way that the bound (2.2) is tractable. This may be achieved, for example, by imposing appropriate parametric or structural constraints on $\mathcal{Q}$ in such a way that the averaging over the codewords y is practically feasible. In our approximate approach to information maximization, we follow the standard iterative variational procedure and maximize $\tilde{I}(\mathsf{x},\mathsf{y})$ with respect to $p(\mathsf{y}|\mathsf{x}) \in \mathcal{P}$ and $q(\mathsf{x}|\mathsf{y}) \in \mathcal{Q}$. A simple recursive optimization procedure performing information maximization (**IM algorithm**) is then given as follows:

1. For a fixed $q^{(t)}_{X|Y} \in \mathcal{Q}$, find $p^{(t+1)}_{Y|X} = \arg\max_{p_{Y|X}\in\mathcal{P}} \tilde{I}(\mathsf{x},\mathsf{y}; p_{Y|X}, q^{(t)}_{X|Y})$ [**M-step**];

2. For a fixed $p^{(t+1)}_{X|Y} \in \mathcal{P}$, find $q^{(t+1)}_{X|Y} = \arg\max_{q_{X|Y}\in\mathcal{Q}} \tilde{I}(\mathsf{x},\mathsf{y}; p^{(t+1)}_{Y|X}, q_{X|Y})$ [**I-step**];

3. Iterate **M-** and **I-** steps until a convergence criterion is met

where we used t for the iteration number and $p_{Y|X}$, $q_{X|Y}$ for the conditional encoder $p(\mathsf{y}|\mathsf{x})$ and decoder $q(\mathsf{x}|\mathsf{y})$ respectively, and implied the distribution constraints on $p_{Y|X}$ and $q_{X|Y}$.

Proposition 2.1. *The variational IM algorithm is guaranteed to maximize or leave unchanged a lower bound on the mutual information.*

Proof. The proof is straight-forward and follows from (2.2) and the algorithm specification. Let t be the iteration number. From the M-step, it is clear that

$$\langle \log q^{(t)}(\mathsf{x}|\mathsf{y}) \rangle_{p^{(t+1)}(\mathsf{y}|\mathsf{x})p(\mathsf{x})} \geq \langle \log q^{(t)}(\mathsf{x}|\mathsf{y}) \rangle_{p^{(t)}(\mathsf{y}|\mathsf{x})p(\mathsf{x})}. \tag{2.3}$$

Analogously, the I-step leads to

$$\langle \log q^{(t+1)}(\mathsf{x}|\mathsf{y}) \rangle_{p^{(t+1)}(\mathsf{y}|\mathsf{x})p(\mathsf{x})} \geq \langle \log q^{(t)}(\mathsf{x}|\mathsf{y}) \rangle_{p^{(t+1)}(\mathsf{y}|\mathsf{x})p(\mathsf{x})}. \tag{2.4}$$

Then from the transitivity we get

$$\tilde{I}(\mathsf{x}, \mathsf{y}|p_{Y|X}^{(t+1)}, q_{X|Y}^{(t+1)}) \geq \tilde{I}(\mathsf{x}, \mathsf{y}|p_{Y|X}^{(t)}, q_{X|Y}^{(t)}), \tag{2.5}$$

where $\tilde{I}(\mathsf{x}, \mathsf{y})$ is defined as in (2.2). Therefore, the IM algorithm is guaranteed to maximize (or leave unchanged) the *lower bound* on the true mutual information given by $\tilde{I}(\mathsf{x}, \mathsf{y})$. □

In general, applicability of proposition 2.1 depends on the specifics of optimization methods used during the *I*- and *M*- steps. For example, if optimization is performed with respect to parameters of $p(\mathsf{y}|\mathsf{x})$ and $q(\mathsf{x}|\mathsf{y})$ by a numerical ascent in the parameter space, one may encounter the usual problem of non-monotonic convergence (e.g. for inappropriate learning rates, etc.) Practically, the proposition holds if it is possible to express the IM steps in terms of closed-form fixed-point updates (i.e. at each iteration the optima are expressed analytically), or if the steps apply numerical optimization procedures which always lead to monotonic changes in the objective functions (Bishop (1995),Galeev and Tihomirov (2000)).

2.1.2.1 Relation to the Blahut-Arimoto algorithm for maximizing channel capacity

In the cases when the exact mutual information is tractable to compute, one may apply any of the known numerical optimization techniques (e.g. Galeev and Tihomirov (2000)) to optimize $I(\mathsf{x}, \mathsf{y})$ directly. Alternatively, one may optimize $I(\mathsf{x}, \mathsf{y})$ by applying an iterative procedure, known as the Blahut-Arimoto algorithm for maximizing the channel capacity (Arimoto (1972), Blahut (1972)). For a fixed distribution of the source variables $p_X(\mathsf{x})$, the algorithm is given as

1. Find $p_{Y|X}^{(t+1)} = \arg\max_{p_{Y|X} \in \mathcal{P}} \langle \log p_{X|Y}^{(t)}(\mathsf{x}|\mathsf{y}) \rangle_{p_{Y|X}(\mathsf{y}|\mathsf{x})p_X(\mathsf{x})}$;
2. Iterate until convergence.

Here it is assumed that $p_{X|Y}^{(t)}(\mathsf{x}|\mathsf{y}) \propto p_{Y|X}^{(t)}(\mathsf{y}|\mathsf{x})p_X(\mathsf{x})$ is the exact Bayes-optimal decoder expressed from the channel distribution at the t^{th} iteration of the training algorithm.

It is easy to see that the Blahut-Arimoto algorithm corresponds to a special case of the variational IM algorithm described above. Indeed, if the decoder is unconstrained, i.e. if $\mathcal{Q}$ defines a family of *all* possible conditional probability density functions $q(\mathsf{x}|\mathsf{y})$ for $\mathsf{x} \in \mathcal{R}_x$ and $\mathsf{y} \in \mathcal{R}_y$, then the information-maximization algorithm reduces to a form of the Blahut-Arimoto algorithm for learning channel distributions. The proof is straight-forward, and follows from the convexity of $\tilde{I}(\mathsf{x}, \mathsf{y})$ in $q(\mathsf{x}|\mathsf{y})$ and the fact that the bound is saturated for $q(\mathsf{x}|\mathsf{y}) \equiv p(\mathsf{x}|\mathsf{y})$

(see expression (2.2)). Clearly, a reduction of the IM algorithm to the Blahut-Arimoto procedure also occurs when $q(\mathsf{x}|\mathsf{y})$ is constrained to be identical to the true posterior $p(\mathsf{x}|\mathsf{y})$ at each iteration of learning, i.e.

$$q^{(t)}(\mathsf{x}|\mathsf{y}) \equiv p^{(t)}(\mathsf{x}|\mathsf{y}). \tag{2.6}$$

In both of the described special cases, the iterative optimization procedure is guaranteed to maximize or leave unchanged the *exact* value of the mutual information. Indeed, a straight-forward substitution of (2.6) into (2.3) and (2.4) implies

$$\langle \log p^{(t+1)}(\mathsf{x}|\mathsf{y}) \rangle_{p^{(t+1)}(\mathsf{y}|\mathsf{x})p(\mathsf{x})} \geq \langle \log p^{(t)}(\mathsf{x}|\mathsf{y}) \rangle_{p^{(t)}(\mathsf{y}|\mathsf{x})p(\mathsf{x})}, \tag{2.7}$$

i.e. $I^{(t+1)}(\mathsf{x},\mathsf{y}) \geq I^{(t)}(\mathsf{x},\mathsf{y})$. An alternative proof shows that the Blahut-Arimoto algorithm does not decrease the true mutual information, and is given here for completeness.

Proposition 2.2. *The Blahut-Arimoto algorithm is guaranteed to maximize or leave unchanged the exact value of mutual information.*

Proof. The iteration step of the algorithm implies

$$\langle \log p^{(t)}(\mathsf{x}|\mathsf{y}) \rangle_{p^{(t+1)}(\mathsf{y}|\mathsf{x})p(\mathsf{x})} \geq \langle \log p^{(t)}(\mathsf{x}|\mathsf{y}) \rangle_{p^{(t)}(\mathsf{y}|\mathsf{x})p(\mathsf{x})}. \tag{2.8}$$

Then, from non-negativity of $\langle KL(p^{(t+1)}(\mathsf{x}|\mathsf{y}) \| p^{(t)}(\mathsf{x}|\mathsf{y})) \rangle_{p(\mathsf{x})}$, we get

$$\langle \log p^{(t+1)}(\mathsf{x}|\mathsf{y}) \rangle_{p^{(t+1)}(\mathsf{y}|\mathsf{x})p(\mathsf{x})} \geq \langle \log p^{(t)}(\mathsf{x}|\mathsf{y}) \rangle_{p^{(t+1)}(\mathsf{y}|\mathsf{x})p(\mathsf{x})}. \tag{2.9}$$

As before, for a fixed $H(\mathsf{x})$, a combination of (2.8) with (2.9) leads to $I(\mathsf{x},\mathsf{y}|p^{(t+1)}_{Y|X}) \geq I(\mathsf{x},\mathsf{y}|p^{(t)}_{Y|X})$. □

It is known empirically that for convex objective functions, iterative optimizers may be slow to converge (see e.g. the empirical results (Minka (2003)) on convergence of the *iterative scaling* algorithms (Darroch and Ratcliff (1972), Berger et al. (1996), Collins et al. (2002)) for log-linear models). If $I(\mathsf{x},\mathsf{y})$ is convex in the encoder's parameters, we may hypothesize that alternative optimization techniques may outperform the Blahut-Arimoto algorithm in terms of the convergence speed; moreover, computational tractability of the Blahut-Arimoto algorithm would typically imply tractability of alternative optimization methods. Generally, it is clear that in tractable channels one should have a flexibility in choosing a specific optimization procedure.

Importantly, we emphasize once again that for many interesting channels, the Blahut-Arimoto algorithm (and other numerical procedures optimizing $I(\mathsf{x},\mathsf{y})$ directly) could be computationally difficult to apply, as it would require optimization of the generally intractable entropy of a mixture in the optimization steps. The proposed IM algorithm (and its numerical variations) address the problems of computational intractability by introducing an additional functional parameter – the variational decoder $q(\mathsf{x}|\mathsf{y})$, which is constrained to lie in a tractable family. Then our goal is to maximize a proper lower bound on the intractable objective $I(\mathsf{x},\mathsf{y})$ with respect to the variational decoder and the adjustable parameters of the channel distribution, subject to the imposed constraints.

2.1.3 Reconstruction of the Source Patterns

As discussed in Section 1.3, by maximizing the exact mutual information for $p(\mathsf{y}|\mathsf{x})$, we try to establish an optimal way of *encoding* the sources for transmitting them across the channel. Our goal remains unchanged in the approximate variational formulation (though the optimization surface for $p(\mathsf{y}|\mathsf{x})$ will now be affected by a specific choice of the variational distribution). Importantly, the bound $\tilde{I}(\mathsf{x},\mathsf{y})$ does not explicitly specify which decoder should be used at the receiver's end. However, just as in variational approaches to likelihood maximization, it provides an option of reconstructing the source patterns with the learned variational posterior $q(\mathsf{x}|\mathsf{y})$. Here we briefly discuss other possible ways to perform the reconstructions.

2.1.3.1 Reconstructions using the exact posterior

Once the optimal encoder is learned, the receiver may use it to compute the Bayes-optimal decoder $p(\mathsf{x}|\mathsf{y})$, and apply it to reconstructing the original sources $\{\mathsf{x}\}$ for the received representations $\{\mathsf{y}\}$. From (2.2) it is easy to see that using any other decoder (for a fixed learned channel) would indeed be $\tilde{I}$-suboptimal. Unfortunately, using the exact posterior $p(\mathsf{x}|\mathsf{y})$ for decoding may not be plausible for many practical cases. For example, if $p(\mathsf{x})$ is the empirical distribution, computation of $p(\mathsf{x}|\mathsf{y})$ makes a recourse to the training data $\{\mathsf{x}^{(i)}|i=1,\ldots,M\}$ (see (1.9)). In practice, this means that in order to perform the reconstructions, the receiver needs to have an access to the empirical distribution $p(\mathsf{x})$ used during training, which requires storing the entire training set. Moreover, even if the training set is available at the receiver's end of the channel, using $p(\mathsf{x}|\mathsf{y})$ for decoding would constrain the reconstructions to lie in the training set, which in some cases may be restrictive (effectively, this case would correspond to extracting training patterns $\{\mathsf{x}^{(m)}|m=1,\ldots,M\}$ from their noisy *encoded* representations $\{\mathsf{y}\}$).

Finally, for many interesting practical problems, computation of $p(\mathsf{x}|\mathsf{y})$ may be intractable *despite* the tractability of averaging in (2.2). One practical case when this may happen is the problem of syndrome decoding in binary symmetric channels (see e.g. McEliece (1977)). Note that if all the source vectors are binary and equiprobable, i.e. $\forall \mathsf{x} \in \{0,1\}^{|\mathsf{x}|}.\ p(\mathsf{x}) = 1/2^{|\mathsf{x}|}$, computation of the exact posterior

$$p(\mathsf{x}|\mathsf{y}) = \frac{p(\mathsf{y}|\mathsf{x})}{\sum_{m=1}^{2^{|\mathsf{x}|}} p(\mathsf{y}|\mathsf{x}^{(m)})} \tag{2.10}$$

involves a summation over an exponential number of states of the source variables $\mathsf{x} \in \{0,1\}^{|\mathsf{x}|}$, which in most cases of interest is prohibitively expensive. Depending on the structure of the graph, the exect posteriors may in some cases be accurately approximated (Gallager (1963), MacKay (1999a)) by iterative algorithms (e.g. Gallager (1963), Pearl (1988), Murphy et al. (1999)). However, for bipartite graphs, the accuracy of approximations will typically drop with an increase in the graph's connectivity (Burshtein and Miller (2002)), and convergence will generally be a problem.

2.1.3.2 Reconstructions using the approximate posterior

It is clear that in the cases when it is impossible to compute the Bayes-optimal decoder $p(\mathsf{x}|\mathsf{y})$, one needs to consider using alternative decoding schemes. For example, one may need to approximate the exact posteriors $p(\mathsf{x}|\mathsf{y})$, e.g. by using (structured) mean field models (Jaakkola (1997), Barber and Wiegerinck (1998), Jordan et al. (1998), Saad and Opper (2001)), or applying iterative algorithms for approximating the posteriors (Pearl (1988), Yedidia et al. (2000a), Yedidia et al. (2000b), Weiss and Freeman (2001)). Generally, these methods would require approximations of a new posterior for each new received vector y, which may be plausible, but expensive to compute in practice. In the case of applying iterative message-passing algorithms, it may be difficult to ensure the convergence (see Ihler et al. (2005), Mooij and Kappen (2005) for the discussion of sufficient conditions), while approximations yielded by the provenly convergent algorithms (Heskes (2002), Yuille (2002)) may not necessarily be accurate estimates of the exact posteriors. On the strong side, in many practical situations when the propagation algorithms do converge, they often lead to accurate approximations of the posteriors (MacKay and Neal (1999), Yedidia et al. (2000a)), as unlike most of the variational approximations they do not impose explicit structural or parametric constraints on the posteriors.

An attractive feature of the bound (2.2) is availability of the variational decoder $q(\mathsf{x}|\mathsf{y})$, which is learned along with the optimal parameters of the encoder $p(\mathsf{y}|\mathsf{x})$. A quick and simple way to reconstruct the transmitted source vectors would be to use $q(\mathsf{x}|\mathsf{y})$ with the optimized variational parameters. This application of the variational distribution $q(\mathsf{x}|\mathsf{y})$ is analogous to using the variational posterior for inference in generative models after training with the variational EM (Neal and Hinton (1998)) algorithm. The principal advantage of variational decoding is the simplicity of the resulting inference procedure. Clearly, subject to the constraints on the variational decoder, the resulting variational reconstructions are optimal, since optimally the objective function (2.2) is maximized when $q(\mathsf{x}|\mathsf{y}) \equiv p(\mathsf{x}|\mathsf{y})$. Moreover, since the bound (2.2) is based on the KL divergence between the true and the approximating posteriors, optimization for the variational decoder is equivalent to a moment matching approximation of $p(\mathsf{x}|\mathsf{y})$ by $q(\mathsf{x}|\mathsf{y})$, averaged over $p(\mathsf{y})$. This fact may potentially be beneficial in terms of decoding, since the more successful decoding algorithms tend to approximate the mean of the posterior $p(\mathsf{x}|\mathsf{y})$ (Saad and Opper, 2001), whilst standard mode matching approaches (such as mean-field theory) typically get trapped in the one of many sub-optimal modes. We stress, however, that using the variational posterior for inference is just one of the possible ways to reconstruct the sources.

2.1.4 Posterior Approximations

There is an interesting relationship between maximizing the bound $\tilde{I}(\mathsf{x},\mathsf{y})$ in a noisy channel, and computing an optimal estimate of an intractable posterior in a generative graphical model with observations y and hidden variables x.

One of the goals of inference in graphical models is to approximate moments of the generally intractable model-specific posterior $p(\mathsf{x}|\mathsf{y})$, where x is a vector of

hidden variables, and y is a vector of observations. In general, this computation is intractable, and approximations are required. A standard mean field approach approximates the posterior marginal by minimizing the KL divergence:

$$KL(q(\mathsf{x}|\mathsf{y})||p(\mathsf{x}|\mathsf{y})) = \sum_{\mathsf{x}} \{q(\mathsf{x}|\mathsf{y}) \log q(\mathsf{x}|\mathsf{y}) - q(\mathsf{x}|\mathsf{y}) \log p(\mathsf{x}|\mathsf{y})\} \tag{2.11}$$

where $q(\mathsf{x}|\mathsf{y}) = \prod_i q(x_i|\mathsf{y})$. In this case, the KL divergence and its functional gradients with respect to $q(x_i|\mathsf{y})$ are usually tractably computable (up to a neglectable prefactor). Assuming that $q(\mathsf{x}_{\setminus i}|\mathsf{y}) \stackrel{\text{def}}{=} \prod_{j \neq i} q(x_j|\mathsf{y})$ is fixed, we may optimize $KL(q(\mathsf{x}|\mathsf{y})||p(\mathsf{x}|\mathsf{y}))$ for $q(x_i|\mathsf{y})$, and express the variational posterior marginal analytically (from the convexity of the KL). Typically, this results in $q(x_i|\mathsf{y})$ approximating any one of a very large number of local modes of the model-specific posterior $p(x_i|\mathsf{y})$. Clearly, if the goal is to approximate moments of the exact posterior, using mean field decoding which optimizes (2.11) is generally suboptimal.

Alternatively, we may consider

$$\begin{aligned} KL(p(\mathsf{x}|\mathsf{y})||q(\mathsf{x}|\mathsf{y})) &= \sum_{\mathsf{x}} (p(\mathsf{x}|\mathsf{y}) \log p(\mathsf{x}|\mathsf{y}) - p(\mathsf{x}|\mathsf{y}) \log q(\mathsf{x}|\mathsf{y})) \\ &= -\sum_{\mathsf{x}} p(\mathsf{x}|\mathsf{y}) \log q(\mathsf{x}|\mathsf{y}) + c, \end{aligned} \tag{2.12}$$

where c is an irrelevant constant (the inverse entropy of $p(\mathsf{x}|\mathsf{y})$, which is not a function of $q(\mathsf{x}|\mathsf{y})$). This is the correct KL divergence in the sense that, optimally, $q(x_i|\mathsf{y}) = p(x_i|\mathsf{y})$, i.e. the posterior mean is correctly calculated. The difficulty lies in performing averages with respect to the exact posterior $p(\mathsf{x}|\mathsf{y})$, which are generally intractable (since the posterior is fully coupled in x). However, it may be possible to minimize the *average* divergence $\langle KL(p(\mathsf{x}|\mathsf{y})||q(\mathsf{x}|\mathsf{y}))\rangle_{p(\mathsf{y})}$, which leads to maximization of the generic bound (2.2) on the mutual information

$$\sum_{\mathsf{y}} \sum_{\mathsf{x}} p(\mathsf{y}) p(\mathsf{x}|\mathsf{y}) \log q(\mathsf{x}|\mathsf{y}) = \sum_{\mathsf{y}} \sum_{\mathsf{x}} p(\mathsf{x}) p(\mathsf{y}|\mathsf{x}) \log q(\mathsf{x}|\mathsf{y}), \tag{2.13}$$

where we ignored the irrelevant entropy of the source vectors. While for any given y, the best posterior mean estimate may be difficult to compute, we may apply the bound $\tilde{I}(\mathsf{x}, \mathsf{y})$ to calculate the best posterior mean estimate *on average*.

2.2 Tractable Choices of Variational Decoders

In Section 2.1 we discussed a simple variational approach to maximization of mutual information in noisy channels, where the problem of optimizing the computationally intractable objective $I(\mathsf{x}, \mathsf{y})$ was transformed to optimization of a proper lower bound on the objective criterion. It is intuitive that the tightness and the tractability of the described lower bound (2.2) depends on a specific choice of the variational distribution $q(\mathsf{x}|\mathsf{y})$. Here we discuss several of such choices, namely structured and Gaussian variational posteriors.

First, we will show that by considering structured variational posteriors, we may indeed improve on simple factorized variational approximations. Then we

re-formulate the bound (2.2) for the special case when the variational decoder is a linear Gaussian, and show that a specific way of optimizing the resulting objective reduces to maximization of Linsker's *as-if Gaussian* criterion (1.16).

2.2.1 Structured Decoders

If the codes y are predictive of the sources x (which is something that we hope to achieve when we maximize the mutual information for the encoder parameters), then $p(\mathsf{x}|\mathsf{y})$ will typically be sharply peaked around its mode. This motivates a choice of simple (for example, uni-modal) approximations $q(\mathsf{x}|\mathsf{y})$ to the posterior, which in practice may significantly reduce computational complexity of optimization. In general, however, it is intuitive that the tightness of the bound may depend on the choice of the decoder $q(\mathsf{x}|\mathsf{y})$. For computational and representational purposes it is often convenient to assume that the approximate posterior $q(\mathsf{x}|\mathsf{y})$ in the variational bound (2.2) is factorized[1] in x. Here we briefly describe effects which such a choice of the decoder structure may have on the bound. We also point out natural extensions of (2.2) to chain-type decoders, which overcome some of the restrictions of the factorized assumption.

2.2.1.1 Effects of the Decoder Structure

First of all, to demonstrate an influence which a choice of the decoder structure may have on the bound (2.2), we consider a simple case of factorized, mean-field type decoders for a model with $|\mathsf{x}|$ input and $|\mathsf{y}|$ output variables. For illustration purposes, we considered three decoders with different structures shown on Figure 2.1, so that

$$q^{(1)}(\mathsf{x}|\mathsf{y}) = \prod_{i=1}^{|\mathsf{x}|} q(x_i|\mathsf{y}), \qquad q^{(2)}(\mathsf{x}|\mathsf{y}) = \prod_{i=1}^{|\mathsf{x}|} q(x_i|y_i), \qquad q^{(3)}(\mathsf{x}|\mathsf{y}) = \prod_{i=1}^{|\mathsf{x}|} q(x_i), \tag{2.14}$$

where the products are computed over all the dimensions of x. Note that in all these cases, different dimensions of the reconstructed vectors are conditionally independent given the codes. Arguably, this makes the considered decoders intrinsically less powerful than the theoretically optimal decoder defined by the true posterior $p(\mathsf{x}|\mathsf{y})$, where the source variables are generally conditionally dependent. However, it is easy to see that the considered structural constraints on $q(\mathsf{x}|\mathsf{y})$ may significantly simplify the resulting optimization procedures.

By computing the functional derivatives of the bound (2.2) with respect to the specified decoders (2.14) subject to the normalizing constraints, it is easy to show that for each of the three cases, the optimal variational distributions are defined as

$$q^{(1)}(x_i|\mathsf{y}) = p(x_i|\mathsf{y}), \quad q^{(2)}(x_i|\mathsf{y}) = q^{(2)}(x_i|y_i) = p(x_i|y_i), \quad q^{(3)}(x_i|\mathsf{y}) = q^{(3)}(x_i) = p(x_i). \tag{2.15}$$

[1]The assumption becomes particularly important when the set of possible vectors generated by $p(\mathsf{x})$ is exponentially large; for example, when $\mathsf{x} \in \{0,1\}^{|\mathsf{x}|}$ and $p(\mathsf{x})$ is uniform.

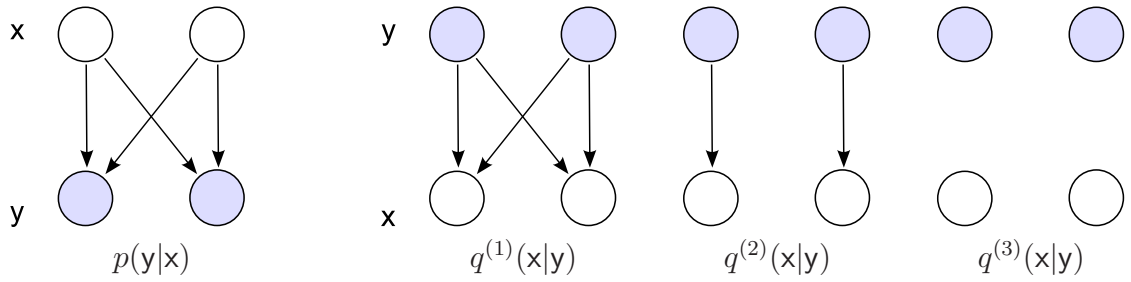


Figure 2.1: Simple variational decoders. *Left:* A noisy encoder $p(\mathsf{y}|\mathsf{x})$; *Right:* constrained variational decoders $q^{(1)}(\mathsf{x}|\mathsf{y})$, $q^{(2)}(\mathsf{x}|\mathsf{y})$, and $q^{(3)}(\mathsf{x}|\mathsf{y})$. The shaded nodes correspond to the hidden encoded representations $\{\mathsf{y}\}$. The transparent nodes correspond to the sources $\{\mathsf{x}\}$ and their reconstructions. (Unless mentioned otherwise, we will use the shaded nodes to indicate unobservable variables).

Then it is clear that for the variational decoder $q^{(1)}(\mathsf{x}|\mathsf{y})$, the optimal achievable bound (2.2) on $I(\mathsf{x},\mathsf{y})$ is given by

$$\tilde{I}^{(1)}(\mathsf{x},\mathsf{y}) = H(\mathsf{x}) - \sum_{i=1}^{|\mathsf{x}|} H(x_i|\mathsf{y}), \tag{2.16}$$

where $H(x_i|\mathsf{y}) \stackrel{\text{def}}{=} -\langle \log p(x_i|\mathsf{y})\rangle_{p(\mathsf{y}|\mathsf{x})p(\mathsf{x})}$. Analogously, for $q^{(2)}(x_i|\mathsf{y})$ and $q^{(3)}(x_i|\mathsf{y})$ we obtain

$$\tilde{I}^{(2)}(\mathsf{x},\mathsf{y}) = H(\mathsf{x}) - \sum_{i=1}^{|\mathsf{x}|} H(x_i|y_i), \qquad \tilde{I}^{(3)}(\mathsf{x},\mathsf{y}) = H(\mathsf{x}) - \sum_{i=1}^{|\mathsf{x}|} H(x_i), \tag{2.17}$$

with the similar definitions of the conditional entropies. It is easy to see that since the conditioning decreases the entropy (see e.g. Cover and Thomas (1991)), and all the variables for the three models are defined on identical domains, we get $H(x_i|\mathsf{y}) \leq H(x_i|y_i) \leq H(x_i)$. This leads to the intuitive relation

$$I(\mathsf{x},\mathsf{y}) \geq \tilde{I}^{(1)}(\mathsf{x},\mathsf{y}) \geq \tilde{I}^{(2)}(\mathsf{x},\mathsf{y}) \geq \tilde{I}^{(3)}(\mathsf{x},\mathsf{y}), \tag{2.18}$$

which demonstrates that simplifications of the decoder structure (see Figure 2.1) result in relaxations of the theoretically achievable lower bounds.

In the case that the codes y contain little information about the sources x (for example, if the channel noise is high or dimensionality of the codewords $|\mathsf{y}|$ is low), factorized uni-modal approximations of the posterior $p(\mathsf{x}|\mathsf{y})$ may be quite restrictive. In this case we may expect that a decrease in the theoretically optimal bound due to a sub-optimal choice of the decoder structure may be quite significant.

Example: To demonstrate the effects numerically, we have considered a simple binary model with $|\mathsf{x}| = 2$ source and $|\mathsf{y}| = 2$ output variables. The encoder mapping was defined in such a way that both units $y_1, y_2 \in \{0, 1\}$ were solving

a noisy XOR problem with the noise levels $\epsilon_1 = 0.2$ and $\epsilon_2 = 0.3$. The channel probabilities for this case were defined as $p(y_i = 1|x_1 = x_2) = \epsilon_i$ and $p(y_i = 1|x_1 \neq x_2) = 1 - \epsilon_i$, where the conditioning indicates whether the source units $x_1, x_2 \in \{0, 1\}$ are set to the same binary states. The source variable marginals were defined as $p(x_1 = 1) = 0.2$ and $p(x_2 = 1) = 0.5$. It turns out that in this case the exact value of the mutual information is given by $I \approx 0.3548$, with the bounds $I^{(1)} \approx 0.1161$, $I^{(2)} \approx 0.0630$, and $I^{(3)} = 0$.

The result that a richer structure of a variational decoder $q(\mathsf{x}|\mathsf{y})$ leads to an improvement of the lower bound on $I(\mathsf{x}, \mathsf{y})$ is in full agreement with our expectations and with (2.18). It is also related to previous results in approximate inference and variational likelihood learning, which show that compared to the fully factorized approximations, structured mean field models (e.g. Barber and Wiegerinck (1998), Saad and Opper (2001)) often lead to tighter bounds on the log partition function. We will now consider a simple choice of the decoder structure which helps to go beyond the simple factorized approximations.

2.2.1.2 Chain Decoders

A possible limitation of the considered variational decoders $q(\mathsf{x}|\mathsf{y})$ is the factorized assumption $q(\mathsf{x}|\mathsf{y}) = \prod_{i=1}^{|\mathsf{x}|} q(x_i|\mathsf{y})$, which may lead to inaccurate approximations of the exact posterior. Specifically, for the factorized channel $p(\mathsf{y}|\mathsf{x}) = \prod_{i=1}^{|\mathsf{y}|} p(y_i|\mathsf{x})$, the posterior $p(\mathsf{x}|\mathsf{y})$ is fully coupled in x (see Figures 2.2 (a), (b)). While in some cases the factorized constraints may be plausible and indeed physically and computationally justifiable, in general they may be too restrictive. Intuitively, the bound on $I(\mathsf{x}, \mathsf{y})$ could be made tighter by retaining some of the structure in the specification of the variational distribution $q(\mathsf{x}|\mathsf{y})$. We will now show how this may be achieved in the context of information maximization.

A straight-forward application of the chain rule for probabilities of multivariate distributions

$$p(\mathsf{x}) = p(x_1) \prod_{i=2}^{|\mathsf{x}|} p(x_i|x_1, \ldots, x_{i-1}) \tag{2.19}$$

leads to the corresponding rule for entropies

$$\begin{aligned} H(\mathsf{x}) &= -\langle \log p(x_1) \rangle_{p(\mathsf{x})} - \sum_{i=2}^{|\mathsf{x}|} \langle \log p(x_i|x_1, \ldots, x_{i-1}) \rangle_{p(\mathsf{x})} \\ &= H(x_1) + \sum_{i=2}^{|\mathsf{x}|} H(x_i|x_1, \ldots, x_{i-1}). \end{aligned} \tag{2.20}$$

Analogously, by conditioning on y and averaging over $p(\mathsf{y})$, we may obtain a simple expression for the conditional entropy $H(\mathsf{x}|\mathsf{y})$:

$$H(\mathsf{x}|\mathsf{y}) = H(x_1|\mathsf{y}) + \sum_{i=2}^{|\mathsf{x}|} H(x_i|x_1, \ldots, x_{i-1}, \mathsf{y}). \tag{2.21}$$

A straight-forward application of the definition of the mutual information (1.6) leads to

$$I(\mathsf{x},\mathsf{y}) \;=\; I(x_1,\mathsf{y}) + \sum_{i=2}^{|\mathsf{x}|} I(x_i,\mathsf{y}|x_1,\ldots,x_{i-1}), \tag{2.22}$$

which is just a chain rule on the mutual information (see e.g. Cover and Thomas (1991)). Equivalently, (2.22) may be expressed as

$$I(\mathsf{x},\mathsf{y}) = H(\mathsf{x}) + \langle \log p(x_1|\mathsf{y})\rangle_{p(x_1,\mathsf{y})} + \sum_{i=2}^{|\mathsf{x}|} \langle \log p(x_i|x_1,\ldots,x_{i-1},\mathsf{y})\rangle_{p(x_1,\ldots,x_{i-1},\mathsf{y})}. \tag{2.23}$$

Clearly, the marginal entropy $H(\mathsf{x})$ in (2.23) is independent of the encoder distribution $p(\mathsf{y}|\mathsf{x})$ and therefore has no effect on the optimization surface. However, the complexity of evaluating each average of $\log p(x_i|x_1,\ldots,x_{i-1},\mathsf{y})$ over the joint distribution $p(\mathsf{x},\mathsf{y})$ is in general exponential in the number of parents of each variable x_i. One obvious way to bound the mutual information in a tractable way, while retaining some of the structure of $p(\mathsf{x}|\mathsf{y})$ in (2.23), would be to limit the number of parental connections in the variational approximation of $p(x_i|x_1,\ldots,x_{i-1},\mathsf{y})$. If $\boldsymbol{\pi}^x(x_i) \subseteq \{x_i|i=1,\ldots,|\mathsf{x}|\}$, $\boldsymbol{\pi}^y(x_i) \subseteq \{y_i|i=1,\ldots,|\mathsf{y}|\}$ are the $\mathsf{x}-$ and $\mathsf{y}-$parents of the i^{th} reconstructed variable x_i, we may define the variational distribution to satisfy

$$q(x_i|\{x\}\backslash x_i,\{y\}) = q(x_i|\boldsymbol{\pi}^x(x_i),\boldsymbol{\pi}^y(x_i)). \tag{2.24}$$

It is clear that the definition (2.24) formally generalizes (2.15). Moreover, in general the resulting conditional $q(\mathsf{x}|\mathsf{y})$ is not factorized in x_i (see Figure 2.2).

It is easy to see that the choice of (2.24) gives rise to the bound

$$I(\mathsf{x},\mathsf{y}) \geq \tilde{I}(\mathsf{x},\mathsf{y}) \stackrel{\text{def}}{=} H(\mathsf{x}) + \sum_i \langle \log q(x_i|\boldsymbol{\pi}^x(x_i),\boldsymbol{\pi}^y(x_i))\rangle_{p(\mathsf{x}_i,\boldsymbol{\pi}^x(x_i),\boldsymbol{\pi}^y(x_i))}, \tag{2.25}$$

Clearly, if the number of parents $|\boldsymbol{\pi}^x(x_i)| + |\boldsymbol{\pi}^y(x_i)|$ is small, the summations in (2.25) may be performed exactly (for discrete domains) or efficiently approximated numerically (e.g. by using standard techniques of numerical integration for continuous domains). A simple structured decoder can be characterized by a sparse mapping from y to x and a chain in the x's (see Figure 2.2), though any choice of parents satisfying the tractability constraints is possible. In general, an extra care must be taken to ensure that $q(\mathsf{x}|\mathsf{y})$ remains a proper distribution (i.e. its graph does not have directed cycles). Clearly, some of the possible choices for the variational decoder $q(\mathsf{x}|\mathsf{y})$ are trees or higher-order Markov chains, with an additional constraint on $|\boldsymbol{\pi}^y(x_i)|$.

By analogy with the simple factorized example (2.15) described above, by computing functional derivatives of the bound (2.25) with respect to $q(x_i|\boldsymbol{\pi}^x(x_i),\boldsymbol{\pi}^y(x_i))$ subject to the normalization constraints, we can easily find that the optimal settings would give rise to $q(x_i|\boldsymbol{\pi}^x(x_i),\boldsymbol{\pi}^y(x_i)) = p(x_i|\boldsymbol{\pi}^x(x_i),\boldsymbol{\pi}^y(x_i))$. This would lead to the optimal theoretically achievable lower bound

$$I(\mathsf{x},\mathsf{y}) \geq H(\mathsf{x}) - \sum_i H(x_i|\boldsymbol{\pi}^x(x_i),\boldsymbol{\pi}^y(x_i)). \tag{2.26}$$

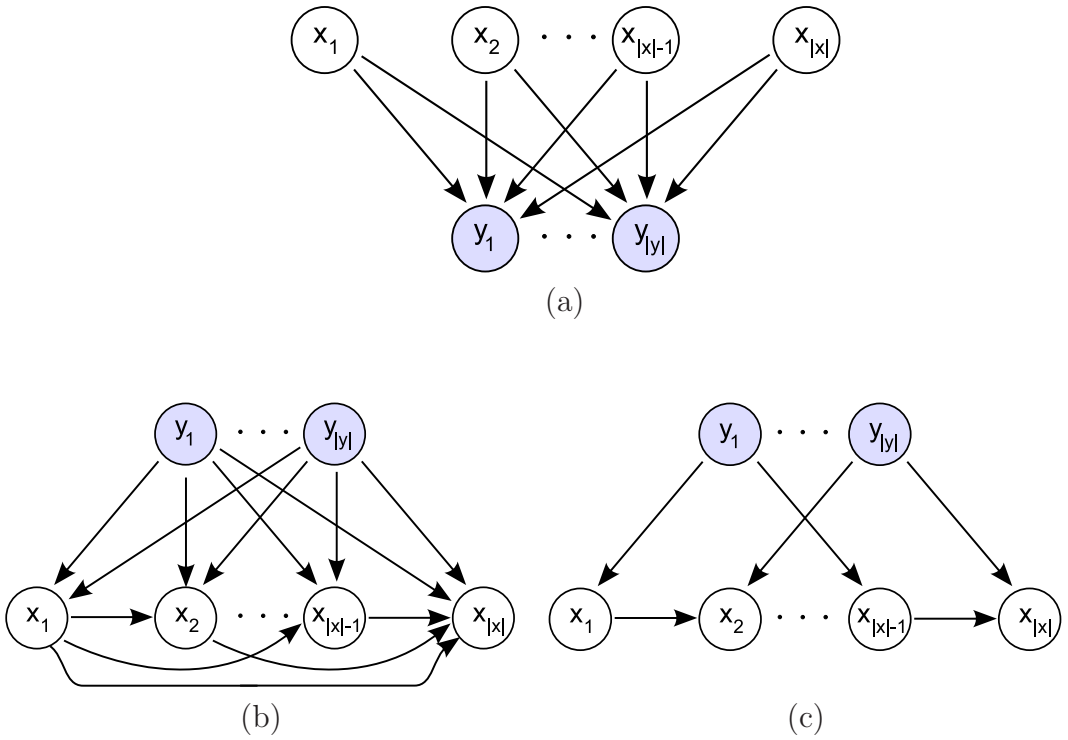


Figure 2.2: Structured variational decoders. *(a)* A noisy encoder $p(\mathsf{y}|\mathsf{x})$; *(b)* the corresponding fully-structured exact decoder $p(\mathsf{x}|\mathsf{y})$; *(c)* a sparse variational decoder $q(\mathsf{x}|\mathsf{y})$. The shaded nodes correspond to the encodings $\{\mathsf{y}\}$. The sparse variational decoder retains some of the structure of the exact posterior $p(\mathsf{x}|\mathsf{y})$.

(compare with (2.23)). Again, from the non-negativity of the KL divergence it is easy to see that unless the variables in $\boldsymbol{\pi}^x(x_i), \boldsymbol{\pi}^y(x_i)$ are independent, the conditioning on additional variables always decreases the entropy, independently of the specifics of channel parameterization. Therefore, by retaining more structure in the variational decoders, we are guaranteed to improve on the theoretically achievable lower bounds on $I(\mathsf{x}, \mathsf{y})$.

Finally, note that even in the cases when the models are defined over discrete spaces with finite alphabets of the code and source variables, it may be difficult to compute the entropic terms in (2.26); for example, this may happen when averaging over $p(\mathsf{x})$ is expensive. However, by imposing sparsity constraints on the *encoding* distribution $p(\mathsf{y}|\mathsf{x})$, we may be able to perform the computations exactly (see the discussion in Section 7.2.2). Obviously, for continuous variable spaces it is necessary to impose additional parametric constraints on $p(\mathsf{y}|\mathsf{x})$ and $q(\mathsf{x}|\mathsf{y})$.

2.2.2 Gaussian Decoders and the Link to Linsker's Criterion

Here we consider a special case of the bound $\tilde{I}(\mathsf{x}, \mathsf{y})$ for linear Gaussian variational decoders. We also show that optimization of Linsker's *as-if* Gaussian objective (Linsker (1992)), approximating the exact mutual information $I(\mathsf{x}, \mathsf{y})$,

corresponds to a special way of optimizing the variational lower bound $\tilde{I}(\mathsf{x},\mathsf{y})$ independently of the choice of the encoder distribution $p(\mathsf{y}|\mathsf{x})$.

Let us define the variational decoder to be a Gaussian $q(\mathsf{x}|\mathsf{y}) \sim \mathcal{N}_{\mathsf{x}}(\mathsf{U}\mathsf{y}, \mathbf{\Sigma})$, where $\mathsf{U} \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{y}|}$. Clearly, in this case the variational lower bound (2.2) is expressed as

$$\begin{aligned}\tilde{I}(\mathsf{x},\mathsf{y}) &= \langle \log q(\mathsf{x}|\mathsf{y})\rangle_{p(\mathsf{y}|\mathsf{x})p(\mathsf{x})} + H(\mathsf{x}) \\ &= -\frac{1}{2}\left\langle \mathrm{tr}\left\{\mathbf{\Sigma}^{-1}(\mathsf{x}-\mathsf{U}\mathsf{y}))(\mathsf{x}-\mathsf{U}\mathsf{y}))^T\right\}\right\rangle_{p(\mathsf{y}|\mathsf{x})p(\mathsf{x})} - \frac{1}{2}\log|\mathbf{\Sigma}| + c, \quad (2.27)\end{aligned}$$

where c is an irrelevant constant. One way to optimize (2.27) would be by applying the iterative IM algorithm for the encoder $p(\mathsf{y}|\mathsf{x})$ and parameters of the decoder U, $\mathbf{\Sigma}$. Alternatively, at the first iteration of the algorithm we may utilize Gaussianity of the decoder $q(\mathsf{x}|\mathsf{y})$ to find analytical expressions of the parameters U and $\mathbf{\Sigma}$, expressing them as functions of the fixed encoder $p(\mathsf{y}|\mathsf{x})$. By substituting the results into (2.27), we may re-define the bound as a function of the encoder alone. Then the new objective function may be optimized numerically with respect to encoder parameters. It turns out that this approach to maximizing the lower bound (2.27) leads to optimization of Linsker's *as-if* Gaussian criterion (1.16).

Proposition 2.3. *For any channel $p(\mathsf{y}|\mathsf{x})$, optimization of Linsker's* as-if *Gaussian criterion $I_G(\mathsf{x},\mathsf{y}) \stackrel{\text{def}}{=} -\log\left|\langle \mathsf{x}\mathsf{x}^T\rangle - \langle \mathsf{x}\mathsf{y}^T\rangle\langle \mathsf{y}\mathsf{y}^T\rangle^{-1}\langle \mathsf{y}\mathsf{x}^T\rangle\right|$ for the encoder parameters corresponds to a specific way of optimizing the variational lower bound on the mutual information $\tilde{I}(\mathsf{x},\mathsf{y}) = \langle \log q(\mathsf{x}|\mathsf{y})\rangle_{p(\mathsf{y}|\mathsf{x})p(\mathsf{x})}$ with linear Gaussian decoders*[2] *$q(\mathsf{x}|\mathsf{y}) \sim \mathcal{N}_{\mathsf{x}}(\mathsf{U}\mathsf{y}, \mathbf{\Sigma})$.*

Proof. Optimizing (2.27) for the decoder covariance $\mathbf{\Sigma}$ (assumed to be non-singular), we obtain the extremum condition

$$\mathbf{\Sigma}^{-1}\left\langle (\mathsf{x}-\mathsf{U}\mathsf{y})(\mathsf{x}-\mathsf{U}\mathsf{y})^T\right\rangle_{p(\mathsf{x})p(\mathsf{y}|\mathsf{x})}\mathbf{\Sigma}^{-1} = \mathbf{\Sigma}^{-1}. \quad (2.28)$$

Clearly, this results in the optimal decoder's covariance

$$\mathbf{\Sigma} = \left\langle (\mathsf{x}-\mathsf{U}\mathsf{y})(\mathsf{x}-\mathsf{U}\mathsf{y})^T\right\rangle_{p(\mathsf{x})p(\mathsf{y}|\mathsf{x})}. \quad (2.29)$$

It is easy to verify that (2.29) is indeed a maximum. Substitution into (2.27) leads to

$$\tilde{I}(\mathsf{x},\mathsf{y}) \propto -\log\left|\left\langle (\mathsf{x}-\mathsf{U}\mathsf{y})(\mathsf{x}-\mathsf{U}\mathsf{y})^T\right\rangle_{p(\mathsf{x},\mathsf{y})}\right|, \quad (2.30)$$

where we ignored irrelevant additive and multiplicative constants. Computing the derivatives for the decoder weights U, we obtain

$$\mathbf{\Sigma}^{-1}\langle \mathsf{x}\mathsf{y}^T\rangle = \mathbf{\Sigma}^{-1}\mathsf{U}\langle \mathsf{y}\mathsf{y}^T\rangle. \quad (2.31)$$

Assuming that $\langle \mathsf{y}\mathsf{y}^T\rangle \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{y}|}$ is non-singular (e.g. this is the case when the number of linearly independent codes exceeds their dimensionality), we may express

[2]Here we will focus on the discussion of the centered data and centered codes, i.e. $\langle \mathsf{x}\rangle = 0$, $\langle \mathsf{y}\rangle = 0$. The more general case may be obtained analogously by considering variational decoders of the form $q(\mathsf{x}|\mathsf{y}) \sim \mathcal{N}_{\mathsf{x}}(\mathsf{U}(\mathsf{y}-\langle \mathsf{y}\rangle), \mathbf{\Sigma})$ (see a brief discussion in Appendix A).

the optimal weights as $\mathsf{U} = \langle \mathsf{x}\mathsf{y}^T \rangle \langle \mathsf{y}\mathsf{y}^T \rangle^{-1} \in \mathbb{R}^{|\mathsf{x}| \times |\mathsf{y}|}$. By substituting the optimal decoder weights into (2.30), we obtain the objective function

$$\tilde{I}(\mathsf{x}, \mathsf{y}) = -\log \left| \langle \mathsf{x}\mathsf{x}^T \rangle - \langle \mathsf{x}\mathsf{y}^T \rangle \langle \mathsf{y}\mathsf{y}^T \rangle^{-1} \langle \mathsf{y}\mathsf{x}^T \rangle \right|, \tag{2.32}$$

which, up to irrelevant constants, is exactly the Linsker's Gaussian criterion (1.16). □

In general the described modification of the objective criterion may complicate the analysis of optimal solutions, as (2.32) may be highly nonlinear in the encoder parameters. Moreover, by expressing the bound as a function of the encoder alone, we may change the optimization surface (for example, if the bound is not a convex function of encoder parameters), which may affect the obtained optima. Nevertheless, by an argument similar to Proposition 2.1, it is easy to see that the described optimization procedure is guaranteed not to weaken $\tilde{I}(\mathsf{x}, \mathsf{y})$, though the obtained solutions will depend on the specifics of the optimization strategies used.

Finally, we stress once again that optimization of the *as-if* Gaussian approximation of the mutual information (Linsker (1992)) corresponds to one specific way of optimizing the variational lower bound $\tilde{I}(\mathsf{x}, \mathsf{y})$ under the specific assumption that the variational decoder is a linear Gaussian. It is clear that the variational formulation is a powerful generalization of Linsker's approach. First, the generic lower bound (2.2) gives us a flexibility in choosing the family of decoder distributions $\mathcal{Q}$. Secondly, for any decoder in the family $q(\mathsf{x}|\mathsf{y}, \boldsymbol{\Theta}) \in \mathcal{Q}$, we a free to choose a specific optimization procedure for the encoder and decoder parameters. Optimization of Linsker's bound corresponds to one of such choices for the family of linear Gaussian decoders. As we show in the following sections, by considering a richer family of variational decoders (and in fact a richer family of bounds on the mutual information) we may significantly improve on simple approximations.

2.3 An Auxiliary Variational Lower Bound on Mutual Information

In section 2.2.1 we showed that by retaining in the variational decoder $q(\mathsf{x}|\mathsf{y})$ some of the local sub-structure of the fully-coupled exact posterior $p(\mathsf{x}|\mathsf{y})$, we could indeed obtain tighter bounds on the mutual information. We also noted that under the specific structural constraints, the computations should indeed remain tractable, thus resulting in simple ways of improving on factorized mean field bounds on the mutual information. Here we discuss a formal generalization of structured bounds on $I(\mathsf{x}, \mathsf{y})$ for mixture-of-experts type decoders.

2.3.1 Representations

A possible way to further improve on the structured bounds (2.25) is to increase the representational power of structured variational decoders $q(\mathsf{x}|\mathsf{y})$ by capturing *global* dependencies between the reconstructed variables. One way to achieve it is

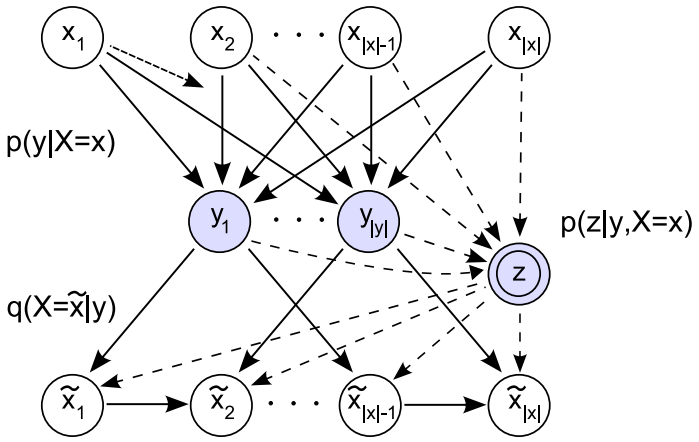


Figure 2.3: A noisy channel $p(\mathsf{y}|\mathsf{x})$ with a structured mixture-type decoder $q(\mathsf{x}|\mathsf{y})$. (The states of the reconstructed variables are denoted by $\tilde{\mathsf{x}}$). The role of the auxiliary variables z is to introduce additional structure into the decoder; they are *not* transmitted across the channel $p(\mathsf{y}|\mathsf{x})$ and do not explicitly constrain $p(\mathsf{x},\mathsf{y})$. The dashed lines show the mappings to and from the auxiliary space. The auxiliary node is shown by the double circle.

to consider multi-modal decoders $q(\mathsf{x}|\mathsf{y}) = \langle q(\mathsf{x}|\mathsf{y},\mathsf{z})\rangle_{q(\mathsf{z}|\mathsf{y})}$, where the introduced *auxiliary variables* z are effectively the unknown mixture labels. Effectively, this choice of the variational decoder corresponds to a form of mixture-of-experts-type models (Jacobs et al. (1991), Neuneier et al. (1994), Bishop (1994), Bishop (1995)) for modeling of the conditional distributions. Clearly, the fully-coupled structure of the resulting variational distribution $q(\mathsf{x}|\mathsf{y})$ will qualitatively agree with the dependency structure of the exact posterior, as different dimensions of the reconstructed vectors x would be coupled through the auxiliary variables z. Moreover, for any interesting choice of the auxiliary space $\{\mathsf{z}\}$, the decoder $q(\mathsf{x}|\mathsf{y})$ will typically be multi-modal, which would agree with the common mixture-type structure of $p(\mathsf{x}|\mathsf{y})$ (see (1.8)). We may therefore intuitively hope that this choice of the variational posterior will generally result in tighter bounds on $I(\mathsf{x},\mathsf{y})$.

Despite the apparent merits of the mixture-type variational decoder $q(\mathsf{x}|\mathsf{y})$, its possible disadvantage relates to the fact that specifying the conditional mixing coefficients $q(\mathsf{z}|\mathsf{y})$ in a principled manner may be rather difficult. Moreover, if the auxiliary variables z are conditionally independent from the *original* source patterns given the codes, any noise in the stochastic encodings y will affect determining of the mixing states. Intuitively, this may have an overwhelming negative effect on decoding, which may cause relaxations of the bound on $I(\mathsf{x},\mathsf{y})$. We may therefore wish to reduce the effects which the noise in the encodings has on the specification of the decoder $q(\mathsf{x}|\mathsf{y})$.

One way to address this matter is by introducing an additional mapping to the auxiliary variable space $p(\mathsf{z}|\mathsf{x},\mathsf{y})$, which may be thought of as an additional variational parameter (see Figure 2.3). Indeed, even when the channel is noisy, the conditional dependence of the auxiliary variables z on the unperturbed source patterns could result in an accurate detection of the states of the auxiliary vari-

ables. Note that the *auxiliary conditional* distribution $p(\mathsf{z}|\mathsf{x},\mathsf{y})$ is defined in a way that does not affect the original noisy channel $p(\mathsf{y}|\mathsf{x})$, as the channel would remain a marginal of the joint distribution of the original sources, codes, and auxiliary variables

$$p(\mathsf{x},\mathsf{y},\mathsf{z}) = p(\mathsf{x})p(\mathsf{y}|\mathsf{x})p(\mathsf{z}|\mathsf{x},\mathsf{y}). \tag{2.33}$$

The role of the auxiliary variables z in this context would be to capture global features of the transmitted sources, and use these features for choosing optimal experts in the mixture-of-experts type[3] decoder. Importantly, the auxiliary variables z are *not* transmitted across the channel. Their purpose here is to define a richer family of bounds on $I(\mathsf{x},\mathsf{y})$ which would generalize over the simpler bounds (2.16)–(2.17), (2.25) with factorial or structured variational decoders.

It is easy to see that for an unconstrained choice of the variational parameters, a straight-forward application of the generic bound (2.2) for the discussed formulation of the mixture-type decoder may result in generally intractable integrals over x and y. A simple way to handle the intractability is to consider further variational relaxations of the lower bound (2.2) by constraining all the variational distributions to be tractable. We will now show that this indeed leads to a tractable generalization over (2.2).

2.3.2 An Auxiliary Variational Lower Bound on $I(\mathsf{x},\mathsf{y})$

An *auxiliary variational* lower bound on $I(\mathsf{x},\mathsf{y})$ may be obtained by considering the general properties of the mutual information. It may also be shown that the resulting bound corresponds to a tractable variational relaxation of the generic form (2.2) for the mixture-of-experts type decoders.

Let $p(\mathsf{x},\mathsf{y},\mathsf{z})$ define a general joint distribution over the sources x, the encodings y, and the auxiliary variables (features) z, where we parameterize the original channel $p(\mathsf{y}|\mathsf{x})$ and the auxiliary variational conditional $p(\mathsf{z}|\mathsf{x},\mathsf{y})$, and assume that $p(\mathsf{x})$ is known and fixed (see expression (2.33) and Figure 2.3). From the chain rule for mutual information (2.22), we may express $I(\mathsf{y},\mathsf{x})$ as

$$I(\mathsf{y},\mathsf{x}) = I(\{\mathsf{z},\mathsf{y}\},\mathsf{x}) - I(\mathsf{x},\mathsf{z}|\mathsf{y}), \tag{2.34}$$

where

$$I(\{\mathsf{z},\mathsf{y}\},\mathsf{x}) \stackrel{\text{def}}{=} H(\mathsf{x}) - H(\mathsf{x}|\mathsf{z},\mathsf{y}) \tag{2.35}$$

is the amount of information that the features z and codes y jointly contain about the sources, and

$$I(\mathsf{x},\mathsf{z}|\mathsf{y}) \stackrel{\text{def}}{=} H(\mathsf{z}|\mathsf{y}) - H(\mathsf{z}|\mathsf{x},\mathsf{y}) \tag{2.36}$$

is the conditional mutual information. Substituting (2.35) and (2.36) into (2.34), we obtain a general expression of the mutual information $I(\mathsf{x},\mathsf{y})$ as a function of conditional entropies of the sources, codes, and auxiliary variables

$$I(\mathsf{y},\mathsf{x}) = H(\mathsf{x}) + H(\mathsf{z}|\mathsf{x},\mathsf{y}) - H(\mathsf{x}|\mathsf{y},\mathsf{z}) - H(\mathsf{z}|\mathsf{y}). \tag{2.37}$$

[3]One may immediately notice that the optimal *unconstrained* variational distribution of the mixing labels $q(\mathsf{z}|\mathsf{y})$ of the decoder $q(\mathsf{x}|\mathsf{y}) = \langle q(\mathsf{x}|\mathsf{y},\mathsf{z})\rangle_{q(\mathsf{z}|\mathsf{y})}$ should itself be a mixture-of-experts model, defined as $q(\mathsf{z}|\mathsf{y}) = \langle p(\mathsf{z}|\mathsf{x},\mathsf{y})\rangle_{p(\mathsf{x}|\mathsf{y})}$. This is confirmed at a later stage (see the objective (2.38)).

Clearly, $H(\mathsf{x})$ in (2.37) is not a function of the channel parameters, and it may be safely ignored for the purpose of learning the optimal encoder $p(\mathsf{x}|\mathsf{y})$. The entropic term $H(\mathsf{z}|\mathsf{x},\mathsf{y}) = -\langle \log p(\mathsf{z}|\mathsf{x},\mathsf{y})\rangle_{p(\mathsf{z},\mathsf{x},\mathsf{y})}$ is a functional of the auxiliary conditional mapping to the feature space $p(\mathsf{z}|\mathsf{x},\mathsf{y})$, which may be chosen to lie in a tractable family (or defined to be a deterministic function of x and y). Computations of the remaining terms in (2.37) are in general problematic, as both posteriors $p(\mathsf{x}|\mathsf{y},\mathsf{z})$ and $p(\mathsf{z}|\mathsf{y})$ are effectively mixture distributions (coupled in x and z respectively). However, by analogy with (2.2) we may bound both terms, which would result in

$$I(\mathsf{x},\mathsf{y}) \geq H(\mathsf{x}) + H(\mathsf{z}|\mathsf{x},\mathsf{y}) + \langle \log q(\mathsf{x}|\mathsf{y},\mathsf{z})\rangle_{p(\mathsf{x},\mathsf{y},\mathsf{z})} + \langle \log q(\mathsf{z}|\mathsf{y})\rangle_{p(\mathsf{y},\mathsf{z})}. \tag{2.38}$$

This may be further recognized as a variational relaxation of the generic criterion (2.2) for a mixture-of-experts decoder $q(\mathsf{x}|\mathsf{y}) = \langle q(\mathsf{x}|\mathsf{y},\mathsf{z})\rangle_{q(\mathsf{z}|\mathsf{y})}$, where the auxiliary variational conditional $p(\mathsf{z}|\mathsf{x},\mathsf{y})$ plays the role of the additional variational parameter.

Again, to ensure that the averages in (2.38) are tractable, we need to constrain the variational decoders $q(\mathsf{x}|\mathsf{y},\mathsf{z})$ and $q(\mathsf{z}|\mathsf{y})$ (in addition to constraining the auxiliary mapping $p(\mathsf{z}|\mathsf{x},\mathsf{y})$). The objective (2.38) needs to be optimized for the channel encoder, variational decoder, and the auxiliary conditional distributions, subject to the imposed constraints. One way to perform the optimization is by considering a straight-forward extension of the variational IM algorithm to include the auxiliary mappings. By analogy with proposition 2.1, it is easy to show that the resulting iterative optimization procedure maximizes or leaves unchanged the bound on the mutual information $I(\mathsf{x},\mathsf{y})$. Note that the role of the auxiliary vectors z in this context is to capture global dependencies in the reconstructed variables; the auxiliary variables are *not* transmitted across the noisy channel, which for our case is defined by $p(\mathsf{y}|\mathsf{x})$.

It is important to note that the objective (2.38) is a formal generalization of the generic lower bound on the mutual information. Indeed, it is clear that if z is a vector of deterministic variables taking a single state, the bound (2.38) reduces to $H(\mathsf{x}) + \langle \log q(\mathsf{x}|\mathsf{y})\rangle_{p(\mathsf{x},\mathsf{y})}$ (*cf* expression (2.2)). Simpler special cases of the objective may be obtained by imposing further constraints on $p(\mathsf{z}|\mathsf{x},\mathsf{y})$ and $q(\mathsf{z}|\mathsf{y})$. For example, by constraining the variational distribution of the coefficients to be $q(\mathsf{z}|\mathsf{y}) \equiv q(\mathsf{z})$, we may obtain a mixture bound on $I(\mathsf{x},\mathsf{y})$ as a special case. By further considering a constrained auxiliary mapping $p(\mathsf{z}|\mathsf{x},\mathsf{y}) \equiv p(\mathsf{z}|\mathsf{y})$, we may transform (2.38) to the Jensen's relaxation of (2.2) for mixture decoders, etc.

In a more general case, the resulting decoder $q(\mathsf{x}|\mathsf{y}) = \langle q(\mathsf{x}|\mathsf{y},\mathsf{z})\rangle_{q(\mathsf{z}|\mathsf{y})}$ is a generalization of the mixture-of-experts model, in the sense that the variational prior over the mixing coefficient $q(\mathsf{z}|\mathsf{y})$ is itself a mixture. Indeed, from (2.38) it is clear that the local unconstrained optimization for the mixture prior leads to $q(\mathsf{z}|\mathsf{y}) = p(\mathsf{z}|\mathsf{y})$, as expressed from the channel $p(\mathsf{y}|\mathsf{x})p(\mathsf{x})$ and the auxiliary variational conditional $p(\mathsf{z}|\mathsf{x},\mathsf{y})$. As the exact decoder $p(\mathsf{x}|\mathsf{y})$ will typically have a mixture form with the data-dependent coefficients (see expression (1.9)), we may informally hope that a choice of a multi-modal variational decoder will generally lead to tighter bounds on $I(\mathsf{x},\mathsf{y})$.

One may further extend (2.38) by considering models with a richer hierarchy, though an extra care should be taken to ensure tractability. For example, in order for the bound (2.38) to be computationally tractable, one may need to impose additional structural constraints on the conditionals, analogous to the ones discussed in section 2.2.1. For example, if $\boldsymbol{\pi}^x(z_i) \subseteq \{\mathsf{x}\}$, $\boldsymbol{\pi}^y(z_i) \subseteq \{\mathsf{y}\}$ are lower-dimensional x- and y- parents of z_i, one may consider a sparse factorized form of the auxiliary conditional

$$p(\mathsf{z}|\mathsf{y},\mathsf{x}) = \prod_i p(z_i|\boldsymbol{\pi}^x(z_i), \boldsymbol{\pi}^y(z_i)). \tag{2.39}$$

In this case, the entropic term $-H(z_i|\mathsf{x},\mathsf{y})$ may be exactly expressed as

$$\begin{aligned}\left\langle \log p\left(z_i|\boldsymbol{\pi}^x(z_i), \boldsymbol{\pi}^y(z_i)\right)\right\rangle_{p(z_i,\boldsymbol{\pi}^x(z_i),\boldsymbol{\pi}^y(z_i))} = \sum_{\boldsymbol{\eta}^x(z_i)} \prod_{k\in\boldsymbol{\eta}^x(z_i)} p(x_k) \sum_{\boldsymbol{\pi}^y(z_i)} \prod_{j\in\boldsymbol{\pi}^y(z_i)} p(y_j|\boldsymbol{\pi}^x(y_j)) \times \\ \sum_{z_i} p(z_i|\boldsymbol{\pi}^x(z_i), \boldsymbol{\pi}^y(z_i)) \log p(z_i|\boldsymbol{\pi}^x(z_i), \boldsymbol{\pi}^y(z_i)),\end{aligned} \tag{2.40}$$

where $\boldsymbol{\eta}^x(z_i) \stackrel{\text{def}}{=} \boldsymbol{\pi}^x(\boldsymbol{\pi}^y(z_i)) \cup \boldsymbol{\pi}^x(z_i)$ and the definition of the parents $\boldsymbol{\pi}$ was extended for sets of variables in the obvious manner. As before, tractability of this computation may be ensured by choosing an appropriate parameterization (or imposing appropriate sparse structural constraints on discrete models). Analogously, we may choose a tractable structure for the remaining variational parameters $q(\mathsf{z}|\mathsf{y})$ and $q(\mathsf{x}|\mathsf{y},\mathsf{z})$.

2.4 Summary

We described a general theoretically justified approach to information maximization in noisy channels, which extends the family of Blahut-Arimoto type algorithms (Arimoto (1972), Blahut (1972)) by considering a constrained tractable family of decoder distributions. We also showed that the constraints are fundamentally important for avoiding intractability of computing and optimizing the mutual information in the presence of noise. By constraining the approximate posterior, we effectively re-define the optimized objective criterion to be a proper variational lower bound on the mutual information. The suggested iterative optimization procedure, the **IM** algorithm, is guaranteed not to weaken the bound, and has a form reminiscent of the variational EM algorithm for approximate likelihood maximization.

Whilst the generic formulation of the objective criterion is straightforward, it appears to have attracted little previous attention as a practical tool for maximizing mutual information. The suggested formulation is conceptually simple and general; moreover, it implies optimization of a proper *lower bound*, rather than an approximation of the true mutual information. The generality of the approach allows a flexibility in the choice of variational decoders or specific optimization procedures, which suggests that the method may naturally generalize

other techniques for approximate information maximization. Indeed, we showed that optimization of Linsker's *as-if* Gaussian objective criterion (Linsker (1992)) corresponds to a specific way of optimizing the variational lower bound on mutual information for a specific choice of linear Gaussian decoders.

Furthermore, we considered extensions of the simple bound on $I(\mathsf{x}, \mathsf{y})$ to the case of structured decoders, and showed that by retaining local dependencies, we could indeed tighten the theoretically achievable lower bounds on $I(\mathsf{x}, \mathsf{y})$.

Finally, we introduced a new *auxiliary variational* approach to information maximization. The key idea of the suggested auxiliary method is to introduce additional variables, which may be used for capturing useful features of the source patterns, and for introducing global dependencies to the decoded sources. Importantly, the variables and the projections to the auxiliary space are defined in a way which does not impose explicit constraints on the original channel. It is intuitive that these richer representations help to retain some of the structure (and, for the auxiliary formulation, multimodality) of the exact decoder. Moreover, for any constrained choice of the variational distributions $q(\mathsf{x}|\mathsf{y}, \mathsf{z})$, they include simpler generic bounds as special cases. We will subsequently confirm that this more general family of bounds may indeed significantly improve on simpler approaches. However, first we will discuss general relations between optimizing the mutual information in encoder models of noisy channels, optimizing the likelihood in generative latent variable models, and optimizing the conditional likelihood in stochastic autoencoders.

Chapter 3

Likelihood, Conditional Likelihood, and Information Maximization

3.1 Introduction

The problem of finding informative lower-dimensional representations of the observed data may be addressed in a number of different ways. One way to model a relationship between hidden and visible variables (codes and sources) y and x is by considering a generative model, and training it by maximizing the likelihood[1] $\mathcal{L}$. By trying to fit the empirical distribution (subject to the modeling constraints), generative models tend to find latent representations which could be useful for generating the observed patterns. The lower-dimensional latent-space representations may then be obtained by applying Bayes rule. A classic alternative to generative models for finding informative codes is given by the self-extracting autoencoder-type networks, where the sources x are passed through the bottleneck of hidden variable representations y, and extracted at the decoding end to yield the reconstructions $\tilde{\mathsf{x}}$. A standard approach is to train such models by minimizing the reconstruction error, which under certain modeling assumptions may be seen as conditional likelihood learning (Bishop (1995)).

In a sense, generative latent variable models $\mathsf{y} \rightarrow \mathsf{x}$ and autoencoders $\mathsf{x} \rightarrow \mathsf{y} \rightarrow \tilde{\mathsf{x}}$ may be viewed as useful frameworks for finding informative hidden variable representations of the data. However, an arguably more principled measure of informativeness is given by the mutual information $I(\mathsf{x}, \mathsf{y})$, defined as an average reduction in uncertainty of one variable given the other in the stochastic encoder model $\mathsf{x} \rightarrow \mathsf{y}$. The fundamental Fano's result (Fano (1961)) states that independently of the specific parameterization of a (discrete) encoder distribution, the probability of incorrect reconstruction of the sources x from the codes y is bounded by the negative mutual information as

$$p_e(\tilde{\mathsf{x}} \neq \mathsf{x}) \geq \left(H(\mathsf{x}) - I(\mathsf{x}, \mathsf{y}) - 1\right) / \log M \in [0, 1], \tag{3.1}$$

where M is the number of possible distinct reconstructions (size of the input alphabet). In other words, in discrete channels the mutual information $I(\mathsf{x}, \mathsf{y})$

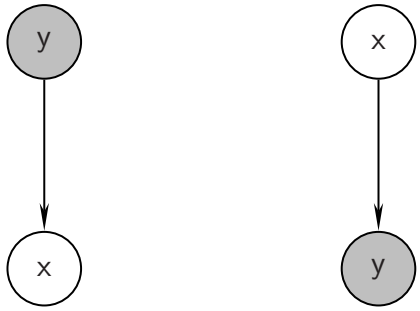
[1] As a monotonic transformation of an objective function does not affect the optimization surface, we will sometimes use the term *likelihood* to refer to the scaled log-likelihood.

is lower-bounded by the probability of the correct reconstruction[2]. Intuitively, this result suggests a relation between information maximization and conditional likelihood training in *stochastic* autoencoders, where the projection into the code space corresponds to the noisy encoding distribution of a channel model. It is interesting to try to understand how all these seemingly different measures relate to each other in completely general settings.

While in some specific cases likelihood and mutual information may indeed have identical optima (see e.g. Cardoso (1997) for a discussion of the noiseless squared ICA case), the optimization surfaces which they define seem to be very different in general. One may understand conceptual differences between the optima provided by likelihood and mutual information maximization for the case of mixture models, where y is a hidden mixture index and x is a visible pattern. It is well known that the likelihood may be trivially maximized by fitting a component to a local segment of the data (as small as a single pattern), provided that the other patterns are explained by the remaining components. One may think of this as an artifact of the definition of the likelihood for under-constrained models (where the theoretically optimal *unconstrained* model is simply the empirical distribution). Intuitively, such solutions should be suboptimal in the information theoretic sense, as they would not be useful for reconstructing most of the source patterns. Analytically, they would result in a decrease in the marginal entropy of the codes $H(y)$, leading to a general reduction in $I(\mathsf{x}, y)$. On the other hand, maximization of mutual information in unconstrained models may result in different forms of trivial optima, which may be characterized by noiseless projections of the training patterns to maximally uniform encodings.

In principle, for any specific parameterization of the generative and encoder (recognition) models, the solutions provided by the likelihood and mutual information maximization may be related by comparing the specific extrema conditions for the parameters. Most of the previous work on relating maximum-likelihood learning in generative models and autoencoders to the information-theoretic learning in encoder models of noisy channels focused primarily on specific invertible mappings (Pearlmutter and Parra (1996), Cardoso (1997), MacKay (1999b)), or constrained linear models (Cottrell et al. (1987), Baldi and Hornik (1989), Oja (1989)). Other more recent work addressed the problem of finding links between a different, but somewhat related family of *Information Bottleneck* approaches and the maximum likelihood methods for a specific tractable mixture model (Slonim and Weiss (2002)). Our goal here is to discuss a relation between maximum-likelihood and information-maximization methods for both exact and approximate cases, independently of the specific parameterizations (see Section 3.2). Then in Section 3.3 we will compare mutual information maximization with the conditional likelihood training of feed-forward models.

[2]Note that originally Fano's inequality was derived under the assumption of discrete source variables x and deterministic reconstructions from the encoded representations, i.e. $H(\tilde{\mathsf{x}}|\mathsf{y}) = 0$, see e.g. Fano (1961), Cover and Thomas (1991).



(a) (b)

Figure 3.1: Generative and encoder models. (a): A generative model $\mathcal{M}_L \stackrel{\text{def}}{=} p(\mathsf{y})p(\mathsf{x}|\mathsf{y})$ trained by maximizing the likelihood $\mathcal{L}$; (b): a model of a noisy channel $\mathcal{M}_I \stackrel{\text{def}}{=} \tilde{p}(\mathsf{x})p(\mathsf{y}|\mathsf{x})$ (trained by maximizing the mutual information $I(\mathsf{x},\mathsf{y})$). The shaded nodes correspond to the hidden variable representations y.

3.2 Information Maximization and Maximum Likelihood

There are fundamental conceptual differences in how we parameterize and train generative and recognition models. In order to define a generative latent variable model $\mathcal{M}_{\mathcal{L}}$, we need to specify the prior on the latent codes $p(\mathsf{y}|\Theta_{\mathsf{y}})$, and the conditional distribution of the observations given the codes $p(\mathsf{x}|\mathsf{y},\Theta_{\mathsf{x}|\mathsf{y}})$ (see Figure 3.1 (a)). Throughout the discussion, we will assume that $p(\Theta_{\mathsf{y}}) \sim \delta(\Theta_{\mathsf{y}} - \Theta^{\star}_{\mathsf{y}})$ and $p(\Theta_{\mathsf{x}|\mathsf{y}}) \sim \delta(\Theta_{\mathsf{x}|\mathsf{y}} - \Theta^{\star}_{\mathsf{x}|\mathsf{y}})$ (for some optimal settings $\Theta^{\star}_{\mathsf{y}}, \Theta^{\star}_{\mathsf{x}|\mathsf{y}}$), and drop the conditioning on the parameters Θ_{y}, $\Theta_{\mathsf{x}|\mathsf{y}}$. The distribution of the data under the model is given by the marginal $p(\mathsf{x}) = \langle p(\mathsf{x}|\mathsf{y})\rangle_{p(\mathsf{y})}$. As discussed earlier, we may train such models by maximizing the log-probability of generating a fixed set of observations $\mathcal{L} \stackrel{\text{def}}{=} \log p(\{\mathsf{x}^{(1)},\ldots,\mathsf{x}^{(M)}\})$, subject to the constraints on $p(\mathsf{y})$ and $p(\mathsf{x}|\mathsf{y})$. Obviously, for independent identically distributed patterns, maximum likelihood training corresponds to minimization of the Kullback-Leibler divergence between the empirical distribution and the model of the observations $p(\mathsf{x})$.

In contrast, to specify a recognition model, we need to define the distribution of the sources $\tilde{p}(\mathsf{x})$ and the generally noisy encoder $\tilde{p}(\mathsf{y}|\mathsf{x})$ (see Figure 3.1 (b)). In a vast number of practical applications of information-theoretic training of such models, it is presumed that $\tilde{p}(\mathsf{x})$ is the empirical distribution (see e.g. Fano (1961), McEliece (1977), Linsker (1989a), Nadal and Parga (1994), Bell and Sejnowski (1995), Linsker (1997), Principe et al. (2000), Torkkola and Campbell (2000), Torkkola (2001), Szummer and Jaakkola (2002)). Here we will focus primarily on the discussion of this case. In principle, we may parameterize the encoder $\tilde{p}(\mathsf{y}|\mathsf{x})$ arbitrarily, as it is a part of the model's specification. However, for the purpose of comparing the maximum-likelihood training in generative models with the mutual information maximization in noisy channels, we will set $\tilde{p}(\mathsf{y}|\mathsf{x})$ to be

the exact posterior of $\mathcal{M}_L$, i.e. $\tilde{p}(\mathsf{y}|\mathsf{x}) \propto p(\mathsf{x}|\mathsf{y})p(\mathsf{y})$. (For simplicity, we will assume that we can compute and average over the posterior $p(\mathsf{y}|\mathsf{x})$, i.e. $p(\mathsf{x},\mathsf{y})$ is in a tractable family – we will discuss a variational extension of this case at a later stage). Effectively, this setting indicates equivalence of the inference under $\mathcal{M}_I$ and $\mathcal{M}_L$ for identical parameter settings; furthermore, both the likelihood[3] $\mathcal{L}$ and the mutual information $I(\mathsf{x},\mathsf{y})$ will be functionals of the same distributions, $p(\mathsf{y})$ and $p(\mathsf{x}|\mathsf{y})$.

To summarize, the distributions defined by the generative and encoder models for the considered case would be given by

$$\mathcal{M}_L \stackrel{\text{def}}{=} p(\mathsf{y})p(\mathsf{x}|\mathsf{y}), \quad \mathcal{M}_I \stackrel{\text{def}}{=} \tilde{p}(\mathsf{x})p(\mathsf{y}|\mathsf{x}), \quad \tilde{p}(\mathsf{x}) \propto \sum_{i=1}^{M} \delta(\mathsf{x} - \mathsf{x}^{(i)}) \tag{3.2}$$

respectively. For these otherwise unconstrained settings, we will be interested in finding a simple relation between maximization of the mutual information $I(\mathsf{x},\mathsf{y})$ in the channel $\mathcal{M}_I$ and maximization of the log-likelihood $\mathcal{L}$ in the corresponding latent variable model $\mathcal{M}_L$. In the rest of this section, by referring to a specific objective function (I or $\mathcal{L}$) we will also implicitly refer to the corresponding model.

There are obvious practical implications of relating these two modeling methods. Indeed, let us assume that we are given parametric constraints on $p(\mathsf{y})$ and $p(\mathsf{x}|\mathsf{y})$, and a set of training patterns $\{\mathsf{x}\}$. If our goal is to find the most informative latent variable representations $\{\mathsf{y}\}$ of the training set $\{\mathsf{x}\}$, an obvious approach would be to maximize the mutual information in $\mathcal{M}_I$. However, as discussed in Section 1.3, the exact optimization would typically be intractable, and approximations or tractable lower bounds on $I(\mathsf{x},\mathsf{y})$ would need to be considered. Such tractable bounds may prove to be weak, which may nullify the initial advantage of trying to optimize a proper measure of information in the first place. In contrast, it is usually significantly easier to maximize the likelihood in the corresponding generative model, and classic probabilistic methods of maximum-likelihood training may eventually prove to be better (in terms of the retained information content) than optimizing any tractable relaxation of $I(\mathsf{x},\mathsf{y})$. We will now show that this is generally *not* the case (at least, not in terms of relaxations of the exact mutual information), and we may indeed consider optimizing a special tractable form of the variational lower bound on $I(\mathsf{x},\mathsf{y})$ in order to maximize the amount of information which $\{\mathsf{y}\}$ contain about $\{\mathsf{x}\}$.

3.2.1 Variational Information Maximization and Exact Likelihood Training

Let us assume that the data is i.i.d., and the models are defined as in expression (3.2). Then it is easy to show that the exact likelihood, expressed from the

[3]We extend Fisher's view of the likelihood (Fisher (1925)) and treat it as a functional of the model $p(\mathsf{x})$ under the specified modeling constraints. In other words, for us it would make perfect sense to integrate the likelihood over distributions of random variables. To indicate the dependence on the functional parameters, we will sometimes denote the likelihood as $\mathcal{L}(p(\mathsf{x}))$.

generative model $\mathcal{M}_L$, is a lower bound on the exact mutual information in the corresponding memoryless communication channel $\mathcal{M}_I$ (up to irrelevant constants which have no effects on the optimization). It is also straight-forward to see that for any tractable choice of $\mathcal{M}_L$, it is possible to define a tractable lower bound on $I(\mathsf{x},\mathsf{y})$, which is *at least* as tight as the likelihood. Intuitively, this means that by maximizing the likelihood in generative models, we indeed maximize a proper lower bound on the mutual information in the corresponding channel, but we do it suboptimally in terms of the retained information content. This also suggests a tractable information-theoretic alternative to maximum-likelihood training.

Proposition 3.1. *For i.i.d. patterns* $\{\mathsf{x}\}$*, maximum likelihood training in the generative model* $\mathcal{M}_L$ *corresponds to maximization of a* **lower bound** *on the mutual information in* $\mathcal{M}_I$*. Up to irrelevant constants, this bound is weaker or as tight as* $\hat{I}(\mathsf{x},\mathsf{y}) = \langle \log p(\mathsf{x}|\mathsf{y}) \rangle_{p(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x})}$*.*

Proof. The results follow immediately from applying straight-forward transformations of the likelihood. Let $\{\mathsf{x}\} \stackrel{\text{def}}{=} \{\mathsf{x}^{(i)} | i = 1, \ldots, M\}$ denote a set of training patterns. For i.i.d. data, the average log-likelihood is given by

$$\begin{aligned} \mathcal{L} \stackrel{\text{def}}{=} \log p(\{\mathsf{x}\})/M &= \langle \log p(\mathsf{x}) \rangle_{\tilde{p}(\mathsf{x})} \\ &= \langle \log p(\mathsf{y}) + \log p(\mathsf{x}|\mathsf{y}) - \log p(\mathsf{y}|\mathsf{x}) \rangle_{\tilde{p}(\mathsf{x})} \,, \end{aligned} \tag{3.3}$$

where $\tilde{p}(\mathsf{x})$ is the empirical distribution, and y is a latent variable in the joint distribution $p(\mathsf{x},\mathsf{y})$. By construction, $p(\mathsf{y})$ and $p(\mathsf{x}|\mathsf{y})$ are parts of the model's specification which need to be learned, and the exact posterior $p(\mathsf{y}|\mathsf{x})$ is simply given by Bayes rule. By averaging both parts of (3.3) over $p(\mathsf{y}|\mathsf{x})$, we obtain

$$\mathcal{L} = \langle \log p(\mathsf{x}|\mathsf{y}) \rangle_{p(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x})} - \langle KL(p(\mathsf{y}|\mathsf{x}) \| p(\mathsf{y})) \rangle_{\tilde{p}(\mathsf{x})} \,. \tag{3.4}$$

Note that the non-negative Kullback-Leibler divergence term in (3.4) would have reduced to the mutual information (expressed from $\mathcal{M}_L$), had the averages over the empirical distribution $\tilde{p}(\mathsf{x})$ been replaced by the averages over the true model-based marginal $p(\mathsf{x})$. (Indeed, by averaging (3.4) over $p(\mathsf{x})$, we would recover the well-known mean of average log-likelihood for i.i.d. data $\langle \mathcal{L}(\mathsf{x}) \rangle_{p(\mathsf{x})} = -H(\mathsf{x})$).

We will now express $I(\mathsf{x},\mathsf{y})$ in the corresponding model $\mathcal{M}_I$ of a noisy memoryless channel. It is easy to see that for the considered specification (3.2), the exact mutual information in $\mathcal{M}_I$ is given as

$$I(\mathsf{x},\mathsf{y}) = H_{\tilde{p}}(\mathsf{x}) - H_{\tilde{p}}(\mathsf{x}|\mathsf{y}) = H_{\tilde{p}}(\mathsf{x}) + \langle \log \tilde{p}(\mathsf{x}|\mathsf{y}) \rangle_{p(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x})}, \tag{3.5}$$

where $H_{\tilde{p}}(\mathsf{x}) \stackrel{\text{def}}{=} -\langle \log \tilde{p}(\mathsf{x}) \rangle_{\tilde{p}(\mathsf{x})}$, $H_{\tilde{p}}(\mathsf{x}|\mathsf{y}) \stackrel{\text{def}}{=} -\langle \log \tilde{p}(\mathsf{x}|\mathsf{y}) \rangle_{\tilde{p}(\mathsf{x},\mathsf{y})}$, and $\tilde{p}(\mathsf{x}|\mathsf{y})$ is the Bayes-optimal decoder for the training patterns

$$\tilde{p}(\mathsf{x}|\mathsf{y}) \propto \tilde{p}(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x}) = p(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x}), \tag{3.6}$$

(as $\tilde{p}(\mathsf{y}|\mathsf{x}) \equiv p(\mathsf{y}|\mathsf{x})$ by construction). By definition, $\tilde{p}(\mathsf{x})$ has a mixture form, which leads to a mixture form for the decoder $\tilde{p}(\mathsf{x}|\mathsf{y})$. Generally, this complicates evaluations of the conditional entropy $H_{\tilde{p}}(\mathsf{x}|\mathsf{y})$ in (3.5), which may be intractable

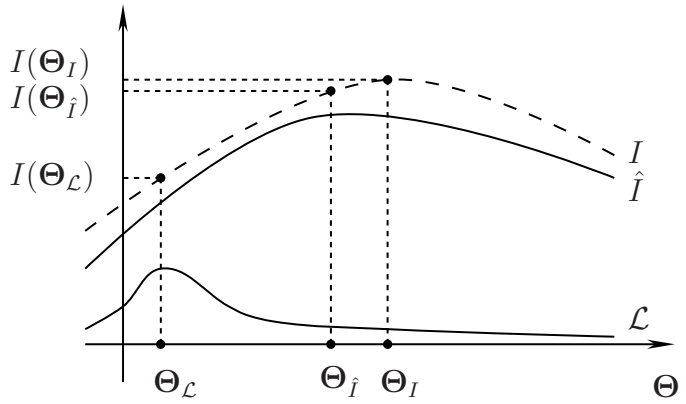


Figure 3.2: Mutual information $I(\mathsf{x},\mathsf{y})$, the variational lower bound $\hat{I}(\mathsf{x},\mathsf{y})$, and the exact log-likelihood score $\mathcal{L}(\mathcal{M}_L)$ in the parameter space $\boldsymbol{\Theta} = \{p(\mathsf{y}), p(\mathsf{x}|\mathsf{y})\}$ (a schematic plot). Intuitively, optimization of the tighter bound $\hat{I}(\mathsf{x},\mathsf{y})$ on the generally intractable mutual information $I(\mathsf{x},\mathsf{y})$ leads to a better estimate $\boldsymbol{\Theta}_{\hat{I}}$ of the I-optimal parameters $\boldsymbol{\Theta}_I$. Heuristically, the tightness of a bound may be treated as an approximate criterion of optimality.

even for the case of a tractable channel distribution $p(\mathsf{y}|\mathsf{x})$. By applying the variational lower bound on the mutual information (2.2), we may bound (3.5) as

$$\begin{aligned} I(\mathsf{x},\mathsf{y}) &\geq H_{\tilde{p}}(\mathsf{x}) + \langle \log p(\mathsf{x}|\mathsf{y}) \rangle_{p(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x})} && (3.7)\\ &\geq H_{\tilde{p}}(\mathsf{x}) + \langle \log p(\mathsf{x}|\mathsf{y}) \rangle_{p(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x})} - \langle KL(p(\mathsf{y}|\mathsf{x}) \| p(\mathsf{y})) \rangle_{\tilde{p}(\mathsf{x})} && (3.8)\\ &= H_{\tilde{p}}(\mathsf{x}) + \mathcal{L} \Rightarrow && (3.9)\\ I(\mathsf{x},\mathsf{y}) &\geq \hat{I}(\mathsf{x},\mathsf{y}) \geq H_{\tilde{p}}(\mathsf{x}) + \mathcal{L}, && (3.10) \end{aligned}$$

where the tighter lower bound on the mutual information is given by

$$\hat{I}(\mathsf{x},\mathsf{y}) \stackrel{\text{def}}{=} H_{\tilde{p}}(\mathsf{x}) + \langle \log p(\mathsf{x}|\mathsf{y}) \rangle_{p(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x})}. \tag{3.11}$$

Note that (3.11) is just the generic lower bound on the mutual information (2.2) in $\mathcal{M}_I$, with the decoder given by the generative conditional $p(\mathsf{x}|\mathsf{y})$. Additionally, since $\tilde{p}(\mathsf{x})$ is the empirical distribution, the entropic term $H_{\tilde{p}}(\mathsf{x})$ in (3.10) and (3.11) has no effect on the optimization surface for the encoder $p(\mathsf{y}|\mathsf{x})$. Thus, up to irrelevant constants, the exact log-likelihood score $\mathcal{L}$ defines a proper lower bound on $I(\mathsf{x},\mathsf{y})$ in the recognition model $\mathcal{M}_I$. This bound is weaker than the variational relaxation given by $\hat{I}(\mathsf{x},\mathsf{y})$. □

Proposition 3.1 shows that by maximizing the likelihood in $\mathcal{M}_L$ we indeed optimize a proper lower bound on the generally intractable mutual information between the source patterns $\{\mathsf{x}^{(i)}|i = 1,\ldots,M\}$ and their encodings $\{\mathsf{y}^{(i)}|i = 1,\ldots,M\}$ generated by the exact posterior $p(\mathsf{y}|\mathsf{x})$. It also suggests that the likelihood bound on $I(\mathsf{x},\mathsf{y})$ may in general be weak, which follows from the non-negativity of the Kullback-Leibler term in (3.8). Intuitively, the weakness of the bound may cause significant differences between solutions obtained by maximizing the likelihood and the mutual information (see the schematic plot of Figure

3.2). More importantly, proposition 3.1 suggests that there exists a proper variational lower bound on the mutual information, which is *(a):* tractable (whenever $p(\mathsf{x},\mathsf{y})$ is in a tractable family); *(b):* tighter than the likelihood. This bound may be used for information-theoretic training of *both* the encoder and the generative models specified by expression (3.2). Specifically, for training the generative model $\mathcal{M}_L$, the idea would be to optimize (3.11) alone, which defines a tighter bound on the exact mutual information. We may consider optimizing the same bound[4] for the encoder model $\mathcal{M}_I$. In fact, it is irrelevant whether we choose to optimize (3.11) in $\mathcal{M}_I$ or in $\mathcal{M}_L$, as in both cases we would be looking at the same objective functional, optimized for the same set of functional parameters.

Note that the non-constant term in the variational lower bound $\hat{I}(\mathsf{x},\mathsf{y})$ may be thought of as an empirical estimate of the conditional entropy $H(\mathsf{x}|\mathsf{y})$, as expressed from $\mathcal{M}_L$. Moreover, from (3.10) and (3.11) we get

$$\begin{aligned}\hat{I}(\mathsf{x},\mathsf{y}) &= \mathcal{L} + \langle KL(p(\mathsf{y}|\mathsf{x})\|p(\mathsf{y}))\rangle_{\tilde{p}(\mathsf{x})} + H_{\tilde{p}}(\mathsf{x}), &(3.12)\\ &= \langle KL(p(\mathsf{y}|\mathsf{x})\|p(\mathsf{y}))\rangle_{\tilde{p}(\mathsf{x})} - KL(\tilde{p}(\mathsf{x})\|p(\mathsf{x})) &(3.13)\end{aligned}$$

where the Kullback-Leibler term explicitly favours large *average* deviations of the posteriors $p(\mathsf{y}|\mathsf{x})$ from the priors $p(\mathsf{y})$ for a fixed set of training patterns $\{\mathsf{x}\}$. (Clearly, the term is *minimized* when the encodings y are independent of the sources x). Note that $\hat{I}$- and $\mathcal{L}$-learning are equivalent when the Kullback-Leibler term in (3.12) is independent of the optimized parameters.

Informally, we can see that optimization of $\hat{I}(\mathsf{x},\mathsf{y})$ corresponds to maximization of the exact log-likelihood, with a certain bias towards deterministic and spread-out encoded representations. Then by expanding (3.12), we get

$$\hat{I}(\mathsf{x},\mathsf{y}) = \mathcal{L} - \langle \log p(\mathsf{y})\rangle_{\tilde{p}(\mathsf{y})} - \sum_{i=1}^{M} H(p(\mathsf{y}|\mathsf{x}^{(i)}))/M + H_{\tilde{p}}(\mathsf{x}), \qquad (3.14)$$

where $\tilde{p}(\mathsf{y}) \stackrel{\text{def}}{=} \langle p(\mathsf{y}|\mathsf{x})\rangle_{\tilde{p}(\mathsf{x})}$ is just the empirical average of the posterior (corresponding to the marginal distribution of the codes in the recognition model $\mathcal{M}_I$). The last term in (3.14) is an empirical average of the entropies of $p(\mathsf{y}|\mathsf{x}^{(i)})$, which favour deterministic encodings (i.e. sharp posteriors in generative models). To show that the objective $\hat{I}(\mathsf{x},\mathsf{y})$ indeed favours spread-out representations in $\{\mathsf{y}\}$, we may once again make use of the non-negativity of the KL-divergence

$$\langle \log \tilde{p}(\mathsf{y})\rangle_{\tilde{p}(\mathsf{y})} - \langle \log p(\mathsf{y})\rangle_{\tilde{p}(\mathsf{y})} \geq 0 \Rightarrow -\langle \log p(\mathsf{y})\rangle_{\tilde{p}(\mathsf{y})} \geq H(\tilde{p}(\mathsf{y})). \qquad (3.15)$$

Substituting (3.15) into the objective (3.14), we obtain a relaxation of $\hat{I}(\mathsf{x},\mathsf{y})$, given by

$$\hat{I}(\mathsf{x},\mathsf{y}) \geq \mathcal{L} + H(\tilde{p}(\mathsf{y})) - \sum_{i=1}^{M} H(p(\mathsf{y}|\mathsf{x}^{(i)}))/M + H(\tilde{p}(\mathsf{x})). \qquad (3.16)$$

[4]Despite the fact that any tractable choice of the variational decoder $q(\mathsf{x}|\mathsf{y})$ could potentially be used for training $\mathcal{M}_I = \tilde{p}(\mathsf{x})p(\mathsf{y}|\mathsf{x})$ (see expression (2.2)), not every such decoder may lead to an improvement over the bound provided by the likelihood score. From proposition 3.1 we see that one choice of the decoder when the improvement does happen is to set $q(\mathsf{x}|\mathsf{y}) = p(\mathsf{x}|\mathsf{y})$.

Clearly, the entropy of $\tilde{p}(\mathsf{y}) = \sum_{i=1}^{M} p(\mathsf{y}|\mathsf{x}^{(i)})/M$ in (3.16) would be maximized for uniform representations of the training patterns in the code space $\{\mathsf{y}\}$. Since a proper lower bound on the objective criterion $\hat{I}(\mathsf{x},\mathsf{y})$ biases the solutions towards spread-out codes, one may informally hope that the objective $\hat{I}(\mathsf{x},\mathsf{y})$ will establish a similar behaviour. As we mentioned earlier, this bias towards a uniform distribution of the latent variables may be useful (for example, for avoiding learning non-informative representations due to singularities and degeneracies of maximum-likelihood solutions).

Example: It is easy to see that there are interesting cases when optimization of the exact mutual information $I(\mathsf{x},\mathsf{y})$ in the encoder model $\mathcal{M}_I$ is intractable *despite* the tractability of both $p(\mathsf{y}|\mathsf{x})$ and $p(\mathsf{x}|\mathsf{y})$. For a quick illustration of this, let $\mathcal{M}_L$ define the factor analysis model (e.g. Bartholomew (1987), Ghahramani and Hinton (1996)), given as $\mathsf{x} = \mathsf{W}\mathsf{y} + \mathsf{e}$; $\mathsf{y} \sim \mathcal{N}(0, \mathbf{1})$, $p(\mathsf{e}) \sim \mathcal{N}(0, \mathbf{\Psi})$. Clearly, for this model we get $p(\mathsf{x}|\mathsf{y}) \sim \mathcal{N}(\mathsf{W}\mathsf{y}, \mathbf{\Psi})$, $p(\mathsf{x}) \sim \mathcal{N}(0, \mathsf{W}\mathsf{W}^T + \mathbf{\Psi})$, and $p(\mathsf{y}|\mathsf{x}) \sim \mathcal{N}((\mathsf{I} + \mathsf{W}^T\mathbf{\Psi}^{-1}\mathsf{W})^{-1}\mathsf{W}^T\mathbf{\Psi}^{-1}\mathsf{x}, (\mathsf{I} + \mathsf{W}^T\mathbf{\Psi}^{-1}\mathsf{W})^{-1})$. Note that for the corresponding encoder model $\mathcal{M}_I$, it is intractable to optimize the exact mutual information (3.5), as this would require evaluation of the entropy of a mixture of Gaussians. On the other hand, it is computationally tractable to optimize the variational lower bound $\hat{I}(\mathsf{x},\mathsf{y})$ on the mutual information, as this would only require computing Gaussian and empirical averages of the quadratic terms $\langle \mathsf{y}\mathsf{x}^T \rangle_{p(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x})}$, $\langle \mathsf{y}\mathsf{y}^T \rangle_{p(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x})}$. By substituting $p(\mathsf{x}|\mathsf{y})$ and $p(\mathsf{y}|\mathsf{x})$ into the expressions for $\mathcal{L}$ and $\hat{I}(\mathsf{x},\mathsf{y})$, we obtain generally different objective criteria. In the special case of $\mathbf{\Psi} = \sigma^2\mathsf{I}$, the relation between the $\hat{I}$- and $\mathcal{L}$-optimal weights may be easily established analytically; in fact, for this special case both objectives would give rise to PCA solutions for W.

Another example, illustrating differences between likelihood and information-theoretic training, is discussed in Section 5.2.1. There we compare $\mathcal{L}$- and $\hat{I}$-maximization for training Gaussian mixture models.

3.2.2 Variational Information Maximization and Variational Likelihood Training

We have shown that optimization of the exact likelihood in the generative model $\mathcal{M}_L = p(\mathsf{y})p(\mathsf{x}|\mathsf{y})$ may be motivated in the variational information-theoretic sense (though the likelihood bounds on $I(\mathsf{x},\mathsf{y})$ may potentially be weak). We have also discussed a simple re-definition of the objective criterion which would lead to a tighter variational lower bound on the mutual information in the corresponding encoder model $\mathcal{M}_I$. These theoretical results are intuitive, but limited to the cases when it is possible to compute and average over the posterior $p(\mathsf{y}|\mathsf{x})$. However, for more general latent variable models, it may be intractable to maximize the likelihood $\mathcal{L}$ or the bound $\hat{I}(\mathsf{x},\mathsf{y})$ directly.

The usual way to handle the intractability of exact maximum likelihood learning in the generative model $\mathcal{M}_L$ is to consider the standard variational lower

bound on the likelihood

$$\begin{aligned}\mathcal{L} &= \left\langle \log \int_{\mathsf{y}} p(\mathsf{x},\mathsf{y})\mathrm{d}\mathsf{y} \right\rangle_{\tilde{p}(\mathsf{x})} \\ &= \left\langle \log \int_{\mathsf{y}} p(\mathsf{x},\mathsf{y}) \frac{q(\mathsf{y}|\mathsf{x})}{q(\mathsf{y}|\mathsf{x})}\mathrm{d}\mathsf{y} \right\rangle_{\tilde{p}(\mathsf{x})} \geq \tilde{\mathcal{L}},\end{aligned}$$

where we have defined

$$\begin{aligned}\tilde{\mathcal{L}} &= \langle \log p(\mathsf{x},\mathsf{y}) \rangle_{q(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x})} - \langle \log q(\mathsf{y}|\mathsf{x}) \rangle_{q(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x})} \\ &= \langle \log p(\mathsf{x}|\mathsf{y}) \rangle_{q(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x})} - \langle KL(q(\mathsf{y}|\mathsf{x}) \| p(\mathsf{y})) \rangle_{\tilde{p}(\mathsf{x})}. \end{aligned} \tag{3.17}$$

Here $\tilde{p}(\mathsf{x})$ is the empirical distribution, and $q(\mathsf{y}|\mathsf{x})$ is an arbitrary variational distribution approximating the true posterior $p(\mathsf{y}|\mathsf{x})$ (see Zemel and Hinton (1994), Zemel and Hinton (1995), Dayan et al. (1995) for the related *Helmholtz machine* formulation, and e.g. Saul et al. (1996), Jaakkola (1997) for the general statement of the variational problem). The standard variational extensions of the expectation-maximizing algorithm (see e.g. Neal and Hinton (1998)) train the generative model $\mathcal{M}_L$ by iteratively optimizing (3.17) with respect to the model parameters $p(\mathsf{y})$, $p(\mathsf{x}|\mathsf{y})$, and the variational posterior $q(\mathsf{y}|\mathsf{x})$. In order to simplify computations of the bound (3.17), the variational distribution $q(\mathsf{y}|\mathsf{x})$ is constrained to ensure tractability of computing the average energy $\langle \log p(\mathsf{x},\mathsf{y}) \rangle_{q(\mathsf{y}|\mathsf{x})}$, which for directed latent variable models implies tractability of computing $\langle \log p(\mathsf{x}|\mathsf{y}) \rangle_{q(\mathsf{y}|\mathsf{x})}$ and $\langle \log p(\mathsf{y}) \rangle_{q(\mathsf{y}|\mathsf{x})}$.

In order to compare variational approaches to maximization of the likelihood and mutual information, we will define a recognition model

$$\tilde{\mathcal{M}}_I \stackrel{\text{def}}{=} q(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x}), \tag{3.18}$$

which ensures equivalence of the approximate *variational* inference[5] in $\mathcal{M}_L$ and $\tilde{\mathcal{M}}_I$. Then by analogy with proposition 3.1, it is easy to see that the lower bound $\tilde{\mathcal{L}}$ on the likelihood in $\mathcal{M}_L$ is in fact a proper lower bound on the mutual information in $\tilde{\mathcal{M}}_I$. Moreover, we can easily find a tractable lower bound on $I(\mathsf{x},\mathsf{y})$ which is tighter than $\tilde{\mathcal{L}}$.

Proposition 3.2. *For i.i.d. patterns* $\{\mathsf{x}\}$*, maximization of the standard variational lower bound on the likelihood in the generative model* $\mathcal{M}_L$ *corresponds to maximization of a lower bound on the mutual information in* $\tilde{\mathcal{M}}_I$*. Up to irrelevant constants, this bound is weaker or as tight as* $\hat{I}_q(\mathsf{x},\mathsf{y}) = \langle \log p(\mathsf{x}|\mathsf{y}) \rangle_{q(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x})}$*, where* $q(\mathsf{y}|\mathsf{x})$ *is the approximate posterior of the generative model.*

Proof. By definition, the exact value of mutual information $I(\mathsf{x},\mathsf{y})$ for model $\tilde{\mathcal{M}}_I$ is given by

$$I(\mathsf{x},\mathsf{y}) = H_{\tilde{p}}(\mathsf{x}) + \langle \log \tilde{p}(\mathsf{x}|\mathsf{y}) \rangle_{q(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x})}, \tag{3.19}$$

[5]Since computing the exact posterior $p(\mathsf{y}|\mathsf{x})$ in $\mathcal{M}_L$ will now be intractable, we will need approximations to perform the inference. The choice of the variational posterior $q(\mathsf{y}|\mathsf{x})$ as a model for the encoder ensures tractability and helps to establish a relationship between variational likelihood and mutual information maximization under the assumption of equivalence of approximate inference.

where $\tilde{p}(\mathsf{x}|\mathsf{y}) \propto \tilde{p}(\mathsf{x})q(\mathsf{y}|\mathsf{x})$ is the exact posterior expressed from the encoder model. Then, by analogy with the proof of proposition 3.1, we get

$$I(\mathsf{x},\mathsf{y}) \geq H_{\tilde{p}}(\mathsf{x}) + \langle \log p(\mathsf{x}|\mathsf{y}) \rangle_{q(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x})} - \langle KL(q(\mathsf{y}|\mathsf{x}) \| p(\mathsf{y})) \rangle_{\tilde{p}(\mathsf{x})} \Rightarrow \quad (3.20)$$

$$I(\mathsf{x},\mathsf{y}) \geq \hat{I}_q(\mathsf{x},\mathsf{y}) \geq H_{\tilde{p}}(\mathsf{x}) + \tilde{\mathcal{L}}, \quad (3.21)$$

where $\tilde{\mathcal{L}}$ is defined as in (3.17), and the tighter lower bound on $I(\mathsf{x},\mathsf{y})$ is defined as

$$\hat{I}_q(\mathsf{x},\mathsf{y}) \stackrel{\text{def}}{=} H_{\tilde{p}}(\mathsf{x}) + \langle \log p(\mathsf{x}|\mathsf{y}) \rangle_{q(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x})}, \quad (3.22)$$

which is tractable by construction. □

Clearly, the relation (3.22) between variational approaches to mutual information and likelihood maximization in $\tilde{\mathcal{M}}_I$ and $\mathcal{M}_L$ is analogous to the similar relation (3.10) for the exact likelihood computations. Effectively, proposition 3.2 states that standard variational approaches to likelihood maximization may indeed be seen as a way to optimize a specific variational lower bound on $I(\mathsf{x},\mathsf{y})$ in the corresponding encoder model $\tilde{\mathcal{M}}_I$. Again, a tighter (yet tractable) bound would be given by $\hat{I}_q(\mathsf{x},\mathsf{y})$, which is exactly the generic variational lower bound on the mutual information in $\tilde{\mathcal{M}}_I$ for the specific parameterization of the decoder distribution. Moreover, from (3.20) it is clear that maximization of the lower bound $\hat{I}_q(\mathsf{x},\mathsf{y})$ on the mutual information and the variational Jensen's bound $\tilde{\mathcal{L}}$ on the likelihood lead to the same fixed points when the Kullback-Leibler divergence in (3.20) is a constant (also see discussion in Appendix B.2).

It is interesting to note that by optimizing the bound on the likelihood $\tilde{\mathcal{L}}$ for the latent space prior $p(\mathsf{y})$ and substituting the optimal functional parameters back into (3.20), we may transform (3.20) to

$$I(\mathsf{x},\mathsf{y}) \geq \underbrace{H_{\tilde{p}}(\mathsf{x}) + \langle \log p(\mathsf{x}|\mathsf{y}) \rangle_{q(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x})}}_{\hat{I}_q(\mathsf{x},\mathsf{y})} - \left\langle KL(q(\mathsf{y}|\mathsf{x}) \| \langle q(\mathsf{y}|\mathsf{x}) \rangle_{\tilde{p}(\mathsf{x})}) \right\rangle_{\tilde{p}(\mathsf{x})} \quad (3.23)$$

$$\geq H_{\tilde{p}}(\mathsf{x}) + \tilde{\mathcal{L}}. \quad (3.24)$$

(Here the second inequality (3.24) reduces to an identity if and only if the current settings of $p(\mathsf{y})$ in the generative model $\mathcal{M}_L$ are indeed $\tilde{\mathcal{L}}$-optimal). Note that the Kullback-Leibler term in (3.23) is generally nonzero. This results in a relaxation of the variational bound on $I(\mathsf{x},\mathsf{y})$, i.e. apart from simple special cases[6], the bound $H_{\tilde{p}}(\mathsf{x}) + \tilde{\mathcal{L}}$ will be strictly weaker than $\hat{I}_q(\mathsf{x},\mathsf{y})$.

Informally, from (3.20) we can see that optimization of the bound on the likelihood in $\mathcal{M}_L$ could reduce to optimizing the generic lower bound $\hat{I}_q(\mathsf{x},\mathsf{y})$ on the mutual information in $\tilde{\mathcal{M}}_I$, if the marginal distribution $p(\mathsf{y}^{(i)})$ of each latent vectors $\mathsf{y}^{(i)}$ was an independent free parameter. Indeed, the variational bound on

[6]It is easy to see that $\hat{I}_q(\mathsf{x},\mathsf{y}) = H_{\tilde{p}}(\mathsf{x}) + \tilde{\mathcal{L}}$ may be achieved when $q(\mathsf{y}|\mathsf{x}) = q(\mathsf{y})$, i.e. both the likelihood and the information-theoretic bound use an extremely weak definition of the encoding distribution. Similarly, both bounds are identical if $M = 1$, i.e. there is a singe pattern to encode. A brief discussion of other special cases when both bounds lead to identical optima is given in Appendix B.2.

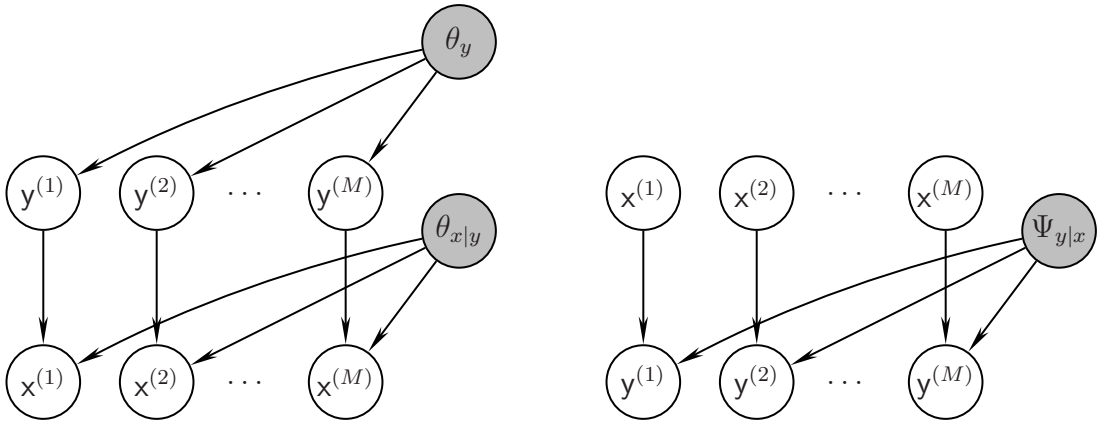


Figure 3.3: Generative and encoder models for i.i.d. patterns. *Left:* generative model $\mathcal{M}_L$ for M i.i.d. training patterns $\{\mathsf{x}^{(1)}, \ldots, \mathsf{x}^{(M)}\}$ with the corresponding latent variable representations $\{\mathsf{y}^{(1)}, \ldots, \mathsf{y}^{(M)}\}$; *Right:* the corresponding model of the memoryless channel $\mathcal{M}_I$. The shaded nodes indicate the implied conditionings on the intrinsically deterministic model parameters.

the likelihood (3.17) could in this case be expressed as

$$\mathcal{L} \propto \sum_{i=1}^{M} \langle \log p(\mathsf{x}^{(i)}|\mathsf{y}) \rangle_{q(\mathsf{y}|\mathsf{x}^{(i)})} - \sum_{i=1}^{M} KL(q(\mathsf{y}|\mathsf{x}^{(i)}) \| p(\mathsf{y}^{(i)})). \tag{3.25}$$

If for all training patterns $i = 1, \ldots, M$ we were free to change each latent-space marginal individually, we could optimally set it to the variational posterior as

$$p(\mathsf{y}^{(i)} = \mathsf{y}) \stackrel{\text{def}}{=} p(\mathsf{y}^{(i)} = \mathsf{y}|\boldsymbol{\Theta}_y^{(i)}) = q(\mathsf{y}|\mathsf{x}^{(i)}, \Psi_{y|x}) \stackrel{\text{def}}{=} q(\mathsf{y}|\mathsf{x}^{(i)}), \tag{3.26}$$

which would lead to $\tilde{\mathcal{L}} = \hat{I}_q(\mathsf{x}, \mathsf{y})$ at the optimum of (3.25). It is clear that (3.26) is exactly the definition of the latent-space marginals $\tilde{p}(\mathsf{y}^{(i)})$ in the encoder model $\tilde{\mathcal{M}}_I$ (see Figure 3.3 (right)). It is also clear that (3.26) does not generally apply for the generative model $\mathcal{M}_L$ and i.i.d observations, where

$$\forall \boldsymbol{\Theta}_y.\ \forall i, j \in \{1, \ldots, M\}.\ p(\mathsf{y}^{(i)} = \mathsf{y}|\boldsymbol{\Theta}_y) = p(\mathsf{y}^{(j)} = \mathsf{y}|\boldsymbol{\Theta}_y) \tag{3.27}$$

(see Figure 3.3 (left)).

Generally, we can say that learning by optimizing the exact likelihood (or standard variational lower bounds on the likelihood) in generative models could be interpreted as optimization of a specific lower bound on the generally intractable mutual information in the corresponding encoder models (3.2), (3.18). However, the optimization surfaces and the obtained solutions will generally be different from those given by the variational approach to information-maximization in the corresponding models of memoryless channels. This difference in the objectives is quantified as a summation of generally non-constant KL divergences between the posteriors $p(\mathsf{y}|\mathsf{x}^{(i)})$ and their empirical average $\langle p(\mathsf{y}|\mathsf{x}) \rangle_{\tilde{p}(\mathsf{x})}$, as expressed from $\mathcal{M}_L$ for i.i.d. data. Arguably, one of the principal causes of the difference is the

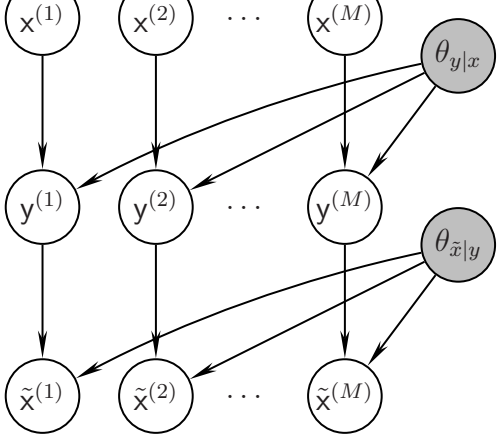


Figure 3.4: An autoencoder model for M training patterns $\{\mathsf{x}^{(1)}, \ldots, \mathsf{x}^{(M)}\}$ and their reconstructions $\{\tilde{\mathsf{x}}^{(1)}, \ldots, \tilde{\mathsf{x}}^{(M)}\}$. The shaded nodes indicate the conditionings on the deterministic model parameters.

identical distribution of the latent vectors in $\mathcal{M}_L$, which is implied by the specific form of the prior on the latent vectors. We may further notice a closer correspondence between optimizing the mutual information in encoder models (Figure 3.3 (right)) and the conditional likelihood training in autoencoders (see Figure 3.4), which are characterized by generally different marginal distributions of the encodings, and the model specification more similar to the encoder paradigm. Presuming that our intuition here is correct, we may perceive the existence of optimization procedures where the conditional training of autoencoders reduces to the variational information maximization.

3.3 Information Maximization and Maximum Conditional Likelihood

A classic method of obtaining informative lower-dimensional representations of higher-dimensional source vectors is to consider an autoencoder with the parameterized mappings from the source vectors to the codes ($\mathsf{x} \to \mathsf{y}$), and from the codes to the reconstructions of the source vectors ($\mathsf{y} \to \tilde{\mathsf{x}}$) (Hinton (1989), Baldi and Hornik (1989), Bishop (1995)). The goal is to learn parameters of encoder and decoder mappings which would lead to accurate reconstructions of the sources from their encoded representations. Most of the conventional approaches to training such models (e.g. minimization of the mean squared reconstruction error, minimization of the cross-entropy for multinomial vectors, etc.) are equivalent to maximizing the conditional likelihood $p(\{\tilde{\mathsf{x}}\}|\{\mathsf{x}\})$ in generalized chains $\mathsf{x} \to \mathsf{y} \to \tilde{\mathsf{x}}$ under specific parametric constraints on the encoder and decoder distributions. Several other methods for optimizing parameters of such networks may be shown to correspond to variational likelihood learning in mixture models (Zemel (1993), Zemel and Hinton (1994)), while in some other cases the optimized objectives do not seem to have a clear probabilistic interpretation (Doi and Lewicki, 2004).

Here we will focus primarily on the discussion of conventional conditional training of feed-forward models, and the relations which the resulting objective functions may have to variational bounds on the mutual information.

While it is commonly presumed that the encoding part of the autoencoder is noiseless (i.e. $p(\mathsf{y}|\mathsf{x}) \sim \delta$), probabilistic extensions may easily be considered. By analogy with the marginal likelihood training of arbitrarily structured graphical models, a proper probabilistic method of training stochastic autoencoders would involve integration of the hidden encodings and maximization of the likelihood with respect to deterministic parameters of the model subject to specific constraints. Typically, the goal of learning would be to find the optimal encoder $p(\mathsf{y}|\mathsf{x})$ and decoder $p(\tilde{\mathsf{x}}|\mathsf{y})$ distributions (which effectively reduces to maximizing the conditional likelihood). Clearly, this formulation is somewhat similar to variational approaches to mutual information maximization, where the objective function is optimized with respect to the encoder and variational decoder. Moreover, as we have mentioned earlier, we may perceive another link between mutual information and conditional likelihood maximization from Fano's inequality (3.1), which relates the mutual information with an *upper bound* on the probability of correct reconstructions. Our goal here is to explore these apparent similarities and establish a possible relation between variational information maximization and conditional likelihood training in generalized chains independently of specific parameterizations.

We will begin the discussion by outlining a general relation between conditional likelihood and conditional mutual information maximization in general feed-forward models $\mathsf{x} \to \mathsf{y} \to \tilde{\mathsf{x}}$. Clearly, this view extends the reconstruction paradigm of autoencoders, since for general chains the sources x and the outputs $\tilde{\mathsf{x}}$ may generally lie in different domains. By analogy with the results of section 3.2, we will show that for i.i.d. patterns, the exact conditional likelihood for the considered chains defines a proper lower bound on the *conditional* mutual information in memoryless channels. Effectively, this formulation corresponds to maximizing the amount of information which the encodings contain about the *outputs* for the given set of source patterns $I(\{\tilde{\mathsf{x}}\}, \{\mathsf{y}\}|\{\mathsf{x}\})$.

Then we will consider a special case of stochastic autoencoders, where the empirical distribution is constrained so that $\tilde{p}(\tilde{\mathsf{x}}|\mathsf{x}) \sim \delta(\mathsf{x} - \tilde{\mathsf{x}})$, i.e. the decoded patterns $\tilde{\mathsf{x}}$ are the exact uncorrupted copies of the sources x. For this specific case, it is easy to establish a simple relation between conditional likelihood training in stochastic autoencoders and maximization of the mutual information in the channel model $\mathsf{x} \to \mathsf{y}$. Specifically, we will show that the exact mutual information $I(\mathsf{x}, \mathsf{y})$ is a proper *lower bound* on the probability of correct reconstructions in stochastic autoencoders where the reconstructing distribution is given by the exact posterior. Moreover, we will show that if the encoding mappings of such autoencoders are noiseless, maximum conditional likelihood training is equivalent to maximizing the exact mutual information in the corresponding noiseless channel.

Finally, we will consider the situation when the conditional likelihood cannot be computed exactly, and show that the standard variational approaches to maximizing the conditional likelihood in stochastic autoencoders reduce to a specific

instance of the generic IM algorithm.

3.3.1 Variational Information Maximization and Exact Conditional Likelihood Training in Feed-Forward Models

To establish a relationship between conditional likelihood and mutual information training, we will now consider a simple feed-forward model, which defines the joint distribution over the latent variables y and outputs $\tilde{\mathsf{x}}$ conditionally on the sources x, i.e.

$$\mathcal{M}_C \stackrel{\text{def}}{=} p(\mathsf{y}|\mathsf{x})p(\tilde{\mathsf{x}}|\mathsf{y}). \tag{3.28}$$

The functional parameters $p(\mathsf{y}|\mathsf{x})$ and $p(\tilde{\mathsf{x}}|\mathsf{y})$ are determined by maximizing the conditional likelihood

$$\mathcal{L}_{\tilde{x}|x} = \log p\left(\{\tilde{\mathsf{x}}\}|\{\mathsf{x}\}\right)/M = \langle \log p(\tilde{\mathsf{x}}|\mathsf{x})\rangle_{\tilde{p}(\mathsf{x},\tilde{\mathsf{x}})}, \tag{3.29}$$

where $\tilde{p}(\tilde{\mathsf{x}},\mathsf{x})$ is the empirical distribution of the input and output pairs. In the general formulation, we will not impose specific constraints on the exact form of the empirical distribution, so the ranges $\mathcal{R}_{\mathsf{x}}$, $\mathcal{R}_{\tilde{\mathsf{x}}}$ of the source and output variables x, $\tilde{\mathsf{x}}$ may in general be different. Once the model is trained, the states of the latent variables y corresponding to any given source-output pair are inferred from the exact posterior $p(\mathsf{y}|\mathsf{x},\tilde{\mathsf{x}}) \propto p(\tilde{\mathsf{x}}|\mathsf{y})p(\mathsf{y}|\mathsf{x})$.

We will now define a recognition model

$$\mathcal{M}_{IC} \stackrel{\text{def}}{=} \tilde{p}(\mathsf{x},\tilde{\mathsf{x}})p(\mathsf{y}|\mathsf{x},\tilde{\mathsf{x}}), \tag{3.30}$$

where $p(\mathsf{y}|\mathsf{x},\tilde{\mathsf{x}})$ is the exact posterior of the feed-forward model $\mathcal{M}_C$. Clearly, the encoder model's specification (3.30) leads to equivalence of the exact inference in $\mathcal{M}_C$ and $\mathcal{M}_{IC}$ for all the visible source-output pairs $\{\mathsf{x},\tilde{\mathsf{x}}\}$. While this formulation might not have an easily interpretable link to communication theory[7], it does have a direct relation to conditional likelihood training in feed-forward models, which generalize the reconstruction paradigm of stochastic autoencoders. Indeed, it is easy to show that the exact conditional likelihood in $\mathcal{M}_C$ is in fact a lower bound on the conditional mutual information $I(\tilde{\mathsf{x}},\mathsf{y}|\mathsf{x})$ in $\mathcal{M}_{IC}$.

Proposition 3.3. *For i.i.d. patterns $\{\mathsf{x},\tilde{\mathsf{x}}\}$, conditional likelihood learning in the feed-forward model $\mathcal{M}_C$ corresponds to maximization of a* **lower bound** *on the conditional mutual information $I(\tilde{\mathsf{x}},\mathsf{y}|\mathsf{x})$ in $\mathcal{M}_{IC}$. Up to irrelevant constants, this bound is weaker or as tight as $\hat{I}_C(\tilde{\mathsf{x}},\mathsf{y}|\mathsf{x}) \stackrel{\text{def}}{=} \langle \log p(\tilde{\mathsf{x}}|\mathsf{x},\mathsf{y})\rangle_{p(\mathsf{y}|\mathsf{x},\tilde{\mathsf{x}})\tilde{p}(\mathsf{x},\tilde{\mathsf{x}})}$.*

We prove and discuss this result in Appendix B.1. Note, however, that in the special case when the chain model $\mathcal{M}_C$ and the empirical distribution $\tilde{p}(\mathsf{x},\tilde{\mathsf{x}})$ define an autoencoder, the bound $\hat{I}_C(\tilde{\mathsf{x}},\mathsf{y}|\mathsf{x})$ gives a *model-specific* upper bound on the probability of correct reconstructions

$$\mathcal{L}_{\tilde{\mathsf{x}}|\mathsf{x}} \leq \hat{I}_C(\tilde{\mathsf{x}},\mathsf{y}|\mathsf{x}) + H_{\tilde{p}}(\tilde{\mathsf{x}}|\mathsf{x}) \tag{3.31}$$

[7] One may potentially view $\mathcal{M}_{IC}$ as a model of the *reverse transmission path* for half-duplex channels (Glover and Grant (2003)), where the received vectors are sent back across a noisy channel whose specific properties may in general be affected by the original sources.

(*cf* Fano's inequality (3.1)).

We will now use the general result of proposition 3.3 to show that for several specific autoencoders, the conditional likelihood training in $\mathcal{M}_C$ has a strong correspondence to mutual information maximization in *unconditional* encoder models $\mathcal{M}_I = \tilde{p}(\mathsf{x})p(\mathsf{y}|\mathsf{x})$, which has a stronger relation to the variational IM formulation (see Section 2.1.2).

3.3.2 Variational Information Maximization and Exact Training of Autoencoders

If $\mathcal{R}_\mathsf{x} \equiv \mathcal{R}_{\tilde{\mathsf{x}}}$ and the outputs $\{\tilde{\mathsf{x}}\}$ are the exact uncorrupted copies of the sources $\{\mathsf{x}\}$ for all the training patterns $(\mathsf{x}^{(i)}, \tilde{\mathsf{x}}^{(i)}) \in \mathcal{X}_C$, $i \in \{1, \ldots, M\}$, then the feed-forward model $\mathcal{M}_C$ reduces to a stochastic autoencoder. We will now outline a few simple relations between mutual information maximization in simple channel models $\mathcal{M}_I$ and conditional likelihood training in autoencoders[8] $\mathcal{M}_C$ for several specific constraints on the encoder and decoder distributions. However, first we will prove a simple lemma which will be extensively used throughout the discussions in the rest of this chapter. Effectively, it will help us to reduce computations of average quantities in chains $\mathsf{x} \to \mathsf{y} \to \tilde{\mathsf{x}}$ to computing specific integrals in the encoder models $\mathsf{x} \to \mathsf{y}$.

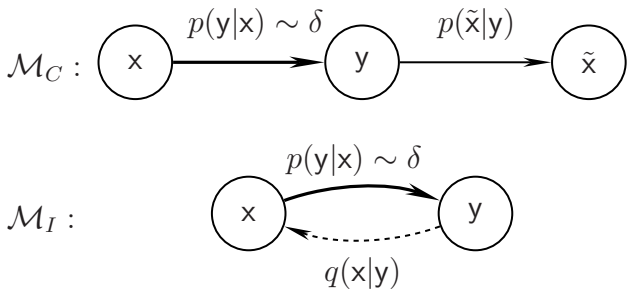
Lemma 3.1. *Let the outputs $\tilde{\mathsf{x}}$ be the exact copies of the sources x, i.e. $\mathcal{R}_\mathsf{x} \equiv \mathcal{R}_{\tilde{\mathsf{x}}}$ and $\tilde{p}(\tilde{\mathsf{x}}|\mathsf{x}) \sim \delta(\mathsf{x}-\tilde{\mathsf{x}})$. Additionally, let $\tilde{p}(\mathsf{x}) = \sum_{i=1}^{M} \delta(\mathsf{x}-\mathsf{x}^{(i)})/M$ be the empirical distribution of the sources, and $q_{\mathsf{x}|\mathsf{y}}(\mathsf{x}=\mathsf{s}|\mathsf{y}) = q_{\tilde{\mathsf{x}}|\mathsf{y}}(\tilde{\mathsf{x}}=\mathsf{s}|\mathsf{y})$ for all patterns $\mathsf{s} \in \mathcal{R}_\mathsf{x}$ and encodings $\mathsf{y} \in \mathcal{R}_\mathsf{y}$. Then $\langle \mathcal{F}\{q_{\tilde{\mathsf{x}}|\mathsf{y}}(\tilde{\mathsf{x}}|\mathsf{y})\}\rangle_{p(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x},\tilde{\mathsf{x}})} = \langle \mathcal{F}\{q_{\mathsf{x}|\mathsf{y}}(\mathsf{x}|\mathsf{y})\}\rangle_{p(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x})}$, where $\mathcal{F}$ is an arbitrary y-integrable functional of the conditionals $q(\tilde{\mathsf{x}}|\mathsf{y})$ and $q(\mathsf{x}|\mathsf{y})$.*

Proof. By the direct substitution, we obtain

$$\begin{aligned}
\langle \mathcal{F}\{q_{\tilde{\mathsf{x}}|\mathsf{y}}(\tilde{\mathsf{x}}|\mathsf{y})\}\rangle_{p(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x},\tilde{\mathsf{x}})} &= \frac{1}{M}\sum_{i=1}^{M} \langle \mathcal{F}\{q_{\tilde{\mathsf{x}}|\mathsf{y}}(\tilde{\mathsf{x}}|\mathsf{y})\}\rangle_{p(\mathsf{y}|\mathsf{x}^{(i)})\tilde{p}(\tilde{\mathsf{x}}|\mathsf{x}^{(i)})} \\
&= \frac{1}{M}\sum_{i=1}^{M} \int_\mathsf{y} p(\mathsf{y}|\mathsf{x}^{(i)}) \int_{\tilde{\mathsf{x}}} \mathcal{F}\{q_{\tilde{\mathsf{x}}|\mathsf{y}}(\tilde{\mathsf{x}}|\mathsf{y})\}\delta(\tilde{\mathsf{x}}-\mathsf{x}^{(i)})\, \mathrm{d}\tilde{\mathsf{x}}\, \mathrm{d}\mathsf{y} \\
&= \frac{1}{M}\sum_{i=1}^{M} \int_\mathsf{y} p(\mathsf{y}|\mathsf{x}^{(i)}) \mathcal{F}\{q_{\mathsf{x}|\mathsf{y}}(\mathsf{x}^{(i)}|\mathsf{y})\}\, \mathrm{d}\mathsf{y} \qquad (3.32) \\
&= \langle \mathcal{F}\{q_{\mathsf{x}|\mathsf{y}}(\mathsf{x}|\mathsf{y})\}\rangle_{p(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x})}, \qquad (3.33)
\end{aligned}$$

where the identity (3.32) follows from the condition that $q_{\mathsf{x}|\mathsf{y}}(\mathsf{x}=\mathsf{s}|\mathsf{y}) = q_{\tilde{\mathsf{x}}|\mathsf{y}}(\tilde{\mathsf{x}}=\mathsf{s}|\mathsf{y}).\forall \mathsf{s} \in \mathcal{R}_\mathsf{x}, \forall \mathsf{y} \in \mathcal{R}_\mathsf{y}$. □

[8] As one-layer autoencoders will have the same graphical structure and parameterization as the general chain $\mathsf{x} \to \mathsf{y} \to \tilde{\mathsf{x}}$, we will refer to them as $\mathcal{M}_C$ (see expression (3.28)). The only things different from the more general feed-forward models are the constraints on the empirical distribution $\tilde{p}(\mathsf{x}, \tilde{\mathsf{x}})$ and the ranges of the source-output variables, which do not form a part of the graphical specification.



$$\boxed{\tilde{I}(\mathsf{x},\mathsf{y}) = \mathcal{L}_{\tilde{\mathsf{x}}|\mathsf{x}} + H_{\tilde{p}}(\mathsf{x}) \text{ for } q(\mathsf{x}=\mathsf{s}|\mathsf{y}) = p(\tilde{\mathsf{x}}=\mathsf{s}|\mathsf{y})}$$

Figure 3.5: Variational information maximization and noiseless autoencoders. Maximization of the conditional likelihood in noiseless autoencoders $\mathcal{M}_C$ reduces to maximization of the variational lower bound $\tilde{I}(\mathsf{x},\mathsf{y})$ in the corresponding noiseless channels $\mathcal{M}_I$, where the variational decoder is given by $q(\mathsf{x}=\mathsf{s}|\mathsf{y}) = p(\tilde{\mathsf{x}}=\mathsf{s}|\mathsf{y})$. Thick arrows show deterministic mappings; the dashed arrow corresponds to the variational distribution.

Specifically, for $\mathcal{F}(q) \equiv \log(q)$, we get

$$\langle \log p(\tilde{\mathsf{x}}|\mathsf{y}) \rangle_{p(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x},\tilde{\mathsf{x}})} = \langle \log q(\mathsf{x}|\mathsf{y}) \rangle_{p(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x})}, \tag{3.34}$$

where we have defined $q(\mathsf{x}=\mathsf{s}|\mathsf{y}) \stackrel{\text{def}}{=} p(\tilde{\mathsf{x}}=\mathsf{s}|\mathsf{y})$ for all patterns $\mathsf{s} \in \mathcal{R}_\mathsf{x} \equiv \mathcal{R}_{\tilde{\mathsf{x}}}$, $\mathsf{y} \in \mathcal{R}_\mathsf{y}$. We will now use the results (3.33) and (3.34) in order to relate the information-theoretic learning in channel models and the conditional learning in autoencoders.

3.3.2.1 Noiseless Autoencoders

In conventional approaches to autoencoder training, it is typically presumed that the encoding part of the autoencoder is deterministic (Hinton (1989), Baldi and Hornik (1989), Bishop (1995)), i.e.

$$p\left(\mathsf{y}|\mathsf{x}^{(i)}\right) \sim \delta\left(\mathsf{y} - \mathsf{f}(\mathsf{x}^{(i)})\right) = \delta\left(\mathsf{y} - \mathsf{y}^{(i)}\right), \tag{3.35}$$

where $\mathsf{y}^{(i)} \stackrel{\text{def}}{=} \mathsf{f}(\mathsf{x}^{(i)})$. With a slight abuse of terminology, we will refer to autoencoders satisfying (3.35) as being *noiseless*, rather than *stochastic*, in order to stress the determinism of the encoding mapping. It is easy to see that for the specific case (3.35), optimization of the exact conditional likelihood $\mathcal{L}_{\tilde{\mathsf{x}}|\mathsf{x}}$ in the autoencoder $\mathcal{M}_C$ is equivalent to maximizing the generic lower bound on the mutual information in the corresponding *noiseless* channel $\mathcal{M}_I$. The model of the noiseless channel is defined similarly to expression (3.2), i.e. $\mathcal{M}_I = \tilde{p}(\mathsf{x})p(\mathsf{y}|\mathsf{x})$.

Proposition 3.4. *For i.i.d. patterns* $\{\mathsf{x}\}$*, exact conditional likelihood learning in* **noiseless autoencoders** $\mathcal{M}_C$ *is equivalent to maximizing the generic lower bound on mutual information* $\tilde{I}(\mathsf{x},\mathsf{y}) = \langle \log q(\mathsf{x}|\mathsf{y})\rangle_{p(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x})} + H_{\tilde{p}}(\mathsf{x})$ *in* **noiseless channels** $\mathcal{M}_I$*, where the variational decoder* $q(\mathsf{x}|\mathsf{y})$ *is constrained to be equivalent to the decoding distribution of the autoencoder.*

Proof. Let $\tilde{\mathsf{x}}$ denote the reconstructed variables of the autoencoder model, so that $\mathcal{R}_{\mathsf{x}} = \mathcal{R}_{\tilde{\mathsf{x}}}$ and $\tilde{p}(\mathsf{x},\tilde{\mathsf{x}}) = \tilde{p}(\mathsf{x})\delta(\tilde{\mathsf{x}}-\mathsf{x})$. From expressions (3.28) and (3.35) it is clear that for noiseless autoencoders the exact posterior of the conditional model $\mathcal{M}_C$ is given by

$$p(\mathsf{y}|\mathsf{x}^{(i)},\tilde{\mathsf{x}}) = \frac{p(\mathsf{y}|\mathsf{x}^{(i)})p(\tilde{\mathsf{x}}|\mathsf{y})}{\int_{\mathsf{y}} p(\mathsf{y}|\mathsf{x}^{(i)})p(\tilde{\mathsf{x}}|\mathsf{y})\,\mathrm{d}\mathsf{y}} = \frac{\delta(\mathsf{y}-\mathsf{y}^{(i)})p(\tilde{\mathsf{x}}|\mathsf{y}^{(i)})}{p(\tilde{\mathsf{x}}|\mathsf{y}^{(i)})}, \tag{3.36}$$

where $\mathsf{y}^{(i)}$ is defined as in expression (3.35). Then we immediately obtain

$$p(\mathsf{y}|\mathsf{x}^{(i)},\tilde{\mathsf{x}}) = \delta(\mathsf{y}-\mathsf{y}^{(i)}) = p(\mathsf{y}|\mathsf{x}^{(i)}), \tag{3.37}$$

i.e. for deterministic encoders, the exact posterior in $\mathcal{M}_C$ is conditionally independent of the source reconstructions. By substituting (3.37) into the general expression (B.3) of the conditional likelihood of feed-forward models (see proposition 3.3), we obtain

$$\begin{aligned} \mathcal{L}_{\tilde{x}|x} &= \langle \log p(\tilde{\mathsf{x}}|\mathsf{x},\mathsf{y})\rangle_{p(\mathsf{y}|\mathsf{x},\tilde{\mathsf{x}})\tilde{p}(\mathsf{x},\tilde{\mathsf{x}})} - \langle KL(p(\mathsf{y}|\mathsf{x},\tilde{\mathsf{x}})\|p(\mathsf{y}|\mathsf{x}))\rangle_{\tilde{p}(\mathsf{x},\tilde{\mathsf{x}})} \\ &= \langle \log p(\tilde{\mathsf{x}}|\mathsf{x},\mathsf{y})\rangle_{p(\mathsf{y}|\mathsf{x},\tilde{\mathsf{x}})\tilde{p}(\mathsf{x},\tilde{\mathsf{x}})} = \langle \log p(\tilde{\mathsf{x}}|\mathsf{y})\rangle_{p(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x},\tilde{\mathsf{x}})}, \end{aligned} \tag{3.38}$$

as the Kullback Leibler term cancels for all training patterns $\mathsf{x}^{(i)} \sim \tilde{p}(\mathsf{x})$. Note that in the last identity in (3.38) we have used the fact that $p(\mathsf{y}|\mathsf{x},\tilde{\mathsf{x}}) = p(\mathsf{y}|\mathsf{x})$ as the consequence of the deterministic assumption (3.37), and $p(\tilde{\mathsf{x}}|\mathsf{x},\mathsf{y}) = p(\tilde{\mathsf{x}}|\mathsf{y})$ from the chain structure of the model $\mathcal{M}_C$ (see Appendix C.3, proposition C.1).

Let us now define the conditional distribution $q(\mathsf{x}|\mathsf{y})$, which is constrained to be equivalent to the decoding mapping of the autoencoder, i.e. $q_{\mathsf{x}|\mathsf{y}}(\mathsf{x}=\mathsf{s}|\mathsf{y}) = p_{\tilde{\mathsf{x}}|\mathsf{y}}(\tilde{\mathsf{x}}=\mathsf{s}|\mathsf{y})$ for all $\mathsf{s}\in\mathcal{R}_{\mathsf{x}}\equiv\mathcal{R}_{\tilde{\mathsf{x}}}$, $\mathsf{y}\in\mathcal{R}_{\mathsf{y}}$. Then from lemma 3.1 and (3.34), the exact conditional likelihood $\mathcal{L}_{\tilde{\mathsf{x}}|\mathsf{x}}$ in the considered noiseless autoencoders reduces to

$$\mathcal{L}_{\tilde{\mathsf{x}}|\mathsf{x}} = \langle \log p(\tilde{\mathsf{x}}|\mathsf{y})\rangle_{p(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x},\tilde{\mathsf{x}})} = \langle \log q(\mathsf{x}|\mathsf{y})\rangle_{p(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x})}, \tag{3.39}$$

which is effectively the varying part of the generic variational lower bound (2.2) on the mutual information $I(\mathsf{x},\mathsf{y})$ for the encoder model $\mathcal{M}_I$. Again, here $\tilde{p}(\mathsf{x})$ is the empirical distribution of the sources, and $p(\mathsf{y}|\mathsf{x})$ is the deterministic encoder given by (3.37). Therefore, from (2.2) and (3.39) we get

$$I(\mathsf{x},\mathsf{y}) \geq \tilde{I}(\mathsf{x},\mathsf{y}) = \mathcal{L}_{\tilde{\mathsf{x}}|\mathsf{x}} + H_{\tilde{p}}(\mathsf{x}). \tag{3.40}$$

□

Thus, optimization of the exact conditional likelihood $\mathcal{L}_{\tilde{\mathsf{x}}|\mathsf{x}}$ in conventional noiseless autoencoders may be seen as a special case of the variational IM algorithm for the corresponding noiseless channel, where the variational decoder

is identical to the autoencoder's decoding distribution (see Figure 3.5). From proposition 3.4 we can see that there is only a constant gap between the surfaces defined by $\tilde{I}(\mathsf{x},\mathsf{y})$ and the conditional likelihood $\mathcal{L}_{\tilde{\mathsf{x}}|\mathsf{x}}$, which for the considered case does not affect the extrema. (Note that this contrasts with the surfaces defined by log-likelihoods $\mathcal{L}$ of generative models $\mathcal{M}_L$, see (3.8)). Moreover, it is easy to see that the bound on $I(\mathsf{x},\mathsf{y})$ is saturated if the decoding distribution of the autoencoder is constrained to be identical to the exact posterior of $\mathcal{M}_I$, i.e.

$$p_{\tilde{\mathsf{x}}|\mathsf{y}}(\tilde{\mathsf{x}}=\mathsf{s}|\mathsf{y}) \propto p_{\mathsf{y}|\mathsf{x}}(\mathsf{y}|\mathsf{x}=\mathsf{s})\tilde{p}(\mathsf{x}=\mathsf{s}).\ \forall \mathsf{s}\in\mathcal{R}_{\mathsf{x}}\equiv\mathcal{R}_{\tilde{\mathsf{x}}},\mathsf{y}\in\mathcal{R}_{\mathsf{y}}. \tag{3.41}$$

Clearly, for this specific case, minimization of the *reconstruction error in noiseless autoencoders* and maximization of the *exact mutual information in noiseless channels* are formally equivalent, if the decoding distribution is the exact posterior.

3.3.2.2 Stochastic Autoencoders with Bayesian Decoders

We have shown that the conventional conditional likelihood training of noiseless autoencoders $\mathcal{M}_C$ may indeed be viewed as a special case of variational information maximization in noiseless channels $\mathcal{M}_I$, where the variational posterior is given by the $\mathcal{M}_C$'s decoding distribution. Additionally, for the special case when the decoder is given by the exact Bayesian posterior, the conditional training of noiseless autoencoders is equivalent to maximizing the *exact* mutual information $I(\mathsf{x},\mathsf{y})$ under the deterministic encoding constraint. Here we relax the noiseless assumption and briefly discuss a general link between $I(\mathsf{x},\mathsf{y})$ and $\mathcal{L}_{\tilde{\mathsf{x}}|\mathsf{x}}$ maximization. Specifically, we will motivate maximization of the mutual information as an approximate method for reducing the reconstruction error.

Proposition 3.5. *If the decoding distribution of an autoencoder model* $\mathcal{M}_C = p_{\tilde{\mathsf{x}}|\mathsf{y}}(\tilde{\mathsf{x}}|\mathsf{y})p_{\mathsf{y}|\mathsf{x}}(\mathsf{y}|\mathsf{x})$ *is given by the exact posterior* $p_{\tilde{\mathsf{x}}|\mathsf{y}}(\tilde{\mathsf{x}}|\mathsf{y}) \propto \tilde{p}(\mathsf{x})p_{\mathsf{y}|\mathsf{x}}(\mathsf{y}|\mathsf{x})$*, then for i.i.d. patterns the conditional likelihood is* **lower-bounded** *by the exact mutual information of the corresponding model of a stochastic memoryless channel* $\mathcal{M}_I = \tilde{p}(\mathsf{x})p_{\mathsf{y}|\mathsf{x}}(\mathsf{y}|\mathsf{x})$ *(up to irrelevant constants).*

Proof. The proof follows from Bayes rule and Jensen's inequality (Jensen (1906) and e.g. Korn and Korn (1968)). By definition, we obtain

$$\begin{aligned}\mathcal{L}_{\tilde{\mathsf{x}}|\mathsf{x}} = \langle \log \langle p_{\tilde{\mathsf{x}}|\mathsf{y}}(\tilde{\mathsf{x}}|\mathsf{y})\rangle_{p_{\mathsf{y}|\mathsf{x}}(\mathsf{y}|\mathsf{x})}\rangle_{\tilde{p}(\tilde{\mathsf{x}},\mathsf{x})} &= \frac{1}{M}\sum_{i=1}^{M}\log\int_{\mathsf{y}} p_{\tilde{\mathsf{x}}|\mathsf{y}}(\tilde{\mathsf{x}}=\mathsf{s}^{(i)}|\mathsf{y})p_{\mathsf{y}|\mathsf{x}}(\mathsf{y}|\mathsf{x}=\mathsf{s}^{(i)})\,\mathrm{d}\mathsf{y} \\ &= \frac{1}{M}\sum_{i=1}^{M}\log\tilde{p}(\mathsf{x}=\mathsf{s}^{(i)})\left[\int_{\mathsf{y}}\frac{p^2_{\mathsf{y}|\mathsf{x}}(\mathsf{y}|\mathsf{x}=\mathsf{s}^{(i)})}{p(\mathsf{y})}\right]\mathrm{d}\mathsf{y},\end{aligned} \tag{3.42}$$

where $p(\mathsf{y}) = \langle p_{\mathsf{y}|\mathsf{x}}(\mathsf{y}|\mathsf{x})\rangle_{\tilde{p}(\mathsf{x})}$. Then from Jensen's inequality we may transform (3.42) to

$$\begin{aligned}\mathcal{L}_{\tilde{\mathsf{x}}|\mathsf{x}} &\geq \frac{1}{M}\sum_{i=1}^{M}\left\langle \log p_{\mathsf{y}|\mathsf{x}}(\mathsf{y}|\mathsf{x}=\mathsf{s}^{(i)}) + \log\tilde{p}(\mathsf{x}=\mathsf{s}^{(i)}) - \log p(\mathsf{y})\right\rangle_{p(\mathsf{y}|\mathsf{x}=\mathsf{s}^{(i)})} \\ &= \left\langle \log p_{\mathsf{y}|\mathsf{x}}(\mathsf{y}|\mathsf{x})/p(\mathsf{y})\right\rangle_{p_{\mathsf{y}|\mathsf{x}}(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x})} + \langle\log\tilde{p}(\mathsf{x})\rangle_{\tilde{p}(\mathsf{x})}.\end{aligned} \tag{3.43}$$

From the definition of the mutual information in the encoder model $\mathcal{M}_I$, we easily get the lower bound on the conditional likelihood

$$\mathcal{L}_{\tilde{\mathsf{x}}|\mathsf{x}} \geq I(\mathsf{x}, \mathsf{y}) - H_{\tilde{p}}(\mathsf{x}). \tag{3.44}$$

□

Unlike propositions 3.3 and 3.4, which effectively define information-theoretic upper bounds on the conditional likelihood in a way which is vaguely reminiscent of Fano's inequality (see expression (3.1) and discussion in Appendix B.1), proposition 3.5 provides a general motivation for maximizing the mutual information as a method of decreasing the reconstruction error. Indeed, inequality (3.44) may be viewed as a proper **lower bound** on the average probability of the correct reconstructions in stochastic autoencoders with the specific Bayesian setting of the decoding distribution. As the exact conditional likelihood (3.42) will typically be intractable due to a generally non-trivial form of the log posterior, one could naively hope to simplify the process of maximizing $\mathcal{L}_{\tilde{\mathsf{x}}|\mathsf{x}}$ by considering optimization of its proper information-theoretic lower bound $I(\mathsf{x}, \mathsf{y})$. However, as we have discussed in previous chapters, in many cases of interest the bound (3.44) would involve computing intractable entropic terms. To handle the intractability we may consider optimizing tractable lower bounds on the mutual information, such as the generic bound (2.2), which would correspond to further relaxations of (3.44) and still define proper bounds on $\mathcal{L}_{\tilde{\mathsf{x}}|\mathsf{x}}$.

Of course, in principle other kinds of lower bounds on the conditional likelihood may be considered. It is interesting and potentially enlightening to see how these objectives and the solutions obtained by the conditional training in completely general, intractable autoencoders could relate to mutual information maximization in stochastic channels. We will now show the fundamental similarity between both approaches for the general variational reformulations.

3.3.3 Variational Information Maximization and Variational Training of Autoencoders

In proposition 3.3 we have shown that for the general case of stochastic autoencoders, the exact conditional likelihood $\mathcal{L}_{\tilde{\mathsf{x}}|\mathsf{x}}$ defines a relaxation of the variational lower bound on the conditional mutual information $I(\tilde{\mathsf{x}}, \mathsf{y}|\mathsf{x})$. We have also shown that conventional approaches to training deterministic autoencoders may indeed be viewed as special cases of variational mutual information maximization in noiseless communication channels. While these results may be useful for understanding a general relation between the exact conditional likelihood and mutual information maximization, they are limited to the cases when the computations are tractable. Of course, in general it may be intractable to optimize the exact conditional likelihood $\mathcal{L}_{\tilde{\mathsf{x}}|\mathsf{x}} = \langle \log \langle p(\tilde{\mathsf{x}}|\mathsf{y}) \rangle_{p(\mathsf{y}|\mathsf{x})} \rangle_{\tilde{p}(\mathsf{x},\tilde{\mathsf{x}})}$, and one needs to consider approximations. Here we handle the intractability of integrating over the hidden states by maximizing the variational Jensen's bound on the objective $\mathcal{L}_{\tilde{\mathsf{x}}|\mathsf{x}}$, and show that variational approaches to the conditional likelihood training in stochastic autoencoders indeed reduce to optimizing the generic lower bound on $I(\mathsf{x}, \mathsf{y})$ in stochastic channels.

Clearly, by analogy with the standard variational approaches of training generative models (see expression (3.17)), we can define the variational lower bound on the conditional likelihood $\mathcal{L}_{\tilde{\mathsf{x}}|\mathsf{x}}$ in $\mathcal{M}_C$ as

$$\mathcal{L}_{\tilde{\mathsf{x}}|\mathsf{x}} \geq \tilde{\mathcal{L}}_{\tilde{\mathsf{x}}|\mathsf{x}} = \underbrace{\langle \log p(\tilde{\mathsf{x}}, \mathsf{y}|\mathsf{x}) \rangle_{q(\mathsf{y}|\mathsf{x},\tilde{\mathsf{x}})\tilde{p}(\mathsf{x},\tilde{\mathsf{x}})}}_{\text{Average energy}} - \underbrace{\langle \log q(\mathsf{y}|\mathsf{x},\tilde{\mathsf{x}}) \rangle_{q(\mathsf{y}|\mathsf{x},\tilde{\mathsf{x}})\tilde{p}(\mathsf{x},\tilde{\mathsf{x}})}}_{\text{Entropy}}. \quad (3.45)$$

We can now expand the energy term to get

$$\tilde{\mathcal{L}}_{\tilde{\mathsf{x}}|\mathsf{x}} = \langle \log p(\tilde{\mathsf{x}}|\mathsf{x},\mathsf{y}) \rangle_{q(\mathsf{y}|\mathsf{x},\tilde{\mathsf{x}})\tilde{p}(\mathsf{x},\tilde{\mathsf{x}})} - \langle KL(q(\mathsf{y}|\mathsf{x},\tilde{\mathsf{x}}) \| p(\mathsf{y}|\mathsf{x})) \rangle_{\tilde{p}(\mathsf{x},\tilde{\mathsf{x}})}, \quad (3.46)$$

where $q(\mathsf{y}|\mathsf{x},\tilde{\mathsf{x}})$ is an arbitrary variational posterior, which we constrain to ensure the tractability of computing (3.46). Clearly, (3.46) is saturated for $q(\mathsf{y}|\mathsf{x},\tilde{\mathsf{x}}) \equiv p(\mathsf{y}|\mathsf{x},\tilde{\mathsf{x}})$. The standard variational EM algorithm optimizes the objective (3.46) with respect to the encoder $p(\mathsf{y}|\mathsf{x})$, decoder $p(\tilde{\mathsf{x}}|\mathsf{x},\mathsf{y}) = p(\tilde{\mathsf{x}}|\mathsf{y})$, and the variational distribution $q(\mathsf{y}|\mathsf{x},\tilde{\mathsf{x}})$, subject to the imposed constraints. Our goal here is to compare the fixed points of the variational EM algorithm for stochastic autoencoders with the fixed points of the IM algorithm for the corresponding encoder model

$$\tilde{\mathcal{M}}_I = q_{\mathsf{y}|\mathsf{x}}(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x}), \quad (3.47)$$

where

$$q_{\mathsf{y}|\mathsf{x}}(\mathsf{y}|\mathsf{x}=\mathsf{s}) \stackrel{\text{def}}{=} q(\mathsf{y}|\mathsf{x}=\mathsf{s},\tilde{\mathsf{x}}=\mathsf{s}) \quad (3.48)$$

for all $\mathsf{s} \in \mathcal{R}_{\mathsf{x}} = \mathcal{R}_{\tilde{\mathsf{x}}}$. Again, during inference in $\mathcal{M}_C$ we approximate the exact posterior $p(\mathsf{y}|\mathsf{x},\tilde{\mathsf{x}})$ by the variational distribution $q(\mathsf{y}|\mathsf{x},\tilde{\mathsf{x}})$. Note that since the channel encoder $q_{\mathsf{y}|\mathsf{x}}(\mathsf{y}|\mathsf{x})$ corresponds to the specific instance of the tractable variational posterior $q(\mathsf{y}|\mathsf{x},\tilde{\mathsf{x}})$, it also lies in a tractable family.

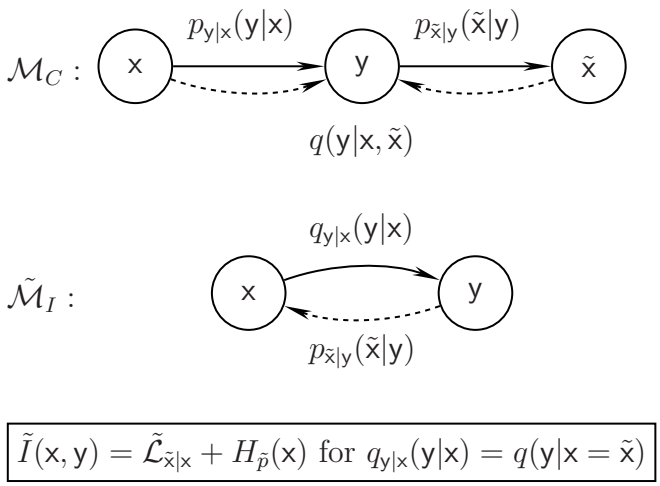
Proposition 3.6. *For i.i.d. patterns* $\{\mathsf{x},\tilde{\mathsf{x}}\}$*, optimization of the standard variational Jensen's bound on the conditional likelihood* $\tilde{\mathcal{L}}_{\tilde{\mathsf{x}}|\mathsf{x}}$ *in* **stochastic autoencoders** $\mathcal{M}_C = p(\mathsf{y}|\mathsf{x})p(\tilde{\mathsf{x}}|\mathsf{y})$ *reduces to maximization of a specific variational lower bound on the mutual information* $I(\tilde{\mathsf{x}},\mathsf{y})$ *in the* **stochastic memoryless channel** $\tilde{\mathcal{M}}_I = q_{\mathsf{y}|\mathsf{x}}(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x})$.

Proof. We will prove the proposition by expressing the fixed point updates of the variational EM on the bound $\tilde{\mathcal{L}}_{\tilde{\mathsf{x}}|\mathsf{x}}$ and comparing them with the variational information maximizing algorithm for the encoder model (3.47).

The autoencoder formulation of the conditional training implies the constraint on the empirical distribution, namely $\tilde{p}(\tilde{\mathsf{x}}|\mathsf{x}) \sim \delta(\tilde{\mathsf{x}}-\mathsf{x})$. This transforms the bound on the conditional likelihood (3.46) to

$$\begin{aligned}\tilde{\mathcal{L}}_{\tilde{\mathsf{x}}|\mathsf{x}} &= \frac{1}{M}\sum_{m=1}^{M} \langle \log p_{\tilde{\mathsf{x}}|\mathsf{y}}(\tilde{\mathsf{x}}=\mathsf{s}^{(m)}|\mathsf{y}) \rangle_{q(\mathsf{y}|\mathsf{x}=\mathsf{s}^{(m)},\tilde{\mathsf{x}}=\mathsf{s}^{(m)})} \\ &- \frac{1}{M}\sum_{m=1}^{M} KL\left(q(\mathsf{y}|\mathsf{x}=\mathsf{s}^{(m)},\tilde{\mathsf{x}}=\mathsf{s}^{(m)}) \| p_{\mathsf{y}|\mathsf{x}}(\mathsf{y}|\mathsf{x}=\mathsf{s}^{(m)})\right), \end{aligned} \quad (3.49)$$

where $\mathsf{s}^{(m)} \in \mathcal{X}$ is the m^{th} training pattern. Note that due to the specific symmetric constraint on $\tilde{p}(\mathsf{x},\tilde{\mathsf{x}})$, the variational posteriors $q(\mathsf{y}|\mathsf{x}=\mathsf{s}^{(m)},\tilde{\mathsf{x}} \neq \mathsf{s}^{(m)})$ do



$$\boxed{\tilde{I}(\mathsf{x},\mathsf{y}) = \tilde{\mathcal{L}}_{\tilde{\mathsf{x}}|\mathsf{x}} + H_{\tilde{p}}(\mathsf{x}) \text{ for } q_{\mathsf{y}|\mathsf{x}}(\mathsf{y}|\mathsf{x}) = q(\mathsf{y}|\mathsf{x}=\tilde{\mathsf{x}})}$$

Figure 3.6: Variational information maximization and stochastic autoencoders. Maximization of the standard lower bound on the conditional likelihood $\tilde{\mathcal{L}}_{\tilde{\mathsf{x}}|\mathsf{x}}$ in $\mathcal{M}_C$ reduces to maximization of the generic lower bound $\tilde{I}(\mathsf{x},\mathsf{y})$ in the corresponding encoder models $\tilde{\mathcal{M}}_I \stackrel{\text{def}}{=} q_{\mathsf{y}|\mathsf{x}}(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x})$ if $q_{\mathsf{y}|\mathsf{x}=\mathsf{s}}(\mathsf{y}|\mathsf{x}) = q(\mathsf{y}|\mathsf{x}=\mathsf{s},\tilde{\mathsf{x}}=\mathsf{s})$. Here $q(\mathsf{y}|\mathsf{x},\tilde{\mathsf{x}})$ is the variational posterior of $\tilde{\mathcal{L}}_{\tilde{\mathsf{x}}|\mathsf{x}}$. Dashed arrows show the graphical structure of the variational distributions.

not affect the bound $\tilde{\mathcal{L}}_{\tilde{\mathsf{x}}|\mathsf{x}}$; in other words, for the optimization surface (3.49), the cardinality of $p(\mathsf{y}|\mathsf{x})$ is greater or equal to the effective cardinality of $q(\mathsf{y}|\mathsf{x},\tilde{\mathsf{x}})$.

At the first step of the variational EM algorithm we will optimize the bound on the conditional likelihood $\tilde{\mathcal{L}}_{\tilde{\mathsf{x}}|\mathsf{x}}$ with respect to the encoder distribution $p_{\mathsf{y}|\mathsf{x}}(\mathsf{y}|\mathsf{x})$. From (3.49), it is easy to see that the optimal encoder must satisfy

$$\forall \mathsf{s}^{(m)} \in \mathcal{X},\ p_{\mathsf{y}|\mathsf{x}}(\mathsf{y}|\mathsf{x}=\mathsf{s}^{(m)}) = q(\mathsf{y}|\mathsf{x}=\mathsf{s}^{(m)},\tilde{\mathsf{x}}=\mathsf{s}^{(m)}) \equiv q_{\mathsf{y}|\mathsf{x}}(\mathsf{y}|\mathsf{x}=\mathsf{s}^{(m)}). \quad (3.50)$$

Clearly, (3.50) is achievable as long as the family $\mathcal{F}_{q_{\mathsf{y}|\mathsf{x}}}$ of the variational distributions $q(\mathsf{y}|\mathsf{x}=\tilde{\mathsf{x}})$ is a subset of the family $\mathcal{F}_{p_{\mathsf{y}|\mathsf{x}}}$ of the exact conditionals $p_{\mathsf{y}|\mathsf{x}}(\mathsf{y}|\mathsf{x})$, which will typically be the case for the simplifying variational approximations.

Substituting the optimal encoder (3.50) into the objective (3.49), we obtain

$$\tilde{\mathcal{L}}^{(1)}_{\tilde{\mathsf{x}}|\mathsf{x}} = \frac{1}{M}\sum_{m=1}^{M} \langle \log p_{\tilde{\mathsf{x}}|\mathsf{y}}(\tilde{\mathsf{x}}=\mathsf{s}^{(m)}|\mathsf{y})\rangle_{q(\mathsf{y}|\mathsf{x}=\mathsf{s}^{(m)},\tilde{\mathsf{x}}=\mathsf{s}^{(m)})} = \langle \log p_{\tilde{\mathsf{x}}|\mathsf{y}}(\tilde{\mathsf{x}}|\mathsf{y})\rangle_{q(\mathsf{y}|\mathsf{x},\tilde{\mathsf{x}})\tilde{p}(\mathsf{x},\tilde{\mathsf{x}})} \quad (3.51)$$

which is formally just the non-constant part of the bound on the conditional mutual information $I(\mathsf{y},\tilde{\mathsf{x}}|\mathsf{x})$, where the conditional encoder is given by $q(\mathsf{y}|\mathsf{x},\tilde{\mathsf{x}})$ (see proposition 3.3). The objective (3.51) will now need to be optimized for $q(\mathsf{y}|\mathsf{x},\tilde{\mathsf{x}})$ and $p_{\tilde{\mathsf{x}}|\mathsf{y}}(\tilde{\mathsf{x}}|\mathsf{y})$, subject to the specific constraints on the decoder and the variational distribution. Without loss of generality, we may assume that the optimum for $q(\mathsf{y}|\mathsf{x}=\tilde{\mathsf{x}})$ is achieved at $q(\mathsf{y}|\mathsf{x}=\mathsf{s},\tilde{\mathsf{x}}=\mathsf{s}) = r(\mathsf{y}|\mathsf{x}=\mathsf{s}) \in \mathcal{F}_{q_{\mathsf{y}|\mathsf{x}}}$, where $r(\mathsf{y}|\mathsf{x}=\mathsf{s})$ is some distribution in the family of the variational posteriors, and

$\mathsf{s} \in \mathcal{R}_{\mathsf{x}}$. This transforms (3.51) into

$$\tilde{\mathcal{L}}^{(2)}_{\tilde{\mathsf{x}}|\mathsf{x}} = \langle \log p_{\tilde{\mathsf{x}}|\mathsf{y}}(\tilde{\mathsf{x}}|\mathsf{y}) \rangle_{r(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x},\tilde{\mathsf{x}})} = \langle \log p_{\tilde{\mathsf{x}}|\mathsf{y}}(\mathsf{x}|\mathsf{y}) \rangle_{r(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x})}, \tag{3.52}$$

where we have applied the results of lemma 3.1. Finally, at the last step of the variational EM we optimize the objective (3.52) for $p_{\tilde{\mathsf{x}}|\mathsf{y}}$, assuming that $r(\mathsf{y}|\mathsf{x})$ is fixed.

It is already easy to see that the iterative optimization of $\tilde{\mathcal{L}}_{\tilde{\mathsf{x}}|\mathsf{x}}$ is strongly related to maximization of the generic lower bound on the mutual information (2.2). Indeed, we may explicitly write the fixed point solutions for the t^{th} iteration of the algorithm[9] as

$$\begin{aligned} p^{(t-1)}_{\mathsf{y}|\mathsf{x}}(\mathsf{y}|\mathsf{x} = \mathsf{s}^{(m)}) &= q^{(t-1)}(\mathsf{y}|\mathsf{x} = \mathsf{s}^{(m)}, \tilde{\mathsf{x}} = \mathsf{s}^{(m)}), & (3.53) \\ q^{(t)}(\mathsf{y}|\mathsf{x} = \mathsf{s}^{(m)}, \tilde{\mathsf{x}} = \mathsf{s}^{(m)}) &= r^{(t)}(\mathsf{y}|\mathsf{x} = \mathsf{s}^{(m)}) \\ &\stackrel{\text{def}}{=} \arg\max_{q} \langle \log p^{(t-1)}_{\tilde{\mathsf{x}}|\mathsf{y}}(\mathsf{x} = \mathsf{s}^{(m)}|\mathsf{y}) \rangle_{q(\mathsf{y}|\mathsf{x}=\mathsf{s}^{(m)}, \tilde{\mathsf{x}}=\mathsf{s}^{(m)})}, & (3.54) \\ p^{(t)}_{\tilde{\mathsf{x}}|\mathsf{y}}(\tilde{\mathsf{x}} = \mathsf{s}^{(m)}|\mathsf{y}) &= \arg\max_{p_{\tilde{\mathsf{x}}|\mathsf{y}}} \langle \log p_{\tilde{\mathsf{x}}|\mathsf{y}}(\mathsf{x} = \mathsf{s}^{(m)}|\mathsf{y}) \rangle_{r^{(t)}(\mathsf{y}|\mathsf{x}=\mathsf{s}^{(m)})}, & (3.55) \end{aligned}$$

where we implied the relevant constraints on the optimized functional parameters. We may further combine (3.53) and (3.54) to get the optimal encoder for the next iteration

$$\begin{aligned} p^{(t)}_{\mathsf{y}|\mathsf{x}}(\mathsf{y}|\mathsf{x} = \mathsf{s}^{(m)}) &= q^{(t)}(\mathsf{y}|\mathsf{x} = \mathsf{s}^{(m)}, \tilde{\mathsf{x}} = \mathsf{s}^{(m)}) \in \mathcal{F}_{q_{\mathsf{y}|\mathsf{x}}} \subseteq \mathcal{F}_{p_{\mathsf{y}|\mathsf{x}}} \\ &= \arg\max_{q_{\mathsf{y}|\mathsf{x}}} \langle \log p^{(t-1)}_{\tilde{\mathsf{x}}|\mathsf{y}}(\mathsf{x} = \mathsf{s}^{(m)}|\mathsf{y}) \rangle_{q_{\mathsf{y}|\mathsf{x}}(\mathsf{y}|\mathsf{x}=\mathsf{s}^{(m)})}. & (3.56) \end{aligned}$$

Clearly, the fixed points (3.55) and (3.56) for the autoencoder's decoding and encoding mappings are equivalent to the ones obtained by the iterative maximization of

$$\tilde{I}(\mathsf{x}, \mathsf{y}) = \langle \log p_{\tilde{\mathsf{x}}|\mathsf{y}}(\mathsf{x}|\mathsf{y}) \rangle_{q_{\mathsf{y}|\mathsf{x}}(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x})} + H_{\tilde{p}}(\mathsf{x}), \tag{3.57}$$

where $q_{\mathsf{y}|\mathsf{x}}(\mathsf{y}|\mathsf{x}) \equiv q(\mathsf{y}|\mathsf{x} = \tilde{\mathsf{x}}) \in \mathcal{F}_{q_{\mathsf{y}|\mathsf{x}}}$ for all $\mathsf{x}, \tilde{\mathsf{x}} \in \mathcal{R}_{\mathsf{x}}$. Note that $\tilde{I}(\mathsf{x}, \mathsf{y})$ is the generic lower bound on the mutual information in the stochastic channel $\tilde{\mathcal{M}}_I$, with the variational decoder given by $p_{\tilde{\mathsf{x}}|\mathsf{y}}(\tilde{\mathsf{x}}|\mathsf{y})$. The iterative optimization of (3.57) for $p_{\tilde{\mathsf{x}}|\mathsf{y}}$ and $q_{\mathsf{y}|\mathsf{x}}$ defines the simplest form of the IM algorithm (see Section 2.1.2). Therefore, the variational conditional likelihood training in autoencoders may be viewed as a special case of the variational IM algorithm for the stochastic channel $\tilde{\mathcal{M}}_I \stackrel{\text{def}}{=} \tilde{p}(\mathsf{x}) q_{\mathsf{y}|\mathsf{x}}(\mathsf{y}|\mathsf{x})$ (see Figure 3.6). □

It is intuitive that the tightness of $\tilde{\mathcal{L}}_{\tilde{\mathsf{x}}|\mathsf{x}}$ (and the equivalent objective (3.57)) will strongly depend on the constraints on the family of the variational distributions $\mathcal{F}_{q_{\mathsf{y}|\mathsf{x}}} \subseteq \mathcal{F}_{p_{\mathsf{y}|\mathsf{x}}}$. Importantly, we may note that in the case when the computations of $\langle \log p_{\tilde{\mathsf{x}}|\mathsf{y}}(\mathsf{x}|\mathsf{y}) \rangle_{p_{\mathsf{y}|\mathsf{x}}(\mathsf{y}|\mathsf{x})}$ are tractable, we may be able to choose the family of the

[9]For the reasons of notational clarity, we will also express the optimal encoder $p^{(t-1)}_{\mathsf{y}|\mathsf{x}}(\mathsf{y}|\mathsf{x})$ for the previous iteration of the algorithm.

variational distributions to satisfy $\mathcal{F}_{q_{\mathsf{y}|\mathsf{x}}} = \mathcal{F}_{p_{\mathsf{y}|\mathsf{x}}}$. This more general choice of the variational distributions transforms (3.57) into the equivalent objective

$$\tilde{I}(\mathsf{x},\mathsf{y}) = \langle \log p_{\tilde{\mathsf{x}}|\mathsf{y}}(\mathsf{x}|\mathsf{y}) \rangle_{p_{\mathsf{y}|\mathsf{x}}(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x})} + H_{\tilde{p}}(\mathsf{x}) \tag{3.58}$$

with $p_{\mathsf{y}|\mathsf{x}}(\mathsf{y}|\mathsf{x}) \in \mathcal{F}_{q_{\mathsf{y}|\mathsf{x}}} = \mathcal{F}_{p_{\mathsf{y}|\mathsf{x}}}$. From (2.2) it is clear that (3.58) is in fact the generic lower bound[10] on $I(\mathsf{x},\mathsf{y})$ in the channel model $\mathcal{M}_I = \tilde{p}(\mathsf{x})p_{\mathsf{y}|\mathsf{x}}(\mathsf{y}|\mathsf{x})$, which is effectively the encoding part of the autoencoder $\mathcal{M}_C$. Thus, we may state the *equivalence of the variational conditional likelihood training of stochastic autoencoders to the simplest form of the variational information maximization* in the corresponding stochastic channels ($\mathcal{M}_I$ or $\tilde{\mathcal{M}}_I$, depending on the tractability of computing (3.58)). This agrees with the intuitive argument of Section 3.2.2, where we hypothesized that our generic approximate approach to information maximization should be relatable to learning in stochastic autoencoders.

3.4 Summary

Maximum likelihood learning in generative models $\mathcal{M}_{\mathcal{L}}$ and mutual information maximization in encoder models of communication channels $\mathcal{M}_I$ may be viewed as alternative frameworks for finding (unknown) informative representations $\{\mathsf{y}\}$ of the source patterns $\{\mathsf{x}\}$. These methods are fundamentally different in several respects. One of the most important differences is parameterization of the corresponding models. Generative models are parameterized by specifying the data generating process and the prior on the latent variable representations. While this parameterization may often be useful and relevant (for example, when there are reasons to believe that the higher-dimensional observations $\{\mathsf{x}\}$ are indeed generated from lower-dimensional latent variable representations by applying specific noisy transformations), the generative framework is arguably more difficult to apply when there is a need to impose specific constraints on the posteriors $p(\mathsf{y}|\mathsf{x})$. Indeed, as the posterior $p(\mathsf{y}|\mathsf{x})$ of a generative model will typically be a highly non-linear function of $\mathcal{M}_{\mathcal{L}}$'s parameters, explicit constraints on its moments will typically be difficult to impose. Clearly, this contrasts with the encoder models $\mathcal{M}_I$, where the channel encoder distribution $p(\mathsf{y}|\mathsf{x})$ is a part of the model's specification. In many cases, the choice of the encoder in $\mathcal{M}_I$ may be intuitive (for example, for clustering applications); alternatively, in some cases it may be determined by the environment (for example, for a known type of a neuro-physiological or communication channel).

The second conceptual difference of the likelihood and information maximization methods follows from the definitions of optimality. As the objective of the generative learning in $\mathcal{M}_{\mathcal{L}}$ is quantified as the cross entropy between the model and the empirical distribution (for i.i.d. patterns), a completely unconstrained

[10] Moreover, it is easy to see that up to irrelevant constants, expression (3.58) is just the Jensen's bound on the conditional likelihood $\mathcal{L}_{\tilde{\mathsf{x}}|\mathsf{x}}$. Since for $\mathcal{F}_{q_{\mathsf{y}|\mathsf{x}}} \subseteq \mathcal{F}_{p_{\mathsf{y}|\mathsf{x}}}$ the bound (3.57) will typically be weaker than (3.58), there are no apparent conceptual gains of using the variational extensions of the Jensen's bound on the conditional likelihood for stochastic autoencoders. However, if the optimization is performed numerically, the specifics of the learning may be different.

$\mathcal{L}$-optimal model would mimic the set of the observations. Since this global optimum is clearly of little interest for any practical inference problem, there is a need for an inductive bias, i.e. constraints on the model distribution $p(\mathsf{x})$ which would hopefully be useful for producing sensible generalizations. Latent variable representations of the observed data, as well as parametric or structured specifications of the models, are useful ways of introducing meaningful constraints, though the resulting models are doomed to produce weaker likelihoods than what one would get by using the empirical distribution $\tilde{p}(\mathsf{x})$ as a model. Divergences and known numerical instabilities and degeneracies of likelihood solutions in certain under-constrained models may themselves be intrinsic artifacts of this definition of optimality. Usually, one may introduce additional constraints by applying Bayesian approaches, which optimize type-2 (marginal) likelihoods.

On the other hand, in the information-maximizing framework, optimal mappings to the hidden space are obtained by maximizing the certainty of reconstructing the source vectors from their latent variable representations. In this case, non-observability of some of the variables is not an inductive modeling assumption; it is a consequence of the channel definition where the information-theoretic learning can make sense. (Indeed, in the complete data case, i.e. when both the sources and the codes are visible, the reduction of uncertainty is specified by the empirical distribution). In general, optimization of mutual information tends to lead to other forms of degeneracies. Indeed, it is intuitive that unconstrained encoding distributions of encoder models will tend to be noiseless, and will tend to produce maximally spread-out representations in the code space. However, in many practical situations, the noise of the encoder $p(\mathsf{y}|\mathsf{x})$ is unavoidable and intrinsic to the environment, which in practice may often simplify model specifications.

Our purpose here was to try to understand possible relations between these approaches, with the specific focus on our variational method for information maximization. We explored the general relation of the generic IM algorithm to maximum likelihood learning in generative models and conditional likelihood learning in stochastic chains. In contrast to much of the previous work which relates the likelihood and the mutual information approaches for relatively simple special cases (e.g. Oja (1989), Pearlmutter and Parra (1996), Cardoso (1997), MacKay (1999b)), we tried to relate the methods for the general variational settings independently of the specific model parameterizations. Specifically, we showed that the likelihood of a generative model in $\mathcal{M}_{\mathcal{L}}$ may be viewed as a lower bound on the mutual information in the corresponding model of the noisy channel $\mathcal{M}_I$, where the encoding distribution is the exact posterior of the generative model. Moreover, specific tractable lower bounds $\hat{I}(\mathsf{x}, \mathsf{y})$ optimized by the information-maximizing algorithm are formally tighter than the corresponding likelihoods in $\mathcal{M}_{\mathcal{L}}$. The generally non-constant gap between $\mathcal{L}$ and $\hat{I}(\mathsf{x}, \mathsf{y})$ suggests the fundamental differences between the induced optimization surfaces and the solutions obtained by both approaches. A practical side-effect of this study is an information-theoretic objective for training generative models (which we discuss further in Section 5.2.1).

We also demonstrated a close relation between optimization of the simple variational bound on mutual information (Section 2.1.1) and conditional likelihood training in stochastic autoencoders. Specifically, we showed that the conventional approaches to maximizing the exact conditional likelihood in noiseless autoencoders $\mathsf{x} \to \mathsf{y} \to \tilde{\mathsf{x}}$ may be viewed as special instances of the generic IM algorithm for the corresponding noiseless channels $\mathsf{x} \to \mathsf{y}$. Optimization of the conditional log-likelihood for *stochastic* autoencoders $\mathcal{M}_C$ is a more difficult computational task, since marginalization of the hidden codes may potentially be intractable. Arguably, one of the most straight-forward and rigorous approaches for training such models is by maximizing the variational Jensen's bound on the conditional likelihood, e.g. by applying the variational EM algorithm. Interestingly, we can show that this procedure for training stochastic autoencoders reduces to the simplest form of the variational IM for a specific noisy channel. Specifically, this happens when the IM is applied to maximizing the generic bound on $I(\mathsf{x}, \mathsf{y})$, with the variational decoder defined by the decoding mapping of the conditionally trained stochastic autoencoder $\mathcal{M}_C$, and the encoding distribution defined by $\mathcal{M}_C$'s variational posterior. Thus, the common methods for training the conditional models may in fact be viewed as special cases of the simple variational IM framework (note that optimization of richer bounds on mutual information may potentially be considered (Section 2.3)). Finally, a curious side-product of our exploration of the general properties of the IM algorithm is a tractable model-specific upper bound on the conditional likelihood, which does not ignore the information about the reconstructing distributions (*cf* Fano's inequality).

Chapter 4

Variational Information Maximization for Linear Dimensionality Reduction

In Chapter 2 and Chapter 3 we described the variational approach to information maximization and discussed how it relates to other methods of training probabilistic graphical models. Similarly to other variational methods, the principal idea was to transform the computationally intractable problem of maximizing the exact mutual information to optimizing tractable bounds on the objective. Our specific focus in this chapter is on applying the *generic* and the *auxiliary variational* bounds on mutual information to extracting informative lower-dimensional projections $\{\mathsf{y}\}$ of higher-dimensional data $\{\mathsf{x}\}$. In order to get an insight into analytical properties of the IM, we will discuss the case when the projections are stochastic, with the encoding distribution $p(\mathsf{y}|\mathsf{x})$ being an isotropic linear Gaussian[1]. This case may be formulated from the communication-theoretic viewpoint, where the goal would be to find the compressed representations $\{\mathsf{t}\}$ of the data $\{\mathsf{x}\}$, for the subsequent transmission of these representations over a non-zero noise Gaussian channel. In this formulation, the goal would be to learn the projections $\mathsf{x} \mapsto \mathsf{t}$ in such a way that maximizes the amount of information which the received patterns $\{\mathsf{y}\}$ contain about the *original* sources $\{\mathsf{x}\}$. Obviously, the problem reduces to maximizing mutual information for an isotropic linear Gaussian channel and a generally non-Gaussian distribution of the source patterns.

We will start the discussion by considering optimization of Linsker's *as-if Gaussian* objective (Linsker (1992)), which corresponds to a specific form of our variational formulation for the case of a linear Gaussian variational decoder (see Chapter 2). We show that for the considered channel, optimization of Linsker's *as-if Gaussian* objective criterion cannot improve on PCA projections. Then we show that by considering a richer family of the auxiliary variational bounds (see Section 2.3), we may significantly improve on achievable lower bounds on mutual

[1]Note that despite the apparent similarity to linear autoencoders (Baldi and Hornik (1989), Oja (1989), Roweis and Ghahramani (1999)), our formulation is significantly more general in several respects. Specifically, in our case the encoder $p(\mathsf{y}|\mathsf{x})$ is constrained to be an isotropic linear Gaussian (rather than a noiseless linear projection), and the variational decoder $q(\mathsf{x}|\mathsf{y})$ is an arbitrary distribution in a tractable family (rather than an isotropic linear Gaussian).

information (under the identical encoding constraints). This result is encouraging, as it suggests a simple way to produce tighter lower bounds on the mutual information (compared with PCA) *without* altering the channel specification or increasing the length of the communicated codewords.

Finally, we discuss a simple and practical reformulation of the dimensionality reduction problem, and show that by storing an additional auxiliary variable z (given by the generalized linear projection to the auxiliary space), we may facilitate reconstructions of the sources from noisy lower-dimensional representations without making a recourse to the stored dataset. Effectively, for this case the objective to optimize would be given by $I(\mathsf{x}, \{\mathsf{y}, z\})$, which results in a simplification of the auxiliary variational bound on $I(\mathsf{x}, \mathsf{y})$. Strictly speaking, the resulting reformulation of the optimization problem does not have a direct mapping to a data transmission problem in a simple Gaussian channel; however, it does give rise to an efficient compression and decompression mechanism. Indeed, we show that by a moderate increase in the size of the compressed representations, we may significantly improve on reconstructions from simple *constrained* encodings.

4.1 Introduction

One of the principal goals of dimensionality reduction is to produce a lower-dimensional representation y of a high-dimensional source vector x, so that useful information which the codes y contain about the sources x is maximally preserved. If it is not known a priori which coordinates of x may be relevant for a specific task, it is sensible to maximize the amount of information which y contains about all the coordinates, for all possible source vectors. As discussed in Chapter 2, the fundamental measure of informativeness in this context is the mutual information

$$I(\mathsf{x}, \mathsf{y}) \equiv H(\mathsf{x}) - H(\mathsf{x}|\mathsf{y}),$$

which quantifies the decrease of uncertainty in the pattern x due to the knowledge of y. Again, $H(\mathsf{x}) \equiv -\langle \log p(\mathsf{x}) \rangle_{p(\mathsf{x})}$ and $H(\mathsf{x}|\mathsf{y}) \equiv -\langle \log p(\mathsf{x}|\mathsf{y}) \rangle_{p(\mathsf{x},\mathsf{y})}$ are marginal and conditional entropies respectively, and the angled brackets represent the averages.

A principal motivation for applying information theoretic techniques for learning informative lower-dimensional representations is the general intuition that the lower dimensional codes should preserve useful information about the higher-dimensional data. Moreover, the information maximizing framework of encoder models is particularly convenient for addressing the problem of *constrained* dimensionality reduction. To demonstrate this, suppose that we are interested in learning an optimal undercomplete *orthonormal projection* from the data in the presence of irreducible Gaussian noise (the stochastic subspace selection view). As discussed in Section 3.4, it would generally be difficult to impose specific constraints on the posterior $p(\mathsf{y}|\mathsf{x})$ in the usual generative formulation. For example, if the considered generative model is a factor analyzer $p(\mathsf{y})p(\mathsf{x}|\mathsf{y})$ with $p(\mathsf{y}) \sim \mathcal{N}(0, s\mathsf{I})$, $p(\mathsf{x}|\mathsf{y}) \sim \mathcal{N}(\mathsf{W}\mathsf{y}, \mathbf{\Psi})$, and $\mathbf{\Psi} = \{\psi_{ij}\delta_{ij}\} \neq c\mathsf{I}$, the orthonormal constraints on $\mathsf{A} : \mathsf{x} \to \langle \mathsf{y}|\mathsf{x} \rangle$ would imply orthonormality of $\mathsf{A} \stackrel{\text{def}}{=} (s\mathsf{I} + \mathsf{W}^T\mathbf{\Psi}^{-1}\mathsf{W})^{-1}\mathsf{W}^T\mathbf{\Psi}^{-1}$

(e.g. von Mises (1964), Bartholomew (1987)). Clearly, by explicitly enforcing (e.g. Arfken (1985), Riley et al. (2002)) the orthonormality constraint, we may significantly complicate the resulting optimization surface. On the other hand, a natural way to address the problem of constrained dimensionality reduction is by considering the information-maximizing paradigm for an encoder model $\tilde{p}(\mathsf{x})p(\mathsf{y}|\mathsf{x})$. In this case we could easily impose the orthonormality constraints (or in fact any other requirements the noisy projections $\mathsf{x} \rightarrow \mathsf{y}$ need to satisfy) by explicitly parameterizing the stochastic mapping $p(\mathsf{y}|\mathsf{x})$. For example, for this specific case we could set $p(\mathsf{y}|\mathsf{x}) \sim \mathcal{N}(\mathsf{W}\mathsf{x}, \mathbf{\Sigma}_{\mathsf{y}|\mathsf{x}})$, where $\mathsf{W}\mathsf{W}^T = \mathsf{I}_{|\mathsf{y}|}$. Effectively, this parameterization would be analogous to specifying the conditionals of the discriminative models; however, in contrast to discriminative models which presume observability of the outputs, the lower-dimensional vectors $\{\mathsf{y}\}$ will in our case be hidden. Our focus in this chapter is on applying the variational information maximizing framework to dimensionality reduction, with specific focus on the linear case.

4.1.1 Optimization of Linsker's Criterion

The principled information theoretic approach to dimensionality reduction would maximize the exact mutual information $I(\mathsf{x}, \mathsf{y})$ with respect to parameters of the encoder $p(\mathsf{y}|\mathsf{x})$. Despite the fact that the dimensionality of the reduced space $|\mathsf{y}|$ will be lower than the dimensionality of the original space, the exact evaluation of $I(\mathsf{x}, \mathsf{y})$ will generally be computationally intractable if $|\mathsf{y}| < |\mathsf{x}|$ is still large. As we mentioned in Section 1.4, the key difficulty lies in the computation of the entropic term $H(\mathsf{x}|\mathsf{y})$, which is tractable only in a few special cases. A computationally tractable alternative to maximizing $I(\mathsf{x}, \mathsf{y})$ is the IM applied to maximizing proper lower bounds on the mutual information.

We will start the discussion by considering optimization of the simple generic lower bound on the mutual information

$$I(\mathsf{x}, \mathsf{y}) \geq \tilde{I}(\mathsf{x}, \mathsf{y}) \stackrel{\text{def}}{=} H(\mathsf{x}) + \langle \log q(\mathsf{x}|\mathsf{y}) \rangle_{p(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x})}, \tag{4.1}$$

where $\tilde{p}(\mathsf{x})$ is the empirical distribution, and the variational approximation $q(\mathsf{x}|\mathsf{y})$ of the exact posterior $p(\mathsf{x}|\mathsf{y})$ is given by a linear Gaussian $q(\mathsf{x}|\mathsf{y}) \sim \mathcal{N}(\mathsf{U}\mathsf{y}, \mathbf{\Sigma})$. As we showed in Section 1.4, optimization of the bound (2.2) for this specific choice of the variational decoder reduces to maximization of Linsker's *as-if Gaussian* criterion

$$2I_G(\mathsf{x}, \mathsf{y}) = \log |\mathbf{\Sigma}_{xx}| - \log |\mathbf{\Sigma}_{xx} - \mathbf{\Sigma}_{xy}\mathbf{\Sigma}_{yy}^{-1}\mathbf{\Sigma}_{xy}^T| + c, \tag{4.2}$$

where $\mathbf{\Sigma}_{xx}$, $\mathbf{\Sigma}_{xy}$, and $\mathbf{\Sigma}_{yy}$ are the partitions of decoder covariance $\mathbf{\Sigma}$, and c is an irrelevant constant. In other words, maximization of Linsker's criterion I_G may be seen as a special case of the variational information-maximization formulation for linear Gaussian decoders, *independently* of the specific encoder parameterization. Clearly, the objective (4.2) may be expressed as a function of the encoder parameters. After training, the learned parameters may be used for producing lower-dimensional representations y for the given sources x by forward-sampling from the encoder $p(\mathsf{y}|\mathsf{x})$ (or computing the conditional mean $\langle \mathsf{y}|\mathsf{x} \rangle_{p(\mathsf{y}|\mathsf{x})}$).

4.1.1.1 Nature of Optimal Solutions

In the considered special case of a linear Gaussian channel, the encoder is given by $p(\mathsf{y}|\mathsf{x}) \sim \mathcal{N}(\mathsf{W}\mathsf{x}, \mathbf{\Sigma}_{\mathsf{y}|\mathsf{x}})$, and $p(\mathsf{x})$ is the empirical distribution of the training patterns. It is now easy to show that the left singular vectors of the optimal projection weights W^T give rise to the $|\mathsf{y}|$-PCA solution on the sample covariance $\mathsf{S} \stackrel{\text{def}}{=} \langle \mathsf{x}\mathsf{x}^T \rangle$ (for clarity, we assume that the data is centered). The remaining moments may be trivially expressed as

$$\begin{aligned} \langle \mathsf{x}\mathsf{y}^T \rangle &= \mathsf{S}\mathsf{W}^T, & (4.3) \\ \langle \mathsf{y}\mathsf{y}^T \rangle &= \mathbf{\Sigma}_{\mathsf{y}|\mathsf{x}} + \mathsf{W}\mathsf{S}\mathsf{W}^T. & (4.4) \end{aligned}$$

Then by substituting into (4.2), we get

$$\begin{aligned} \tilde{I}(\mathsf{x}, \mathsf{y}) &= \log |\mathsf{S} - \mathsf{S}\mathsf{W}^T(\mathsf{W}\mathsf{S}\mathsf{W}^T + \mathbf{\Sigma}_{\mathsf{y}|\mathsf{x}})^{-1}\mathsf{W}\mathsf{S}|^{-1} + c \\ &= \log |\mathsf{S}^{-1} + \mathsf{W}^T \mathbf{\Sigma}_{\mathsf{y}|\mathsf{x}} \mathsf{W}| + c, \end{aligned} \tag{4.5}$$

where we used the matrix inversion lemma (e.g. Press et al. (1992)) and let c be a constant.

To explore spectral properties of the optimal solutions, we will now consider the singular value decomposition of the weights, i.e. $\mathsf{W}^T = \mathsf{V}\mathsf{L}\mathsf{R}^T \in \mathbb{R}^{|\mathsf{x}| \times |\mathsf{y}|}$. Here $\mathsf{V} \in \mathbb{R}^{|\mathsf{x}| \times |\mathsf{y}|}$ is a matrix of orthonormal columns, i.e. $\mathsf{V}^T\mathsf{V} = \mathsf{I}_{|\mathsf{y}|}$, $\mathsf{L} \in \mathbb{R}^{|\mathsf{y}| \times |\mathsf{y}|}$ is a diagonal, and $\mathsf{R} \in \mathbb{R}^{|\mathsf{y}| \times |\mathsf{y}|}$ is a rigid rotation matrix. By assuming that the irreducible channel noise is white (i.e. $\mathbf{\Sigma}_{\mathsf{y}|\mathsf{x}} = \sigma^2\mathsf{I}$), we may transform the objective (4.5) into

$$\tilde{I}(\mathsf{x}, \mathsf{y}) = \log |\mathsf{S}^{-1} + \sigma^{-2}\mathsf{V}\mathsf{L}^2\mathsf{V}^T| + c. \tag{4.6}$$

Clearly, the orthonormality constraint on the projection weights $\mathsf{W}\mathsf{W}^T = \mathsf{I}_{|\mathsf{y}|}$ implies $\mathsf{L} = \mathsf{I}$ (which also ensures convergence of the objective (4.6)). This leads to

$$\tilde{I}(\mathsf{x}, \mathsf{y}) = \log |\mathsf{S}^{-1} + \sigma^{-2}\mathsf{V}\mathsf{V}^T| - \operatorname{tr}\left\{\mathsf{M}(\mathsf{V}^T\mathsf{V} - \mathsf{I})\right\}. \tag{4.7}$$

A straight-forward matrix optimization for $\mathsf{V} \in \mathbb{R}^{|\mathsf{x}| \times |\mathsf{y}|}$ leads to

$$(\mathsf{S}^{-1} + \sigma^{-2}\mathsf{V}\mathsf{V}^T)\mathsf{V} = \mathsf{V}\mathsf{M}^{-1}\sigma^{-2}, \tag{4.8}$$

i.e.

$$\mathsf{S}^{-1}\mathsf{V} = \mathsf{V}\tilde{\mathsf{M}}, \quad \text{where} \quad \tilde{\mathsf{M}} \stackrel{\text{def}}{=} (\mathsf{M}^{-1} - \mathsf{I})\sigma^{-2}, \tag{4.9}$$

where $\mathsf{M} \in \mathbb{R}^{|\mathsf{y}| \times |\mathsf{y}|}$ is a matrix of Lagrange multipliers. Without loss of generality, we may assume that $\tilde{\mathsf{M}}$ is symmetric (since $\operatorname{tr}\{\mathsf{M}\} = \operatorname{tr}\{\mathsf{M}^T\}$), i.e. the weights V satisfying the extremum criterion (4.9) correspond to eigenvectors of S^{-1} and their rigid rotations. Since the objective (4.6) is not influenced by the rotation factor, we may as well ignore them in our analysis.

Finally, by substituting (4.9) into (4.6) and expressing the objective in terms of the eigenspectra of the sample covariance S, it is easy to see that the optimal weights V correspond to *principal* components of S. Indeed, (4.9) implies that the objective (4.6) may be expressed as

$$\tilde{I}(\mathsf{x}, \mathsf{y}) = \log \left| \mathsf{V}(\Lambda_S^{-1} + \sigma^{-2}\mathsf{I})\mathsf{V}^T + \tilde{\mathsf{V}}\tilde{\Lambda}_S^{-1}\tilde{\mathsf{V}}^T \right| + c. \tag{4.10}$$

Here $\tilde{\mathsf{V}} \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{x}|-|\mathsf{y}|}$ are the eigenvectors of S which are orthonormal to the space spanned by $\mathsf{V} \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{y}|}$, i.e.

$$\mathsf{V}\mathsf{V}^T + \tilde{\mathsf{V}}\tilde{\mathsf{V}}^T = \mathsf{I}_{|\mathsf{x}|}, \tag{4.11}$$

and $\Lambda_S \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{y}|}$, $\tilde{\Lambda}_S \in \mathbb{R}^{|\mathsf{x}|-|\mathsf{y}|\times|\mathsf{x}|-|\mathsf{y}|}$ are the eigenvalues of S corresponding to V and $\tilde{\mathsf{V}}$ respectively. From (4.9) it is clear that $\tilde{\Lambda}_S$ and $\tilde{\mathsf{V}}$ define the spectrum discarded by the projection weights W. The new objective (4.10) may be equivalently expressed as

$$\tilde{I}(\mathsf{x},\mathsf{y}) = \sum_{i\in\mathcal{R}} \log(\lambda_i^{-1} + \sigma^{-2}) + \sum_{j\notin\mathcal{R}} \log \lambda_j^{-1} + c, \tag{4.12}$$

where λ_i is the i^{th} eigenvalue of the sample covariance, and $\mathcal{R}$ specifies the spectrum retained in W. Equivalently, we may get

$$\begin{aligned}\tilde{I}(\mathsf{x},\mathsf{y}) &= \sum_{i\in\mathcal{R}} \log(\lambda_i^{-1} + \sigma^{-2}) - \sum_{j\notin\mathcal{R}} \log \lambda_j + c \\ &= \sum_{i\in\mathcal{R}} \log(\lambda_i + \sigma^2) - |\mathsf{y}| \log \sigma^2 - \log|\mathsf{S}| + c. \end{aligned} \tag{4.13}$$

From (4.13) it is clear that the bound $\tilde{I}(\mathsf{x},\mathsf{y})$ is maximized when the spectrum $\mathcal{R}$ retained in the projection weights W indeed corresponds to the principal components of the sample covariance. This may be further related to the special case of noiseless linear autoencoders discussed in the early work of Baldi and Hornik (1989) and Oja (1989), which may be obtained from (4.13) by computing the limit at $\sigma^2 \to 0$ and using the result of proposition 3.4.

4.2 Optimization of the Auxiliary Variational Bound

Importantly, the result of Section 4.1.1 suggests that optimization of Linsker's criterion for isotropic linear Gaussian channels cannot improve on the simple $|\mathsf{y}|$-PCA projections. Since optimization of Linkser's criterion may be viewed as a special instance of the IM algorithm with linear Gaussian variational decoders, a natural question to explore is how the linear projections could be affected by using more complex decoder types. A principal conceptual difficulty of applying the bound (2.2) is in specifying a powerful yet tractable variational decoder $q(\mathsf{x}|\mathsf{y})$. Here we consider a richer family of tractable auxiliary variational bounds on $I(\mathsf{x},\mathsf{y})$ (see Section 2.3), which may overcome the fundamental limitations of Linsker's criterion.

The key idea of the auxiliary variational method (Agakov and Barber (2005a)) is to introduce mappings to the auxiliary space $\{\mathsf{z}\}$ in a way which does not affect the original channel $p(\mathsf{y}|\mathsf{x})$, and learn the resulting joint distribution $p(\mathsf{x},\mathsf{y},\mathsf{z}) = p(\mathsf{x},\mathsf{y})p(\mathsf{z}|\mathsf{x},\mathsf{y})$ of the sources x, encodings y, and auxiliary ("feature") variables z (see Section 2.3 for a detailed discussion). By applying straight-forward algebraic manipulations, we may obtain a tractable lower bound on $I(\mathsf{x},\mathsf{y})$

$$I(\mathsf{y},\mathsf{x}) \geq \tilde{I}(\mathsf{y},\mathsf{x}) \stackrel{\text{def}}{=} H(\mathsf{x}) + H(\mathsf{z}|\mathsf{x},\mathsf{y}) + \langle \log q(\mathsf{x}|\mathsf{y},\mathsf{z})\rangle_{p(\mathsf{x},\mathsf{y},\mathsf{z})} + \langle \log q(\mathsf{z}|\mathsf{y})\rangle_{p(\mathsf{y},\mathsf{z})} \tag{4.14}$$

The variational distributions $q(\mathsf{x}|\mathsf{y},\mathsf{z})$ and $q(\mathsf{z}|\mathsf{y})$, as well as the auxiliary conditional $p(\mathsf{z}|\mathsf{x},\mathsf{y})$ are chosen to ensure tractability of the objective (4.14). Then it is tractable to optimize $\tilde{I}(\mathsf{x},\mathsf{y})$ for the channel encoder, variational decoder, and the auxiliary conditional distributions. Effectively, we will still be learning an optimal constrained encoder ($p(\mathsf{y}|\mathsf{x}) \sim \mathcal{N}(\mathsf{W}\mathsf{x},\mathbf{\Sigma})$ in the special case we consider here), but for a richer family of variational distributions. Note that if the auxiliary space is discrete (for example, if a single auxiliary variable z is multinomial), the variational decoder $q(\mathsf{x}|\mathsf{y}) = \langle q(\mathsf{x}|\mathsf{y},z)\rangle_{q(z|\mathsf{y})}$ is indeed defined as a multi-modal distribution.

4.2.1 Representations

Here we discuss a tractable choice of the variational parameters of $\tilde{I}(\mathsf{x},\mathsf{y})$ for a linear Gaussian channel $p(\mathsf{y}|\mathsf{x})$ with constrained projection weights. In the model which we consider, the auxiliary space is given by the multinomial variable z which takes one of $|z|$ possible states $\{z_1,\ldots,z_{|z|}\}$. The encoder and the auxiliary conditional distributions are given by $p(\mathsf{y}|\mathsf{x}) \sim \mathcal{N}_{\mathsf{y}}(\mathsf{W}\mathsf{x};\mathbf{\Sigma})$ and

$$p(z_j|\mathsf{x},\mathsf{y}) = p(z_j|\mathsf{x}) \propto \exp\{-(\mathsf{v}_j^T\mathsf{x} + b_j)\} \tag{4.15}$$

respectively. Here $\mathsf{W} \in \mathbb{R}_C^{|\mathsf{y}|\times|\mathsf{x}|}$, $\mathsf{V} = \{\mathsf{v}_1,\ldots,\mathsf{v}_{|z|}\} \in \mathbb{R}^{|\mathsf{x}|\times|z|}$, and $\mathsf{b} \in \mathbb{R}^{|z|}$ are weights and biases to be learned, where we have assumed that $\mathbb{R}_C^{|\mathsf{y}|\times|\mathsf{x}|} \subseteq \mathbb{R}^{|\mathsf{y}|\times|\mathsf{x}|}$ is a subspace of encoder weights satisfying specific constraints (with $\mathbb{R}_C^{|\mathsf{y}|\times|\mathsf{x}|} \equiv \mathbb{R}^{|\mathsf{y}|\times|\mathsf{x}|}$ if the weights are unconstrained). For computational convenience, we assumed that the auxiliary variables z are conditionally independent of the codes y, i.e. $I(\mathsf{z},\{\mathsf{x},\mathsf{y}\}) = I(\mathsf{z},\mathsf{x})$. Also, during learning we constrained the variational distribution used for reconstruction of the data x and auxiliary variables z to satisfy

$$q(\mathsf{x},z|\mathsf{y}) \propto q(\mathsf{x}|\mathsf{y},z)q(z), \tag{4.16}$$

which facilitates computations of the integrals in $\tilde{I}(\mathsf{x},\mathsf{y})$.

Clearly, for the considered parameterization the objective (4.14) is transformed into

$$I(\mathsf{x},\mathsf{y}) \geq H(z,\mathsf{x}) + \langle \log q(z|\mathsf{y})\rangle_{p(z,\mathsf{y})} + \langle \log q(\mathsf{x}|\mathsf{y},z)\rangle_{p(\mathsf{x},\mathsf{y},z)}. \tag{4.17}$$

Note that for an arbitrary mixture decoder $q(\mathsf{x}|\mathsf{y})$, the computational complexity of evaluating the first two terms in the bound is linear in the number of states $|z|$. In general, computation of $\langle \log q(\mathsf{x}|\mathsf{y},z)\rangle_{p(\mathsf{x},\mathsf{y},z)}$ is more problematic, since it requires averaging of a non-factorized function of the codes over the channel distribution $p(\mathsf{y}|\mathsf{x})$. For the special case when each component is a Gaussian with a constrained mean $q(\mathsf{x}|\mathsf{y},z_j) \sim \mathcal{N}_{\mathsf{x}}(\mathsf{U}_j\mathsf{y},\mathsf{S}_j)$, the rightmost term in (4.17) is expressed as

$$\begin{aligned}\langle \log q(\mathsf{x}|\mathsf{y},z)\rangle_{p(\mathsf{x},\mathsf{y},z)} &= -\frac{1}{2M}\sum_{j=1}^{|z|}\sum_{i=1}^{M} p(z_j|\mathsf{x}^{(i)})\mathrm{tr}\left\{\mathsf{S}_j^{-1}\left(\mathsf{d}_j^{(i)}\mathsf{d}_j^{(i)T} + \mathsf{U}_j\mathbf{\Sigma}\mathsf{U}_j^T\right)\right\} \\ &\quad -\frac{1}{2M}\sum_{j=1}^{|z|}\log|\mathsf{S}_j|\sum_{i=1}^{M} p(z_j|\mathsf{x}^{(i)}).\end{aligned} \tag{4.18}$$

Here we ignored the irrelevant constants and defined

$$\mathsf{d}_j^{(i)} \stackrel{\text{def}}{=} \mathsf{x}^{(i)} - \mathsf{U}_j \mathsf{W} \mathsf{x}^{(i)} \in \mathbb{R}^{|\mathsf{x}|} \tag{4.19}$$

to be the distortion between the i^{th} pattern and its reconstruction from a noiseless code at the mean of $q(\mathsf{x}|\mathsf{y}, z_j)$. From (4.17) and (4.18) it is easy to see that small values of the distortion terms $\mathsf{d}_j^{(i)}$ lead to improvements in the bound on the mutual information. This agrees with the intuition that the trained model should favour accurate reconstructions of the source patterns from their compressed representations passed through a noisy channel.

4.2.2 Learning Optimal Parameters

From (4.17) it is clear that the optimal settings of the auxiliary conditionals are given by

$$q(z|\mathsf{y}) = p(z|\mathsf{y}) \propto \sum_{i=1}^{M} p(z|\mathsf{x}^{(i)}) p(\mathsf{y}|\mathsf{x}^{(i)}). \tag{4.20}$$

It is now possible to derive an iterative learning rule for the parameters of $q(\mathsf{x}|\mathsf{y}, z)$, $p(\mathsf{y}|\mathsf{x})$ and $p(z|\mathsf{x})$. These results are obtained by computing matrix derivatives of $\tilde{I}(\mathsf{x}, \mathsf{y})$ and (where possible) deriving the closed-form fixed-point updates. The updates are performed iteratively, assuming parameter independence at each iteration. We will state and briefly discuss the results, omitting the straight-forward derivations.

Optimal decoder: It is easy to see that the considered variational decoder defines a constrained mixture of Gaussians $q(\mathsf{x}|\mathsf{y}) = \langle q(\mathsf{x}|\mathsf{y}, z) \rangle_{q(z|\mathsf{y})}$, where $q(\mathsf{x}|\mathsf{y}, z_j) \sim \mathcal{N}_{\mathsf{x}}(\mathsf{U}_j \mathsf{y}, \mathsf{S}_j)$. For each component j, the optimal weights U_j parameterizing the component's mean are given by

$$\mathsf{U}_j^{(new)} = \left(\sum_{i=1}^{M} p(z_j|\mathsf{x}^{(i)}) \mathsf{x}^{(i)} \mathsf{x}^{(i)T} \mathsf{W}^T \right) \left(\sum_{i=1}^{M} p(z_j|\mathsf{x}^{(i)}) \left(\mathsf{W} \mathsf{x}^{(i)} \mathsf{x}^{(i)T} \mathsf{W}^T + \mathbf{\Sigma} \right) \right)^{-1} \tag{4.21}$$

where we have assumed that at the current iteration of the algorithm the encoder weights W are fixed. Similarly, the update for the components' covariances is given by

$$\mathsf{S}_j^{(new)} = \sum_{i=1}^{M} p(z_j|\mathsf{x}^{(i)}) \left(\mathsf{d}_j^{(i)} \mathsf{d}_j^{(i)T} + \mathsf{U}_j^{(new)} \mathbf{\Sigma} \left(\mathsf{U}_j^{(new)} \right)^T \right) \frac{1}{\sum_{i=1}^{M} p(z_j|\mathsf{x}^{(i)})}. \tag{4.22}$$

Here $\mathbf{\Sigma}$ is the covariance of the channel noise (which is presumed to be fixed and independent of x and y), and the distortion $\mathsf{d}_j^{(i)} \in \mathbb{R}^{|\mathsf{x}|}$ is given by (4.19) computed for the new decoder weights $\mathsf{U}_j^{(new)}$. Perhaps not very surprisingly, the fixed point equations (4.21) and (4.22) resemble maximum-likelihood updates for mixtures of constrained Gaussian distributions. The constraints are influenced by the specific choice of encoder and decoder distributions, as well as the mapping to the auxiliary space. Note that for noisy channels and non-singular decoder weights U_j,

computation of the inverses in (4.21) should *not* be numerically unstable. However, in order to ensure numerical stability of optimization for the case of small datasets or near-singular decoder weights U_j, it may be practical to increment S_j by a multiple of the identity matrix with a small positive scaling factor (which would correspond to learning the constrained covariances $\mathsf{S}_j + \epsilon \mathsf{I}_{|\mathsf{y}|}$ for $\epsilon > 0$ and $j = 1, \ldots, |z|$).

Optimal auxiliary mappings:
One way to learn the optimal auxiliary conditional $p(z|\mathsf{x})$ is by performing numerical ascent on $\tilde{I}(\mathsf{x}, \mathsf{y})$ with respect to the parameters of the conditional. The gradients are computed from (4.17) as

$$\frac{\partial \tilde{I}}{\partial \mathsf{v}_j} = \frac{1}{M} \sum_{i=1}^{M} \mathsf{x}^{(i)} p(z_j|\mathsf{x}^{(i)}) \left(\langle e(z, \mathsf{x}^{(i)}) \rangle_{p(z|\mathsf{x}^{(i)})} - e(z_j, \mathsf{x}^{(i)}) \right), \tag{4.23}$$

where

$$e(z_j, \mathsf{x}^{(i)}) \stackrel{\text{def}}{=} \tilde{q}_{ji} + \log q(z_j) - \left(1 + \log p(z_j|\mathsf{x}^{(i)})\right), \tag{4.24}$$

and $\tilde{q}_{ji} \stackrel{\text{def}}{=} \langle \log q(\mathsf{x}^{(i)}|\mathsf{y}, z_j) \rangle_{p(\mathsf{y}|\mathsf{x}^{(i)})}$ is given by expression (4.18) computed for the new settings of the decoder parameters U_j and S_j. Again, z_j is the state of the auxiliary variable z, and i is an index of a training pattern. The gradients for the biases $\partial \tilde{I} / \partial b_j$ have a form similar to (4.23), with an omitted pre-multiplication by $\mathsf{x}^{(i)}$ in the summation. Note that throughout the iterations for $\mathsf{V} \in \mathbb{R}^{|\mathsf{x}| \times |z|}$ and $\mathsf{b} \in \mathbb{R}^{|z|}$, the posterior $p(z|\mathsf{x})$ and the average $\tilde{q}_{ji}$ may be kept fixed. For this case, evaluation of the gradient (4.23) is computationally efficient, as the complexity of computing the averages is just $O(|z|M)$. Having computed the updated parameters of the auxiliary conditional, we can use (4.15) to obtain $p^{(new)}(z_j|\mathsf{x}^{(i)})$ for all $j = \{1, \ldots, |y|\}$ and $i = \{1, \ldots, M\}$.

Optimal encoder:
Throughout the learning, we assumed that the covariance $\mathbf{\Sigma}$ of the Gaussian encoder was fixed (which corresponds to a fixed channel noise distribution). In this case, optimization for the encoder $p(\mathsf{y}|\mathsf{x})$ reduces to learning the projection weights W. From (4.17), it is easy to get

$$\frac{\partial \tilde{I}}{\partial \mathsf{W}} = \frac{1}{M} \sum_{j=1}^{|z|} \sum_{i=1}^{M} \left[\left(\mathsf{U}_j^{(new)} \right)^T \left(\mathsf{S}_j^{(new)} \right)^{-1} \left(\mathsf{I}_{|\mathsf{x}|} - \mathsf{U}_j^{(new)} \mathsf{W} \right) \right] \mathsf{x}^{(i)} \mathsf{x}^{(i)T} p^{(new)}(z_j|\mathsf{x}^{(i)}), \tag{4.25}$$

where we *implied* that the weights $\mathsf{W} \in \mathbb{R}_C^{|\mathsf{y}| \times |\mathsf{x}|} \subseteq \mathbb{R}^{|\mathsf{y}| \times |\mathsf{x}|}$ lie in the matrix space $\mathbb{R}_C^{|\mathsf{y}| \times |\mathsf{x}|}$ satisfying any of the required constraints on the encoding weights. In general, if the number of mixture components $|z| > 1$, it is not easy to find a closed form expression for W. Instead, we may perform numerical optimization of the bound $\tilde{I}(\mathsf{x}, \mathsf{y})$ for W (subject to the constraints).

Generally, as we mentioned in Section 4.1, the encoder formulation facilitates the choice of constraints on the encoder distribution $p(\mathsf{y}|\mathsf{x})$. For example, in our case it is easy to incorporate norm constraints on W into the objective function

(4.17). In the simplest case when the constraints are soft, (i.e. when the Lagrange multipliers are fixed), this would result in the regularization penalty $-\mathsf{DW}$ on the gradient (4.25), where $\mathsf{D} = \{D_{ij}\delta_{ij}|D_{ii} \geq 0\} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{y}|}$ is a fixed diagonal matrix. Alternatively, we can impose hard constraints on the encoder weights by considering a specific construction in a way which always results in the desired singular spectrum. For example, we may impose the othonormality constraints on the rows of W by parameterizig the weights as

$$\mathsf{W} = (\tilde{\mathsf{W}}\tilde{\mathsf{W}}^T)^{-1/2}\tilde{\mathsf{W}}, \tag{4.26}$$

where $\tilde{\mathsf{W}} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{x}|}$ is an arbitrary rank-$|\mathsf{y}|$ real-valued matrix. Clearly, (4.26) implies the desired orthonormality constraint on the projection weights, as $\mathsf{WW}^T = \mathsf{I}_{|\mathsf{y}|}$. Other kinds of constraints on the encoder parameters may potentially be considered. For example, if $\mathsf{W} = \{w_{ij}\} \in \mathbb{R}_C^{|\mathsf{y}|\times|\mathsf{x}|}$ is constrained to a hyper-cube (*cf* expression (4.26)), we may parameterize each weight component as $w_{ij} = f(\tilde{w}_{ij})$, where $f : \mathbb{R} \mapsto [-\omega, \omega]$ defines a mapping from the real space to a closed line segment. Learning the encoder parameters $\mathsf{W} \in \mathbb{R}_C^{|\mathsf{y}|\times|\mathsf{x}|}$ would then involve unconstrained optimization of $\tilde{I}(\mathsf{x}, \mathsf{y})$ for $\tilde{\mathsf{W}} = \{\tilde{w}_{ij}\} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{x}|}$, with the corresponding gradients obtained from (4.14) and (4.25) by the chain rule. Note that while the mean of the considered encoder $p(\mathsf{y}|\mathsf{x})$ is linear in the constrained encoder parameters $\mathsf{W} \in \mathbb{R}_C^{|\mathsf{y}|\times|\mathsf{x}|}$, it is generally nonlinear in $\tilde{\mathsf{W}} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{x}|}$.

Finally, we note that at the end of each iteration of the optimization procedure, it is necessary to re-compute the moments $\tilde{q}_{ji}$ and the prior on the auxiliary variables $q(z)$ according to (4.18) and (4.20) respectively. The iterations of (4.20) – (4.25) continue until convergence or meeting a termination criterion.

4.2.2.1 Role of the Auxiliary Variables

By analogy with Section 2.3, we may interpret the information maximization framework from the communication-theoretic viewpoint. As the objective function which we optimize is the lower bound on $I(\mathsf{x}, \mathsf{y})$ (bounding the channel capacity in the stochastic communication channel $\mathsf{x} \rightarrow \mathsf{y}$), we presume that the auxiliary variables z are not transmitted across the channel. Their purpose in this context is to define a richer family of variational bounds on the true mutual information in $\mathsf{x} \rightarrow \mathsf{y}$. This auxiliary variational family of lower bounds includes the generic bound (4.1) as a special case. (Indeed, we can easily see that by allowing a flexibility in the mappings to the auxiliary space, we may obtain the bounds which are at least as tight as (4.1). Generally, the simple bound (4.1) arises as a special case of (4.14) when z takes a single state). The role of the auxiliary variables z in this context is to capture regularities in the training data and introduce additional dependencies and multi-modality to the structure of the variational decoder.

4.2.3 Learning Optimal Representations in the Augmented $\{\mathsf{y}, z\}$-space

Now suppose that the auxiliary variables z are actually observable at the receiver's end of the channel. Under this assumption, we may consider optimizing an alternative bound $\tilde{I}_H(\mathsf{x}, \{\mathsf{y}, \mathsf{z}\}) \geq I(\mathsf{x}, \mathsf{y})$, defined by analogy with (2.2). (We will use the index I_H to indicate that the channel $\mathsf{x} \to \{\mathsf{y}, \mathsf{z}\}$ is generally heterogeneous; for example, z may be a vector of generally unknown class labels, while $\mathsf{y} \in \mathbb{R}^{|\mathsf{y}|}$ may define a lower-dimensional projection). It will lead to a slight simplification of (4.14), which effectively reduces to

$$\tilde{I}_H(\mathsf{x}, \{\mathsf{y}, \mathsf{z}\}) = H(\mathsf{x}) + \langle \log q(\mathsf{x}|\mathsf{y}, \mathsf{z}) \rangle_{\tilde{p}(\mathsf{x})p(\mathsf{y}|\mathsf{x})p(\mathsf{z}|\mathsf{x})}, \tag{4.27}$$

where the cross-entropic term is given by (4.18). For our specific case, this leads to a change in the updates for parameters of the auxiliary conditional $p(z|\mathsf{x})$, which in this case leads to

$$\frac{\partial \tilde{I}}{\partial \mathsf{v}_j} = \frac{1}{M} \sum_{i=1}^{M} \mathsf{x}^{(i)} p(z_j|\mathsf{x}^{(i)}) \left(\langle \log q(\mathsf{x}^{(i)}|\mathsf{y}, z) \rangle_{p(\mathsf{y},z|\mathsf{x}^{(i)})} - \langle \log q(\mathsf{x}^{(i)}|\mathsf{y}, z_j) \rangle_{p(\mathsf{y}|\mathsf{x}^{(i)})} \right) \tag{4.28}$$

(again, the gradients $\partial \tilde{I}/\partial b_j$ are given by expression (4.28) without the $\mathsf{x}^{(i)}$ factor inside the summation). Clearly, the averages in (4.28) are easy to compute, as both $q(\mathsf{x}^{(i)}|\mathsf{y}, z_j)$ and $p(\mathsf{y}|\mathsf{x}^{(i)})$ in this case are Gaussians. The updates (4.21), (4.22), and (4.25) for the remaining parameters are not affected by the change in the channel definition.

Note that in the communication-theoretic interpretation of the considered case, the auxiliary variables z will need to be communicated over the channel (generally, at a small increase in the communication cost, which in this case is of the order of $|\mathsf{y}| + |z|$). For the model parameterization described in Section 4.2.2, this would correspond to sending an additional natural number z, which would effectively index the decoder used at the reconstruction. Generally, the compressed representations of $\{\mathsf{x}\}$ will include not only the codes $\{\mathsf{y}\}$, but also the auxiliary labels z. Finally, we may note that unless $p(z|\mathsf{x})$ is strongly constrained, the mapping $\mathsf{x} \to z$ will typically tend to be nearly noiseless, as this would decrease $H(z|\mathsf{x})$ and maximize $I(\mathsf{x}, \{\mathsf{y}, z\})$.

4.2.3.1 Comparison with Mixtures of Probabilistic Principal Component Analyzers

It is interesting to compare the IM framework with the likelihood-based training for mixtures of latent variable models (where the auxiliary variables z are the mixture components, and y are the latent variable representations of the data patterns). For such models the likelihood may be expressed as

$$\mathcal{L} \stackrel{\text{def}}{=} \left\langle \log \langle q(\mathsf{x}|\mathsf{y}, z) \rangle_{q(\mathsf{y})q(z)} \right\rangle_{\tilde{p}(\mathsf{x})}, \tag{4.29}$$

where $\tilde{p}(\mathsf{x})$ is the empirical distribution. Effectively, $q(\mathsf{x})$ defines a mixture of latent variable models. While the objective (4.29) is generally different from

$\tilde{I}_H(\mathsf{x}, \{\mathsf{y}, \mathsf{z}\})$ (see expression (4.27)), for several special cases we may observe a close relation between maximizing (4.29) in a mixture model and learning the optimal variational *decoders* in the context of variational information maximization.

In particular, for the special case of learning linear projections of the data patterns to a lower-dimensional space, we can compare the I-step of the IM algorithm in our formulation with the M-step of the EM algorithm applied to training mixtures of probabilistic PCA models $\mathcal{M}_L \stackrel{\text{def}}{=} q(\mathsf{y})q(z)q(\mathsf{x}|\mathsf{y}, z)$, where $q(z)$ is a multinomial distribution, $q(\mathsf{y}) \sim \mathcal{N}_{\mathsf{y}}(0, \mathbf{1})$, and $q(\mathsf{x}|\mathsf{y}, z_j) \sim \mathcal{N}_{\mathsf{x}}(\mathsf{U}_j\mathsf{y}, s_j^2)$ (Tipping and Bishop (1999a), Tipping and Bishop (1999b)). For mixtures of PPCAs, the updates for parameters of the mixture components may be expressed as

$$\mathsf{U}_j^{(new)} = \left(\sum_{i=1}^{M} p(z_j|\mathsf{x}^{(i)})\mathsf{x}^{(i)}\langle\mathsf{y}^T\rangle_{q_j^{(i)}}\right)\left(\sum_{i=1}^{M} p(z_j|\mathsf{x}^{(i)})\left(\langle\mathsf{y}\rangle_{q_j^{(i)}}\langle\mathsf{y}^T\rangle_{q_j^{(i)}} + \mathbf{\Sigma}\right)\right)^{-1} \tag{4.30}$$

and

$$\left(s_j^{(new)}\right)^2 = \sum_{i=1}^{M} \frac{p(z_j|\mathsf{x}^{(i)})}{|\mathsf{x}|\sum_{i=1}^{M} p(z_j|\mathsf{x}^{(i)})}\text{tr}\left\{(\mathsf{x} - \mathsf{U}_j\langle\mathsf{y}\rangle_{q_j^{(i)}})(\mathsf{x} - \mathsf{U}_j\langle\mathsf{y}\rangle_{q_j^{(i)}})^T + \mathsf{U}_j^{(new)}\mathbf{\Sigma}\left(\mathsf{U}_j^{(new)}\right)^T\right\} \tag{4.31}$$

(see Tipping and Bishop (1999a)), where the expectations of the codes are computed over the exact component-based posteriors $\langle\mathsf{y}\rangle_{q_j^{(i)}} \equiv \langle\mathsf{y}\rangle_{q(\mathsf{y}|\mathsf{x}^{(i)}, z_j)}$. The posterior is easily expressed from the mixture model $\mathcal{M}_L$ by Bayes rule $q(\mathsf{y}|\mathsf{x}^{(i)}, z_j) \propto q(\mathsf{x}^{(i)}|\mathsf{y}, z_j)q(\mathsf{y})$, which for the considered case gives a Gaussian with the mean $\langle\mathsf{y}\rangle_{q_j^{(i)}} = (s_j^2\mathsf{I}_{|\mathsf{y}|} + \mathsf{U}_j^T\mathsf{U}_j)^{-1}\mathsf{U}_j^T\mathsf{x}^{(i)}$ (see e.g. von Mises (1964)).

It is easy to see that under the assumption of isotropic Gaussian mixture components, learning the decoder at the I-step of the variational information-maximizing algorithm given by (4.21) and (4.22) has the same form as the PPCA updates (4.30) and (4.31), with the difference that expectations of the codes $\langle\mathsf{y}\rangle$ are computed over the explicitly *constrained encoder* $p(\mathsf{y}|\mathsf{x}^{(i)}, z_j)$ rather than the posterior $q(\mathsf{y}|\mathsf{x}^{(i)}, z_j)$ expressed from the generative model $\mathcal{M}_L$ by applying Bayes rule. These distributions are generally different; particularly, as we mentioned in Section 4.1, the explicit parameterization of $p(\mathsf{y}|\mathsf{x}^{(i)}, z_j)$ makes it easier to impose the required constraints on the communication channel, while the exact analytical form of $q(\mathsf{y}|\mathsf{x}^{(i)}, z_j)$ might not necessarily satisfy such constraints. Generally, the information-theoretic learning of the optimal encoder involves optimization of the bound $\tilde{I}_H(\mathsf{x}, \{\mathsf{y}, z\})$ in the space of constrained encoder parameters, rather than probabilistic inference in the generative model $\mathcal{M}_L$.

Interestingly, we can demonstrate that while the considered variational framework for maximizing the bound on mutual information is generally different from maximizing the likelihood in a mixture of probabilistic PCA models, under the assumption of a fixed Gaussian channel noise it gives rise to the same fixed points as a variational EM applied to fitting a mixture of constrained Gaussians with the components $q(\mathsf{x}|\mathsf{y}, z_j) \sim \mathcal{N}_{\mathsf{x}}(\mathsf{U}_j\mathsf{y}, s_j^2)$ and the uniform (rather than the Gaussian) distribution of the hidden variables (i.e. $q(\mathsf{y}) \sim \mathcal{U}_{\mathsf{y}}$). In other words, this specific

application of our framework leads to the same solutions as a variational approach to fitting a mixture of factor analysis-like models with the uniform (rather than the Gaussian) distribution of the hidden factors. We discuss the link between these methods in Appendix B.2.

4.2.3.2 Comparison with Mixtures of Latent Variable Models

By analogy with the results of Chapter 3, we may point out general differences between maximizing the likelihood in mixture models and maximizing the bound on the mutual information in hybrid channels. In particular, the variational IM algorithm applied to training the encoder model of a hybrid channel $\mathcal{M}_I \stackrel{\text{def}}{=} \tilde{p}(\mathsf{x})p(\mathsf{y}, z|\mathsf{x})$ may be compared with the conventional fitting of a mixture of latent variable models $\mathcal{M}_L = p(\mathsf{y})p(z)p(\mathsf{x}|\mathsf{y}, z)$ to a data set given by the empirical distribution $\tilde{p}(\mathsf{x})$. Not surprisingly, the solutions obtained by training the generative and the encoding models would generally be different. By analogy with proposition 3.1, we may show that for the special case when the posteriors of $\mathcal{M}_L$ and $\mathcal{M}_I$ are constrained to be equivalent, we may define tractable variational lower bounds on $I(\mathsf{x}, \{\mathsf{y}, z\})$ which are at least as tight as (and sometimes significantly tighter than) the bound given by the exact likelihood. Effectively, this means that if the goal is to maximize information content which the hidden codes y and component labels z contain about the data patterns x, the variational information-maximizing framework should generally be more preferable than maximum-likelihood approaches (at least, in terms of the resulting bounds on $I(\mathsf{x}, \{\mathsf{y}, z\})$). Some of the intersections of the *approximate* approaches to likelihood and mutual information maximization are described in Appendix B.2.

Furthermore, we note once again that the encoder and the generative frameworks are different conceptually. Fundamentally, the goal of maximizing the mutual information in the hybrid channel $\mathsf{x} \rightarrow \{\mathsf{y}, z\}$ is to learn the *encoder* model (a specific tractable choice of the variational decoder is something which facilitates the generally intractable computations). In contrast, the conceptual goal of likelihood training would be to fit the generative model to data, which in our framework would correspond to learning the variational *decoder* (and the marginal distribution of the hidden variables). As discussed in Chapter 3 and Section 4.1, the fundamental feature of the encoder framework is the possibility of specifying explicit constraints on the encoder model $p(\mathsf{y}|\mathsf{x})$. Thus, the suggested variational information-maximizing framework is particularly convenient in situations when our goal is to learn unknown *constrained* encodings of the visible patterns (in practice, such constraints might be artifacts of biophysical or engineering requirements). In general, this is different from generative models, where the desired constraints on the posterior (encoding) mapping $p(\mathsf{y}|\mathsf{x})$ would need to follow from the explicit parameterization of the priors $p(\mathsf{y})$ and conditionals $p(\mathsf{x}|\mathsf{y})$. It is therefore intuitive that it may be rather difficult to apply generative models for addressing constrained encoding problems exactly. For example, apart from relatively simple cases (e.g. when $\mathcal{M}_L$ is the probabilistic PCA model (Tipping and Bishop, 1999b)), it is difficult to apply an exact generative framework for learning an $\mathcal{L}$-optimal orthonormal projection of the data. As

we showed, the problem may be conveniently addressed within the information-maximization framework, where the encoder constraints may be satisfied as a part of the model's specification.

4.3 Demonstrations

Here we demonstrate a few applications of the method to extracting optimal orthonormal subspaces for the digits dataset (which is a sub-sampled lower-dimensional version of MNIST, LeCun and Cortes (1998)). As mentioned in Section 4.2.3, apart from simple cases the problem cannot be easily addressed by exact applications of generative models, as we would require specific orthonormal constraints on the posterior distributions. In all cases, it was assumed that $|\mathsf{y}| < |\mathsf{x}|$. We also assumed that $p(\mathsf{x}) = \tilde{p}(\mathsf{x})$ is the empirical distribution.

4.3.1 Hand-Written Digits: Comparing the Bounds

In the first set of experiments, we compared optimal lower bounds on the mutual information $I(\mathsf{x}, \mathsf{y})$ obtained by maximizing the as-if Gaussian $I_G(\mathsf{x}, \mathsf{y})$ and the auxiliary variational $\tilde{I}(\mathsf{x}, \mathsf{y})$ objectives for hand-written digits. The dataset contained $M = 30$ gray-scaled instances of 14-by-14 digits 1, 2, and 8 (10 of each class), which were centered and normalized. The goal was to find an orthogonal projection of the $|\mathsf{x}| = 196$-dimensional training data into a $|\mathsf{y}| = 6$-dimensional space, so that the bounds $I_G(\mathsf{x}, \mathsf{y})$ and $\tilde{I}(\mathsf{x}, \mathsf{y})$ were maximized. By analogy with Section 4.1.1, we considered a linear Gaussian channel with an irreducible white noise, which in this case leads to the encoder distribution $p(\mathsf{y}|\mathsf{x}) \sim \mathcal{N}_y(\mathsf{W}\mathsf{y}, s^2\mathsf{I})$ with $\mathsf{W} \in \mathbb{R}^{6\times 196}$. Our interest was in finding optimal orthogonal projections, so the weights were normalized to satisfy $\mathsf{W}\mathsf{W}^T = \mathsf{I}_{|\mathsf{y}|}$ (by considering the parameterization $\mathsf{W} = (\tilde{\mathsf{W}}\tilde{\mathsf{W}}^T)^{-1/2}\tilde{\mathsf{W}}$ with $\tilde{\mathsf{W}} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{x}|}$). Effectively, this case corresponds to finding the most informative compressed representations of the source vectors for improving communication of the *non-Gaussian* data over a noisy Gaussian channel (by maximizing lower bounds on the channel capacity). Our specific interest here was to find whether we may indeed improve on Linsker's as-if Gaussian bound on the mutual information (with the optima given in this case by the PCA projection) by considering a richer family of auxiliary variational bounds with multi-modal mixture-type decoders.

Figure 4.1 shows typical changes in the auxiliary variational bound $\tilde{I}(\mathsf{x}, \mathsf{y})$ as a function of the IM's iterations T for $|z| \in \{2, \ldots, 5\}$ states of the discrete auxiliary variable. (On the plot, we ignored the irrelevant constants $H(\mathsf{x})$ identical for both $\tilde{I}(\mathsf{x}, \mathsf{y})$ and $I_G(\mathsf{x}, \mathsf{y})$, and interpolated $\tilde{I}(\mathsf{x}, \mathsf{y})$ for the consecutive iterations). The mappings were parameterized as described in Section 4.2, with the random initializations of the parameters v_j and b_j around zero, and the initial settings of the variational prior $q(z) = 1/|z|$. The encoder weights W were initialized at 6 normalized principal components $\mathsf{W}_{pca} \in \mathbb{R}^{6\times 196}$ of the sample covariance $\langle \mathsf{x}\mathsf{x}^T \rangle$, and the variance of the channel noise was fixed at $s^2 = 1$. For each choice of the auxiliary space dimension $|z|$, Figure 4.1 (a) shows the results averaged over 30 random initializations of the IM algorithm. As we see from the plot, the IM

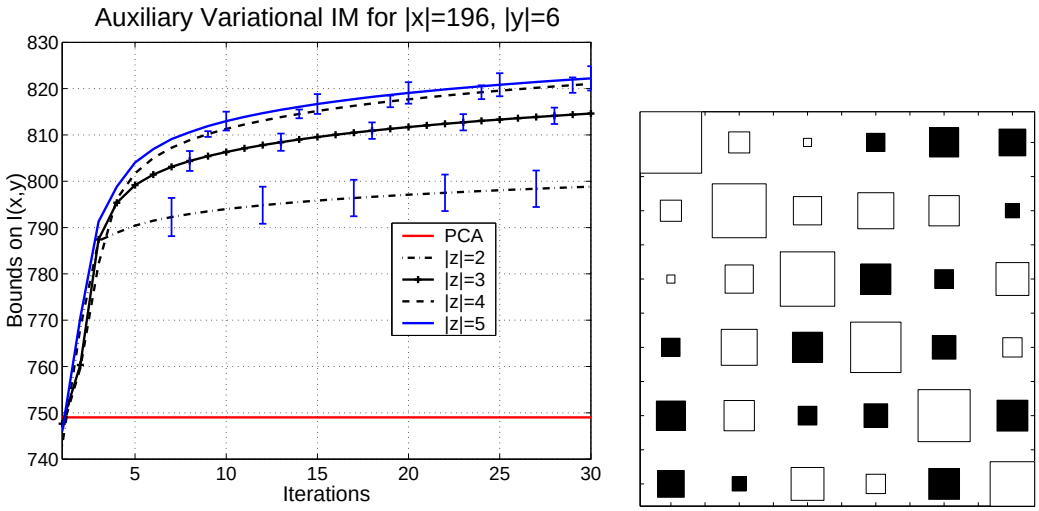


Figure 4.1: Variational information maximization for noisy constrained dimensionality reduction. **(a):** *Top curves:* Average values of the variational auxiliary bounds $\tilde{I}(\mathsf{x},\mathsf{y})$, obtained by the IM algorithm started at 10 random model initializations (shown for $|z| = 2,\ldots,5$); *bottom line*: the *as-if* Gaussian $I_G(\mathsf{x},\mathsf{y})$ bound (computed numerically). The results are shown for the digits data with $|\mathsf{x}| = 196$, $|\mathsf{y}| = 6$ for $M = 30$ patterns and $T = 30$ iterations of the IM. **(b):** Hinton diagram for $\mathsf{W}\mathsf{W}_{pca}^T(\mathsf{W}\mathsf{W}_{pca}^T)^T \in \mathbb{R}^{6\times 6}$ for $|z| = 3$, $T = 30$. For orthonormal weights spanning identical subspaces, we would expect to see the identity matrix.

learning leads to a consistent improvement in the auxiliary variational bound, which (on average) varies from $\tilde{I}_0(\mathsf{x},\mathsf{y}) \approx 745.7$ to $\tilde{I}_T(\mathsf{x},\mathsf{y}) \approx 822.2$ at $T = 30$ for $|z| = 5$. Small variances in the obtained bounds ($\sigma_T \approx 2.6$ for $T = 30$, $|z| = 5$) indicate a stable increase in the information content independently of the model initializations. From Figure 4.1 (a) we can also observe a consistent improvement in the average $\tilde{I}(\mathsf{x},\mathsf{y})$ with $|z|$, changing as $\tilde{I}_{10}(\mathsf{x},\mathsf{y}) \approx 793.9$, ≈ 806.3, ≈ 811.2, and ≈ 812.9 for $|z| = 2,\ldots,5$ after $T = 10$ IM's iterations. In comparison, the PCA projection weights W_{pca} result in $I_G(\mathsf{x},\mathsf{y}) \approx 749.0$, which is visibly worse than the auxiliary bound with the optimized parameters, and is just a little better than $\tilde{I}(\mathsf{x},\mathsf{y})$ computed at a random initialization of the variational decoder for $|z| \geq 2$.

Importantly, we stress once again that the auxiliary variables z are not passed through the channel. In the specific case which we considered here, the auxiliary variables were used to define a more powerful family of variational bounds which we used to extract the $\tilde{I}$-optimal orthonormal subspace. The results are encouraging, as they show that for a specific constrained channel distribution we may indeed obtain tighter lower bounds on the mutual information $I(\mathsf{x},\mathsf{y})$ *without* communicating more data than in the conventional case. Specifically, for Gaussian channels with orthonormal projections to the code space, we do improve on simple *as-if Gaussian* solutions (leading to the PCA projections) by considering optimization of the auxiliary variational bounds (4.14).

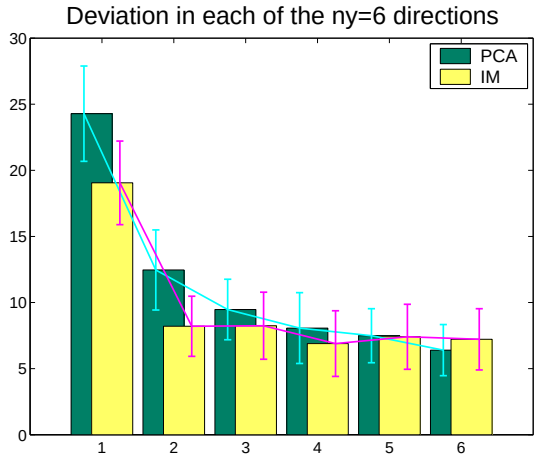
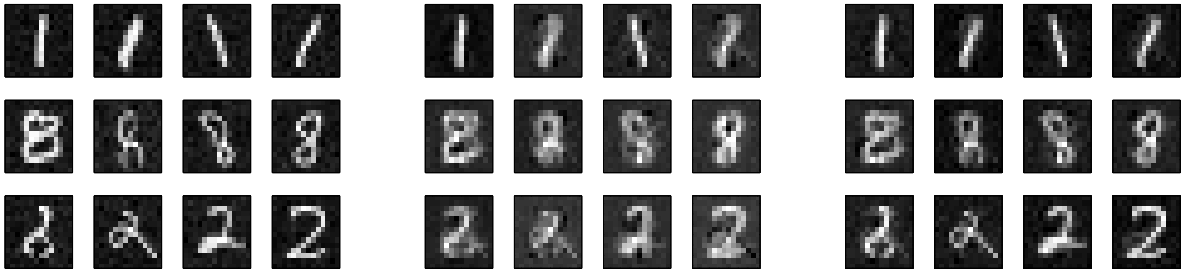


Figure 4.2: *Histogram bars:* marginal variances of the orthonormal noisy projections of the data patterns $\langle (y_i - \langle y_i \rangle)^2 \rangle_{p(y_i|\mathsf{x})\tilde{p}(\mathsf{x})}$ along each of $i = 1, \ldots, 6$ dimensions of the code space. The code space $\mathcal{R}^{|y|}$ is spanned by W_{pca} and W, for Linsker's and the auxiliary variational objectives respectively. *Vertical lines:* variances of the average distances for $M = 30$ data patterns. The results are shown for $|\mathsf{x}| = 196$, $|\mathsf{y}| = 6$; the auxiliary space size $|z| = 3$; the number of iterations $T = 30$.

As expected, we may also note that the $\tilde{I}$-optimal *encoder* weights W are in general different from rotations of W_{pca}. This is easy to see by computing $\mathsf{W}\mathsf{W}_{pca}^T(\mathsf{W}\mathsf{W}_{pca}^T)^T$, which in our case is visibly different from the identity matrix (see Fig. 4.1 (b) for $|\mathsf{y}| = 6$ and $|z| = 3$), which we would have expected to obtain otherwise. Effectively, this means that by allowing a greater flexibility in the choice of the *variational decoder* distributions, the $\tilde{I}(\mathsf{x}, \mathsf{y})$-optimal constrained *encoders* become different from the optimal encoders of simpler models. This result is intuitive: it is natural to expect that a richer structure of the decoder model may change our notion of the optimal codes (at least, at the stage of optimizing $\tilde{I}(\mathsf{x}, \mathsf{y})$). And vice versa, a choice of simple variational decoders (such as linear Gaussians) may impose severe constraints on the types of codes which they can decode efficiently, which may lead to a loss in the coding efficiency and result in a general reduction in the retained information content (see Chapter 5.3 for a more detailed discussion).

Finally, figure 4.2 shows average distances of the noisy linear projections of the testing data from the y-space mean (i.e. the marginal variance of the codes) for $M = 30$ patterns. The histogram indicates the variances of the projections for each of the $|\mathsf{y}| = 6$ dimensions of the code space, spanned by W_{pca} and W (for Linsker's and the auxiliary variational bounds respectively). The results are shown for $M = 30$ digit patterns after $T = 30$ iterations of learning (again, we assumed that the size of the auxiliary space $|z| = 3$). We can see that the encoded representations produced by the auxiliary variational method result in a more uniform spectrum of the projection variances, with roughly equal error bars. This result is not unexpected, as under the fixed channel noise assumption, the information content between the data and the codes increases with an increase



(a) (b) (c)

Figure 4.3: Reconstructions of the source patterns from encoded representations. **(a):** A subset of the generic patterns used to generate the source vectors; **(b):** the corresponding reconstructions from 6 principal components; **(c):** the corresponding $\tilde{I}_H$-optimal reconstructions at $\langle \mathsf{x} \rangle_{q(\mathsf{x}|\mathsf{y},z)} = \mathsf{U}_z \mathsf{y}$ for the hybrid $\{\mathsf{y}, z\}$ representations ($|\mathsf{y}| = 6$, $|z| = 3$).

in the entropy of the encoded representations. Whilst it is not easy to compute the exact entropy of the mixture distribution $p(\mathsf{y})$, qualitatively it is natural to expect that more uniform encodings would typically result in higher values of the mutual information between the encodings and the source patterns.

4.3.2 Hand-Written Digits: Reconstructions

Additionally, for the problem settings described in Sec. 4.3.1, we have computed reconstructions of the source patterns $\{\mathsf{x}\}$ from their noisy encoded representations. First, we generated source vectors by adding an isotropic Gaussian noise to the generic patterns (see Fig. 4.3 (a)), where the variance of the source noise was set as $s_s^2 = 0.5$. Then we computed noisy linear projections $\{\mathsf{y}\}$ of the source vectors by using the I_G- and the $\tilde{I}_H$- optimal encoder weights (in the latter case, we also computed the auxiliary label z by sampling from the learned $p(z|\mathsf{x})$). This stage corresponds to passing encoded representations over the noisy channels, where the noise variance for the Gaussian part of the channel was fixed at $s^2 = 1$. Finally, we have used the optimal *approximate* decoders to perform the reconstructions from $\{\mathsf{y}\}$ (for I_G-optimal PCA projections) and $\{\mathsf{y}, z\}$ (for $\tilde{I}_H$-optimal hybrid channels).

As we see from Figure 4.3 (b), (c), by a slight modification of the channel (due to encoding and communicating a multinomial variable z), we may achieve a visible improvement in the reconstruction of the sources by using the $\tilde{I}_H$- optimal projections[2]. The results are shown for $|\mathsf{y}| = 6$, $|z| = 3$ after $T = 3$ iterations, and the reconstructions are computed at the analytical mean of the decoder's component $q(\mathsf{x}|\mathsf{y}, z)$ indexed by the auxiliary variable z. Even though the resulting hybrid channel may be difficult to justify from the communication viewpoint, the

[2]Similar reconstructions could be obtained by maximizing the auxiliary bound $\tilde{I}(\mathsf{x}, \mathsf{y})$ *without* communicating z. However, the approximate decoder for this case would be given as $q(\mathsf{x}|\mathsf{y}) = \sum_z q(\mathsf{x}|\mathsf{y}, z) \frac{\langle p(z|\mathsf{x}) p(\mathsf{y}|\mathsf{x}) \rangle_{p(\mathsf{x})}}{\langle p(z|\mathsf{x}) \rangle_{p(\mathsf{x})}}$, which requires knowing $p(\mathsf{x})$.

results suggest that maximization of the bound on $I(\mathsf{x}, \{\mathsf{y}, z\})$ provides a sensible way to reduce dimensionality of the sources for the purpose of reconstructing inherently noisy non-Gaussian patterns. Importantly, the variational decoder $q(z|\mathsf{x}, \mathsf{y})$ which maximizes $\tilde{I}_H(\mathsf{x}, \{\mathsf{y}, z\})$ makes no recourse to $p(\mathsf{x})$. Therefore, just like in the PCA case, we do not need to store the training instances in order to perform an accurate reconstruction from noisy lower-dimensional projections. We note once again that the weights of the optimal encoder were chosen to satisfy the specific orthonormality constraint (though other kinds of constrained encoders may easily be considered). This contrasts with the exact approaches to training generative models, where encoding constraints may be more difficult to enforce.

4.4 Summary

Here we considered an application of the variational information maximizing approach to linear orthonormal dimensionality reduction in the presence of irreducible Gaussian noise. We showed that the well-known *as-if* Gaussian approximation of the mutual information (Linsker (1992)), which may be seen as a special case of the variational bound for correlated linear Gaussian variational distributions, leads to the PCA solution for isotropic linear Gaussian channels. Importantly, this means that by using linear Gaussian variational decoders under the considered Gaussian channel, maximization of the generic lower bound (2.2) on the mutual information cannot improve on the PCA projections.

The situation becomes strikingly different if we consider a richer family of variational auxiliary lower bounds on $I(\mathsf{x}, \mathsf{y})$ under the same encoding constraints. In particular, we showed that in the cases when the source distribution was non-Gaussian, we could significantly improve on the PCA projections by considering multi-modal variational decoders. This confirms the conceptually simple idea that by allowing a greater flexibility in the choice of variational decoders, we may get significant improvements over simple bounds on the mutual information at a limited increase in the computational cost. This result is interesting from the communication-theoretic perspective, as it demonstrates a simple and computationally efficient way to produce better bounds on the capacity of communication channels without altering channel properties (e.g. without communicating more data across the channels).

Finally, we discussed a simple information-theoretic approach to constrained dimensionality reduction for hybrid representations $\mathsf{x} \to \{\mathsf{y}, z\}$, which may significantly improve reconstructions of the sources $\{\mathsf{x}\}$ from their lower-dimensional representations $\{\mathsf{y}\}$ at a small increase in the transmission cost (given by $|z|$). We applied the hybrid framework for extracting an informative orthonormal projection subspace of the data. While being vaguely related to the exact maximum-likelihood fitting of mixtures of latent variable models (at least, in terms of the specified variable domains), our variational information-maximizing framework is fundamentally different in terms of model specifications. One of the important features of the IM framework is the explicit parameterization of the encoder model, which is particularly convenient in situations when the goal is to learn unknown encodings of the visible patterns for a known family of constrained

encoder distributions. Usually, such constraints would be difficult to impose in generative models. On the other hand, they may be easily introduced in the suggested information-maximizing framework, while a tractable choice of variational decoders simplifies the computations for large-scale stochastic channels.

Generally, we have confirmed the intuition that we may improve on simple generic bounds on $I(\mathsf{x}, \mathsf{y})$ by considering a richer family of the *auxiliary variational* bounds, which effectively increase the power of the variational decoders. On the other hand, it is natural to expect that we may maximize the information content by considering richer families of *encoder* distributions. In the next chapter we consider specific applications of the information-maximizing framework to the case of nonlinear channels. We will describe theoretical properties of $\tilde{I}$-optimal nonlinear Gaussian encoder distributions, and focus specifically on tractable practical applications to information-theoretic clustering.

Chapter 5

Variational Information Maximization for Nonlinear Dimensionality Reduction

In previous chapters we introduced a simple variational lower bound on the mutual information, which resolves some of the computational difficulties of computing the exact mutual information. Then we compared our framework with the conventional learning in generative and autoencoder-type models, and considered specific applications of the variational information maximization to the constrained dimensionality reduction. As we demonstrated in Chapter 4, the variational lower bounds on the mutual information can be made tighter by considering more powerful families of variational decoders. It is also intuitive that one may hope to obtain tighter bounds on the mutual information by increasing the power of the stochastic *encoder*. Intuitively, by appropriately choosing nonlinear encoder distributions so that the encoder satisfies specific local constraints, we may hope to obtain better reconstructions of the transmitted sources, more anthropomorphically sensible visualizations in the compressed variable spaces, etc. Here we explore these matters by considering the problem of maximizing information content for stochastic non-linear channels.

We will focus primarily on the problem of information-theoretic clustering, where the encoder distribution is defined by a generally stochastic, nonlinear mapping from the source patterns to discrete cluster labels. For this case, we will consider two principally different learning techniques. First, we will consider optimizing the specific lower bound on the mutual information, where the encoder distribution of the channel is given by the exact posterior of the corresponding generative model (see expression (3.11) and the discussion in Section 3.2.1). Effectively, this approach may be viewed as an information-theoretic method of training generative models, which we study for the case of Gaussian mixtures. Then we consider a different information-theoretic clustering technique, where we maximize the exact mutual information in explicitly parameterized encoder models. For most of the clustering applications, this may be done reasonably easily, since the cardinality of the code space will typically be low, so that maximization of the exact objective will typically be computationally tractable. For

this case, we describe a simple and practical algorithm for unsupervised discriminative learning of cluster allocations. By allowing the flexibility in the choice of the encoder structure, we may consider a variety of encoder model parameterizations, including those involving nonlinear projections of the source vectors into high-dimensional feature spaces. Empirically, we demonstrate that the resulting information-theoretic clustering approach favorably compares with the common generative clustering methods.

In the second part of the chapter we will briefly review some of the theoretical properties of the IM for higher-dimensional code spaces, and show that some of the popular dimensionality reduction techniques may be seen as special instances of the variational information maximization procedure.

5.1 Introduction

The limiting case of a dimensionality reduction problem, where the code y of a high-dimensional pattern x is defined by a single multinomial hidden variable, may be viewed as a simple instance of *clustering*. The codes $\{y\}$ in this case may be interpreted as the unknown cluster labels corresponding to the data patterns. If the family of the generally stochastic mappings $\mathsf{x} \to y$ satisfies specific local constraints, the resulting cluster allocations may provide a simple anthropomorphically sensible visualization of generally high-dimensional patterns.

Arguably, there are at least two simple requirements which a cluster allocation procedure should satisfy. First of all, we may require the resulting clusters of data patterns to be locally smooth. There may be several ways to interpret local smoothness in the clustering context; for example, it may be sensible to require that each two patterns have a high probability of being assigned to the same cluster if the corresponding vectors satisfy specific geometric constraints. Secondly, we may require the clustering method to avoid assigning unique cluster labels to outliers (or other constrained regions in the data space), so that local regions in the data space are not over-represented in the code space.

Generally, we may expect the clusters to be well-separated if patterns x are predictive of cluster labels y, for all the data patterns $\{\mathsf{x}\}$. Here we argue that a reasonable way of learning optimal cluster allocations is by maximizing the exact mutual information $I(\mathsf{x}, y) = H(y) - H(y|\mathsf{x})$. Indeed, in the clustering context $I(\mathsf{x}, y)$ may be interpreted as a measure of predictability of cluster allocations from training patterns. Intuitively, it is clear that in contrast to the generative mixture model clustering, optimization of $I(\mathsf{x}, y)$ does not intrinsically favour outliers. (Indeed, for the fixed cardinality $|y| < M$ of the code space, assignments of unique cluster labels to under-represented training patterns would have resulted in a decrease in the marginal entropy $H(y)$. In other words, regions under-represented in the data space would have been over-represented in the code space, which for the fixed noise levels would have led to a reduction in the mutual information). Generally, we may expect that the mutual information maximization favours hard assignments of cluster labels to regions of roughly identical sizes, resulting in the growth in $H(y)$ and reduction in $H(y|\mathsf{x})$.

In the first part of the chapter we will discuss simple and efficient information-theoretic clustering algorithms, which may be used for *discriminative unsupervised*[1] learning of nonlinear cluster allocations. First we will consider optimizing the lower bound $\hat{I}(\mathsf{x}, y) = \langle \log p(\mathsf{x}|y) \rangle_{p(y|\mathsf{x})\tilde{p}(\mathsf{x})}$, where the encoder distribution of the channel is given by the exact posterior of the corresponding Gaussian mixture model $\mathcal{M}_{\mathcal{L}} = p(y)p(\mathsf{x}|y)$, where $p(y)$ are mixture coefficients, and $p(\mathsf{x}|y)$ are the Gaussian components. Effectively, this approach may be viewed as an information-theoretic method of training Gaussian mixture models. We compare this approach with the standard likelihood maximization for Gaussian mixtures, and show that by analogy with the EM training, the IM algorithm reduces to the k-means clustering in the limiting case of zero-variance components.

Then we consider a different information-theoretic clustering technique, where we impose specific constraints on the encoding distribution of the channel model $\mathcal{M}_I$, and maximize $I(\mathsf{x}, y)$ directly. Unlike most of the existing information-theoretic approaches applicable in the clustering context (see e.g. Fisher and Principe (1998), Torkkola and Campbell (2000), Gokcay and Principe (2002), Dhillon et al. (2002), Corduneanu and Jaakkola (2003), Jenssen et al. (2003)), we consider optimization of the exact mutual information in an encoder model, which for the purpose of learning cluster allocations may be computed exactly. Furthermore, as a straight-forward extension of our method, we consider clustering in the feature space, with a constrained kernelized representation of the encoder for nonlinearly transformed source patterns. We show that for this case we may apply the information-maximization framework to learn the optimal kernel parameters, which often leads to sensible solutions of the clustering problem.

In the last part of the chapter we extend the discussion of the information-maximizing framework by considering other types of nonlinear encoders. We will focus particularly on the discussion of *non-linear Gaussian* channels with the fixed isotropic noise. Since in this case maximization of the mutual information cannot be performed exactly, we consider optimizing the generic lower bound (2.2) on the mutual information with several specific choices of the variational decoder, and describe properties of the obtained optimal solutions. While our analysis of the analytical properties of the IM solutions is mainly theoretical, it shows interesting links of our variational procedure to other popular dimensionality reduction techniques, inducing PCA, kernel PCA (Schoelkopf et al. (1998), Mika et al. (1999)), and Gaussian Process Latent Variable Models (Lawrence (2003)) as special cases. Some of these theoretical findings extend previous results for noiseless autoencoders (Bourlard and Kamp (1988), Baldi and Hornik (1989), Diamantaras and Kung (1996)); however, they are derived for *stochastic* communication channels and general nonlinear mappings. Most of the derivations for this chapter can be found in Appendix C.

[1] As mentioned above, the encoder models are specified by explicitly parameterizing the stochastic mapping $p(\mathsf{y}|\mathsf{x})$ to the code space, which corresponds to a similar parameterization of a discriminative model. However, in contrast to discriminative models, the encodings y of the sources x are presumed to be hidden in our case.

5.2 Nonlinear Channels for Information-Theoretic Clustering

One of the practical side-effects of Section 3.2 is a practical information-theoretic method of training generative models (see expression (3.11)). Here we apply it to training Gaussian mixture models. With a slight abuse of terminology, we will refer to the method as the IM for Gaussian mixtures; by this we mean that the encoder distribution of the channel model trained by the IM has the exact form of the posterior $p(y|\mathsf{x})$ of a Gaussian mixture model. We show that the IM applied for training the considered models indeed provides a sensible clustering technique. Moreover, the results help to empirically verify the theoretical findings of Chapter 3.

5.2.1 Variational Information Maximization for Gaussian Mixtures

In proposition 3.1 we showed that maximization of the exact likelihood $\mathcal{L}$, expressed from the generative model $\mathcal{M}_L \stackrel{\text{def}}{=} p(\mathsf{y})p(\mathsf{x}|\mathsf{y})$ for i.i.d. patterns, may be viewed as maximization of a lower bound on the mutual information in the encoder model $\mathcal{M}_I \stackrel{\text{def}}{=} \tilde{p}(\mathsf{x})p(\mathsf{y}|\mathsf{x})$, where y is a latent code, x is a source vector, and $\tilde{p}(\mathsf{x})$ is the empirical distribution. However, as we noted in Section 3.2.1, the likelihood bound on $I(\mathsf{x},\mathsf{y})$ may potentially be weak, and much tighter bounds on $I(\mathsf{x},\mathsf{y})$ may be obtained by considering a special instance of the variational information-maximizing framework. As we showed in Section 3.2.1, the bound provided by the likelihood is weaker or as tight as

$$\hat{I}(\mathsf{x},\mathsf{y}) = \langle \log p(\mathsf{x}|\mathsf{y}) \rangle_{\tilde{p}(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x})} + H_{\tilde{p}}(\mathsf{x}) = \langle \log p(\mathsf{x}|\mathsf{y}) \rangle_{p(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x})} + H_{\tilde{p}}(\mathsf{x}), \tag{5.1}$$

where we defined the encoding distribution of the encoder model

$$\tilde{p}(\mathsf{y}|\mathsf{x}) \propto p(\mathsf{y})p(\mathsf{x}|\mathsf{y}) \tag{5.2}$$

to be the exact posterior of the generative model $\mathcal{M}_L$.

As in the conventional approaches to latent variable modeling, our goal would be to produce informative latent variable representations of the sources, so optimization of the tighter bound (5.1) on $I(\mathsf{x},\mathsf{y})$ may be a reasonable strategy to consider. Moreover, as discussed in Section 3.2.1, under the considered constraint on the encoder, expression (5.1) may be viewed as an information-theoretic objective for training both the generative $\mathcal{M}_L$ and the encoder $\mathcal{M}_I$ models[2]. Effectively, this is our main motivation for considering optimization of the bound $\hat{I}(\mathsf{x},y)$. If it is indeed the case that optimization of objective (5.1) helps to avoid common degeneracies of the likelihood solutions (such as overfitting to local data segments), we may hope that the data generated from $\mathcal{M}_L$ (trained by maximizing the bound $\hat{I}(\mathsf{x},\mathsf{y})$) could in some sense be more representative of the underlying process than the data sampled from the likelihood-trained mixture.

[2]This contrasts with the exact likelihood and the exact mutual information training which are commonly applied to training generative and encoder models respectively.

Here we will consider a simple case when $\mathcal{M}_L$ is a Gaussian mixture, where $y \in \{y_j | j = 1, \ldots, |y|\}$ is the mixture label, and $p(\mathsf{x}|y_j) \sim \mathcal{N}_{\mathsf{x}}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, where $\boldsymbol{\mu}_j \in \mathbb{R}^{|\mathsf{x}|}$, $\boldsymbol{\Sigma}_j \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{x}|}$ are the mean and the covariance corresponding to the j^{th} component y_j. The objective $\hat{I}(\mathsf{x}, \mathsf{y})$ may be easily expressed as

$$\hat{I}(\mathsf{x}, \mathsf{y}) = \frac{1}{M} \sum_{m=1}^{M} \sum_{j=1}^{|y|} \frac{p(y_j) p(\mathsf{x}^{(m)}|y_j)}{p(\mathsf{x}^{(m)})} \log p(\mathsf{x}^{(m)}|y_j), \tag{5.3}$$

which needs to be optimized for $p(y_j)$ and $\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j$ for all $j = 1, \ldots, |y|$. For simplicity, we will start the discussion of the resulting optimization procedure by *implying* the constraints on the optimized parameters. Effectively, this implication allows us to treat all the parameters (including the mixture coefficients) as unconstrained vectors in high-dimensional real spaces; we will impose meaningful construction constraints shortly.

5.2.1.1 Learning Optimal Parameters

By computing the matrix derivatives of (5.3) for $\boldsymbol{\Sigma}_j$ and $\boldsymbol{\mu}_j$ (see e.g. Magnus and Neudecker (1999), Minka (2000)), it is easy to find that

$$\frac{\partial \hat{I}(\mathsf{x}, \mathsf{y})}{\partial \boldsymbol{\mu}_j} = -\frac{1}{M} \sum_{m=1}^{M} \boldsymbol{\Sigma}_j^{-1} (\mathsf{x}^{(m)} - \boldsymbol{\mu}_j) p(y_j|\mathsf{x}^{(m)}) (\hat{\alpha}_j^{(m)} + 1) \tag{5.4}$$

$$\frac{\partial \hat{I}(\mathsf{x}, \mathsf{y})}{\partial \boldsymbol{\Sigma}_j} = -\frac{1}{2M} \sum_{m=1}^{M} \boldsymbol{\Sigma}_j^{-1} \left((\mathsf{x}^{(m)} - \boldsymbol{\mu}_j)(\mathsf{x}^{(m)} - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1} - \mathsf{I}_{|\mathsf{x}|} \right) p(y_j|\mathsf{x}^{(m)}) (\hat{\alpha}_j^{(m)} + 1), \tag{5.5}$$

where

$$\begin{aligned} \hat{\alpha}_j^{(m)} &\stackrel{\text{def}}{=} \log p(\mathsf{x}^{(m)}|y_j) - \sum_{l=1}^{|y|} p(y_l|\mathsf{x}^{(m)}) \log p(\mathsf{x}^{(m)}|y_l) \\ &= \log \frac{p(\mathsf{x}^{(m)}|y_j)}{p(\mathsf{x}^{(m)})} - KL\left(p(y|\mathsf{x}^{(m)}) \| p(y)\right), \end{aligned} \tag{5.6}$$

and $p(y_j|\mathsf{x}^{(m)})$ is the exact posterior expressed from the generative model $\mathcal{M}_L$. Note that in the trivial case when $\hat{\alpha}_j^{(m)} = const$ for all $m = 1, \ldots, M$ and $j = 1, \ldots, |y|$, the gradients (5.4) and (5.5) are identical to those obtained by maximizing the log-likelihood of a Gaussian mixture model (up to irrelevant constant pre-factors). We may further express the functional derivatives (e.g. Gelfand and Fomin (1963), Weinstock (1974), Smith (1998)) for the mixture coefficients $p(y_j)$ to get

$$\begin{aligned} \frac{\partial \hat{I}(\mathsf{x}, \mathsf{y})}{\partial p(y_j)} &= \frac{1}{M} \sum_{m=1}^{M} \frac{p(y_j|\mathsf{x}^{(m)})}{p(y_j)} \left(\log p(\mathsf{x}^{(m)}|y_j) - \sum_{l=1}^{|y|} p(y_l|\mathsf{x}^{(m)}) \log p(\mathsf{x}^{(m)}|y_l) \right) \\ &= \frac{1}{M} \sum_{m=1}^{M} \frac{p(y_j|\mathsf{x}^{(m)})}{p(y_j)} \hat{\alpha}_j^{(m)}. \end{aligned} \tag{5.7}$$

(again, we have implied that the mixing coefficients $p(y_j)$ lie in a convex space (e.g. Boyd and Vandenberghe (2004)), i.e. there is no need to explicitly incorporate Lagrange multipliers into the objective (5.3)). If the coefficients $\hat{\alpha}_j^{(m)}$ were constant, we could use (5.7) to derive the fixed-point updates for $p(y_j)$ (subject to the normalizing constraints), which in this case would lead to the updates of the iterative EM algorithm for Gaussian mixtures (e.g. Bishop (1995)).

From (5.4) – (5.7) we see that for Gaussian mixture models, optimization of the likelihood $\mathcal{L}$ and the bound $\hat{I}(\mathsf{x}, y)$ result in generally different learning rules, as specified by the model-dependent weighting coefficients $\hat{\alpha}_j^{(m)}$. Note that $\hat{\alpha}_j^{(m)}$ includes the Kullback-Leibler divergence between the posteriors $p(y|\mathsf{x})$ and the prior $p(y)$, which quantifies the difference between the likelihood and the information-theoretic objectives (see expression (3.4)). It is easy to see that in the degenerate case when the latent variables are independent of the data (i.e. for $p(y|\mathsf{x}) = p(y)$), the factor $\hat{\alpha}_j^{(m)}$ will become irrelevant, and both EM and IM will result in simple data averaging for $\boldsymbol{\mu}_j$, $\boldsymbol{\Sigma}_j$ with arbitrary setting of $p(y_j)$ – clearly, this is not an interesting case to consider. Generally, however, $\hat{\alpha}_j^{(m)}$ is a function of the optimized mixture coefficients $p(y)$ and the parameters $\boldsymbol{\mu}_j$, $\boldsymbol{\Sigma}_j$, which makes it awkward to derive the closed-form iterative updates for the model parameters. This contrasts with the standard EM algorithms (Dempster et al. (1977)), which may be viewed as maximizing the expected complete data log-likelihood (see e.g. Bishop (1995)).

A simple alternative to the closed-form iterative optimization could be given by any of the known non-linear optimization methods (e.g. Bishop (1995), Dennis and Schnabel (1996), Bertsekas (1999), Galeev and Tihomirov (2000)), where the optimized parameters lie in appropriately constrained spaces. In this specific case we would need to constrain the covariance matrices to ensure their positive-definiteness; we would also need to ensure that the mixing coefficients $p(y_j)$ follow the probability requirements. Such constraints are easy to impose by considering meaningful non-restrictive constructions. For instance, to ensure positive-definiteness, we may parameterize the covariances $\boldsymbol{\Sigma}_j$ as

$$\boldsymbol{\Sigma}_j = \mathsf{A}_j \mathsf{A}_j^T + \mathsf{I}_{|x|}\epsilon, \tag{5.8}$$

where $\mathsf{A}_j \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{x}|}$ is an arbitrary real-valued matrix and $\epsilon \gtrsim 0$ is a small constant, which ensures that $\boldsymbol{\Sigma}_j$ is non-singular. Analogously, we may parameterize the mixing coefficients as

$$p(y_j) = \exp\{a_j\} / \sum_{l=1}^{|y|} \exp\{a_l\}, \tag{5.9}$$

which ensures satisfaction of the probability constraints. From the existence of the singular value decomposition of $\boldsymbol{\Sigma}_j$ (see e.g. Golub and Loan (1996)) and continuity of $a_j \in \mathbb{R}$, it is clear that parameterizations (5.8), (5.9) are non-restrictive[3].

[3]In what follows we assume that $\epsilon \to 0^+$, which implies that eigenvalues of $\boldsymbol{\Sigma}_j$ are the squares of the non-negative singular vectors of A_j. Clearly, in the considered limit the construction (5.8)

From the definition of the objective (5.1) and the construction (5.8), (5.9), it is easy to see that the constrained optimization of $\hat{I}(\mathsf{x},\mathsf{y})$ with respect to the covariances $\mathbf{\Sigma}_j$ and mixture coefficients $p(y_j)$ may be transformed to the unconstrained optimization problem for A_j and a_j as

$$\frac{\partial \hat{I}(\mathsf{x},y)}{\partial a_j} = p(y_j)\left(\frac{\partial \hat{I}(\mathsf{x},y)}{\partial p(y_j)} - \sum_l \frac{\partial \hat{I}(\mathsf{x},y)}{\partial p(y_l)} p(y_l)\right), \quad \frac{\partial \hat{I}(\mathsf{x},y)}{\partial \mathsf{A}_j} = 2\frac{\partial \hat{I}(\mathsf{x},y)}{\partial \mathbf{\Sigma}_j}\mathsf{A}_j, \tag{5.10}$$

where $\partial \hat{I}(\mathsf{x},y)/\partial \mathbf{\Sigma}_j$ and $\partial \hat{I}(\mathsf{x},y)/\partial p(y_j)$ are given by (5.5), (5.7), and $j = 1,\ldots,|y|$. The gradients (5.4) and (5.10) for $\boldsymbol{\mu}_j$, A_j, and a_j may then be used by a numerical optimization procedure performing an ascent on the bound $\hat{I}(\mathsf{x},y)$.

Finally, we note once again that we can view the procedure of optimizing the bound (5.3) as a way of training the generative model $\mathcal{M}_L$. Once the model is trained, we can use it for generating new training data. If optimization of the bound $\hat{I}(\mathsf{x},y)$ helps to avoid degeneracies of the $\mathcal{L}$-optimal solutions, we may expect that the generated samples may be more representative of the underlying clusters than samples from the mixture model trained by maximizing the likelihood. Specifically, we may hope that the information-theoretic training may limit over-representations of the local segments of the training data, which may occur due to possible singularities of the likelihood solutions.

5.2.1.2 Relation to the K-means Algorithm

Interestingly, we may note that in the limiting case of deterministic *decoders* $p(\mathsf{x}|y_j) \sim \mathcal{N}_{\mathsf{x}}(\boldsymbol{\mu}_j, \mathbf{\Sigma}_j)$ with $\mathbf{\Sigma}_j = \sigma^2\mathsf{I} \to 0$, optimization of $\hat{I}(\mathsf{x},y)$ reduces to the well-known k-means algorithm. Indeed, for this case the bound on the mutual information may be expressed as

$$\begin{aligned}\hat{I}(\mathsf{x},y) &= \langle \log p(\mathsf{x}|y)\rangle_{p(y|\mathsf{x})\tilde{p}(\mathsf{x})} \propto -\sum_{m=1}^{M}\sum_{j=1}^{|y|} \|\mathsf{x}^{(m)} - \boldsymbol{\mu}_j\|^2 p(y_j|\mathsf{x}^{(m)}) \\ &= -\sum_{m=1}^{M}\sum_{j=1}^{|y|} \|\mathsf{x}^{(m)} - \boldsymbol{\mu}_j\|^2 \frac{p(y_j)}{p(y_j) + \sum_{k\neq j}^{|y|} p(\mathsf{x}^{(m)}|y_k)p(y_k)/p(\mathsf{x}^{(m)}|y_j)},\end{aligned} \tag{5.11}$$

where we have ignored the entropy of the empirical distribution since it has no effect on the optimization surface for $p(y|\mathsf{x})$. By analogy with the well-known limiting case of the Gaussian mixtures (see e.g. Bishop (1995)), we can see that in the considered limit the second factor in the summation (5.11) reduces to the Kronecker delta $\delta(j - \arg\min_k \|\mathsf{x}^{(m)} - \boldsymbol{\mu}_k\|)$, which leads to the k-means updates for $\boldsymbol{\mu}_j$ (Hartigan and Wong (1979)). In other words, both the likelihood and the variational information maximization reduce to the k-means algorithm in the limit of zero-variance spherical components of a Gaussian mixture model. This is also a limiting case of a specific Information Bottleneck formulation (e.g. Tishby

does not restrict the covariance space. Non-restrictiveness of the parameterization of $p(y_j)$ trivially follows from the range of (5.9).

et al. (1999), Slonim (2002)) for the model $\mathsf{x} \leftarrow i \rightarrow y$, where $i = 1, \ldots, M$ is the training pattern indicator, $\mathsf{x} : i \rightarrow \mathsf{x}$ is a one-to-one deterministic "pattern-extracting" mapping, and $p(y|i) = p(y|\mathsf{x}^{(i)})$ is the exact posterior of a flat spherical Gaussian mixture model (Still et al. (2004), Still and Bialek (2004)). We may therefore conclude that while likelihood maximization, information bottleneck, and the variational IM define very different methodologies for training different graphical models, their point of intersection for the limiting case of Gaussian mixture models in the noiseless limits is the k-means algorithm.

5.2.2 Information Maximization for Clustering with Encoder Models

We will now discuss a more intuitive approach to information-theoretic clustering by considering encoder models $\mathcal{M}_I \stackrel{\text{def}}{=} \tilde{p}(\mathsf{x})p(y|\mathsf{x})$, where $\{\mathsf{x}\}$ is the set of data patterns, and $\{y\}$ is the set of the corresponding (and generally unknown) cluster labels. As in the standard approaches to information maximization in intrinsically tractable channels, we will maximize the exact mutual information with respect to parameters of the explicitly parameterized encoder $p(y|\mathsf{x})$. In many practical clustering applications, we will have the flexibility of choosing the encoder's parameterization. For clustering, we may be particularly interested in encoders which favour consistent cluster allocations to locally smooth neighborhoods of the training patterns (by this we mean that data patterns which are close to each other according to some distance measure would need to be clustered similarly). As discussed in Chapter 3, one of the fundamental advantages of the information-maximization principle for the encoder models $\mathcal{M}_I$ is the simplicity of imposing such *smoothness* constraints on the encoder distribution; moreover, by aiming to maximize the capacity of the resulting channels, we obtain a proper information-theoretic framework for performing fully unsupervised training. Another potential advantage of clustering with encoder (rather than generative) models is the observation that we do not need to specify the data-generating process to be able to cluster the data properly. Indeed, generative approaches suffer from the fact that $p(\mathsf{x}|y)$ needs to be normalized in x, which in high dimensions restricts the class of the generative models to mixtures of simple distributions (such as Gaussians). Usually data will lie on low dimensional curved manifolds embedded in the high dimensional x-space. If we are restricted to using mixtures of Gaussians to model this curved manifold, typically a very large number of mixture components will be required. No such restrictions apply in the information maximization case, so that the mappings $p(y|\mathsf{x})$ may be very complex, subject only to sensible local constraints (Agakov and Barber (2005b)).

Before discussing applications of the IM to specific nonlinear channels, it is important to outline fundamental differences of the suggested formulation from the existing approaches to information-theoretic clustering and feature extraction. We stress that our approach to clustering is a special case of the general formulation of the information-maximization problem for a specific choice of the code variables. In particular, our goal here is to obtain informative codes (cluster labels) by optimizing the information content $I(\mathsf{x}, y)$ in the undercom-

plete stochastic channel. While the information-maximizing framework seems to be somewhat related to the existing information-theoretic approaches to feature extraction (Principe et al. (2000), Torkkola and Campbell (2000), Gokcay and Principe (2002)), it is in many ways fundamentally different. Specifically, these methods presume a complete observability of the data patterns and the corresponding class labels $\{\mathsf{x}^{(i)}, y^{(i)} | i = 1, \ldots, M\}$, and suggest to maximize the information content $I(\mathsf{z}, y)$ between the *known* cluster labels $\{y\}$ and the hidden transformations $\{\mathsf{z}\}$ of the training patterns $\{\mathsf{x}\}$. Since the resulting mutual information is generally computationally intractable, it is typically approximated by a somewhat heuristic *"information potential"*, which may be related to the quadratic Renyi entropy $H_R(\mathsf{z}, y) \stackrel{\text{def}}{=} -\log\langle p(\mathsf{z}, y)\rangle_{p(\mathsf{z},y)}$. The joint distribution $p(\mathsf{z}, y)$ is usually evaluated by applying Parzen density estimation techniques (see e.g. Principe et al. (1998), Fisher and Principe (1998), Torkkola (2000)). Effectively, these methods may be interpreted as a way of training discriminative models $\mathsf{x} \rightarrow \mathsf{z} \rightarrow y$ by maximizing an approximation of $I(\mathsf{z}, y)$, which is computationally and conceptually different from the information-maximizing formulation for encoder models $\mathsf{x} \rightarrow y$.

Other clustering techniques, which make a recourse to information theory, presume partial observability of the cluster labels (Szummer and Jaakkola (2002), Corduneanu and Jaakkola (2003)). The key idea there is to learn the encoder distribution $p(y|\mathsf{x})$ by maximizing the conditional likelihood for the labeled part of the training set $\{\mathsf{x}^{(i)}, y^{(i)} | i = 1, \ldots, L\}$, *penalized* by the mutual information $I(\mathsf{x}, y)$ for all (labeled and unlabeled) source patterns $\{\mathsf{x}^{(i)} | i = 1, \ldots, M\}$, where $M \geq L$. Clearly, this framework is also different from maximizing mutual information in encoder models; specifically, when applied to clustering of the unlabeled data (i.e. $L = 0$), it would result in *minimization* of the information content between the clusters and the patterns. Of course, this would give rise to independent cluster assignments $p(y|\mathsf{x}) = p(y)$ for unconstrained non-parametric settings. Other methods applied in the vaguely related regularized contexts (Tishby et al. (1999), Slonim et al. (2001), Chechik and Tishby (2002), Dhillon et al. (2002)) are also different from the information-maximizing framework in terms of the conceptual definitions of optimality, channel definitions, observability of the modeled domain, and availability of the additional observable *variables of relevance*.

Our motivation here is to explore whether optimization of the exact mutual information for nonlinear channels may give rise to an anthropomorphically sensible clustering technique. As a simple choice of the encoder for the information-theoretic clustering method, we could trivially consider

$$p(y_j|\mathsf{x}^{(i)}) \propto \exp\{-\|\mathsf{x}^{(i)} - \mathsf{w}_j\|^2/s_j + b_j\}, \tag{5.12}$$

where the cluster centers w_j, the dispersions s_j, and the biases b_j are the encoder parameters to be learned, and we use the notation y_j to indicate that the code variable y is in state j. Clearly, the encoder constraint (5.12) favors local smoothness of the underlying clusters, so that patterns x lying close to specific centers w_j in the *data* space will tend to be clustered together. We may note that parameterization of the channel (5.12) has a strong relation to the posteriors expressed from the isotropic Gaussian mixture models. Indeed, for the specific setting of

$b_j = \log p(y_j) - (|\mathsf{x}|/2)\log s_j$, expression (5.12) would reduce to the channel defined by the posterior of the corresponding Gaussian mixture (see Section 5.2.1). In general, for encoder distributions parameterized by (5.12) we will assume that $\mathsf{b} \stackrel{\text{def}}{=} \{b_j\} \in \mathbb{R}^{|\mathsf{y}|}$ is an unconstrained free parameter, and the optimized objective is the exact mutual information rather than the specific bound $\hat{I}(\mathsf{x},\mathsf{y})$ discussed in Section 5.2.1. In principle, we could consider other choices of the encoder distributions; however (5.12) will prove to be particularly convenient for the kernelized representations which we will introduce at a later stage.

As the dimensionality of the code space $|y|$ in clustering problems is usually relatively small, the summations over the $|y|$ states may be performed exactly, and we may proceed by optimizing the exact mutual information $I(\mathsf{x},y)$. In our case it may be expressed analytically as

$$I(\mathsf{x},y) \propto \sum_{m=1}^{M}\sum_{j=1}^{|y|} p(y_j|\mathsf{x}^{(m)})\log p(y_j|\mathsf{x}^{(m)}) - \sum_{k=1}^{M}\sum_{j=1}^{|y|} p(y_j|\mathsf{x}^{(k)})\log\frac{1}{M}\sum_{l=1}^{M} p(y_j|\mathsf{x}^{(l)}). \tag{5.13}$$

Clearly, $I(\mathsf{x},y)$ may be computed in $O(M^2|y|)$.

5.2.2.1 Learning Optimal Parameters

Objective (5.13) needs to be optimized with respect to the encoder parameters (w_j, s_j, and b_j for the channels defined by (5.12)). By computing the functional derivatives of (5.13) for the encoders $p(y|\mathsf{x})$, we get

$$\frac{\partial I(\mathsf{x},y)}{\partial p(y_j|\mathsf{x}^{(m)})} = \frac{1}{M}\log\frac{p(y_j|\mathsf{x}^{(m)})}{p(y_j)}. \tag{5.14}$$

This implies

$$\frac{\partial I(\mathsf{x},y)}{\partial\theta_k} = \frac{1}{M}\sum_{m=1}^{M}\sum_{j=1}^{|y|}\log\frac{p(y_j|\mathsf{x}^{(m)})}{p(y_j)}\frac{\partial p(y_j|\mathsf{x}^{(m)})}{\partial\theta_k}, \tag{5.15}$$

where θ_k parameterizes $p(y_k|\mathsf{x})$ (again, by construction we have presumed that $p(y|\mathsf{x})$ lies in the convex space of the conditional distributions). Note that due to the normalization, the parameters θ_k would occur in the normalizing constants of $p(y_j|\mathsf{x})$ for all $j = 1,\ldots,|y|$, which leads to

$$\frac{\partial p(y_j|\mathsf{x}^{(m)})}{\partial\theta_k} = -p(y_j|\mathsf{x}^{(m)})\left[\frac{\partial f_j(\mathsf{x}^{(m)})}{\partial\theta_k} - \sum_{l=1}^{|y|} p(y_l|\mathsf{x}^{(m)})\frac{\partial f_l(\mathsf{x}^{(m)})}{\partial\theta_k}\right]. \tag{5.16}$$

Here we assumed that $p(y_j|\mathsf{x}^{(m)}) \propto \exp\{-f_j(\mathsf{x}^{(m)})\}$ with the potentials defined as $f_j \stackrel{\text{def}}{=} \|\mathsf{x}^{(i)} - \mathsf{w}_j\|^2/s_j - b_j$. Now we can easily express the gradients of the exact mutual information $I(\mathsf{x},y)$ for the means w_j and the isotropic variances s_j as

$$\frac{\partial I(\mathsf{x},y)}{\partial\mathsf{w}_j} = \frac{1}{M}\sum_{m=1}^{M} p(y_j|\mathsf{x}^{(m)})\frac{(\mathsf{x}^{(m)} - \mathsf{w}_j)}{s_j}\alpha_j^{(m)}, \tag{5.17}$$

$$\frac{\partial I(\mathsf{x},y)}{\partial s_j} = \frac{1}{M}\sum_{m=1}^{M} p(y_j|\mathsf{x}^{(m)})\frac{\|\mathsf{x}^{(m)} - \mathsf{w}_j\|^2}{2s_j^2}\alpha_j^{(m)}, \tag{5.18}$$

where $\alpha_j^{(m)}$ is again the increment of the specific Kullback-Leibler term

$$\alpha_j^{(m)} \stackrel{\text{def}}{=} \alpha_j(\mathsf{x}^{(m)}) \stackrel{\text{def}}{=} \log \frac{p(y_j|\mathsf{x}^{(m)})}{p(y_j)} - KL\left(p(y|\mathsf{x}^{(m)}) \| \langle p(y|\mathsf{x}) \rangle_{\tilde{p}(\mathsf{x})}\right). \tag{5.19}$$

Note that in contrast to the weighting coefficients (5.6) of the IM training of Gaussian mixture models, the KL divergence in (5.19) is computed between the posterior $p(y|\mathsf{x})$ and their empirical (rather than model-based) average. Analogously, we get

$$\frac{\partial I(\mathsf{x}, y)}{\partial b_j} = \frac{1}{M} \sum_{m=1}^{M} p(y_j|\mathsf{x}^{(m)}) \alpha_j^{(m)}. \tag{5.20}$$

It is easy to see that apart from the variable factors, the fixed point updates (5.17) and (5.18) are identical to (5.4) and (5.5) with the assumed isotropic covariances. A simple construction constraint ensuring that $s_j > 0$ may be given by $s_j \stackrel{\text{def}}{=} \exp\{\tilde{s}_j\}$ where $\tilde{s}_j \in \mathbb{R}$. For this case, we may re-express the gradients for the variances as

$$\partial I(\mathsf{x}, y)/\partial \tilde{s}_j = s_j \partial I(\mathsf{x}, y)/\partial s_j. \tag{5.21}$$

Expressions (5.17) – (5.21) could then be used by a numerical optimization procedure performing an ascent on $I(\mathsf{x}, y)$ for w_j, $\tilde{s}_j$, and b_j, where $j = 1, \ldots, |y|$, as discussed in Agakov and Barber (2005c).

5.2.3 Information Maximization for Clustering with Kernelized Encoder Models

We will now extend (5.12) by considering kernelized (Aizerman et al. (1964), Boser et al. (1992)) parameterizations of encoders $p(y_j|\mathsf{x})$. Let us assume that the source patterns $\mathsf{x}^{(i)}$, $\mathsf{x}^{(j)}$ should have a high probability of being assigned to the same cluster, if they lie close to a specific cluster center in some *feature* space. A reasonable choice of the encoder distribution for this case is given by

$$p(y_j|\mathsf{x}^{(i)}) \propto \{-\|\boldsymbol{\phi}(\mathsf{x}^{(i)}) - \mathsf{w}_j\|^2/s_j + b_j\}, \tag{5.22}$$

where $\boldsymbol{\phi}(\mathsf{x}^{(i)}) \in \mathbb{R}^{|\boldsymbol{\phi}|}$ is the feature vector corresponding to the source pattern $\mathsf{x}^{(i)}$, and $\mathsf{w}_j \in \mathbb{R}^{|\boldsymbol{\phi}|}$ is the (unknown) cluster center in the feature space. Typically, we will presume intractability of explicit computations in feature spaces (which may occur, for example, when $|\boldsymbol{\phi}| \to \infty$). We will also presume that it is impossible to store vectors in the feature space, which enforces constraints on the encoder parameters $\mathsf{w}_j \in \mathbb{R}^{|\boldsymbol{\phi}|}$.

Note that since each cluster center w_i in the feature space $\mathbb{R}^{|\boldsymbol{\phi}|}$ has the same dimensionality as the projected source patterns $\boldsymbol{\phi}(\mathsf{x}^{(i)})$, it is representable in the basis of the projections as

$$\mathsf{w}_j = \sum_{m=1}^{M} \alpha_{mj} \boldsymbol{\phi}(\mathsf{x}^{(m)}) + \mathsf{w}_j^{\perp}, \tag{5.23}$$

where $\tilde{\mathsf{w}}_i^{\perp} \in \mathbb{R}^{|\boldsymbol{\phi}|}$ is orthogonal to the span of $\boldsymbol{\phi}(\mathsf{x}_1), \ldots, \boldsymbol{\phi}(\mathsf{x}_M)$, and $\{\alpha_{mj}\}$ is a set of coefficients (here j and m index $|y|$ codes and M patterns respectively). Then we may transform the encoder distribution (5.22) to

$$\begin{aligned} p(y_j|\mathsf{x}^{(m)}) &\propto \exp\left\{-\left(K_{mm} - 2\mathsf{k}^T(\mathsf{x}^{(m)})\mathsf{a}_j + \mathsf{a}_j^T\mathsf{K}\mathsf{a}_j + c_j\right)/s_j\right\} \\ &\stackrel{\text{def}}{=} \exp\{-f_j(\mathsf{x}^{(m)})\}, \end{aligned} \tag{5.24}$$

where $\mathsf{k}(\mathsf{x}^{(m)})$ corresponds to the m^{th} column (or row) of the Gram matrix $\mathsf{K} \stackrel{\text{def}}{=} \{K_{ij}\} \stackrel{\text{def}}{=} \{\boldsymbol{\phi}(\mathsf{x}^{(i)})^T\boldsymbol{\phi}(\mathsf{x}^{(j)})\} \in \mathbb{R}^{M\times M}$, $\mathsf{a}_j \in \mathbb{R}^M$ is the j^{th} column of the matrix of the coefficients $\mathsf{A} \stackrel{\text{def}}{=} \{a_{mj}\} \in \mathbb{R}^{M\times|y|}$, and $c_j = (\mathsf{w}_j^{\perp})^T\mathsf{w}_j^{\perp} - s_j b_j$. Without loss of generality, we may assume that $\mathsf{c} = \{c_j\} \in \mathbb{R}^{|y|}$ is a free unconstrained parameter, as the assumption does not limit the parameter's domain. Also, by analogy with the previous case (see expressions (5.12), (5.21)), we will ensure positivity of the dispersions s_j by assuming $s_j = \exp\{\tilde{s}_j\}$. Learning in the encoder model may now be seen as unconstrained optimization of $I(\mathsf{x}, y)$ for the encoder parameters.

5.2.3.1 Learning Optimal Parameters

It is easy to see that for a *fixed* and *known* Gram matrix $\mathsf{K} \in \mathbb{R}^{M\times M}$, the considered construction (5.23) helps us to avoid computations in high-dimensional feature spaces $\mathbb{R}^{|\boldsymbol{\phi}|}$. Indeed, if we know K, the encoder (5.24) is a function of the coefficients $\mathsf{A} \stackrel{\text{def}}{=} \{a_{jm}\} \in \mathbb{R}^{M\times|\mathsf{y}|}$, $\mathsf{c} \in \mathbb{R}^{|y|}$, and s_j for $j = 1, \ldots, |y|$, i.e. evaluation of the encoder potential $f_j(\mathsf{x}^{(m)})$ does not require explicit computations in high-dimensional feature spaces.

The exact mutual information (5.13) should be optimized with respect to the log-dispersions $\tilde{s}_j \equiv \log(s_j)$, biases c_j, and coordinates A in the space spanned by the feature vectors $\{\boldsymbol{\phi}(\mathsf{x}^{(i)})|i = 1, \ldots, M\}$. From expressions (5.15) and (5.24) we get

$$\frac{\partial I(\mathsf{x}, y)}{\partial \mathsf{a}_j} = \frac{1}{s_j}\left\langle p(y_j|\mathsf{x})\left(\mathsf{k}(\mathsf{x}) - \mathsf{K}\mathsf{a}_j\right)\alpha_j(\mathsf{x})\right\rangle_{\tilde{p}(\mathsf{x})} \in \mathbb{R}^M, \tag{5.25}$$

$$\frac{\partial I(\mathsf{x}, y)}{\partial \tilde{s}_j} = \frac{1}{2s_j}\left\langle p(y_j|\mathsf{x}) f_j(\mathsf{x})\alpha_j(\mathsf{x})\right\rangle_{\tilde{p}(\mathsf{x})}. \tag{5.26}$$

Also, by analogy with (5.20) we obtain

$$\partial I(\mathsf{x}, y)/\partial c_j = \langle\alpha_j(\mathsf{x})\rangle_{\tilde{p}(\mathsf{x})}, \tag{5.27}$$

where the coefficients $\alpha_j(\mathsf{x})$ are given by (5.19). For a known Gram matrix K, the gradients $\partial I/\partial \mathsf{a}_j$, $\partial I/\partial \tilde{s}_j$, and $\partial I/\partial c_j$ given by expressions (5.25) – (5.27) may be used for numerical optimization of the model parameters (e.g. Luenberger (1973), Bertsekas (1999)).

Note that so far we have assumed that the Gram matrix $\mathsf{K} \in \mathbb{R}^{M\times M}$ is fixed and known. Clearly, in general it needs to be recomputed for each new dataset. Moreover, in addition to learning the feature space coordinates A, we could potentially consider learning optimal features, which would generally lead to changes in the matrices of scalar products. Since we presume that explicit computations

in $\mathbb{R}^{|\phi|}$ are expensive, we cannot compute the Gram matrix by trivially applying its definition $\mathsf{K} = \{\boldsymbol{\phi}(\mathsf{x}_i)^T\boldsymbol{\phi}(\mathsf{x}_j)\}$. Instead, we may interpret scalar products in feature spaces as *kernel functions*

$$\boldsymbol{\phi}(\mathsf{x}^{(i)})^T\boldsymbol{\phi}(\mathsf{x}^{(j)}) = \mathcal{K}_{\boldsymbol{\Theta}}(\mathsf{x}^{(i)}, \mathsf{x}^{(j)}; \boldsymbol{\Theta}), \quad \forall \mathsf{x}^{(i)}, \mathsf{x}^{(j)} \in \mathcal{R}_{\mathsf{x}}, \tag{5.28}$$

where $\mathcal{K}_{\boldsymbol{\Theta}} : \mathcal{R}_{\mathsf{x}} \times \mathcal{R}_{\mathsf{x}} \to \mathbb{R}$ satisfies Mercer's kernel properties (symmetry and non-negative definiteness, see e.g. Mercer (1909), Courant and Hilbert (1953), Vapnik (1998), Smola (1998), Cristianini and Shawe-Taylor (2000)). We may therefore evaluate scalar products in feature spaces $K_{ij} = \mathcal{K}_{\boldsymbol{\Theta}}(\mathsf{x}^{(i)}, \mathsf{x}^{(j)})$ by performing computations in the data space $\mathbb{R}^{|\mathsf{x}|}$; the projections into $\mathbb{R}^{|\boldsymbol{\phi}|}$ will in this case be implied.

A number of nonlinear extensions of common models (applicable in a variety of contexts: from unsupervised dimensionality reduction to supervised classification) typically consider specific kernel functions with fixed parameters (e.g. see a review in Smola (1998), Cristianini and Shawe-Taylor (2000), Scholkopf and Smola (2002)). This raises a natural question of finding optimal settings of kernel parameters, which would implicitly correspond to finding optimal nonlinearities for constrained function spaces. The problem of learning optimal kernels may be conveniently addressed in the (supervised) context of probabilistic classification or regression, with Gaussian process priors over regression functions (Williams and Rasmussen (1996), Williams (1997), Williams (1998)) or the functions' arguments (Barber and Williams (1997), Williams and Barber (1998), Gibbs and MacKay (2000)). Optimal kernel parameters for these cases are typically obtained by maximizing the likelihoods with respect to parameters of covariance functions. Generally, the likelihoods will be computed for a set of labeled patterns. An important difference of our information-theoretic formulation is that it offers a convenient and simple way to learn parameters of kernel functions $\mathcal{K}_{\boldsymbol{\Theta}}$ in the unsupervised context. (Additionally, in contrast to Bishop, Svensen and Williams (1998a), Bishop, Svensen and Williams (1998b), we do not need to impose explicit constraints on the code space representations). For the considered channel definitions (5.22), learning the kernel function parameters is particularly computationally convenient; moreover, choosing appropriate constraints on the kernel matrices in this case is generally less complicated than for some other communication channels (*cf* Agakov and Barber (2004c)).

5.2.3.2 Learning Optimal Kernels

We can apply our information-maximizing framework to learning of optimal kernel parameters. Indeed, from (5.13) and (5.15) we can derive the gradients of the exact mutual information with respect to the parameters $\boldsymbol{\Theta}$ of the kernel function $\mathcal{K}_{\boldsymbol{\Theta}}$. After some algebraic manipulations, we get

$$\begin{aligned}\frac{\partial I(\mathsf{x}, y)}{\partial \boldsymbol{\Theta}} &= \frac{1}{M}\sum_{m=1}^{M} KL(p(y|\mathsf{x}^{(m)})\|p(y)) \sum_{k=1}^{|y|} \frac{\partial f_k(\mathsf{x}^{(m)})}{\partial \boldsymbol{\Theta}} p(y_k|\mathsf{x}^{(m)}) - \\ &\quad \frac{1}{M}\sum_{m=1}^{M}\sum_{j=1}^{|y|} \frac{\partial f_j(\mathsf{x}^{(m)})}{\partial \boldsymbol{\Theta}} p(y_j|\mathsf{x}^{(m)}) \log \frac{p(y_j|\mathsf{x}^{(m)})}{p(y_j)},\end{aligned} \tag{5.29}$$

where $f_k(\mathsf{x}^{(m)})$ is given by (5.24). Note that the computational complexity of computing the updates for $\boldsymbol{\Theta}$ is $O(M|y|^2)$, where M is the number of training patterns and $|y|$ is the number of clusters. It is easy to see that generally (5.29) does not require further approximations. Also, we may note that neither computation of the objective (5.13), nor computation of its gradients (5.25) – (5.27), (5.29) requires inversion of the Gram matrix $\mathsf{K} \in \mathbb{R}^{M \times M}$.

In the special case of the radial basis function (RBF) kernels

$$\mathcal{K}_\beta(\mathsf{x}^{(i)}, \mathsf{x}^{(j)}) = \exp\{-\beta \|\mathsf{x}^{(i)} - \mathsf{x}^{(j)}\|^2\}, \tag{5.30}$$

the gradients of the encoder potentials are given by

$$\frac{\partial f_j(\mathsf{x}^{(m)})}{\partial \beta} = \frac{1}{s_j} \left(\mathsf{a}_j^T \tilde{\mathsf{K}} \mathsf{a}_j - 2\tilde{\mathsf{k}}^T(\mathsf{x}^{(m)}) \mathsf{a}_j \right), \tag{5.31}$$

where $\tilde{\mathsf{K}} \stackrel{\text{def}}{=} \{\tilde{K}_{ij}\} \stackrel{\text{def}}{=} K(\mathsf{x}^{(i)}, \mathsf{x}^{(j)})(1 - \delta(\mathsf{x}^{(i)} - \mathsf{x}^{(j)}))$, and δ is the Kronecker delta. By substituting (5.31) into the general expression (5.29), we obtain the gradient of the mutual information with respect to the RBF kernel parameters (Agakov and Barber (2005b)). As usual, the kernel parameter may be learned by performing a numerical ascent on the objective (5.13). Similarly, we may learn continuous parameters of other kernel functions, or use the mutual information (5.13) as a criterion for kernel comparison.

The discussed framework suggests a proper information-theoretic approach to clustering by projecting the data to potentially high-dimensional feature spaces. Importantly, we stress that in contrast to other techniques performing clustering in feature spaces (such as the kernelized k-means, e.g. Dhillon et al. (2004), Wang et al. (2004)), the information-maximizing framework suggests a principled way of learning optimal kernels. Moreover, the proper information-theoretic interpretation may facilitate extensions of the method to richer channel distributions.

5.3 Nonlinear Gaussian Channels

In Section 5.2.3 we assumed that the code space was low-dimensional and discrete (i.e. $y \in \{y_1, \ldots, y_{|y|}\}$), which made it possible to optimize the exact mutual information $I(\mathsf{x}, y)$ analytically. For that case, we considered several kinds of nonlinear Gaussian encoders (see (5.2), (5.12), (5.22)) and outlined the corresponding IM learning rules. We also showed that if the nonlinear channel was defined by the exact posterior of a Gaussian mixture model, the limiting noiseless case of the bound (5.3) gave rise to the well-known k-means clustering algorithm. Here we consider a different family of communication channels, where the code space is high-dimensional and continuous (i.e. $\mathsf{y} \in \mathbb{R}^{|\mathsf{y}|}$). Specifically, we focus on applying the variational information-maximizing framework to high-dimensional nonlinear channels $\mathsf{x} \to \mathsf{y}$ with the independent isotropic Gaussian noise. The encoder distribution we consider is given by $p(\mathsf{y}|\mathsf{x}) \sim \mathcal{N}_\mathsf{y}(\mathsf{W}\boldsymbol{\phi}(\mathsf{x}), s^2\mathsf{I}_{|\mathsf{y}|})$, where $\mathsf{W} \in \mathbb{R}^{|\mathsf{y}| \times |\boldsymbol{\phi}|}$. By analogy with Section 5.2.3, we presume that explicit computations in $\mathbb{R}^{|\boldsymbol{\phi}|}$ are expensive, and work with the kernelized encoders instead. For this case, we

consider optimizing the generic bound (2.2) for several choices of the variational decoder distributions $q(\mathsf{x}|\mathsf{y})$.

Our objective here is to outline several important properties of the IM approach for Gaussian channels and tractable decoding distributions. Specifically, as the first obvious choice of $q(\mathsf{x}|\mathsf{y})$, we consider using linear Gaussian decoders, which simplifies computations of the bound $\tilde{I}(\mathsf{x},\mathsf{y})$ and facilitates the analysis of optimal solutions for encoder parameters. We show that for the case of isotropic channel noise, nothing is gained by using nonlinear encoders and linear decoders in the context of variational information maximization. This result extends the work of Bourlard and Kamp (1988) and Bourlard (2000) to stochastic communication channels with arguably more general choices of encoding nonlinearities.

Then we consider variational information maximization for the case of nonlinear Gaussian variational distributions. Generally, such choices of $q(\mathsf{x}|\mathsf{y})$ may significantly increase the complexity of computing the generic lower bound $\tilde{I}(\mathsf{x},\mathsf{y})$, which motivates further reformulations of the variational procedure. We find that the generic lower bound $\tilde{I}(\mathsf{x},\mathsf{y})$ may indeed be formally modified to ensure tractable computations, which leads to kernel PCA (Schoelkopf et al. (1998)) as the optimal solution for *encoder* weights $\mathsf{W} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{x}|}$. By analogy with a simpler case of discrete nonlinear channels (see Section 5.2.3), the IM suggests a formal procedure for learning kernel parameters. However, we point out that in order to avoid degenerate solutions for parameters of the encoder model, it may generally be important to impose constraints on the Gram matrix $\mathsf{K} \in \mathbb{R}^{M\times M}$, or carefully choose the feature-to-data decoding mappings.

Finally, we outline a simple relation of our framework to the recent work on Gaussian Process Latent Variable Models (Lawrence (2003)), which may be interpreted as the variational information-maximization procedure in the *noiseless* limit of a nonlinear channel. The presentation in this section summarizes the obtained theoretical results. For their derivations and extended discussions we refer the reader to Appendix C.

5.3.1 Analytic Properties of IM Solutions

As discussed in Section 1.4, optimization of the exact mutual information for high-dimensional code spaces is generally computationally intractable. Here we describe optimization of the bound

$$I(\mathsf{x},\mathsf{y}) \geq \tilde{I}(\mathsf{x},\mathsf{y}) \stackrel{\text{def}}{=} H(\mathsf{x}) + \langle \log q(\mathsf{x}|\mathsf{y}) \rangle_{p(\mathsf{x})p(\mathsf{y}|\mathsf{x})}$$

for the case of a nonlinear Gaussian encoder $p(\mathsf{y}|\mathsf{x}) \sim \mathcal{N}_{\mathsf{y}}(\mathsf{W}\boldsymbol{\phi}(\mathsf{x}), s^2\mathsf{I}_{|\mathsf{y}|})$. Note that for the data patterns $\{\mathsf{x}^{(m)}|m = 1,\ldots,M\}$, the set of encodings $\{\mathsf{y}^{(m)}\}$ is given by a noisy linear projection from the (potentially high-dimensional) feature space $\{\boldsymbol{\phi}(\mathsf{x}^{(m)})\}$. We also assume that $|\mathsf{y}| \leq |\mathsf{x}| \ll |\boldsymbol{\phi}|$ and $|\mathsf{x}| \leq M$, so that y is a compressed representation of x, and the number of training points is sufficient to ensure invertibility of the sample covariance (for linearly independent patterns). We will focus primarily on the discussion of the IM framework for the cases when variational decoders $q(\mathsf{x}|\mathsf{y})$ are linear and nonlinear Gaussians.

5.3.1.1 Linear Gaussian Decoders

Let us assume that $p(\mathsf{y}|\mathsf{x}) \sim \mathcal{N}_{\mathsf{y}}(\mathsf{W}\boldsymbol{\phi}(\mathsf{x}), s^2\mathsf{I}_{|\mathsf{y}|})$, and the variational decoder is a linear Gaussian $q(\mathsf{x}|\mathsf{y}) \sim \mathcal{N}_{\mathsf{x}}(\mathsf{U}\mathsf{y}, \sigma^2\mathsf{I}_{|\mathsf{x}|})$. By analogy with (5.23), we note that the encoder weights are representable in the basis of feature vectors $\{\boldsymbol{\phi}(\mathsf{x}^{(i)})^T | i = 1, \ldots, M\}$ as

$$\mathsf{W} = \mathsf{A}\mathsf{F}^T + \mathsf{W}^{\perp} \in \mathbb{R}^{|\mathsf{y}|\times|\boldsymbol{\phi}|}, \quad \mathsf{F} \stackrel{\text{def}}{=} \left[\boldsymbol{\phi}(\mathsf{x}^{(1)}),\ \ldots,\ \boldsymbol{\phi}(\mathsf{x}^{(M)})\right] \in \mathbb{R}^{|\boldsymbol{\phi}|\times M}, \tag{5.32}$$

where $\mathsf{A} = \{\alpha_{ij}\} \in \mathbb{R}^{|\mathsf{y}|\times M}$ is the matrix of coefficients, and the rows of $\mathsf{W}^{\perp}$ are the corresponding orthogonal compliments (see expression (5.23)). Also, by analogy with Section 5.2.3 we may define the Gram matrix of scalar products in $\mathbb{R}^{|\boldsymbol{\phi}|}$ as

$$\mathsf{K} \stackrel{\text{def}}{=} \{K_{ij}\} \stackrel{\text{def}}{=} \{\boldsymbol{\phi}(\mathsf{x}^{(i)})^T\boldsymbol{\phi}(\mathsf{x}^{(j)})\} = \mathsf{F}^T\mathsf{F} \in \mathbb{R}^{M\times M}. \tag{5.33}$$

After some straight-forward algebraic manipulations (see Appendix C.2), we may express the bound (2.2) on $I(\mathsf{x}, \mathsf{y})$ as

$$\tilde{I}(\mathsf{x}, \mathsf{y}) \propto \frac{1}{M\sigma^2}\mathrm{tr}\left\{\mathsf{U}\mathsf{A}\mathsf{B}^T\right\} - \frac{s^2}{\sigma^2}\mathrm{tr}\left\{\mathsf{U}^T\mathsf{U}\right\} - \frac{1}{2M\sigma^2}\mathrm{tr}\left\{\mathsf{U}^T\mathsf{U}\mathsf{A}\mathsf{K}^2\mathsf{A}^T\right\} - \frac{1}{2\sigma^2}\mathrm{tr}\left\{\mathsf{S}\right\} + c, \tag{5.34}$$

where $\mathsf{S} \stackrel{\text{def}}{=} \langle\mathsf{x}\mathsf{x}^T\rangle = \sum_m \mathsf{x}^{(m)}(\mathsf{x}^{(m)})^T/M \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{x}|}$ is the sample covariance of the centered data, $\mathsf{B} \stackrel{\text{def}}{=} \sum_{m=1}^{M} \mathsf{x}^{(m)}\mathsf{k}(\mathsf{x}^{(m)})^T \in \mathbb{R}^{|\mathsf{x}|\times M}$, $\mathsf{k}(\mathsf{x}^{(m)}) \in \mathbb{R}^{M\times 1}$ is the m^{th} column of the Gram matrix, and c is an irrelevant constant. If $\mathsf{K} \in \mathbb{R}^{M\times M}$ is fixed, the objective (5.34) needs to be optimized for the encoder coefficients and decoder weights.

By optimizing the bound (5.34) with respect to the encoder coefficients $\mathsf{A} \in \mathbb{R}^{|\mathsf{y}|\times M}$ and plugging the optimal values back into (5.34), we may express the objective $\tilde{I}(\mathsf{x}, \mathsf{y})$ as a function of the *decoder* weights $\mathsf{U} \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{y}|}$. Interestingly, in our case this leads to

$$\tilde{I}(\mathsf{x}, \mathsf{y}) = \frac{1}{2\sigma^2}\mathrm{tr}\left\{\mathsf{U}(\mathsf{U}^T\mathsf{U})^{-1}\mathsf{U}^T\mathsf{S}\right\} - \frac{s^2}{2\sigma^2}\mathrm{tr}\left\{\mathsf{U}\mathsf{U}^T\right\} - \frac{1}{2\sigma^2}\mathrm{tr}\left\{\mathsf{S}\right\} + c, \tag{5.35}$$

which is exactly the expression for the optimal bound for *linear* Gaussian communication channels (see Appendix C.1 and expression (C.7)). The result is derived in the context of variational information maximization for a stochastic channel $\mathsf{x} \to \mathsf{y}$, and it holds independently of the specific form of the nonlinearity $\boldsymbol{\phi}(\mathsf{x})$. Indeed, we have made no assumptions about the mappings to the feature space. Hence, we reach an important conclusion: in the considered variational framework, *nothing is gained by using a nonlinear encoder and a linear Gaussian variational decoder for isotropic channel noise*. This extends the related work of Bourlard and Kamp (1988) and Bourlard (2000), who showed that for noiseless 1-layer autoencoders, there are no gains of using specific nonlinearities[4].

By analogy with Appendix C.1 we may note that the optimal left singular vectors of $\mathsf{U} \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{y}|}$ span the subspace defined by the $|\mathsf{y}|$ principal components

[4]To show this, Bourlard (2000) considers noiseless autoencoders with the encoding functions approximately linear around $x = 0$, i.e. $\phi(x) \approx a_0 + a_1 x$.

of S. Moreover, for both linear and nonlinear Gaussian encoders, the bound (5.35) is maximized when the weights are unconstrained (see the discussion in Appendix C.1). By imposing norm constraints on $\mathsf{U} \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{y}|}$, both channels result in the same optimization surface for the decoder weights independently of the choice of nonlinearity.

The result suggests that in order to improve the power of the method, we need to consider both nonlinear encoders and decoders. However, from (2.2) it is clear that in the stochastic context, the naive approach of using a nonlinear decoder will typically result in intractable averages over y in the expression for the variational bound $\tilde{I}(\mathsf{x},\mathsf{y})$. In order to avoid the computational problems, we derive a modified bound on the mutual information by considering further relaxations of the generic bound and performing decoding in the feature space.

5.3.1.2 Nonlinear Gaussian Decoders

The considered nonlinear Gaussian channel may be represented by the Markov chain $\mathsf{x} \to \mathsf{f} \to \mathsf{y}$, where $\mathsf{f} \in \mathbb{R}^{|\boldsymbol{\phi}|}$ and $p(\mathsf{f}|\mathsf{x}) \sim \boldsymbol{\delta}(\mathsf{f} - \boldsymbol{\phi}(\mathsf{x}))$, $p(\mathsf{y}|\mathsf{f}) \sim \mathcal{N}_{\mathsf{y}}(\mathsf{W}\mathsf{f}, \boldsymbol{\Sigma}_y)$. Intuitively, since projections to the feature space are deterministic, the codes y are as predictable from the feature vectors $\mathsf{f} \stackrel{\text{def}}{=} \boldsymbol{\phi}(\mathsf{x}) \in \mathbb{R}^{|\boldsymbol{\phi}|}$ as they are from the source variables x (see Appendix C.3 and proposition C.1). Then the bound on $I(\mathsf{x},\mathsf{y})$ may be expressed as

$$I(\mathsf{x},\mathsf{y}) = I(\mathsf{f},\mathsf{y}) \geq \tilde{I}(\mathsf{f},\mathsf{y}), \text{ where } \tilde{I}(\mathsf{f},\mathsf{y}) \stackrel{\text{def}}{=} \langle \log q(\mathsf{f}|\mathsf{y}) \rangle_{\tilde{p}(\mathsf{x})p(\mathsf{f}|\mathsf{x})p(\mathsf{y}|\mathsf{f})} + H(\mathsf{f}). \quad (5.36)$$

Evaluation of the bound (5.36) is complicated by the need of computing the entropy of the features $H(\mathsf{f})$. Despite the fact that the mapping to the feature space is deterministic, we have assumed that we do not generally know explicit feature space representations of the training patterns, i.e. numeric approximations due to Brunel and Nadal (1998), Shriki et al. (2002), Corduneanu and Jaakkola (2003) are not directly applicable. Moreover, even if explicit computations in the feature space were possible, the existing numerical methods would not generally retain a proper bound on $I(\mathsf{x},\mathsf{y})$, i.e. other ways of handling the intractability may need to be considered.

In Appendix C.3.1 and C.3.2 we discuss several ways of addressing the problem of evaluating the feature space entropy in the context of the optimization problem. As one of such methods, we suggest optimizing a proper variational relaxation of (5.36) given by

$$I(\mathsf{x},\mathsf{y}) \geq \tilde{I}(\mathsf{x},\mathsf{y}) \geq \langle \log q(\mathsf{f}|\mathsf{y}) \rangle_{\tilde{p}(\mathsf{x})p(\mathsf{f}|\mathsf{x})p(\mathsf{y}|\mathsf{f})} + \langle \log q(\mathsf{x}|\mathsf{f}) \rangle_{\tilde{p}(\mathsf{x})p(\mathsf{f}|\mathsf{x})} + c, \quad (5.37)$$

where c is an irrelevant constant (note that it corresponds to a sequential application of the generic bound for the chain $\mathsf{x} \to \mathsf{f} \to \mathsf{y}$). As in the conventional applications of the variational IM algorithm (see Section 2.1.2), we will be optimizing (5.37) with respect to parameters of the encoder $p(\mathsf{y}|\mathsf{f})$, the *feature decoder* $q(\mathsf{f}|\mathsf{y})$, and the *data decoder* $q(\mathsf{x}|\mathsf{f})$. Note that objective (5.37) is a proper bound on $I(\mathsf{x},\mathsf{y})$ independently of the specific parameterization of the encoder and the variational decoder distributions.

By assuming that the *feature space* decoder is an isotropic linear Gaussian $q(\mathsf{f}|\mathsf{y}) \sim \mathcal{N}_{\mathsf{f}}(\mathsf{U}\mathsf{y}, \sigma_f^2\mathsf{I})$ and considering kernelized representations of $p(\mathsf{y}|\mathsf{f})$ and $q(\mathsf{f}|\mathsf{y})$, it is easy to see that the optimal encoder and decoder weights $\mathsf{W}^T, \mathsf{U} \in \mathbb{R}^{|\boldsymbol{\phi}|\times|\mathsf{y}|}$ span the principal eigen-subspace of

$$\mathsf{S}_F \;=\; \frac{1}{M}\sum_{i=1}^{M} \boldsymbol{\phi}(\mathsf{x}^{(i)})\boldsymbol{\phi}(\mathsf{x}^{(i)})^T. \tag{5.38}$$

Hence, *for a fixed kernel function $\mathcal{K}_{\boldsymbol{\Theta}}$ and data decoder $q(\mathsf{x}|\mathsf{f})$ of the considered nonlinear Gaussian channel, the variational lower bound $\tilde{I}(\mathsf{x},\mathsf{y})$ is maximized by the kernel PCA solutions for encoder and feature decoder weights* (see Appendix C.3.3 for derivations and discussions of non-centered representations). As expected, we obtain the linear PCA for the special case of linear mappings $\boldsymbol{\phi}(\mathsf{x}) \equiv \mathsf{x}$.

A few further comments about the objective (5.37) are in order. First of all, we note that analytical properties of the *data decoder* depend on the specific parameterization of $q(\mathsf{x}|\mathsf{f})$. Since integration over x and f reduces to evaluations of the empirical averages, the average $\langle \log q(\mathsf{x}|\mathsf{f})\rangle_{p(\mathsf{x},\mathsf{f})}$ may typically be easily computed (provided that $q(\mathsf{x}|\mathsf{f})$ is appropriately kernelized, so that no explicit computations in the feature space are performed). Having specified the data decoder, we may optimize (5.37) to learn parameters of the kernel function $\mathcal{K}_{\boldsymbol{\Theta}}$, which parameterizes scalar products in $\mathbb{R}^{|\boldsymbol{\phi}|}$. While this strategy may indeed lead to nontrivial optimization surfaces for kernel parameters, our current experience shows that the results may be strongly influenced by the specific definitions of the feature-to-data mappings[5]. Nevertheless, while any practical application of the framework may require a careful consideration of appropriate choices for $q(\mathsf{x}|\mathsf{f})$ and $\mathcal{K}_{\boldsymbol{\Theta}}$, the considered objective (5.2.3) indeed defines a theoretically rigorous tractable way of optimizing the bound on the mutual information in large scale nonlinear Gaussian channels.

5.3.1.3 Noiseless Channels and Gaussian Process Latent Variable Models

The discussed variational framework of maximizing lower bounds on $I(\mathsf{x},\mathsf{y})$ in nonlinear Gaussian channels is particularly useful in situations when the channel $\mathsf{x} \to \mathsf{y}$ is intrinsically noisy. However, in many practical situations when the stochasticity of $p(\mathsf{y}|\mathsf{x})$ is not a necessary modeling requirement, simplifications of the IM approach may be possible. One practical example of such situation is an application of the variational information-maximizing framework to nonlinear dimensionality reduction, where the code space $\{\mathsf{y}\}$ is continuous, and the encoder is deterministic, i.e. $p(\mathsf{y}|\mathsf{x}) \sim \delta$. Interestingly, for a specific parameterization of the variational decoder distribution, this noiseless limit of the variational information-maximizing framework is closely related to the recently introduced (Lawrence (2003)) Gaussian Process Latent Variable Models (GPLVMs).

[5]Intuitively, it is clear that unless the feature vectors are constrained, the bound $\tilde{I}(\mathsf{x},\mathsf{y})$ may diverge due to the contributions of the feature decoder (see expression (5.37) and Appendix C.3.5). The situation is analogous to the case of linear Gaussian channels, where the divergent weights lead to the diminishing noise effects (see e.g. expression (C.8)), which may generally result in the divergence of $I(\mathsf{x},\mathsf{y})$.

In order to show this, let us consider a conventional noiseless autoencoder $\mathsf{x} \mapsto \mathsf{y} \to \tilde{\mathsf{x}}$, with $p(\mathsf{y}|\mathsf{x}) \sim \delta(\mathsf{y} - \mathsf{y}(\mathsf{x}))$ and the empirical distribution $\tilde{p}(\mathsf{x}, \tilde{\mathsf{x}}) \propto \sum_m \delta(\mathsf{x} - \mathsf{x}^{(m)})\delta(\mathsf{x} - \tilde{\mathsf{x}})$. For i.i.d. patterns the exact conditional likelihood training in such models reduces to maximizing

$$\mathcal{L}_{\tilde{\mathsf{x}}|\mathsf{x}} = \left\langle \log \langle q(\tilde{\mathsf{x}}|\mathsf{y}) \rangle_{p(\mathsf{y}|\mathsf{x})} \right\rangle_{\tilde{p}(\mathsf{x},\tilde{\mathsf{x}})} = \langle \log q(\mathsf{x}|\mathsf{y}) \rangle_{p(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x})}, \tag{5.39}$$

where $p(\mathsf{y}|\mathsf{x})$ and $q(\mathsf{x}|\mathsf{y})$ define the encoding and decoding mappings respectively (also see lemma 3.1 and proposition 3.4). If $\{\mathsf{x}\} \stackrel{\text{def}}{=} \{\mathsf{x}^{(m)} | m = 1, \ldots, M\}$ defines the dataset of training patterns with the corresponding codes $\{\mathsf{y}\} \stackrel{\text{def}}{=} \{\mathsf{y}(\mathsf{x}^{(m)}) | m = 1, \ldots, M\}$, the objective (5.39) may be transformed to

$$\mathcal{L}_{\tilde{\mathsf{x}}|\mathsf{x}} = \langle \log \hat{q}(\{\mathsf{x}\}|\{\mathsf{y}\}) \rangle_{\hat{p}(\{\mathsf{y}\}|\{\mathsf{x}\})}. \tag{5.40}$$

Here the decoder $\hat{q}(\{\mathsf{x}\}|\{\mathsf{y}\})$ and the encoder

$$\hat{p}(\{\mathsf{y}\}|\{\mathsf{x}\}) = \prod_{m=1}^{M} \delta\left(\mathsf{y}^{(m)} - \mathsf{y}(\mathsf{x}^{(m)})\right) \tag{5.41}$$

are defined for the whole set of the observed patterns $\{\mathsf{x}\}$ and their deterministic projections $\{\mathsf{y}\}$.

We can now see that if the decoding mapping is defined as

$$\hat{q}(\{\mathsf{x}\}|\{\mathsf{y}\}) \propto \prod_{i=1}^{|\mathsf{x}|} \exp\left\{-\frac{1}{2}\left(\tilde{\mathsf{x}}^{(i)}\right)^T \mathsf{K}^{-1}\tilde{\mathsf{x}}^{(i)}\right\} = \exp\left\{-\frac{1}{2}\mathrm{tr}\left\{\mathsf{X}\mathsf{K}^{-1}\mathsf{X}^T\right\}\right\}, \tag{5.42}$$

the conditional likelihood (5.40) reduces to Neil Lawrence's Gaussian Process Latent Variable Models, optimizing the objective

$$\mathcal{L}_{\tilde{\mathsf{x}}|\mathsf{x}} = -\frac{1}{2}\mathrm{tr}\left\{\mathsf{X}\mathsf{K}^{-1}\mathsf{X}^T\right\} - \frac{|\mathsf{x}|}{2}\log|\mathsf{K}| + const. \tag{5.43}$$

In (5.42) we assumed that $\mathsf{K} \stackrel{\text{def}}{=} \{\mathcal{K}_{\boldsymbol{\Theta}}(\mathsf{y}^{(i)}, \mathsf{y}^{(j)})\} \in \mathbb{R}^{M\times M}$ is the symmetric positive-semidefinite covariance defined by the covariance function $\mathcal{K}_{\boldsymbol{\Theta}}$, and $\left(\tilde{\mathsf{x}}^{(i)}\right)^T \in \mathbb{R}^{1\times M}$ is the i^{th} row of the data matrix $\mathsf{X} = \{\mathsf{x}^{(1)}, \ldots, \mathsf{x}^{(M)}\} \in \mathbb{R}^{|\mathsf{x}|\times M}$. In this formulation learning corresponds to optimizing the conditional likelihood (5.40) for the encoded representations $\{\mathsf{y}^{(i)} | i = 1, \ldots, M\}$ and parameters of the covariance function $\mathcal{K}_{\boldsymbol{\Theta}}$.

On the other hand, it is easy to see that the objective (5.40) may be interpreted as the variational bound on $I(\{\mathsf{x}\}, \{\mathsf{y}\})$ for the **noiseless encoder** (5.41) and the variational decoder $\hat{q}(\{\mathsf{x}\}|\{\mathsf{y}\})$ given by expression (5.42). Effectively, this interpretation offers an alternative, information-theoretic justification of GPLVMs for noiseless encoder models. Note that in contrast to Lawrence (2003), our derivation of (5.40) is independent of the specific assumptions about the model – the GPLVM may be viewed as a special case of the variational information-maximizing framework for noiseless channels with the specific choice of the variational decoder.

5.4 Demonstrations

Here we demonstrate experimental results of maximizing the information content for the discussed discrete nonlinear channels. We will focus specifically on the discussion of information-theoretic clustering (see Section 5.2.1 and Section 5.2.2). First we consider clustering by maximizing the alternative objective function $\hat{I}(\mathsf{x},\mathsf{y})$ for training Gaussian mixture models (see expression (5.1)). As we showed in proposition 3.1, this objective corresponds to a generic lower bound on $I(\mathsf{x},\mathsf{y})$ for the encoder model $\mathcal{M}_I = \tilde{p}(y|\mathsf{x})\tilde{p}(\mathsf{x})$, where the encoder distribution is given by the exact posterior of a Gaussian mixture model $\mathcal{M}_L = p(\mathsf{y})p(\mathsf{x}|\mathsf{y})$, i.e.

$$\tilde{p}(y|\mathsf{x}) \propto p(y)p(\mathsf{x}|y), \quad y \in \{y_1,\ldots,y_{|y|}\},\ p(\mathsf{x}|y_j) \sim \mathcal{N}_{\mathsf{x}}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j).$$

For this case we will compare maximization of the bound $\hat{I}(\mathsf{x},\mathsf{y})$ with maximization of the exact likelihood for Gaussian mixtures, and show that the variational approach may indeed favor more uniform cluster assignments. As the method may be used for training both encoder and generative models (due to the specifics of the channel definition), we will refer to it as IM for Gaussian mixtures.

Then we will apply the information-theoretic clustering method to several intrinsically non-Gaussian datasets. For these tasks, we consider the specific definition of the encoder distribution (5.12) and its kernelized extension (5.24) described in Sections 5.2.2, 5.2.3. We compare both approaches to Gaussian mixture, k-means, and the kernelized k-means clustering, and show that for the considered datasets our method may indeed lead to visible anthropomorphic improvements over the common clustering techniques. Interestingly, we show that by *learning* parameters of kernel functions (see expression (5.31)), we may indeed obtain better visualizations of cluster assignments.

5.4.1 IM for Gaussian Mixture Models

5.4.1.1 $\mathcal{L}$- and $\hat{I}(\mathsf{x},y)$-maximization for Gaussian mixtures

Figure 5.1 shows typical soft cluster allocations produced by a mixture model trained by maximizing the exact likelihood (*left*) and the bound $\hat{I}(\mathsf{x},y)$ on the mutual information (*middle*). The dataset consisted of $M = 150$ training patterns, generated at uniform random from the convex area bounded by a flat triangle (for $|\mathsf{x}| = 2$). The number of mixture components was $|y| = 3$. The initial mixture coefficients were set as $p_0(y_j) = 1/|y|$, the initial covariances were set to be spherical with $\boldsymbol{\Sigma}_j = \mathsf{I}_{|\mathsf{x}|}$, and the components' means were initialized by applying the k-means algorithm. Then the model was trained until convergence by applying the two methods (with the learning rate $\eta = 0.01$ for the IM). Not surprisingly, maximization of the likelihood for the mixture model often resulted in near-singular components' covariances, which led to locally constrained cluster segments (Figure 5.1 (*left*)). On the other hand, under the identical initializations, the IM on $\hat{I}(\mathsf{x},y)$ typically led to more uniform cluster allocations with non-degenerate covariance matrices (Figure 5.1 (*middle*)).

Figure 5.1 (*right*) shows the means and the variances for the proportions of source patterns allocated to each of the $|y|$ clusters for 10 runs of the EM and

Figure 5.1: $\mathcal{L}$- and $\hat{I}$- maximization for a three-component Gaussian mixture ($|\mathsf{y}| = 3$). *Left:* clustering with the EM algorithm on $\mathcal{L}$ (the squares show the cluster centers); *Middle:* clustering with the IM on $\hat{I}(\mathsf{x}, \mathsf{y})$; *Right:* Proportion of the testing points assigned to each of $|\mathsf{y}| = 3$ clusters under $T = 10$ different runs of the optimization procedure (light gray bars indicate the variance). The exact mutual information computed by the EM- and IM-trained models was $I_L(\mathsf{x}, y) \approx 0.75$ and $I_I(\mathsf{x}, y) \approx 1.00$ respectively; the corresponding log-likelihoods are $\mathcal{L}_L \approx -4.74$ and $\mathcal{L}_I \approx -4.99$.

IM algorithms, where the training set was re-sampled from the underlying distribution at each run. For both EM and IM-trained[6] models, the cluster allocation $y(\mathsf{x})$ for a pattern x was given by $y(\mathsf{x}) = \arg\max_y p(y|\mathsf{x})$. Again, the results suggest that on average the models trained by the IM typically result in more uniform class assignments and smaller variances on cluster sizes. This empirically confirms the informal argument of section 3.2.1 that maximization of the bound $\hat{I}(\mathsf{x}, y)$ may favor more spread-out representations in the latent space. We also note that the likelihoods $\mathcal{L}_L$ computed for EM-trained models typically exceed the likelihoods $\mathcal{L}_I$ of Gaussian mixture models trained by maximizing the bound $\hat{I}(\mathsf{x}, y)$ (for the illustrated case we had $\mathcal{L}_L \approx -4.74$ and $\mathcal{L}_I \approx -4.99$). At the same time, the exact mutual information computed for a conventionally trained mixture was typically inferior to that of an IM-trained mixture ($I_L \approx 0.7467$ *vs.* $I_I \approx 0.9950$ for the considered case). Additionally, we see that cluster allocations produced by both training procedures suggest that the models with higher likelihoods do not necessarily lead to better representations of the underlying distribution, which once again illustrates the conceptual problems of maximizing the likelihoods of under-constrained models for learning informative representations of the data.

5.4.1.2 Samples from the Trained Models

As mentioned in Section 5.2.1, our motivation for maximizing the bound $\hat{I}(\mathsf{x}, y)$ was to produce informative latent variable representations y of the sources $\{\mathsf{x}\}$. Importantly, due to the specific parameterization of the encoder distribution, the suggested framework of maximizing $\hat{I}(\mathsf{x}, y)$ could be used for training both encoder and generative models. Our hope was that by maximizing the bound,

[6]Throughout the discussion in this section, we consider a special instance of the IM algorithm for the specific definition of the encoder distribution (5.2).

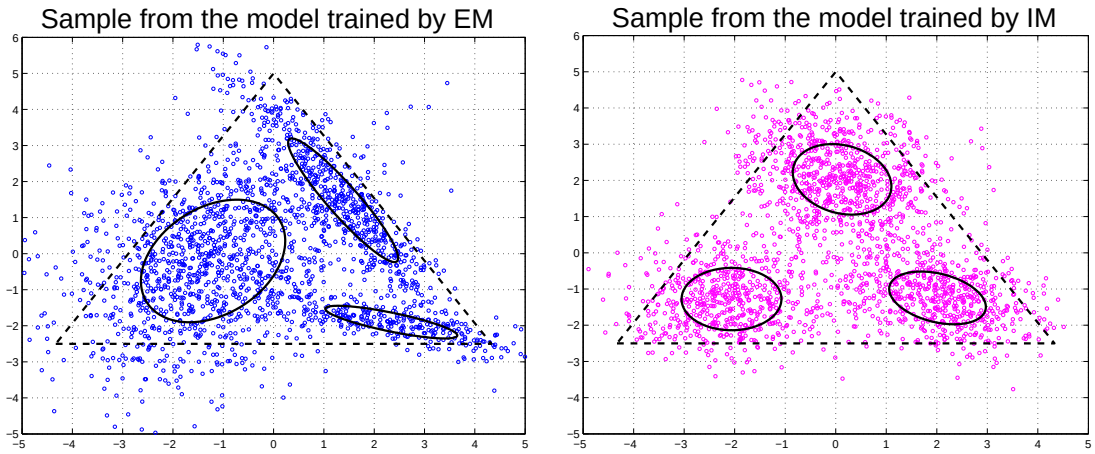


Figure 5.2: Samples from the Gaussian mixture model $\mathcal{M}_L$ trained by the EM- and IM-algorithms with $|\mathsf{x}| = 2$, $|y| = 3$. The number of training patterns was $M = 250$. *Left*: $\mathcal{M}_L$ is trained by maximizing the likelihood $\mathcal{L}$ (resulting in $I_L(\mathsf{x}, y) \approx 0.654$); *Right*: $\mathcal{M}_L$ is trained by maximizing the bound $\hat{I}(\mathsf{x}, \mathsf{y})$ (resulting in $I_I(\mathsf{x}, y) \approx 0.911$). The ellipses show probability contours of the mixture components at $1/2$ Mahalanobis distance from the means $\boldsymbol{\mu}_j$ for $j = 1, \ldots, |y|$. The dashed lines indicate the boundary of the convex region used to generate the underlying data.

the resulting generative models could potentially give rise to samples which were somewhat more representative of the underlying data distribution. Figure 5.2 illustrates typical samples produced by the three-component Gaussian mixture model trained by maximizing the likelihood (*left*) and the bound $\hat{I}(\mathsf{x}, y)$ (*right*). As in the previous case, we assumed that $|\mathsf{x}| = 2$, $|y| = 3$, and both models were trained by starting from identical initializations at the flat component weights, the initial means given by the k-means algorithm, and the initial covariances set as $\boldsymbol{\Sigma}_j = \mathsf{I}_{|\mathsf{x}|}$. For illustration purposes, the training data was sampled from the equiangular triangle, which is more difficult to model by a Gaussian mixture distribution than the data shown on Figure 5.1 (*top left*) (for example, it is clear that in the limiting case when one of the angles approaches zero, the data could be trivially modeled by a single near-singular Gaussian).

We can see that in the considered case the Gaussian mixture model trained by maximizing the likelihood leads to the components fitting local segments of the dataset, with a wide flat component responsible for the remaining training patterns (see Figure 5.2 (*left*)). While the resulting samples have a vague resemblance to the underlying equiangular area, they are arguably less representative of the convex region than the samples produced by the IM-trained mixture (Figure 5.2 (*right*)). Note that the latter model is characterized by flatter spectra of the components' covariances (allowing more uniform space coverage within the constrained area). Moreover, for the IM-trained model the components' means lie close to the corners of the underlying shape, which in this case preserves the higher-level relations between the angles.

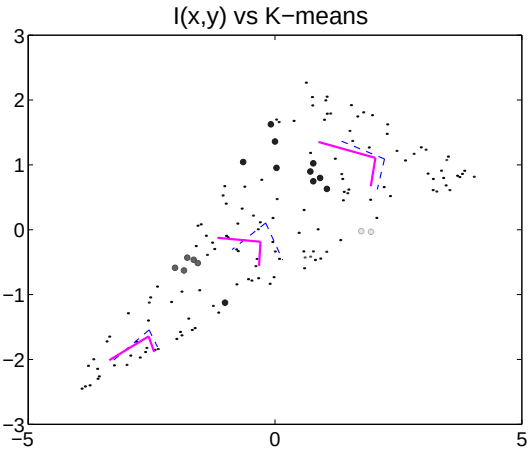


Figure 5.3: $\hat{I}(\mathsf{x}, y)$-maximization *vs.* K-means. *Solid lines:* eigenvectors of the covariances $\boldsymbol{\Sigma}_j$ obtained by the IM. *Dashed lines:* eigenvectors of the sample covariances $\boldsymbol{\Sigma}_j^{KM}$ associated with each of the clusters obtained by the deterministic k-means algorithm. All the eigenvectors are weighted by the corresponding eigenvalues and centered at the components' means ($\boldsymbol{\mu}_j$ and $\boldsymbol{\mu}_j^{KM}$ for the IM and k-means respectively). The IM was initialized at $\boldsymbol{\mu}_j^{KM}$ and $\boldsymbol{\Sigma}_j^{KM}$. The filled circles show the training patterns which are classified differently by the k-means and the IM started at the considered initialization.

5.4.1.3 $\hat{I}(\mathsf{x}, y)$-maximization for Gaussian Mixtures and the K-means Algorithm

In Section 5.2 we showed that in the limiting noiseless case of the decoder distributions $p(\mathsf{x}|y)$, optimization of the specific bound $\hat{I}(\mathsf{x}, y)$ for the considered models reduces to the k-means algorithm. However, in the non-limiting cases, maximization of the specific bound $\hat{I}(\mathsf{x}, y)$ for Gaussian mixtures is different from the k-means, which may be seen analytically from (5.11) or illustrated by the deviations of the IM from the k-means initializations. Figure 5.3 compares the means and covariances $\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j$ of the IM-trained mixture model with the cluster centers and the associated sample covariances $\boldsymbol{\mu}_j^{KM}$, $\boldsymbol{\Sigma}_j^{KM}$ obtained by the k-means algorithm. In the illustrated case, the IM-trained mixture was initialized at $\boldsymbol{\mu}_j^{KM}$ and $\boldsymbol{\Sigma}_j^{KM}$, with the initial mixing coefficients set according to the cluster sizes. Figure 5.3 indicates the changes in the parameters and the resulting cluster allocations. We see that for nonzero covariances, the IM diverges from the initial settings, though for the considered dataset the cluster assignments produced by the k-means and the IM are not very different.

Generally, the results of our experiments in this section have confirmed the intuition that optimization of the bound $\hat{I}(\mathsf{x}, y)$ may result in more uniform cluster assignments than the likelihood-based training of generative models. We also showed that samples from the mixture models trained by maximizing the bound may indeed be qualitatively more representable of the underlying distributions. Extensions of this form of the IM to training other kinds of generative models may potentially be considered, provided that the encoder may be computed exactly.

As demonstrated in this section, the resulting models may be used for clustering and data generation.

5.4.2 Kernelized Information Theoretic Clustering

5.4.2.1 Gaussian Mixture, K-means, and Kernelized Information-Theoretic Clustering for Spiral Data

Here we consider a direct application of the encoder models (5.12) and their kernelized extensions (5.24) to clustering of intrinsically non-Gaussian data. In contrast to the bound $\hat{I}(\mathsf{x}, y)$ optimized in Section 5.4.1, the encoder models considered here were trained by maximizing the exact mutual information $I(\mathsf{x}, y)$ as discussed in Section 5.2.2. For the kernelized encoders (5.22), we also learned optimal kernel parameters (focusing specifically on the RBF kernel, see expression (5.30)). We compare both information-theoretic approaches to the k-means and the Gaussian mixture clustering (assuming the standard EM training).

As discussed earlier in Section 5.4.1, clustering by maximizing the mutual information corresponds to learning of an optimal encoder, where the code space $\{y\}$ defines the generally unknown cluster labels. In the first set of experiments, we considered the discrete code space with $|y| = 3$ states of the output variables. (Effectively, the size of the code space $|y|$ gives an upper bound on the number of clusters; in what follows we assume that this number is fixed). The unlabeled training data was generated as $x_1(t) = t\cos(t)/4$, $x_2(t) = t\sin(t)/4$, with t changing uniformly in $[0, 3.4\pi]$ (here x_1 and x_2 correspond to different coordinates of the sources $\mathsf{x} \in \mathbb{R}^{|\mathsf{x}|}$, $|\mathsf{x}| = 2$). The total number of patterns was $M = 70$.

Figure 5.4 illustrates typical allocations of cluster labels to the considered dataset. The data was clustered by a Gaussian mixture model trained by maximizing the likelihood (Figure 5.4 *left*), the k-means algorithm (Figure 5.4 *medium*), and the kernelized encoder model (5.24) trained by maximizing $I(\mathsf{x}, y)$ (Figure 5.4 *right*). For the considered RBF kernel (5.31), we learned the inverse variance parameter β. Color intensity of each pattern shown on Figure 5.4 is given as an average of the cluster colors weighted by the probability of cluster allocations $p(y_j|\mathsf{x})$. The "pure" cluster colors corresponding to deterministic cluster assignments are shown by light-, medium-, and dark-gray squares.

As we see from Figure 5.4 (*left*), Gaussian mixture models trained by maximizing the likelihood resulted in largely stochastic cluster allocations; from the corresponding plot, it is qualitatively difficult to distinguish between patterns which belong to "medium-gray" and "dark-gray" clusters. Moreover, for the illustrated case we see a large disproportion in cluster sizes. Specifically, we see that 3 or 4 patterns (out of 70) form a unique cluster (shown in light-gray on Figure 5.4 (*left*)). This observation is further confirmed by Figure 5.5 (*left*)), which shows responsibilities of the three mixture components for all 70 data patterns. Figures 5.4 (*middle*) and 5.5 (*middle*) show the corresponding cluster assignments for the k-means algorithm. We see that while the similarly clustered points indeed lie close to each other in $\mathbb{R}^2$ according to the L_2-norm, the allocated clusters are *not* locally smooth in t. (We also note that maximization of the bound (5.1) on $I(\mathsf{x}, \mathsf{y})$ for the encoding distribution corresponding to the exact posterior of a

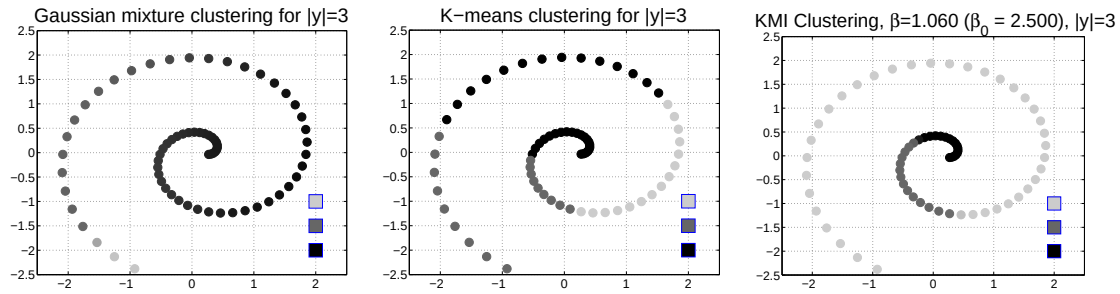


Figure 5.4: Clustering for $|y| = 3$. *Left:* Gaussian mixtures; *Middle:* K-means; *Right:* information-maximization for the (RBF-)kernelized encoder. Light, medium, and dark-gray squares show the cluster colors corresponding to deterministic cluster allocations. The color intensity of each training point $\mathsf{x}^{(m)}$ is the average of the pure cluster intensities, weighted by the responsibilities $p(y_j|\mathsf{x}^{(m)})$. Nearly indistinguishable dark colors for the majority of patterns under the Gaussian mixture clustering indicate soft cluster assignments (see also Figure 5.5). Note that by applying the kernelized IM algorithm, we obtain nearly deterministic cluster assignments to locally smooth data regions.

Gaussian mixture model led to the results similar to a soft form of the k-means clustering. This agrees with previous observations that the cluster allocations produced by a model maximizing (5.1) are qualitatively not very different from the k-means (see Section 5.4.1)).

Figures 5.4 and 5.5 (*right*) illustrate cluster allocations produced by maximizing the mutual information for the kernelized encoder (5.22). As mentioned above, in the considered set of experiments we have applied the RBF kernels (5.31) with adaptable parameters β. The kernel parameters β were initialized at $\beta_0 = 2.5$. The initial settings of the coefficients $\mathsf{A} \in \mathbb{R}^{M\times|y|}$ in the feature space were samples from $\mathcal{N}_{A_{ij}}(0, 0.1)$. The log-variances $\tilde{s}_1, \ldots, \tilde{s}_{|y|}$ were initialized at zeros. The encoder parameters A and $\{\tilde{s}_j|j = 1, \ldots, |y|\}$ (along with the RBF kernel parameter β) were optimized by applying the scaled conjugate gradients procedure (see e.g. Bishop (1995)). The training stopped at the 46^{th} iterations of the algorithm, after the changes in the consecutive evaluations of the objective $I(\mathsf{x}, y)$ remained lower than 10^{-15} for three consecutive iterations. We see that in the considered case we indeed obtain non-degenerate well-separated clusters which are locally smooth in t (see Figure 5.4 (*right*)). The local smoothness of cluster allocations is further confirmed by Figure 5.5 (*right*) (note a slight decrease in coding certainty for the patterns lying close to the cluster boundaries). The results are shown for $\beta \approx 0.825$ obtained by the non-linear ascent on $I(\mathsf{x}, y)$.

Note that in the experiments described here we have made no strong assumptions about the choice of the kernel function which could be particularly suitable for clustering this specific dataset. Undoubtedly, a careful choice of the kernel could potentially lead to a better visualization of the locally smooth, non-degenerate structure. Instead, we considered the commonly used RBF kernel with adaptable kernel parameters. The results suggest that maximization of mu-

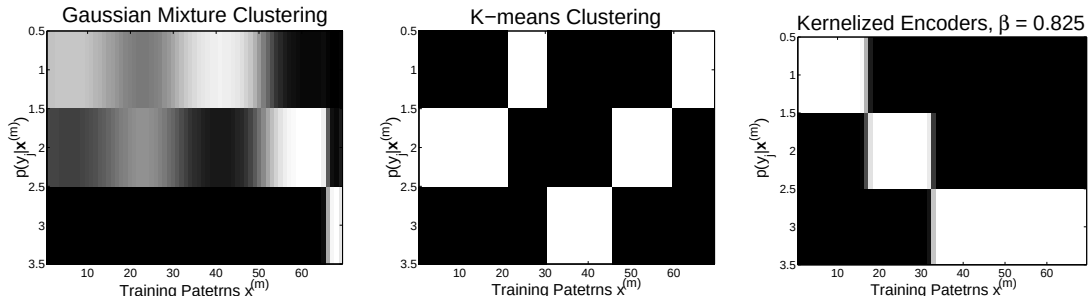


Figure 5.5: Probabilities $p(y_j|\mathsf{x}^{(m)})$ for the spiral data, $\mathsf{x} \in \mathbb{R}^2$, $y \in \{y_1, y_2, y_3\}$. The total number of patterns $M = 70$. *Left:* Gaussian mixtures; *Middle:* K-means; *Right:* information-maximization for the kernelized encoder model (RBF kernel with the learned parameter $\beta = 0.825$).

tual information in discrete nonlinear channels may indeed be a useful technique to consider in unsupervised clustering applications. We will now consider another application of the method, and show that by learning kernel parameters we may indeed obtain better visualizations than by using fixed kernel functions or applying more common clustering techniques.

5.4.2.2 Kernelized Information-Theoretic Clustering for Spatially Translated Letters

Figures 5.6 and 5.7 show an application of the the mutual information maximization to clustering of spatially translated letters, for the code sizes $|y| = 2$ and $|y| = 3$ respectively. As in the previous case shown on Figure 5.4, patterns allocated to different clusters are shown by different color intensities. The data consisted of $M = 210$ patterns, sampled from models of Latin letters[7] (70 per letter). The training set was constructed in such a way that the distance between the neighboring letters was roughly constant (by this we mean that the distances between the sample means for the neighboring letter models were fixed to be the same). In the considered case, we assumed that the neighboring letters were relatively close to each other (compared to the letter sizes), so that points sampled from different letters could in fact be geometrically closer to each other than points sampled from a single letter model. As in the previous set of experiments, we have compared information-theoretic clustering in kernelized encoder models with the k-means and Gaussian mixture clustering. We have also explored the effects of introducing projections into the feature space and applied the IM procedure for both encoder models described in Section 5.2.2 and 5.2.3 (which we will refer to as *simple*, or *non-kernelized* encoders and *kernelized* encoders respectively).

By analogy with the previously discussed experiments (see Figure 5.4), we can see that for the considered case the IM framework indeed results in anthro-

[7]The generative model for each letter is given by a mixture of constrained uniform line segments with additive spherical Gaussian noise.

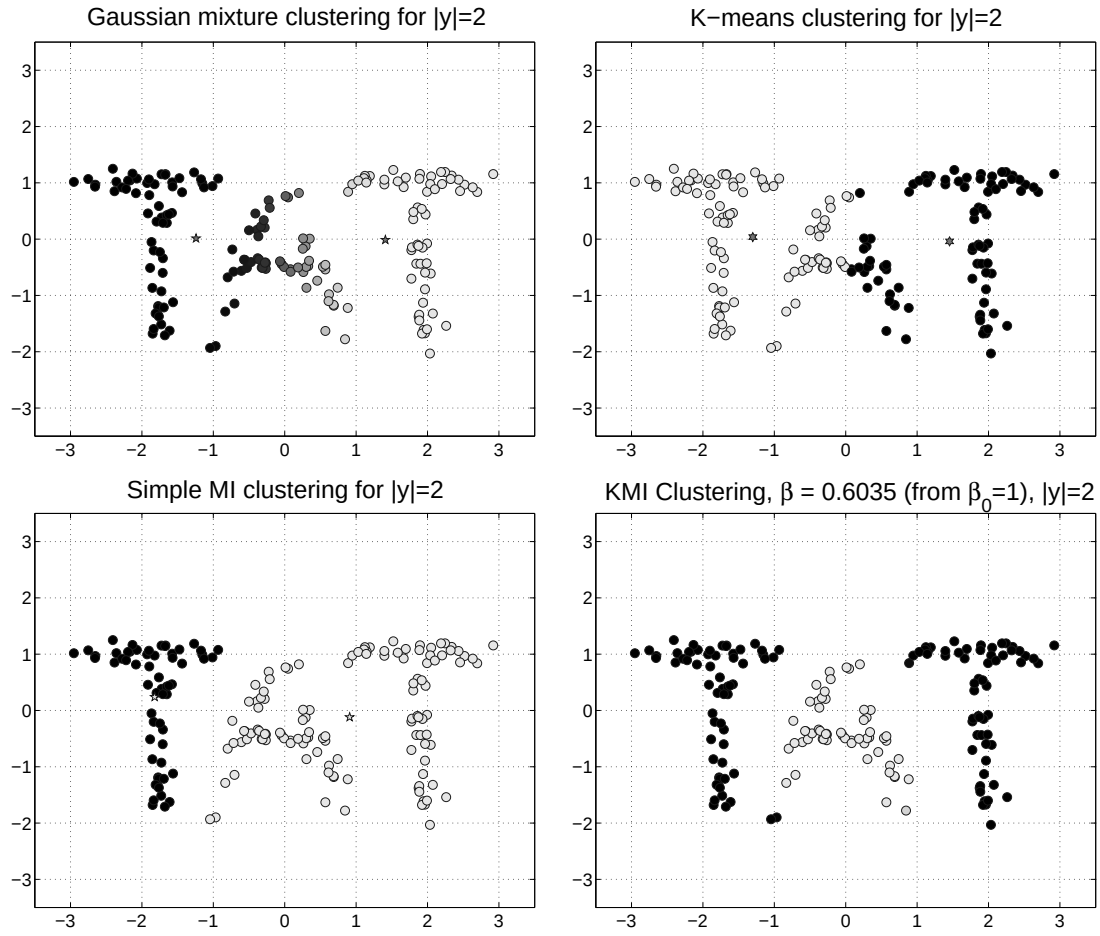


Figure 5.6: Learning cluster allocations for $|y| = 2$. Where appropriate, the stars show the cluster centers. *Top left:* clustering with the two-component Gaussian mixture trained by the EM algorithm on $\mathcal{L}$; *Top right:* clustering with the k-means; *Bottom left:* clustering with the encoder model $p(y_j|\mathsf{x}) \propto \exp\{-\|\mathsf{x} - \mathsf{w}_j\|^2/s_j\}$ trained by maximizing mutual information $I(\mathsf{x}, y)$; *Bottom right:* clustering with the kernelized encoder model $p(y_j|\mathsf{x}) \propto \exp\{-\|\boldsymbol{\phi}(\mathsf{x}) - \mathsf{w}_j\|^2/s_j\}$ trained by maximizing $I(\mathsf{x}, y)$ (the results are shown for the RBF kernel). For the kernelized model, the inverse variance β of the RBF kernel varied from $\beta_0 = 1$ (at the initialization) to $\beta \approx 0.604$ after convergence.

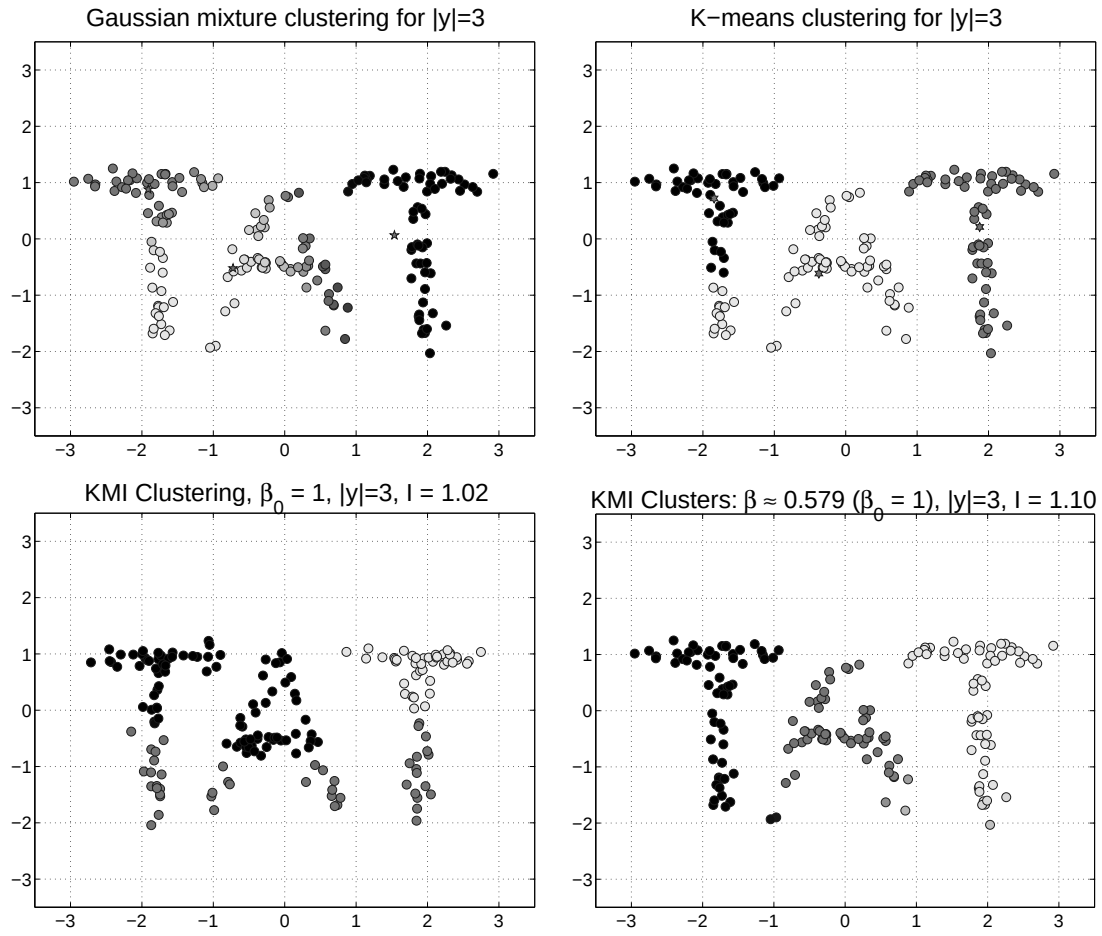


Figure 5.7: Learning cluster allocations for $|y| = 3$. Where appropriate, the stars show the cluster centers. *Top left:* clustering with the three-component Gaussian mixture trained by the EM algorithm on $\mathcal{L}$; *Top right:* clustering with the k-means algorithm; *Bottom left:* clustering in the kernelized encoder model (the results are shown for the RBF kernel with the inverse variance fixed at $\beta_0 = 1$); *Bottom right:* clustering with the kernelized encoder model with the adaptable kernel parameter (the inverse variance of the RBF kernel varied from $\beta_0 = 1$ (at the initialization) to $\beta \approx 0.579$ after convergence). Note that by learning the kernel parameter β we obtain higher values of the mutual information ($I \approx 1.10$ *vs* $I \approx 1.02$) and more intuitive cluster assignments.

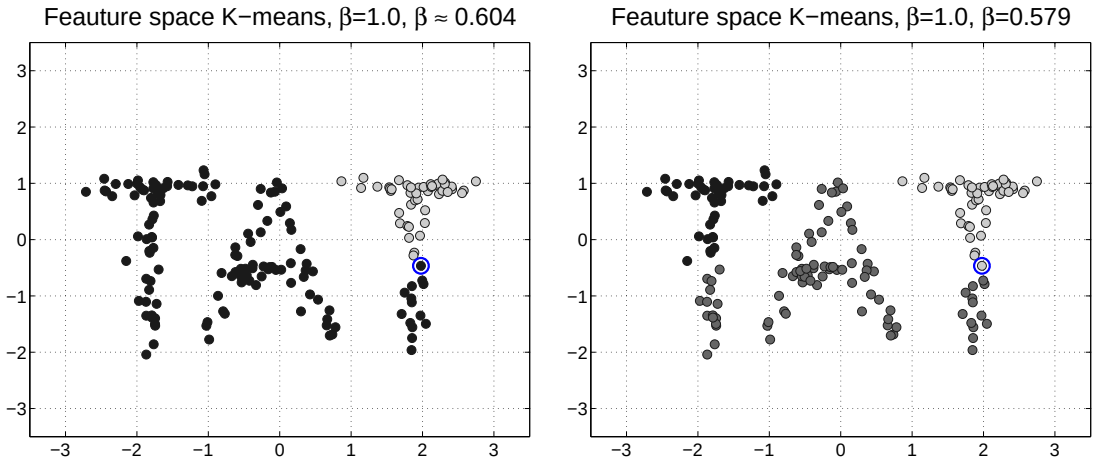


Figure 5.8: Learning cluster allocations by the feature-space k-means. The results are shown for the fixed RBF kernels with the inverse variance parameter β. *Left:* $|y| = 2$, $\beta_1 = 1$ and $\beta_2 = 0.604$; *Right:* $|y| = 3$, $\beta_1 = 1$ and $\beta_2 = 0.579$. The number of training patterns was $M = 210$. The patterns shown by the double circles ◎ are the only ones clustered differently for the two settings of the kernel parameters (β_1 and β_2), assuming identical initializations. The parameters β_2 were set to the converged values of β for the corresponding encoder models (*cf* Figure 5.6 (*bottom left*), Figure 5.7 (*bottom left*)).

pomorphically sensible cluster allocations. Indeed, in contrast to the k-means or Gaussian mixture clustering, the information-theoretic approach appears to favour smoother local representations, which results in arguably more intuitive cluster assignments. Figure 5.6 shows typical cluster assignments produced by the k-means algorithm (*top right*) and Gaussian mixtures trained by the EM (*top left*) for $|y| = 2$. Again, two different clusters are illustrated by different color intensities (light- and dark-gray), and soft cluster assignments are shown by the intermediate intensities. As usual, the EM for the mixture model was started at the k-means initialization, with the initial covariances set to the sample covariances of the associated k-means clusters, and the mixture coefficients proportional to the cluster sizes. As we see from the plots, both methods result in roughly identical cluster allocations, and for this case divide the dataset into two roughly symmetric parts, consisting of the letter T and half of the letter A. (Not surprisingly, the cluster allocations produced by the Gaussian mixture model are soft – other than that, the resulting clusters are not too different from those given by the k-means). Generally, other initializations of the mixture model resulted in more degenerate cluster assignments.

On the other hand, we can see that for the considered case the clusters produced by the encoder models are arguably more intuitive. Figure 5.6 (*bottom left*) shows two typical clusters generated by the non-kernelized encoder model trained by the IM for $|y| = 2$. The initial weights of the simple encoder model (5.12) were set according to $W_{ij}^{(0)} \sim \mathcal{N}_{W_{ij}}(0, 0.1)$, $\mathsf{W}^{(0)} = \{W_{ij}^{(0)}\} \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{y}|}$, with the log-variances $\tilde{s}_1^{(0)}, \tilde{s}_2^{(0)} \sim \mathcal{N}(0, 0.1)$. We used a similar initialization for the

kernelized encoder (5.22), where instead of the matrix of the weights W we have learned the matrix of the feature space coefficients $\mathsf{A}^{(0)} = \{A_{ij}^{(0)}\} \in \mathbb{R}^{M \times |\mathsf{y}|}$. For the kernelized model, we considered the RBF kernel with the initial settings of $\beta_0 = 1$. Both the non-kernelized and kernelized encoder models were trained by performing a numerical ascent on $I(\mathsf{x}, y)$ (by the scaled conjugate gradients). As we see from the plot, the simpler model extracts a single letter T as a separate cluster, and allocates the remaining training patterns to the other cluster (Figure 5.6 (*bottom left*)). The clusters produced by the trained kernelized encoder model are typically different (see Figure 5.6 (*bottom right*)). As we see from the plot, the kernelized encoder model clusters both T's similarly, separating them from the model of A, which may arguably be an intuitive coding scheme. Note that both Ts are clustered together despite being located geometrically far from each other in the data space. The results of (Figure 5.6 (*bottom right*)) are shown for $\beta \approx 0.6035$ obtained after convergence of the SCG learning procedure. Importantly, the obtained results appear to be stable for different samples from the underlying distribution and different initializations of the IM learning procedure (provided that the RBF kernel parameter β is not too large or too low, i.e. $\mathsf{K} \in \mathbb{R}^{M \times M}$ has a non-degenerate spectrum).

Figure 5.7 shows typical cluster allocations produced by the three methods for the codesize $|y| = 3$. Again, Gaussian mixtures and the k-means (Figure 5.7 (*top left*), Figure 5.7 (*top right*)) result in an arguably inferior performance to the kernelized IM clustering (Figure 5.7 (*bottom right*)). Moreover, we can see that by learning the RBF kernel parameter β, we may obtain better visualizations than for the case when β is fixed (Figure 5.7 (*bottom left*)). Also, by learning kernel parameters we typically get higher values of the mutual information $I(\mathsf{x}, y)$ ($I \approx 1.10$ *vs* $I \approx 1.02$ when β is fixed for the illustrated case). The results were confirmed repeatedly for different samples from the underlying distribution and different initializations of the learning algorithms. Generally, the kernelized encoder with the adaptable RBF kernel parameters was the only method (out of the considered ones) which resulted in an almost perfect separation of the letter models.

Finally, we have compared the information-theoretic clustering method with the kernel k-means algorithm (see e.g. Zhang and Rudnicky (2002), Dhillon et al. (2004), Wang et al. (2004)), where for a *fixed kernel function* each pattern is deterministically assigned to a cluster depending on the distance from the cluster mean in the feature space. Figure 5.8 illustrates a typical application of the kernelized k-means to clustering of the considered dataset for $|y| = 2$ (*left*), and $|y| = 3$ (*right*), started at the same initializations as the previous experiments shown on Figure 5.6 and Figure 5.7. Again, for comparison with the kernelized IM clustering, we show the results for the RBF kernels with the inverse variance parameter β. For both choices of the code space sizes $|y|$, the kernel parameter was set to the initial and the converged value of β obtained during the *kernelized encoder* clustering (see Figure 5.6 (*bottom right*), 5.7 (*bottom right*)); for example, for $|y| = 2$ we have considered $\beta = 1$ and $\beta = 0.6035$. We see that while the settings of β could influence the resulting cluster allocations, the clusters were not very different (assuming identical initializations for both choices of β). As

mentioned in Section 5.2.3, the principle advantage of our method is that it may indeed be viewed as an optimization procedure (namely, mutual information maximization) for learning kernel parameters. This contrasts with the majority of other approaches to unsupervised nonlinear clustering in feature spaces, which typically consider fixed similarity measures. As we have shown, learning kernel parameters may indeed lead to more intuitive visualizations of the underlying structure (see also Agakov and Barber (2005b) for an empirical comparison with the spectral clustering method of Ng et al. (2001)).

5.5 Summary

In this chapter we described several possible extensions of the IM framework to nonlinear encoder models. First, we considered optimizing the specific lower bound on the mutual information, where the encoder distribution of the channel was given by the exact posterior of the corresponding generative model. We explored this case for Gaussian mixtures, and compared the EM and the IM learning for a dataset which could not be easily modeled by a Gaussian mixture distribution. For this case we empirically demonstrated that both methods give rise to generally different optimization surfaces, and optimization of the bound $\hat{I}(\mathsf{x}, y)$ may favour more uniform representations in the code space, which may in some cases lead to an arguably more informative representation of the underlying distribution. Additionally, we pointed out that in the limiting special case of asymptotically noiseless spherical Gaussian decoders, the IM algorithm optimizing the considered bound $\hat{I}(\mathsf{x}, y)$ reduces to the k-means algorithm.

Then we focused on the problem of information-theoretic clustering in encoder models, where the generally stochastic nonlinear mapping from the sources to the discrete codes was learned by maximizing the exact mutual information $I(\mathsf{x}, y)$. For this case, we described a simple and practical algorithm applicable to unsupervised clustering, and explored its extensions by considering kernelized encoder models. Empirically, we demonstrated that the resulting information-theoretic clustering approach favorably compares with the common clustering techniques, and the option of learning kernel functions may indeed be of a practical benefit for visualizing the underlying structure of the data.

Finally, we reviewed some of the theoretical properties of the IM for higher-dimensional code spaces, and showed that some of the popular dimensionality reduction techniques may be interpreted as special instances of the variational information maximizing procedure. Specifically, we extended the work of Bourlard and Kamp (1988) and Bourlard (2000) to *arbitrary kernelizable* feature mappings applied in the context of stochastic encoding, and showed the nothing could be gained by using nonlinear encoders and linear variational decoders in the context of variational information maximization in Gaussian channels. To handle the intrinsic constraints of linear Gaussian variational decoders applied in the context of nonlinear encodings, we suggested a proper variational relaxation of the bound on $I(\mathsf{x}, \mathsf{y})$ which could be optimized for the encoder, *data decoder*, and *feature decoder* distributions under the assumption of a nonlinear variational decoding distribution. For *nonlinear Gaussian* encoding distributions, this led to

kernel PCA (Schoelkopf et al. (1998)) as the optimal solution for encoder weights. Additionally, we outlined a simple relation of the variational framework to the recent work on Gaussian Process Latent Variable Models (Lawrence (2003)), which may be interpreted as the variational information-maximization procedure in the *noiseless* limit of a nonlinear channel.

As an extension of the work described in this chapter, we note that a further study of the bound (5.37) on mutual information for nonlinear Gaussian encoders may need to be considered. As discussed in Section 5.3 and Appendix C.3.5, the bound indeed provides a proper information-theoretic objective for learning optimal kernel parameters. However, our current experience suggests that the results may be strongly influenced by the choice of constraints on the Gram matrix $\mathsf{K} \in \mathbb{R}^{M \times M}$ and specific definitions of the feature-to-data mappings. By choosing appropriate constraints on the kernel functions and the variational decoder distributions, one may hope to adapt the nonlinear Gaussian IM framework to visualizing high-dimensional data[8]. Another application field to explore is communication of discrete-valued data over channels with Gaussian noise, where a specific practical application may include code division multiple access in cellular telephony (see e.g. Viterbi (1995)).

Additionally, a comparison of the IM clustering with other families of clustering techniques will need to be considered. Specifically, it is interesting to see how information-maximization in nonlinear channels with the considered definition of the encoding distribution (5.22) could be related to the common spectral clustering approaches (see e.g. Perona and Freeman (1998), Kannan et al. (2000), Ng et al. (2001), Yu and Shi (2003)). Most of such approaches use eigenvalues of the fixed similarity matrices to transform the original data set, and apply any of the known clustering techniques on the transformations (see Weiss (1999) for a unifying discussion). It is intuitive that it may indeed be possible to interpret some of the spectral clustering methods from the information-theoretic viewpoint. Indeed, it is well known that the standard k-means algorithm (Hartigan and Wong (1979)) may be viewed as the trace maximization problem of the Gram matrix of the original data patterns (Zha et al. (2001)). Recently this result has been generalized to show that popular multi-class spectral clustering techniques using *normalized cuts* (e.g. Shi and Malik (2000), Ng et al. (2001)) may in fact be interpreted as a form of the weighted k-means algorithm (Bach and Jordan (2003), Dhillon et al. (2004)). The results of Section 5.2.1 suggest that it may potentially be possible to relate the spectral clustering methods to a specific form of the variational information-maximizing procedure, though a direct relation between the methods currently remains unclear.

Fundamentally, one of the principal advantages of our approach to nonlinear dimensionality reduction is a simple way to learn parameters of the kernel function in the unsupervised context. As we showed, the resulting procedure is numerically and conceptually simple; specifically, in the considered cases it did not require

[8]Note that if the projection noise may be reduced to zero, one may consider optimizing simpler objective functions, such as (5.40) – *cf* Gaussian Process Latent Variable Models. It is therefore believed that a further study of nonlinear Gaussian channels for dimensionality reduction may be of a practical interest mainly in situations when the encoding noise is unavoidable.

computations of the matrix inversions or eigenvalue decompositions of the Gram matrices. Also, once the channel distributions (and the corresponding kernels) are parameterized, the method does not require complex problem-specific algorithmic heuristics. Furthermore, one may expect that the information-theoretic view of clustering could potentially offer common advantages over the algorithmic approaches; specifically, it may be possible to extend the method to richer encoding distributions.

Chapter 6

Variational Information Maximization for Learning High-Dimensional Discrete Representations

In Chapter 5 we described applications of the information-maximizing framework to learning optimal parameters of nonlinear encoder models. One specific application of the framework which we considered was information-theoretic clustering. Due to the low cardinality of the codes (cluster labels), optimization of mutual information between the discrete codes y and the continuous sources x could in that case be performed exactly, and a number of nonlinear extensions of simple channels could easily be considered (see Section 5.2).

While being practically useful for visualizing unlabeled data, information-theoretic clustering may be viewed as a limiting case of discrete encoding. In many domains much richer encoding schemes must be considered. For example, in the neurophysiological domain it may often be a matter of interest to estimate how much information about the source stimuli may be contained in a (generally unknown) *population* of neural spikes. Additionally, the problem domain may impose constraints on the channel distributions, which may generally lead to the need of learning stochastic high-dimensional encoding mappings. As discussed in Section 1.4, maximization of the exact mutual information in this case may potentially be problematic, as it will generally require an explicit integration over the high dimensional encodings.

The primary goal of this chapter is to explore applicability of the variational information-maximizing framework in the context of learning high-dimensional binary representations of continuous source patterns. While the obtained results are general, we believe that the application area where they may be particularly useful is stochastic neural coding. We show that the simplest instance of our variational information-maximizing formulation provides a convenient framework for population coding in stochastic point neuron models. By analogy with Linsker (1997), we demonstrate that it is possible to derive an information-maximizing procedure which only requires local computations; however, our results are signifi-

cantly more general, as they are applicable in the case of stochastic, non-invertible encoding mappings. Moreover, we demonstrate that for the considered choice of a conditionally factorized encoder model, our variational method favorably compares with two approximate techniques maximizing Fisher Information-based *approximations* of lower bounds on the generally intractable $I(\mathsf{x}, \mathsf{y})$.

6.1 Introduction

The problem of encoding real-valued stimuli $\mathsf{x} \in \mathbb{R}^{|\mathsf{x}|}$ by a population of binary spikes $\mathsf{y} \in \{-1, 1\}^{|\mathsf{y}|}$ may be addressed in many different ways. Essentially, the goal is to adapt the parameters of any mapping $p(\mathsf{y}|\mathsf{x})$ to make a desirable population code for a given set of input stimuli $\{\mathsf{x}\}$. There are many possible ways to address this problem. One could be that *any* reconstruction based on the population should be accurate. This is typically handled by appealing to the Fisher Information (e.g. Cramer (1946), Cover and Thomas (1991)) which, with care, can be used in order to bound mean square reconstruction error (see e.g. Johnson (2003) for an introductory discussion).

Here we consider maximizing the amount of information which the spiking patterns contain about the stimuli. In this framework, a population coding formulation may be viewed as a special case of the information-maximizing problem, where continuous source signals are mapped into a discrete high-dimensional space. While much of the earlier work focuses on simple channels (e.g. Linsker (1988), Pouget et al. (1998), Zhang and Sejnowski (1999), Bethge et al. (2002)) or noiseless invertible mappings (Nadal and Parga (1994), Bell and Sejnowski (1995), Linsker (1997)), it is particularly interesting to address the problem of high-dimensional stochastic coding. Many previous attempts to apply mutual information[1] to population coding have been made (e.g. Brunel and Nadal (1998), Stocks and Mannella (2001), Kang and Sompolinsky (2001), Samengo and Treves (2001)). However, for large population sizes and non-invertible mappings to the code space, maximization of the mutual information is generally a computationally intractable task (see Section 1.4). Most current studies address the problem of computational intractability of maximizing $I(\mathsf{x}, \mathsf{y})$ by considering alternative objective criteria, e.g. those based on numerical approximations of $I(\mathsf{x}, \mathsf{y})$ derived under specific asymptotic assumptions (see e.g. Brunel and Nadal (1998), Kang and Sompolinsky (2001), Hoch et al. (2003)). While these approximations may result in asymptotic efficiency for many encoding channels, their applicability for specific models may be strongly affected by the form of encoder distributions, which may lead to instability of the resulting optimization procedure for specific encoding schemes (Agakov and Barber (2004b)).

In Chapter 2 we described a simple and general variational procedure optimizing a proper bound on mutual information $I(\mathsf{x}, \mathsf{y})$ between the sources x and the codes y. In contrast to most of the existing techniques, the procedure maximizes

[1] Many other methods suggesting to optimize alternative objectives, e.g. *redundancy* (Barlow (1989), Atick (1992), Redlich (1993), Field (1994)), may be shown to correspond to specific relaxations (Nadal et al. (1998)) of the exact mutual information, which also involves generally intractable computations.

a proper bound on $I(\mathsf{x},\mathsf{y})$, rather than an asymptotic approximation of a bound (Brunel and Nadal (1998)) or approximation of the exact mutual information for a local region of the source space (Szummer and Jaakkola (2002), Corduneanu and Jaakkola (2003)). The primary goal of this chapter is to apply the variational IM method to learning *high-dimensional stochastic* binary representations of continuous source patterns. As we believe that this procedure offers a convenient framework for addressing sub-goals of predictive population coding, we focus specifically on a biologically inspired channel parameterization.

6.2 Variational Learning of Population Codes

The principled information theoretic approach to learning neural codes involves maximization of the mutual information with respect to parameters of the encoder $p(\mathsf{y}|\mathsf{x})$. In what follows we assume a conditionally factorized decoder, i.e. $p(\mathsf{y}|\mathsf{x}) = \prod_{i=1}^{|\mathsf{y}|} p(y_i|\mathsf{x})$, which is arguably the simplest case of a feed-forward mapping (each unit y_i is defined by a simple *point-neuron* model). This assumption also facilitates comparisons with other approximate information-maximizing techniques (such as the common approximations of Brunel and Nadal (1998), or reformulations of the local criteria derived in a different context by Corduneanu and Jaakkola (2003)).

Since for large-scale stochastic systems exact evaluation of $I(\mathsf{x},\mathsf{y})$ is in general computationally intractable, we consider optimizing the generic lower bound $\tilde{I}(\mathsf{x},\mathsf{y}) = const + \langle \log q(\mathsf{x}|\mathsf{y}) \rangle_{p(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x})}$, where $\tilde{p}(\mathsf{x})$ is the empirical distribution of the high-dimensional continuous stimuli $\mathsf{x} \in \mathbb{R}^{|\mathsf{x}|}$. To learn optimal stochastic representations of the continuous training patterns $\mathsf{x}^{(1)}, \ldots, \mathsf{x}^{(M)}$ according to the lower bound $\tilde{I}(\mathsf{x},\mathsf{y})$, we need to choose a continuous density function for the decoder $q(\mathsf{x}|\mathsf{y})$. Computationally, it is convenient to assume that the decoder is given by the isotropic Gaussian $q(\mathsf{x}|\mathsf{y}) \sim \mathcal{N}(\mathsf{U}\mathsf{y}, s^2\mathsf{I})$, where $\mathsf{U} \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{y}|}$. Other (e.g. non-linear) variational decoders may potentially be considered and relaxations analogous to the ones described in Section 5.3.1 may potentially be used. However, we show that in situations when the variational decoder is a constrained linear Gaussian, the resulting iterative optimization procedure is particularly convenient, which results in a local (and arguably more biologically plausible) learning scheme.

We note that the proposed method describes a theoretically rigorous framework for maximizing information content which the high-dimensional spikes contain about (generally high-dimensional) continuous stimuli. The encoder distribution $p(\mathsf{y}|\mathsf{x})$ may in this case be interpreted as a stochastic mapping from the physical to the neural domain. One particularly interesting biological example where such mappings may occur is a mammalian retina, where the continuous-valued sources $\mathsf{x} \in \mathbb{R}^{|\mathsf{x}|}$ define activations at the photoreceptor layer, and the high-dimensional outputs $\mathsf{y} \in \{-1, 1\}^{|\mathsf{y}|}$ correspond to the on-off encodings in the ganglion cells (e.g. Watanabe and Rodieck (1989), Lee et al. (1998)). While the neurophysiological interpretation of the variational decoder $q(\mathsf{x}|\mathsf{y})$ is not properly understood, effectively the distribution defines the stochastic perceptual projection of the retinal encoder models (Eckmiller et al. (2005)). Under the considered

parameterization, the local field of each perceptual unit (x_i under the variational distribution $q(x_i|\mathsf{y})$) is a linear combination of the ganglion firings, with an added white noise. Effectively, this choice of the decoder indicates that small changes in the post-synaptic firings do not significantly vary our guesses about the generating stimuli. It is clear that for the considered choice of the decoder distribution, the percepts are conditionally independent (i.e. $q(\mathsf{x}|\mathsf{y}) = \prod_{i=1}^{|\mathsf{x}|} q(x_i|\mathsf{y})$), i.e. the reconstruction at each perceptual unit is independent from the neighboring percepts for a given pattern of ganglion firings. Conveniently, this facilitates derivations of a local learning rule. Note that for the considered case it is straightforward to evaluate the generic bound $\tilde{I}(\mathsf{x},\mathsf{y})$ exactly, since it only involves computations of the second-order moments of y over the factorized distribution $p(\mathsf{y}|\mathsf{x})$.

6.2.1 Local Iterative Learning

Here we consider the case of high-dimensional continuous patterns $\mathsf{x} \in \mathbb{R}^{|\mathsf{x}|}$ represented by stochastic firings of the post-synaptic neurons $\mathsf{y} \in \{-1,+1\}^{|\mathsf{y}|}$. For each neuron y_i, we assume the logistic parameterization of the encoder $p(y_i|\mathsf{x})$, so that the probability of firing monotonically increases with an increase in the membrane potential (using any other parameterization of $p(y_i|\mathsf{x})$ will lead to a straight-forward re-formulation of the model). For conditionally independent activations, we obtain

$$p(\mathsf{y}|\mathsf{x}) = \prod_{i=1}^{|\mathsf{y}|} p(y_i|\mathsf{x}) \stackrel{\text{def}}{=} \prod_{i=1}^{|\mathsf{y}|} \sigma(y_i(\mathsf{w}_i^T\mathsf{x} + b_i)) \tag{6.1}$$

where $\mathsf{w}_i \in \mathbb{R}^{|\mathsf{x}|}$ is a vector of the synaptic weights for neuron y_i, b_i is the corresponding threshold, and $\sigma(a) \stackrel{\text{def}}{=} 1/(1+e^{-a})$. It is obvious that (6.1) indeed defines a properly normalized conditional distribution.

By utilizing the factorial assumptions $p(\mathsf{y}|\mathsf{x}) = \prod_{i=1}^{|\mathsf{y}|} \sigma(y_i(\mathsf{w}_i^T\mathsf{x} + b_i))$ and $q(\mathsf{x}|\mathsf{y}) \sim \mathcal{N}(\mathsf{U}\mathsf{y}, s^2\mathsf{I})$, the straight-forward substitution into (2.2) leads to

$$\tilde{I}(\mathsf{x},\mathsf{y}) \propto \sum_{m=1}^{M} \mathrm{tr}\left\{\mathsf{U}\langle\mathsf{y}\rangle_{p(\mathsf{y}|\mathsf{x}_m)}\mathsf{x}_m^T - \frac{1}{2}\mathsf{U}^T\mathsf{U}\langle\mathsf{y}\mathsf{y}^T\rangle_{p(\mathsf{y}|\mathsf{x}_m)}\right\} + const, \tag{6.2}$$

which needs to be optimized for the encoder and decoder weights $\mathsf{W} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{x}|}$, $\mathsf{U} \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{y}|}$ and the biases $\mathsf{b} \in \mathbb{R}^{|\mathsf{y}|}$ (we have assumed that s^2 is a constant incorporated into the proportionality factor). We may now compute the matrix derivatives of (6.2) to obtain

$$\frac{\partial\tilde{I}(\mathsf{x},\mathsf{y})}{\partial\mathsf{U}} \;\propto\; \sum_{m=1}^{M}\left[\mathsf{x}^{(m)}\boldsymbol{\lambda}^T(\mathsf{x}^{(m)}) - \mathsf{U}\left(\boldsymbol{\lambda}(\mathsf{x}^{(m)})\boldsymbol{\lambda}^T(\mathsf{x}^{(m)}) - \mathsf{D}_{\lambda_x}(\mathsf{x}^{(m)}) + \mathsf{I}\right)\right] \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{y}|} \tag{6.3}$$

where $\mathsf{x}^{(m)} = \{x_m^1, \ldots, x_m^{|\mathsf{x}|}\} \in \mathbb{R}^{|\mathsf{x}|}$ is the m^{th} vector of input stimuli, and

$$\lambda_i(\mathsf{x}^{(m)}) \stackrel{\text{def}}{=} \langle y_i\rangle_{p(y_i|\mathsf{x}^{(m)})} = 2\sigma(\mathsf{w}_i^T\mathsf{x}^{(m)} + b_i) - 1 \tag{6.4}$$

is the conditional mean. The diagonal matrix $\mathsf{D}_{\lambda_x}(\mathsf{x}_m) \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{y}|}$ is defined by

$$\mathsf{D}_{\lambda_x} \stackrel{\text{def}}{=} \text{diag}\left(\lambda_1^2(\mathsf{x}), \ldots, \lambda_{|\mathsf{y}|}^2(\mathsf{x})\right) = \mathsf{I} - \text{cov}(\mathsf{y}|\mathsf{x}) \in [0,1]^{|\mathsf{y}|\times|\mathsf{y}|}, \tag{6.5}$$

which is a measure of consistency of neural firings. Note that evaluation of the gradient (6.3) requires computing the expected firings $\lambda_i(\mathsf{x})$ of the output units $\{y_i | i = 1, \ldots, |y|\}$, which only involves local feed-forward computations. Specifically, we note that computations of the gradients for the decoder weights $\partial\tilde{I}/\partial\mathsf{U}$ do not require global calculations (such as evaluations of matrix inverses, etc.), which would have been difficult to justify biologically.

Analogously, by differentiating the bound $\tilde{I}(\mathsf{x},\mathsf{y})$ with respect to the weights and biases of the i^{th} encoding unit, we get

$$\frac{\partial\tilde{I}(\mathsf{x},\mathsf{y})}{\partial w_{ij}} \propto \sum_{m=1}^{M} x_j^{(m)} \left[1 - \lambda_i^2(\mathsf{x}^{(m)})\right] \mathsf{u}_i^T \left[\mathsf{x}^{(m)} + \mathsf{u}_i\lambda_i(\mathsf{x}^{(m)}) - \mathsf{U}\boldsymbol{\lambda}(\mathsf{x}^{(m)})\right], \tag{6.6}$$

$$\frac{\partial\tilde{I}(\mathsf{x},\mathsf{y})}{\partial b_i} \propto \sum_{m=1}^{M} \left[1 - \lambda_i^2(\mathsf{x}^{(m)})\right] \mathsf{u}_i^T \left[\mathsf{x}^{(m)} + \mathsf{u}_i\lambda_i(\mathsf{x}^{(m)}) - \mathsf{U}\boldsymbol{\lambda}(\mathsf{x}^{(m)})\right] \tag{6.7}$$

where u_i corresponds to the i^{th} column of the decoder weights $\mathsf{U} \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{y}|}$, and w_{ij} links the j^{th} pre-synaptic and the i^{th} post-synaptic units. Note that the updates for the biases b_i and encoder weights w_{ij} involve computations of the third and fourth power terms $x_k^{(m)}\langle y_j|\mathsf{x}^{(m)}\rangle\langle y_l|\mathsf{x}^{(m)}\rangle$ and $x_i^{(m)}x_k^{(m)}\langle y_j|\mathsf{x}^{(m)}\rangle\langle y_l|\mathsf{x}^{(m)}\rangle$, which may also be performed locally.

An arguably more biologically plausible interpretation of the learning rule may be obtained by performing the stochastic updates. If $\tilde{y}_i^{(m)} = 1$ with probability $\sigma(\mathsf{w}_i^T\mathsf{x}^{(m)} + b_i)$ (and $\tilde{y}_i^{(m)} = -1$ otherwise), the gradients (6.6) and (6.7) may be transformed to

$$\Delta w_{ij}^{(m)} = \eta \text{var}(y_i|\mathsf{x}^{(m)}) x_j^{(m)} \mathsf{u}_i^T \left[\mathsf{x}^{(m)} + \mathsf{u}_i\tilde{y}_i^{(m)} - \mathsf{U}\tilde{\mathsf{y}}^{(m)}\right] \tag{6.8}$$

$$\Delta b_i^{(m)} = \eta \text{var}(y_i|\mathsf{x}^{(m)}) \mathsf{u}_i^T \left[\mathsf{x}^{(m)} + \mathsf{u}_i\tilde{y}_i^{(m)} - \mathsf{U}\tilde{\mathsf{y}}^{(m)}\right], \tag{6.9}$$

where η is the learning rate, and $\Delta w_{ij}^{(m)}$, $\Delta b_i^{(m)}$ are the encoder updates for the m^{th} observation. The pre-factor $\text{var}(y_i|\mathsf{x}^{(m)}) = 1 - \lambda_i(\mathsf{x}^{(m)}) \in [0,1]$ indicates that optimally, the training should slow down once the weights saturate and the firings become more deterministic. (Suboptimally and conventionally, the term is ignored, which corresponds to its first-order Taylor expansion around zero field of y_i). Similarly, we may derive the stochastic updates for the parameter of the perceptual mapping, leading to

$$\Delta\mathsf{U}^{(m)} = \eta_U \left[\mathsf{x}^{(m)}\tilde{\mathsf{y}}^{(m)} - \mathsf{U}\tilde{\mathsf{Y}}^{(m)}\right] \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{y}|}, \tag{6.10}$$

where η_U is the learning rate, $\tilde{\mathsf{Y}}^{(m)} \stackrel{\text{def}}{=} \{\tilde{y}_i^{(m)}\tilde{y}_j^{(m)}(1-\delta_{ij}) + \delta_{ij}\} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{y}|}$, and δ_{ij} is Kronecker delta. Note that the updates in (6.8) – (6.10)) are easily representable as weighted Hebbian and anti-Hebbian terms, where the receptive field of the i^{th}

post-synaptic unit y_i is affected not only by the activations at the pre-synaptic layer x, but also (implicitly) by the stochastic firings of the neighboring post-synaptic units $\tilde{y}_j$ $(j \neq i)$.

Generally, expressions (6.3), (6.6) – (6.7) define the updates of the IM algorithm on $\tilde{I}(\mathsf{x}, \mathsf{y})$ and may be used in any standard numerical optimization procedure (see e.g. Bishop (1995)). An arguably more biologically plausible alternative is to perform the stochastic ascent (expressions (6.8) – (6.10)). Finally, we note once again that the explicit parameterization of the encoder $p(\mathsf{y}|\mathsf{x})$ and the variational decoder $q(\mathsf{x}|\mathsf{y})$ makes it easy to incorporate additional constraints on the encoder and decoder parameters $\mathsf{W} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{x}|}$, $\mathsf{b} \in \mathbb{R}^{|\mathsf{y}|}$, and $\mathsf{U} \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{y}|}$, which may be used to transform the variational IM learning so that the solutions satisfy additional requirements.

6.2.2 Optimal Encoder Models

In order to simplify the analysis of the variational bound (6.2), we will follow the strategy of Section 2.2.2 and re-define the objective $\tilde{I}(\mathsf{x}, \mathsf{y})$ as a function of the encoder parameters. As mentioned in Chapter 2, for non-convex functions this may lead to a different optimization surface and a generally more complex (non-local) learning rule. However, the analysis is still interesting, as the resulting bound may be more easily compares with the common numerical approximations of the generally intractable mutual information $I(\mathsf{x}, \mathsf{y})$.

Expressing the bound (6.2) as a function of the encoder $p(\mathsf{y}|\mathsf{x})$ alone, we get

$$\mathsf{U} = \langle \mathsf{x}\mathsf{y}^T \rangle \langle \mathsf{y}\mathsf{y}^T \rangle^{-1}, \quad \tilde{I}(\mathsf{x}, \mathsf{y}) \propto \mathrm{tr}\left\{ \langle \mathsf{x}\mathsf{y}^T \rangle \langle \mathsf{y}\mathsf{y}^T \rangle^{-1} \langle \mathsf{y}\mathsf{x}^T \rangle \right\} + const \tag{6.11}$$

(from now on we will ignore the constant which has no effect on the optimization surface). The objective (6.11) is a proper bound on $I(\mathsf{x}, \mathsf{y})$ for any choice of the stochastic mapping $p(\mathsf{y}|\mathsf{x})$. We may therefore[2] use it for optimizing a variety of channels with continuous source vectors.

For the considered parameterization of the encoding distribution (6.1), we may transform (6.11) to obtain

$$\tilde{I}(\mathsf{x}, \mathsf{y}) \propto \mathrm{tr}\left\{ \langle \mathsf{x}\boldsymbol{\lambda}_x^T \rangle \langle \boldsymbol{\lambda}_x \boldsymbol{\lambda}_x^T - \mathsf{D}_{\lambda_x} + \mathsf{I} \rangle^{-1} \langle \boldsymbol{\lambda}_x \mathsf{x}^T \rangle \right\}, \tag{6.12}$$

where the averages are computed over $\tilde{p}(\mathsf{x})p(\mathsf{y}|\mathsf{x})$, and $\tilde{p}(\mathsf{x}) \propto \sum_{m=1}^{M} \delta(\mathsf{x} - \mathsf{x}^{(m)})$ is the empirical distribution. Again, $\boldsymbol{\lambda}_x \in [-1, 1]^{|\mathsf{y}|}$ is a vector whose elements $\lambda_i(\mathsf{x}) \stackrel{\text{def}}{=} \langle y_i \rangle_{p(y_i|\mathsf{x})} = 2\sigma(\mathsf{w}_i^T \mathsf{x} + b_i) - 1$ correspond to expected firings of the i^{th} unit for a fixed stimulus x, and $\mathsf{D}_{\lambda_x}(\mathsf{x}^{(m)}) \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{y}|}$ is defined by expression (6.5). The expectations in (6.12) are computed over the empirical distribution.

Since the lower bound (6.12) depends only on the thresholds and synaptic weights, the learning rule is easily obtained by differentiating (6.12) with respect

[2]From (6.11) it is clear that if $\langle \mathsf{y}\mathsf{y}^T \rangle$ is near-singular, the varying part of the objective $\tilde{I}(\mathsf{x}, \mathsf{y})$ may be infinitely large. However, if the mapping $\mathsf{x} \mapsto \mathsf{y}$ is probabilistic and the number of training stimuli M exceeds the dimensionality of the neural codes $|\mathsf{y}|$, the optimized criterion is typically positive and finite.

to the encoder parameters $\mathsf{b} \in \mathbb{R}^{|\mathsf{y}|}$ and $\mathsf{W} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{x}|}$ (where rows of W are given by $\mathsf{w}_i^T \in \mathbb{R}^{1\times|\mathsf{x}|}$). This leads to

$$\Delta\mathsf{W} \;\propto\; \sum_{m=1}^{M} \left(\mathsf{I} - \mathsf{D}_{\lambda(\mathsf{x}^{(m)})}\right) \left(\tilde{\mathsf{D}}\boldsymbol{\lambda}(\mathsf{x}^{(m)}) + \boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}\left[\mathsf{x}^{(m)} - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\lambda}(\mathsf{x}^{(m)})\right]\right) \left(\mathsf{x}^{(m)}\right)^T \tag{6.13}$$

where $\Delta\mathsf{W} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{x}|}$ is the weight update, $\boldsymbol{\Sigma}_{yy} \stackrel{\text{def}}{=} \langle \mathsf{y}\mathsf{y}^T \rangle$, $\boldsymbol{\Sigma}_{yx} \equiv \boldsymbol{\Sigma}_{xy}^T \stackrel{\text{def}}{=} \langle \mathsf{y}\mathsf{x}^T \rangle$ are the second-order moments, and $\tilde{\mathsf{D}}$ is the diagonal matrix

$$\tilde{\mathsf{D}} \stackrel{\text{def}}{=} \operatorname{diag}\left\{\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}\left(\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}\right)^T\right\} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{y}|}, \tag{6.14}$$

where $\operatorname{diag}(\mathsf{A}) = \{a_{ij}\delta_{ij}\}$, and δ_{ij} is Kronecker delta. The update for the threshold $\Delta\mathsf{b} \in \mathbb{R}^{|\mathsf{y}|}$ has the same form as (6.13) without the post-multiplication of each term by the training stimulus $(\mathsf{x}^{(m)})^T$.

From (6.13) it is clear that the magnitude of each weight update $\Delta\mathsf{w}_i \in \mathbb{R}^{|\mathsf{x}|}$ should decrease with an increase in the corresponding conditional variance $\operatorname{var}(y_i|\mathsf{x}_m)$. Similarly to the case of the local iterative learning (described in Section 6.2.1), the pre-factor

$$\left(\mathsf{I} - \mathsf{D}_{\lambda(\mathsf{x}^{(m)})}\right) = \operatorname{cov}(\mathsf{y}|\mathsf{x}^{(m)}) = \operatorname{diag}\left\{\operatorname{var}(y_1|\mathsf{x}^{(m)}), \ldots, \operatorname{var}(y_{|\mathsf{y}|}|\mathsf{x}^{(m)})\right\} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{y}|} \tag{6.15}$$

may be interpreted as a variable learning rate. It is intuitive that as training continues, the magnitudes of the synaptic weights would typically increase (as this would lead to a decrease in the conditional entropy $H(\mathsf{y}|\mathsf{x})$). For each unit y_i, this would typically result in a decrease in the conditional variance $\operatorname{var}(y_i|\mathsf{x}^{(m)})$, leading to the parameter saturation effect. The sufficient condition for the training to stop is the noiseless limit of the encoding projection $\operatorname{cov}(\mathsf{y}|\mathsf{x}^{(m)}) \approx \mathbf{0}_{|\mathsf{y}|}$; for the considered parameterization of the encoding distribution, this may only happen for the divergent membrane potentials of the post-synaptic units. By analogy with a discussion of the Gaussian channel in Section 4.1.1, this stipulates the need of additional constraints[3] on the encoder weights.

It is clear that by analogy with (6.8) – (6.10) we may consider stochastic approximations of (6.13), which leads to an arguably more biologically plausible formulation. Nevertheless, the fundamental criticism of the learning rule (6.13) still applies: the rule (6.13) has an intrinsically non-local nature, and a careful approximation of the second-order moments $\boldsymbol{\Sigma}_{yy}^{-1}$, $\boldsymbol{\Sigma}_{yx}$ may need to be considered. The learning rule (6.13) may then be decomposed as a weighted combination of Hebbian and anti-Hebbian terms, though the nature of dependencies of the weighting coefficients on the firing patterns and encoder parameters is non-local.

Finally, it is interesting to note that by considering the first-order expansion of each pre-factor $\operatorname{var}(y_i|\mathsf{x}^{(m)})$ around $\mathsf{w}_i^T\mathsf{x}^{(m)} + b_i \approx 0$, we may transform (6.13) to

$$\Delta\mathsf{W} \propto \tilde{\mathsf{D}}\langle\boldsymbol{\lambda}_x\mathsf{x}^T\rangle + \boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}\left(\boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\langle\boldsymbol{\lambda}_x\mathsf{x}^T\rangle\right) \tag{6.16}$$

[3]Possibly the simplest of such constraints could be introduced by considering a soft constraint on the variances of the conditional firings, which amends (6.13) with an anti-Hebbian term $-\mathsf{A}\langle\boldsymbol{\lambda}_x\mathsf{x}^T\rangle \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{x}|}$ such that $\mathsf{A} \succeq 0 \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{y}|}$.

where $\mathbf{\Sigma}_{xx} \stackrel{\text{def}}{=} \langle \mathsf{x}\mathsf{x}^T \rangle$. Clearly, (6.16) is decomposable as a combination of the stochastic Hebbian and anti-Hebbian terms, with the weighting coefficients determined by the second-order moments of the firings and input stimuli. Additionally, it is easy to see that the parenthesized factor at the correction of the Hebbian term in (6.16) corresponds to the Schur compliment of the joint correlation $\tilde{\mathbf{\Sigma}} \stackrel{\text{def}}{=} \langle [\mathsf{x}\ \mathsf{y}]\,[\mathsf{x}\ \mathsf{y}]^T \rangle_{p(\mathsf{x},\mathsf{y})}$ (see e.g. von Mises (1964)). This may be viewed as a conditional covariance $\tilde{\mathbf{\Sigma}}_{\mathsf{x}|\mathsf{y}}$ of the decoder expressed from the *joint* Gaussian model $\tilde{p}_{\mathsf{x}\mathsf{y}} \sim \mathcal{N}_{xy}(\mathbf{0}, \tilde{\mathbf{\Sigma}})$, resulting in

$$\Delta \mathsf{W} \propto \tilde{\mathsf{D}} \langle \boldsymbol{\lambda}_x \mathsf{x}^T \rangle + \mathbf{\Sigma}_{yy}^{-1} \mathbf{\Sigma}_{yx} \tilde{\mathbf{\Sigma}}_{\mathsf{x}|\mathsf{y}}, \quad \tilde{\mathbf{\Sigma}}_{\mathsf{x}|\mathsf{y}} \stackrel{\text{def}}{=} \mathbf{\Sigma}_{xx} - \mathbf{\Sigma}_{xy} \mathbf{\Sigma}_{yy}^{-1} \mathbf{\Sigma}_{yx}. \tag{6.17}$$

(*cf* Linsker's *as-if Gaussian* bound (1.16)). From (6.13) and (6.17) we see that a simple fixed-rate Hebbian update would generally be a suboptimal approximation of both the exact gradient of the bound $\tilde{I}(\mathsf{x}, \mathsf{y})$ and its more conventional fixed-rate approximation.

6.3 Fisher Information and Mutual Information

Now we will briefly review two classes of methods using the Fisher Information criterion in order to approximate the exact mutual information. The first class of methods which we consider proposes to optimize an approximation of a specific bound on $I(\mathsf{x}, \mathsf{y})$, which is conveniently derived by applying the Cramer-Rao and the data processing inequalities (Brunel and Nadal (1998)), and computing the limit for $|\mathsf{y}| \to \infty$. The same result may be obtained by considering a numerical approximation of the "local mutual information" $I_{\mathsf{x}} = \langle \log p(\mathsf{y}|\mathsf{x}) / p(\mathsf{y}) \rangle_{p(\mathsf{y}|\mathsf{x})}$ for $|\mathsf{y}| \to \infty$, and integrating the approximation over the sources $\tilde{p}(\mathsf{x})$ (Kang and Sompolinsky (2001)).

The second method is inspired by Szummer and Jaakkola (2002) and Corduneanu and Jaakkola (2003), who also propose to approximate the local information $I_{\mathcal{R}}(\mathsf{x}, \mathsf{y})$ for each small region of the source space $\mathcal{R} \subset \mathcal{R}_{\mathsf{x}}, \mathcal{R} \to 0$ (*cf* Kang and Sompolinsky (2001)), but use different numerical relaxations, which generally leads to a different optimization surface. Specifically, their approximations do not depend on the population sizes $|\mathsf{y}|$, which suggests better convergence properties for low and intermediate values of $|\mathsf{y}|$. Though the original results of Szummer and Jaakkola (2002) and Corduneanu and Jaakkola (2003) are applied in the different context of semi-supervised learning, we show that their approach may be considered as a general alternative for computing an approximate lower bound on the exact mutual information $I(\mathsf{x}, \mathsf{y})$.

It turns out that both approximations we discuss here involve computations of the Fisher Information criterion. To differentiate between the methods, we will refer to the method of Szummer and Jaakkola (2002) and Corduneanu and Jaakkola (2003) as the *local approximation* of mutual information, since the fundamental assumption used in the derivations is the locality of the regions (the Fisher Information matrix arises as a result of a specific expansion of the encoder distribution). In contrast, we will refer to the approach of Brunel and Nadal (1998) as *Fisher approximation* of mutual information (or *Fisher criterion*), since

the term arises as the result of applying the fundamental Cramer-Rao inequality. For both of the considered methods, we outline the corresponding learning rules for the considered case of a sigmoidal network.

6.3.1 Fisher Approximation of Mutual Information

Let us assume that the stochastic firings of the output units are conditionally independent given the source pattern, i.e. $p(\mathsf{y}|\mathsf{x}) = \prod_{i=1}^{|\mathsf{y}|} p_i(y_i|\mathsf{x})$, where all $p_i(y_i|\mathsf{x}) \equiv p(y_i|\mathsf{x})$ are in the same parametric family. If we view $\mathsf{x} \in \mathbb{R}^{|\mathsf{x}|}$ as a fixed parameter, and the vector of binary spikes $\mathsf{y} \in \{\pm 1\}^{|\mathsf{y}|}$ – as $|\mathsf{y}|$ independent identically distributed samples from $p(y_i|\mathsf{x})$, we may apply the results of the theory of statistical parameter estimation (see e.g. Cramer (1946)) to bound the mutual information $I(\mathsf{x},\mathsf{y})$ (Brunel and Nadal (1998)).

Indeed, let $\hat{\mathsf{x}} \in \mathbb{R}^{|\mathsf{x}|}$ be a statistical estimator of the input stimulus x obtained from the stochastic neural firings y. Generally, we may assume that the sources, the encodings, and the estimators form a Markov chain $\mathsf{x} \to \mathsf{y} \mapsto \hat{\mathsf{x}}$, where $p(\hat{\mathsf{x}}|\mathsf{y}) \sim \delta(\hat{\mathsf{x}} - \hat{\mathsf{x}}(\mathsf{y}))$ and $p(\mathsf{y}|\mathsf{x})$ is factorized in y. It is well known that if the estimator $\hat{\mathsf{x}}(\mathsf{y})$ is unbiased, i.e. $\langle \hat{\mathsf{x}}(\mathsf{y}) \rangle_{p(\mathsf{y}|\mathsf{x})} = \mathsf{x}$, then its covariance may be bounded according to the Cramer-Rao inequality, i.e.

$$\left\langle (\hat{\mathsf{x}}(\mathsf{y}) - \mathsf{x}) (\hat{\mathsf{x}}(\mathsf{y}) - \mathsf{x})^T \right\rangle_{p(\mathsf{y}|\mathsf{x})} \succeq \mathsf{F}_{\mathsf{x}} \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{x}|}, \tag{6.18}$$

where $\mathsf{A}_1 \succeq \mathsf{A}_2$ indicates that $\mathsf{A}_1 - \mathsf{A}_2$ is positive semi-definite, and

$$\mathsf{F}_{\mathsf{x}} = \{F_{ij}(\mathsf{x})\} \stackrel{\text{def}}{=} -\left\{ \langle \partial^2 \log p(\mathsf{y}|\mathsf{x}) / \partial x_i \partial x_j \rangle_{p(\mathsf{y}|\mathsf{x})} \right\} \tag{6.19}$$

$$\equiv \left\{ \langle \partial \log p(\mathsf{y}|\mathsf{x}) / \partial x_i \cdot \partial \log p(\mathsf{y}|\mathsf{x}) / \partial x_j \rangle_{p(\mathsf{y}|\mathsf{x})} \right\} \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{x}|} \tag{6.20}$$

is the Fisher Information matrix (see e.g. Cramer (1946), Cover and Thomas (1991), Johnson (2003)). It is easy to see that we may equivalently express the bound (6.18) as

$$\mathrm{cov}(\hat{\mathsf{x}}|\mathsf{x}) \succeq \mathsf{F}_{\mathsf{x}} \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{x}|}. \tag{6.21}$$

Here $\mathrm{cov}(\hat{\mathsf{x}}|\mathsf{x})$ is the conditional covariance of $p(\hat{\mathsf{x}}|\mathsf{x})$, and we have used the fact that for unbiased estimators $\langle \hat{\mathsf{x}} \rangle_{p(\hat{\mathsf{x}}|\mathsf{x})} = \langle \hat{\mathsf{x}}(\mathsf{y}) \rangle_{p(\mathsf{y}|\mathsf{x})} = \mathsf{x}$.

If the estimator $\hat{\mathsf{x}}(\mathsf{y})$ is also *efficient*, its covariance saturates the Cramer-Rao bound, which results in an upper bound on the entropy of the conditional distribution $H\left(p(\hat{\mathsf{x}}|\mathsf{x})\right) \leq H(\mathcal{N}_{\mathsf{x}}(\mathbf{0}, \mathsf{F}_{\mathsf{x}}^{-1}))$. The bound follows from the well-known fact that for a fixed covariance, the maximum entropy distribution is a Gaussian (see e.g. Feller (1971), McEliece (1977)). We may now obtain a lower bound on the mutual information

$$I(\mathsf{x},\mathsf{y}) \geq H(\hat{\mathsf{x}}) + \frac{1}{2} \langle \log |\mathsf{F}_{\mathsf{x}}| \rangle_{\tilde{p}(\mathsf{x})} + const, \tag{6.22}$$

where we have used the data processing inequality $I(\mathsf{x},\mathsf{y}) \geq I(\mathsf{x},\hat{\mathsf{x}})$ for the considered chain $\mathsf{x} \to \mathsf{y} \mapsto \hat{\mathsf{x}}$.

Unfortunately, despite the fact that the mapping $\mathsf{y} \mapsto \hat{\mathsf{x}}$ is deterministic, exact computation of the entropy of statistical estimates $H(\hat{\mathsf{x}})$ in the objective (6.22) is in general computationally intractable. Brunel and Nadal (1998) show that under some assumptions we may assume $H(\hat{\mathsf{x}}) \approx H(\mathsf{x})$, which leads to the approximation

$$I(\mathsf{x},\mathsf{y}) \gtrsim \tilde{I}_F(\mathsf{x},\mathsf{y}) \stackrel{\text{def}}{=} H(\mathsf{x}) + \frac{1}{2}\langle \log|\mathsf{F}_\mathsf{x}| \rangle_{\tilde{p}(\mathsf{x})} + const \tag{6.23}$$

(specifically, this approximation applies for $|\mathsf{y}| \to \infty$ when the estimators are sharply peaked around the mean values). Brunel and Nadal also show that under similar assumptions, (6.23) may be used as an approximation of $I(\mathsf{x},\mathsf{y})$ independently of the bias of the estimator. Note that since $H(\mathsf{x})$ is independent of the parameters of $p(\mathsf{y}|\mathsf{x})$, maximization of (6.23) is equivalent to maximization of (6.22) where the generally intractable term $H(\hat{\mathsf{x}})$ (entropy of the mixture) is ignored.

6.3.1.1 Sigmoidal Activations

It is straight-forward to see that for the considered sigmoidal activations (6.1), each element $F_{ij}(\mathsf{x})$ of the Fisher Information matrix $\mathsf{F}_\mathsf{x} = \{F_{ij}(\mathsf{x})\}$ is given by

$$F_{ij}(\mathsf{x}) = \sum_{l=1}^{|\mathsf{y}|} w_{li} w_{lj} \sigma(\mathsf{w}_l^T \mathsf{x} + b_l)(1 - \sigma(\mathsf{w}_l^T \mathsf{x} + b_l)), \quad i,j = 1,\ldots,|\mathsf{y}|, \tag{6.24}$$

where w_l^T is the l^{th} row of the synaptic weights $\mathsf{W} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{x}|}$. Then we may express the Fisher approximation of the mutual information (6.23) as

$$\tilde{I}_F(\mathsf{x},\mathsf{y}) \propto \sum_{m=1}^{M} \log\left|\mathsf{W}^T\left(\mathsf{I} - \mathsf{D}_{\lambda_{x_m}}\right)\mathsf{W}\right| + const, \tag{6.25}$$

where the constant incorporates the remaining terms which have no effect on the optimization. Again, $\mathsf{I} - \mathsf{D}_{\lambda_{x_m}} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{y}|}$ is the conditional covariance of the stochastic spikes (given by expression (6.5)).

By computing the matrix derivatives of (6.25) for $\mathsf{W} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{x}|}$ and $\mathsf{b} \in \mathbb{R}^{|\mathsf{y}|}$ and performing some straight-forward algebraic manipulations, we get

$$\frac{\partial \tilde{I}_F}{\partial \mathsf{W}} = \frac{1}{4M}\sum_{m=1}^{M}\left\{2(\mathsf{I} - \mathsf{D}_{\lambda_{x_m}})\mathsf{W}\mathsf{A}_{x_m} - \mathsf{C}_{x_m}(\mathsf{I} - \mathsf{D}_{\lambda_{x_m}})\boldsymbol{\lambda}(\mathsf{x}^{(m)})\left(\mathsf{x}^{(m)}\right)^T\right\} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{x}|}. \tag{6.26}$$

Here $\boldsymbol{\lambda}(\mathsf{x}^{(m)}) \stackrel{\text{def}}{=} \langle \mathsf{y} \rangle_{p(\mathsf{y}|\mathsf{x}^{(m)})} \in \mathbb{R}^{|\mathsf{y}|}$ is a vector of the expected firing at the encoding layer (see expression (6.4)), and

$$\mathsf{A}_{x_m} = \left(\mathsf{W}^T\left(\mathsf{I} - \mathsf{D}_{\lambda_{x_m}}\right)\mathsf{W}\right)^+ \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{x}|} \tag{6.27}$$

$$\mathsf{C}_{x_m} = \operatorname{diag}\left(\mathsf{W}\mathsf{A}_{x_m}\mathsf{W}^T\right) \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{y}|}, \tag{6.28}$$

where A^+ is the pseudo-inverse[4] (so that $\mathsf{A}\mathsf{A}^+\mathsf{A} = \mathsf{A}$). Analogously,

$$\frac{\partial \tilde{I}_F}{\partial \mathsf{b}} = -\frac{1}{4M}\sum_{m=1}^{M} \mathsf{C}_{x_m}(\mathsf{I} - \mathsf{D}_{\lambda_{x_m}})\boldsymbol{\lambda}(\mathsf{x}^{(m)}) \in \mathbb{R}^{|\mathsf{y}|} \tag{6.29}$$

with the similar definitions of $\boldsymbol{\lambda}(\mathsf{x}^{(m)}) \in \mathbb{R}^{|\mathsf{y}|}$ and $\mathsf{C}_{x_m} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{y}|}$.

It is interesting to note that for the square model with $|\mathsf{x}| = |\mathsf{y}|$, optimization of (6.25) leads to

$$\Delta\mathsf{W} = 2\mathsf{W}^{-T} - \langle \boldsymbol{\lambda}_x \mathsf{x}^T \rangle, \tag{6.30}$$

which (apart from the coefficient at the inverse weight – *redundancy* term) has the same form as the learning rule of Bell and Sejnowski (1995) derived for the invertible channel with $p(\mathsf{y}|\mathsf{x}) \sim \delta(\mathsf{y} - \sigma(\mathsf{x}))$. If the encoded representations are overcomplete (i.e. $|\mathsf{x}| < |\mathsf{y}|$) and the variable rate $\mathrm{cov}(\mathsf{y}|\mathsf{x}^{(m)}) = \mathsf{I} - \mathsf{D}_{\lambda_{x_m}}$ is approximated by the 1^{st}-order Taylor expansion around zero receptive fields, the redundancy term in (6.30) is replaced by the transposed pseudo-inverse $\mathsf{W}(\mathsf{W}^T\mathsf{W})^{-1} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{x}|}$. Notably, the resulting learning rule (6.26), (6.29) is non-local (as it involves computations of the inverses), and the weight updates (6.26), (6.30) have no Hebbian terms.

More importantly, one can see that optimization of the Fisher approximation of the mutual information (6.25) may be problematic when $\mathsf{W}^T(\mathsf{I} - \mathsf{D}_{\lambda_{x_m}})\mathsf{W} \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{x}|}$ is rank-deficient, which complicates applicability of the method for a variety of tasks involving relatively low-dimensional encodings of high-dimensional input stimuli. Apart from instability of numerical optimization on $\tilde{I}_F(\mathsf{y},\mathsf{x})$ (which we have partially addressed by computing the pseudo-inverses, rather than the exact inverses in (6.27), (6.28)), the conceptual problem of optimizing the criterion $\tilde{I}_F(\mathsf{y},\mathsf{x})$ is the weakness of the approximate bound for $|\mathsf{y}| < |\mathsf{x}|$ (in most cases, such channel choices would violate the fundamental assumptions of the approximation). Additionally, from (6.25) we may note that the approximate bound $\tilde{I}_F(\mathsf{x},\mathsf{y})$ may become loose with the decrease in the conditional variance of the stochastic firings. Intuitively, this happens because the directions of low variations in the code space swamp the optimized volume of the $|\mathsf{x}|$-dimensional manifold (see (6.25)), thus leading to a drop in the determinant.

6.3.2 Local Approximation of Mutual Information

Here we describe an alternative approximation of mutual information $I(\mathsf{x},\mathsf{y})$. It is inspired by recent work of Szummer and Jaakkola (2002) and Corduneanu and Jaakkola (2003), who approximate information content $I(\mathsf{x} \in \mathcal{R}, \mathsf{y})$ in infinitely small local regions $\mathcal{R} \subset \mathcal{R}_\mathsf{x}$, and apply the approximations as regularizers for semi-supervised classification. Here we show that their method may in fact be used to obtain an alternative approximate lower bound on mutual information.

Consider the model $r \to \mathsf{x} \to \mathsf{y}$, where r defines a local region in the data space, so that $\mathcal{R}_r \subset \mathcal{R}_\mathsf{x}$. The local information $I_r(\mathsf{x},\mathsf{y})$ in the region r may be

[4]We make a recourse to computing the pseudo-inverse in order to handle the singularity of the gradient $\partial\tilde{I}_F(\mathsf{x},\mathsf{y})/\partial\mathsf{W}$ for rank-deficient Fisher Information matrices (which happens for example when $|\mathsf{x}| > |\mathsf{y}|$). Note that $\mathsf{A}^+ = \mathsf{A}^{-1} \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{x}|}$ when $\mathrm{rank}(\mathsf{A}) = |\mathsf{x}|$.

defined as

$$I_r(\mathsf{x},\mathsf{y}) \stackrel{\text{def}}{=} \int_{\mathcal{R}_\mathsf{y}} \int_{\mathcal{R}_r} p(\mathsf{x}|r)p(\mathsf{y}|\mathsf{x}) \log \frac{p(\mathsf{y}|\mathsf{x})}{p(\mathsf{y}|r)} \, \mathrm{d}\mathsf{x} \, \mathrm{d}\mathsf{y}, \tag{6.31}$$

where $\mathcal{R}_\mathsf{y}$ is the code space. The conditional $p(\mathsf{x}|r)$ may generally be computed as $p(\mathsf{x}|r) \propto p(\mathsf{x})\mathbb{I}(\mathsf{x} \in \mathcal{R}_r)$ [$\mathbb{I}$ defines an indicator variable], where $p(\mathsf{x})$ is the distribution of the source variables. By computing the 2^{nd}-order Taylor expansion of $p(\mathsf{y}|\mathsf{x})$ around $p(\mathsf{y}|\langle\mathsf{x}\rangle_{p(\mathsf{x}|r)})$ and performing some algebraic manipulations, Corduneanu and Jaakkola (2003) show that the local information (6.31) may be approximated as

$$I_r(\mathsf{x},\mathsf{y}) = \frac{1}{2}\mathrm{tr}\left\{\mathrm{cov}(\mathsf{x}|r)\mathsf{F}_{\langle\mathsf{x}\rangle_{p(\mathsf{x}|r)}}\right\} + O(\mathcal{R}_r^3), \tag{6.32}$$

where $\mathsf{F}_{\langle\mathsf{x}\rangle_{p(\mathsf{x}|r)}}$ is the Fisher Information matrix (6.19) computed at the local mean $\langle\mathsf{x}\rangle_{p(\mathsf{x}|r)}$.

If the local regions $\mathcal{R}_r$ are symmetric and centered at the training patterns $\{\mathsf{x}\}$ then we may define $\mathrm{cov}(\mathsf{x}|r) \stackrel{\text{def}}{=} v_{\mathsf{x}|r}\mathsf{I}_{|\mathsf{x}|}$, and up to a constant pre-factor we obtain[5]

$$\langle I_r(\mathsf{x},\mathsf{y})\rangle_{p(r)} \approx v_{\mathsf{x}|r}\langle \mathrm{tr}\,\{\mathsf{F}_\mathsf{x}\}\rangle_{\tilde{p}(\mathsf{x})} \tag{6.33}$$

(approximation is accurate for small regions $\mathcal{R}_r$, i.e. small variances $v_{\mathsf{x}|r}$). Again, $\tilde{p}(\mathsf{x})$ is the empirical distribution. We also note that when expression (6.33) is coupled with the minimization of the sum-of-squared error, the resulting objective is equivalent (to first order) to minimization of the sum-of-squared error with noise, i.e. when the source distribution is given by $p(\mathsf{x}) = \sum_{m=1}^{M} \mathcal{N}_\mathsf{x}(\mathsf{x}^{(m)}, v_{\mathsf{x}|r}\mathsf{I}_{|\mathsf{x}|})$ for the training set $\{\mathsf{x}^{(m)}|m = 1,\ldots,M\}$, and the encoding distribution $p(\mathsf{y}|\mathsf{x})$ is approximated around $p(\mathsf{y}|\mathsf{x}^{(m)})$ (see Bishop (1995)).

We will now show that (6.33) may be used to derive an approximate lower bound on the global mutual information $I(\mathsf{x},\mathsf{y})$. To do so, we will outline a general relation between $I(\mathsf{x},\mathsf{y})$ and the local average $\langle I_r(\mathsf{x},\mathsf{y})\rangle_{p(r)}$ for the considered model $p(r,\mathsf{x},\mathsf{y}) = p(r)p(\mathsf{x}|r)p(\mathsf{y}|\mathsf{x})$. From the chain rule on mutual information (see e.g. Cover and Thomas (1991)), we may express the joint mutual information between the code and the source vectors (with the corresponding local partitionings) as

$$I(\{r,\mathsf{x}\},\mathsf{y}) = I(r,\mathsf{y}) + I(\mathsf{x},\mathsf{y}|r). \tag{6.34}$$

On the other hand, for the considered chain model we get

$$I(\{r,\mathsf{x}\},\mathsf{y}) = H(\mathsf{y}) - H(\mathsf{y}|\mathsf{x},r) = I(\mathsf{x},\mathsf{y}), \tag{6.35}$$

where $I(\{r,\mathsf{x}\},\mathsf{y})$ is the amount of information which the codes y contain about the sources and the regions x, r jointly. From (6.31), (6.33), and the definition of the conditional mutual information

$$I(\mathsf{x},\mathsf{y}|r) \stackrel{\text{def}}{=} H(\mathsf{y}|r) - H(\mathsf{y}|\mathsf{x},r) = \langle I_r(\mathsf{x},\mathsf{y})\rangle_{p(r)}, \tag{6.36}$$

[5]This result may be generalized for the case of infinitely small local regions $\mathcal{R}_r$, provided that the covering of the data space satisfies specific properties. For example, (6.33) applies if $\mathcal{R}_\mathsf{x}$ consists of axis-parallel cubes $\mathcal{R}_r \to 0$ centered at the axis-parallel lattice points, see Corduneanu and Jaakkola (2003) for details.

it is clear that

$$I(\mathsf{x},\mathsf{y}) = I(r,\mathsf{y}) + \langle I_r(\mathsf{x},\mathsf{y})\rangle_{p(r)} \gtrapprox v_{\mathsf{x}|r}\langle \mathrm{tr}\left\{\mathsf{F}_{\mathsf{x}}\right\}\rangle_{\tilde{p}(\mathsf{x})} \stackrel{\mathrm{def}}{=} \tilde{I}_L(\mathsf{x},\mathsf{y}), \tag{6.37}$$

where the slackness of the bound is stipulated by the non-negativity of $I(r,\mathsf{y})$. Note that the variance $v_{\mathsf{x}|r}$ is a constant parameter which does not affect the optimization surface for the encoding distribution $p(\mathsf{y}|\mathsf{x})$. Note that for the special case of one-dimensional input stimuli, the approximation (6.37) reduces to the well-known scalar Fisher criterion optimized by a number of the currently used approaches to population coding (e.g. Pouget et al. (1998), Zhang and Sejnowski (1999), Bethge et al. (2002)).

Despite the apparent similarity of the Fisher approximation criterion $\tilde{I}_F(\mathsf{x},\mathsf{y})$ (expression (6.23)) and the local criterion $\tilde{I}_L(\mathsf{x},\mathsf{y})$ (expression (6.37)), one can see that in general the methods result in different fixed points. Indeed, if $\lambda_i(\mathsf{F}_{\mathsf{x}})$ is the i^{th} eigenvalue of $\mathsf{F}_{\mathsf{x}} \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{x}|}$ then the extrema of $\tilde{I}_F(\mathsf{x},\mathsf{y})$ must satisfy

$$\partial\tilde{I}_F(\mathsf{x},\mathsf{y})/\partial\mathsf{W} = \left\langle \sum_{i=1}^{|\mathsf{x}|} \lambda_i^{-1}(\mathsf{F}_{\mathsf{x}}) \frac{\partial\lambda_i(\mathsf{F}_{\mathsf{x}})}{\partial\mathsf{W}} \right\rangle_{\tilde{p}(\mathsf{x})} = \mathbf{0} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{x}|}. \tag{6.38}$$

On the other hand,

$$\partial\tilde{I}_L(\mathsf{x},\mathsf{y})/\partial\mathsf{W} = \left\langle \sum_{i=1}^{|\mathsf{x}|} \frac{\partial\lambda_i(\mathsf{F}_{\mathsf{x}})}{\partial\mathsf{W}} \right\rangle_{\tilde{p}(\mathsf{x})} = \mathbf{0} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{x}|}. \tag{6.39}$$

Clearly, both (6.38) and (6.39) hold simultaneously if $\partial\lambda_i(\mathsf{F}_{\mathsf{x}})/\partial\mathsf{W} = \mathbf{0}$ for all eigenvalues $\{\lambda_i | i = 1,\ldots,|\mathsf{x}|\}$, and all training patterns $\{\mathsf{x}^{(m)} \in \mathbb{R}^{|\mathsf{x}|} | m = 1,\ldots,M\}$. However, it is clear that despite positive semi-definiteness of $\mathsf{F}_{\mathsf{x}} \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{x}|}$, the criteria $\tilde{I}_F(\mathsf{x},\mathsf{y})$ and $\tilde{I}_L(\mathsf{x},\mathsf{y})$ will generally give rise to different solutions.

Finally, it is important to note that in contrast to the Fisher approximation $\tilde{I}_F(\mathsf{x},\mathsf{y})$ of Brunel and Nadal (1998), the local approximation $\tilde{I}_L(\mathsf{x},\mathsf{y})$ is also defined for $|\mathsf{x}| > |\mathsf{y}|$. This suggests that in general numerical optimization of (6.37) may be more stable than optimization of (6.23) for small population sizes. Optimization of the local approximation is also less computationally demanding, since by optimizing $\tilde{I}_L(\mathsf{x},\mathsf{y})$ we avoid computing the inverse of the Fisher Information matrix. Moreover, as we will see, for the considered logistic parameterization of the encoder distribution, numerical ascent on $\tilde{I}_L(\mathsf{x},\mathsf{y})$ may often lead to better values of the exact mutual information $I(\mathsf{x},\mathsf{y})$ (which we will compute for small models for the purpose of testing).

6.3.2.1 Sigmoidal Activations

By analogy with Section 6.3.1, we may derive the local approximation of $I(\mathsf{x},\mathsf{y})$ for the considered sigmoidal encoder model (6.1). Up to the constant pre-factor $v_{\mathsf{x}|r}$, the local approximation is given by

$$\tilde{I}_L(\mathsf{x},\mathsf{y}) \propto \sum_{m=1}^{M} \mathrm{tr}\left\{\mathsf{W}^T\left(\mathsf{I} - \mathsf{D}_{\lambda_{x_m}}\right)\mathsf{W}\right\}, \tag{6.40}$$

where $\mathsf{I} - \mathsf{D}_{\lambda_{x_m}} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{y}|}$ is given by expression (6.5). Note that in contrast to the previously discussed Fisher approximation for sigmoidal activations (6.25), the objective (6.40) is finite even when $\mathsf{F}_\mathsf{x} \in \mathbb{R}^{|\mathsf{x}|\times\mathsf{x}}$ is rank-deficient. This generally leads to a more stable behavior of the numerical optimization procedures performing an ascent on $\tilde{I}_L(\mathsf{x},\mathsf{y})$.

The gradients for the encoder parameters in this case are given by

$$\frac{\partial \tilde{I}_L}{\partial \mathsf{W}} \propto \frac{1}{4M}\sum_{m=1}^{M}\left\{2(\mathsf{I}-\mathsf{D}_{\lambda_{x_m}})\mathsf{W} - \mathrm{diag}(\mathsf{W}\mathsf{W}^T)(\mathsf{I}-\mathsf{D}_{\lambda_{x_m}})\boldsymbol{\lambda}(\mathsf{x}^{(m)})\left(\mathsf{x}^{(m)}\right)^T\right\} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{x}|} \tag{6.41}$$

$$\frac{\partial \tilde{I}_L}{\partial \mathsf{b}} \propto -\frac{1}{4M}\sum_{m=1}^{M}\mathrm{diag}(\mathsf{W}\mathsf{W}^T)(\mathsf{I}-\mathsf{D}_{\lambda_{x_m}})\boldsymbol{\lambda}(\mathsf{x}^{(m)}) \in \mathbb{R}^{|\mathsf{y}|}, \tag{6.42}$$

where $\mathrm{diag}(\mathsf{W}\mathsf{W}^T) \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{y}|}$ is the diagonal matrix of the squares of L2-norms $\|\mathsf{w}_i\|^2$ of the encoder weight vectors. Note that symbolically the only difference of the gradients of the local approximation $\tilde{I}_L(\mathsf{x},\mathsf{y})$ (expressions (6.41) and (6.42)) from the gradients of the Fisher approximation $\tilde{I}_F(\mathsf{x},\mathsf{y})$ ((6.26) and (6.29)) is multiplication by the non-local term $\mathsf{A}_{x_m} \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{x}|}$ in $\partial\tilde{I}_F/\partial\mathsf{W}$ and $\mathsf{C}_{x_m} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{y}|}$ (see expressions (6.27), (6.28)).

6.3.3 Constraints on the Encoder Parameters

Interestingly, from the definitions (6.25), (6.40) of the Fisher and local approximations of the mutual information for the considered encoder model, one can see that unconstrained optimization for the encoder weights $\mathsf{W} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{x}|}$ does not necessarily lead to their divergence. To see this, consider the simple case of a scalar input stimulus (e.g. $x \in \mathbb{R}$ may correspond to the angle of saccadic eye movements or direction of head motion) encoded by a single spiking neuron $y \in \{-1, 1\}$. The only synaptic weight for this simple model would be given by $w \in \mathbb{R}$ (for clarity, we ignore the bias b). From (6.40) and the definition of the conditional variance (6.5), it is clear that

$$\tilde{I}_L(\mathsf{x},\mathsf{y}) \propto \left\langle w^2/(2+e^{-wx}+e^{wx})\right\rangle_{\tilde{p}(\mathsf{x})} \geq 0, \tag{6.43}$$

where $\tilde{p}(\mathsf{x})$ is the empirical distribution. Assume that $\exists\epsilon > 0$ such that for all training patterns $|x^{(m)}| > \epsilon$. Then by computing the limit of (6.43) for $|w| \to \infty$ we can see that the approximation $\tilde{I}_L(\mathsf{x},\mathsf{y})$ is *minimized* for the divergent weight norm $|w|$. (Note that finding the analytical solution for w which maximizes (6.43) would require solving a transcendental equation even for this simple model with $|\mathsf{x}| = |\mathsf{y}| = 1$). Similar intuition applies in the case of Brunel and Nadal's approximation[6] $\tilde{I}_F(\mathsf{x},\mathsf{y})$.

Analogously, for the high-dimensional case we may note that the conditional variances of the stochastic firings are functions of the singular values $\mathsf{L} \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{x}|}$

[6]Consider the case when $|x| = |y| = 1$. For the Fisher approximation we note that from Jensen's inequality $\tilde{I}_F(x,y) = \langle\log|\mathsf{F}_x|\rangle_{\tilde{p}(x)} \leq \log\langle|\mathsf{F}_x|\rangle_{\tilde{p}(x)} \stackrel{\text{def}}{=} \hat{I}_F(x,y)$. It is easy to see that $\lim_{|w|\to\infty}\hat{I}_F(x,y) \to -\infty$. Since $\tilde{I}_F(x,y) \leq \hat{I}_F(x,y)$, we get $\lim_{|w|\to\infty}\tilde{I}_F(x,y) \to -\infty$.

of the encoding weights $\mathsf{W} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{x}|}$. For the considered choice of the encoding distributions, higher weight magnitudes would generally lead to lower conditional variances of the firings, i.e. the growth in $\|\mathsf{W}\|_F \equiv \mathrm{tr}^{1/2}\{\mathsf{W}\mathsf{W}^T\}$ in the approximation $\tilde{I}_L(\mathsf{x},\mathsf{y})$ would be compensated by a decrease in $\mathsf{I} - \mathsf{D}_{\lambda_{x_m}} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{y}|}$. This suggests generally non-trivial changes in $\hat{I}_L(\mathsf{x},\mathsf{y})$ with an increase in the weight magnitudes (*cf* Gaussian channels with the isotropic noise).

Despite the observation that unconstrained optimization of $\tilde{I}_F(x,y)$ and $\tilde{I}_L(x,y)$ would not necessarily lead to indefinite growth of the encoder weights W (at least, for the considered channel), in many cases we may still be interested in constraining the magnitudes of the synaptic weights or the conditional variances of the firings. This may be motivated, for example, by neuro-physiological constraints on the channel, or technical limitations on precision. Possibly one of the simplest ways of introducing implicit constraints on the encoder parameters for the considered channels is to penalize large Frobenius norms $\|\mathsf{W}\|_F$, or small conditional variances $\mathrm{var}(y_j|\mathsf{x}) = 1 - \lambda_j^2(\mathsf{x}^{(m)})$ of the firing units. It is straight-forward to see that the gradients of the penalty terms would in this case be given as $-\mathsf{M}\mathsf{W}$ and $-\mathsf{M}\langle\boldsymbol{\lambda}_x\mathsf{x}^T\rangle \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{x}|}$ (which is the anti-Hebbian term) respectively, where $\mathsf{M} \succeq 0 \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{y}|}$ is a fixed positive semi-definite matrix. However, while being conceptually and computationally simple, the choice of the penalty constants M may be rather heuristic. A more rigorous approach would involve optimization of the dual Lagrangian

$$\tilde{\mathbb{L}} \stackrel{\mathrm{def}}{=} \sup_{\mathsf{W}}\{\mathbb{L}(\mathsf{W},\mathsf{M})\} \stackrel{\mathrm{def}}{=} \sup_{\mathsf{W}}\{\tilde{I}_F(\mathsf{x},\mathsf{y}) - \mathsf{m}^T\mathsf{f}(\mathsf{W})\}, \tag{6.44}$$

where $\mathsf{f}(\mathsf{W}) \geq \mathbf{0}$ defines a set of the inequality constraints, and $\mathsf{m} \geq \mathbf{0}$ is a vector of Lagrange multipliers (see e.g. Bertsekas (1996), Boyd and Vandenberghe (2004)), which requires further numerical approximations when the dual cannot be expressed as an analytical function of W alone (e.g. Rubinov and Yang (2003)).

Alternatively, we may impose explicit construction constraints on the encoder parameters $\mathsf{W} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{x}|}$, for example by constraining the weights to a unit hypersphere. For example, for $|\mathsf{y}| \geq |\mathsf{x}|$ we may define $\mathsf{W} = \tilde{\mathsf{W}}(\tilde{\mathsf{W}}^T\tilde{\mathsf{W}})^{-1/2}$, where $\tilde{\mathsf{W}} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{x}|}$ is an arbitrary matrix such that $\mathrm{rank}(\tilde{\mathsf{W}}) = |\mathsf{y}|$, and re-compute the derivatives for $\tilde{\mathsf{W}}$ (the case when $|\mathsf{x}| \geq |\mathsf{y}|$ is analogous). We may also bound each synaptic weight, so that $w_{ij} \in [-\omega,\omega]$, for example by imposing parametric constraints $w_{ij} \stackrel{\mathrm{def}}{=} \phi(a_{ij})$, where $\phi(a) : \mathbb{R} \to [-\omega,\omega]$. Presumably, such constraints on the individual synaptic weights are more biologically intuitive; moreover, they do not require non-local computations. Optimization of $\tilde{I}_F(\mathsf{x},\mathsf{y})$ or $\tilde{I}_L(\mathsf{x},\mathsf{y})$ with respect to $\mathsf{W} = \{w_{ij}\} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{x}|}$ would then need to be replaced by optimization with respect to $\mathsf{A} = \{a_{ij}\} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{x}|}$, with the gradients expressed e.g. by

$$\partial\tilde{I}_F(\mathsf{x},\mathsf{y})/\partial\mathsf{A} = \partial\tilde{I}_F(\mathsf{x},\mathsf{y})/\partial\mathsf{W} \circ \nabla_{\mathsf{A}}\phi(\mathsf{A}) \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{x}|}. \tag{6.45}$$

Here $\mathsf{A}_1 \circ \mathsf{A}_2$ denotes the element-wise product, and $\nabla_{\mathsf{A}}\phi(\mathsf{A}) \stackrel{\mathrm{def}}{=} \{\partial\phi(a_{ij})/\partial a_{ij}\} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{x}|}$. Interestingly, we can empirically demonstrate that in some cases (e.g. when the input stimuli $\mathsf{x} \in \mathbb{R}^{|\mathsf{x}|}$ are sufficiently close to $\mathbf{0}_{|\mathsf{x}|}$), this choice of the constraint on $\mathsf{W} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{x}|}$ pushes the optimal weights towards the corners of the

hypercube $[-\omega, \omega]^{|\mathsf{y}|\times|\mathsf{x}|}$, thus suggesting a straight-forward approach to unsupervised information-theoretic training of ising perceptrons (see e.g. Penney and Sherrington (1993), Rosen-Zvi and Kanter (2001)).

6.4 Variational Lower Bound vs. Fisher and Local Approximations of Mutual Information

Since $\tilde{I}_F(\mathsf{x},\mathsf{y})$ and $\tilde{I}_L(\mathsf{x},\mathsf{y})$ are not proper lower bounds on the mutual information, it is difficult to analyze their tightness or compare them with the generic variational bound (2.2). To illustrate a relation between the approaches, we may consider a Gaussian decoder $q(\mathsf{x}|\mathsf{y}) \sim \mathcal{N}_{\mathsf{x}}(\boldsymbol{\mu}_{\mathsf{y}}; \boldsymbol{\Sigma})$, which transforms the variational bound into

$$\tilde{I}(\mathsf{x},\mathsf{y}) = -\frac{1}{2}\left\langle \mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}(\mathsf{x}-\boldsymbol{\mu}_y)(\mathsf{x}-\boldsymbol{\mu}_y)^T\right\}\right\rangle_{p(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x})} + \frac{1}{2}\log|\boldsymbol{\Sigma}^{-1}| + \acute{c}. \tag{6.46}$$

Here $\acute{c}$ incorporates $H(\mathsf{x})$ and other constants which do not affect the optimization surface for the encoder parameters, and $\boldsymbol{\Sigma} \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{x}|}$ is a function of parameters of the conditional $p(\mathsf{y}|\mathsf{x})$.

Clearly, if the log eigenspectrum of the inverse covariance of the decoder is constrained to satisfy

$$\sum_{i=1}^{|\mathsf{x}|}\log l_i(\boldsymbol{\Sigma}^{-1}) = \sum_{i=1}^{|\mathsf{x}|}\langle \log l_i(\mathsf{F}_{\mathsf{x}})\rangle_{\tilde{p}(\mathsf{x})}, \tag{6.47}$$

where $\{l_i(\boldsymbol{\Sigma}^{-1})\}$ and $\{l_i(\mathsf{F}_{\mathsf{x}})\}$ are eigenvalues of $\boldsymbol{\Sigma}^{-1}$ and F_{x} respectively, then the lower bound (6.46) reduces to the objective (6.23) amended with the average quadratic reconstruction error

$$\tilde{I}(\mathsf{x},\mathsf{y}) = -\frac{1}{2}\underbrace{\left\langle \mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}(\mathsf{x}-\boldsymbol{\mu}_y)(\mathsf{x}-\boldsymbol{\mu}_y)^T\right\}\right\rangle_{p(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x})}}_{\text{reconstruction error}} + \frac{1}{2}\underbrace{\langle \log|\mathsf{F}_{\mathsf{x}}|\rangle_{\tilde{p}(\mathsf{x})}}_{\text{Fisher criterion}} + \acute{c}. \tag{6.48}$$

Arguably, it is due to the subtraction of the non-negative quadratic term that (6.46) remains a general lower bound independently of the parameterization of the model and spectral properties of $\mathsf{F}_{\mathsf{x}} \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{x}|}$. We may find an analogous relation for the local approximation of mutual information $\tilde{I}(\mathsf{x},\mathsf{y})$ by constraining the log eigenspectrum of $\boldsymbol{\Sigma}^{-1}$ to satisfy

$$\sum_{i=1}^{|\mathsf{x}|}\log l_i(\boldsymbol{\Sigma}^{-1}) = v_{\mathsf{x}|r}\sum_{i=1}^{|\mathsf{x}|}\langle l_i(\mathsf{F}_{\mathsf{x}})\rangle_{\tilde{p}(\mathsf{x})}, \tag{6.49}$$

where $v_{\mathsf{x}|r}$ is a constant variance of the symmetric regions (see the discussion in Section 6.3.2). It is easy to see that in this case we get

$$\tilde{I}(\mathsf{x},\mathsf{y}) = -\frac{1}{2}\underbrace{\left\langle \mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}(\mathsf{x}-\boldsymbol{\mu}_y)(\mathsf{x}-\boldsymbol{\mu}_y)^T\right\}\right\rangle_{p(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x})}}_{\text{reconstruction error}} + \frac{1}{2}\underbrace{\langle \mathrm{tr}\{\mathsf{F}_{\mathsf{x}}\}\rangle_{\tilde{p}(\mathsf{x})}}_{\text{local criterion}} + \acute{c}. \tag{6.50}$$

One potential advantage of the variational approach over the Fisher approximation of mutual information is the fact that the optimized objective $\tilde{I}(\mathsf{x}, \mathsf{y})$ remains a proper bound on $I(\mathsf{x}, \mathsf{y})$ independently of the modeling assumptions. This contrasts with the approximations of Brunel and Nadal (1998) and the related methods of Kang and Sompolinsky (2001) and Hoch et al. (2003), where the accuracy is strongly influenced by the size of the code space. While the method based on local approximations of Szummer and Jaakkola (2002), Corduneanu and Jaakkola (2003) helps to avoid some of the constraints on the encoding mappings (specifically, it is applicable for $|\mathsf{y}| < |\mathsf{x}|$), it may be rather difficult to justify from the information-theoretic viewpoint (as by introducing the noise into the empirical distribution we decrease the information content between the codes and the original sources).

Another principal advantage of the variational approach to information maximization is the flexibility in the choice of the variational decoder. Intuitively, if the Fisher Information matrices are nearly singular, both (6.23) and (6.48) may be quite weak. However, by relaxing (6.47) and imposing full-rank constraints on the covariances of the variational decoders, the variational bound may be significantly strengthened. Moreover, as we showed in Section 6.2.1, by imposing specific constraints on the variational decoder, we can derive an iterative learning rule, which only requires local computations (such as evaluations of weighted Hebbian and anti-Hebbian terms). As discussed in Section 6.2.1, this may lead to an arguably more biologically plausible learning.

6.5 Demonstrations

Variational IM vs Fisher and Local Approximations

As we mentioned in Section 6.4, due to the fact that the Fisher and the local approximation criteria are not proper bounds on $I(\mathsf{x}, \mathsf{y})$, it is difficult to justify comparisons of the generic lower bound $\tilde{I}(\mathsf{x}, \mathsf{y})$ with $\tilde{I}_F(\mathsf{x}, \mathsf{y})$ and $\tilde{I}_L(\mathsf{x}, \mathsf{y})$. Indeed, while we can generally justify comparisons of proper bounds (as a tighter bound would imply a smaller gap from the true unknown functional for a specific parameter subspace), it is not always easy to judge on accuracy of approximations of unknown underlying functionals. Generally, this complicates empirical comparisons of the described learning methods for large scale problems.

In order to gain intuition on how the approximate information-maximizing methods influence the true underlying objective $I(\mathsf{x}, \mathsf{y})$, we considered a low-scale problem (so that the mutual information $I(\mathsf{x}, \mathsf{y})$ could be computed exactly). We were particularly interested to see whether or not maximization of the generic lower bound $\tilde{I}(\mathsf{x}, \mathsf{y})$ could lead to consistent improvements in the true mutual information $I(\mathsf{x}, \mathsf{y})$ for the considered conditionally factorized channel (6.1). To answer this question, we computed the exact value of $I(\mathsf{x}, \mathsf{y})$ at each iteration of the variational IM learning, where the encoder parameters $\mathsf{W} \in \mathbb{R}^{|\mathsf{y}| \times |\mathsf{x}|}$, $\mathsf{b} \in \mathbb{R}^{|\mathsf{y}|}$ were obtained by maximizing the bound $\tilde{I}(\mathsf{x}, \mathsf{y})$. We compared this with the changes in the exact mutual information under maximization of the Fisher approximation criterion $\tilde{I}_F(\mathsf{x}, \mathsf{y})$ of Brunel and Nadal (1998), and the local approx-

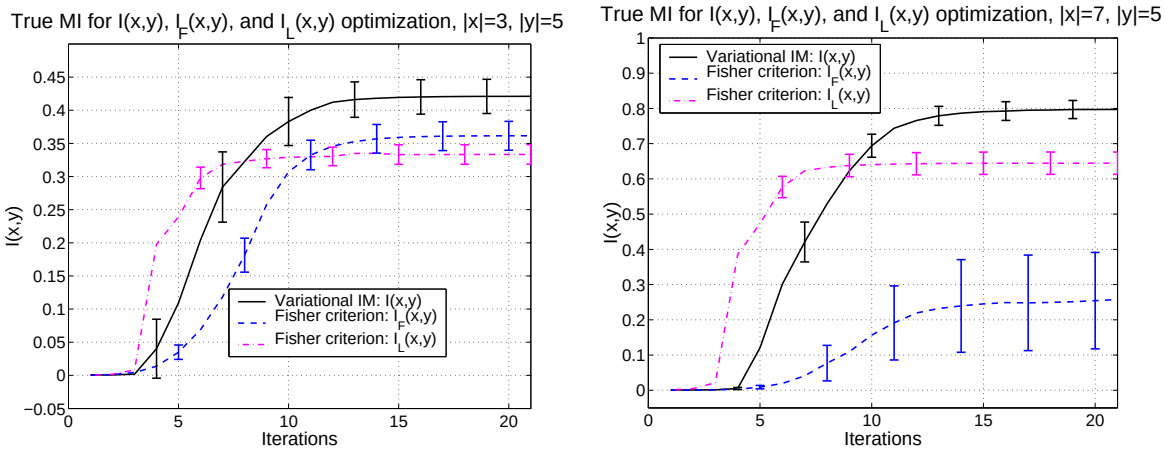


Figure 6.1: Changes in the exact mutual information $I(\mathsf{x},\mathsf{y})$ for parameters of the encoder $p(\mathsf{y}|\mathsf{x})$ obtained by maximizing the variational lower bound $\tilde{I}(\mathsf{x},\mathsf{y})$, the Fisher information criterion $\tilde{I}_F(\mathsf{x},\mathsf{y})$, and the local approximation criterion $\tilde{I}_L(\mathsf{x},\mathsf{y})$. The number of training stimuli $M = 20$. The results are averaged over 30 runs from random parameter initializations. *Left*: $|\mathsf{x}| = 3$, $|\mathsf{y}| = 5$ *Right*: $|\mathsf{x}| = 7$, $|\mathsf{y}| = 5$.

imation criterion $\tilde{I}_L(\mathsf{x},\mathsf{y})$ motivated in Section 6.2.1 and inspired by Corduneanu and Jaakkola (2003). To ensure that the true mutual information $I(\mathsf{x},\mathsf{y})$ could indeed be computed exactly, we restricted the dimensionality of the response variables $|\mathsf{y}|$. In order to illustrate the effects which the codesize $|\mathsf{y}|$ could have on the performance, we considered the cases when $|\mathsf{y}| < |\mathsf{x}|$ and $|\mathsf{y}| > |\mathsf{x}|$.

Figure 6.1 illustrates changes in the exact mutual information $I(\mathsf{x},\mathsf{y})$ with iterations of the scaled conjugate gradients (SCG) optimization on $\tilde{I}(\mathsf{x},\mathsf{y})$, $\tilde{I}_F(\mathsf{x},\mathsf{y})$, and $\tilde{I}_L(\mathsf{x},\mathsf{y})$. The variational decoder $q(\mathsf{x}|\mathsf{y})$ of the generic lower bound was chosen to be an isotropic linear Gaussian with the unconstrained optimal weights (see expression (6.11) and the discussion in Section 6.2.2). The plot shows the mean values and error bars on $I(\mathsf{x},\mathsf{y})$ for 30 runs from different initializations. The initial settings of the encoder parameters were the same for all the considered learning methods. The dimensionality of the code space was chosen to be $|\mathsf{y}| = 5$. The training stimuli $\mathsf{x} \in \mathbb{R}^{|\mathsf{x}|}$ were sampled from $\mathsf{x} \sim \mathcal{N}_{\mathsf{x}}(\mathbf{0}, \mathsf{I}_{|\mathsf{x}|})$ and centered, where we considered $|\mathsf{x}| = 3$ (Figure 6.1 (*left*)) and $|\mathsf{x}| = 7$ (Figure 6.1 (*right*)). The results are shown for $M = 20$ training patterns. Moreover, we imposed additional parametric constraints on each synaptic weight, so that $w_{ij} = \sigma(a_{ij}) - 0.5 \in (-0.5, 0.5)$. (The gradients for this case could be easily obtained from (6.45)). The auxiliary parameters $\{a_{ij}\}$ were initialized at random as $a_{ij} \sim \mathcal{N}_a(0, 0.1)$.

As we can see from (Figure 6.1 (*left*)), all three methods tended to lead to consistent improvements in the true mutual information for $|\mathsf{x}| \leq |\mathsf{y}|$. The variational approach usually resulted in higher values of $I(\mathsf{x},\mathsf{y})$ after just a few training iterations, and typically converged to higher values of $I(\mathsf{x},\mathsf{y})$. The local approximation criterion described in Section 6.2.1 may be characterized by a rapid convergence and low variance of the mutual information estimates, which may be explained by

the smoothness of the optimized surface defined by $\tilde{I}_L(\mathsf{x},\mathsf{y})$. For the considered case of overcomplete representations, the Fisher approximation criterion $\tilde{I}_F(\mathsf{x},\mathsf{y})$ typically resulted in a stable convergence to relatively high values of $I(\mathsf{x},\mathsf{y})$.

In the case of undercomplete binary representations (i.e. $|\mathsf{x}| > |\mathsf{y}|$), optimization of the Fisher criterion $\tilde{I}_F(\mathsf{x},\mathsf{y})$ was numerically unstable and often led to no visible improvements of mutual information $I(\mathsf{x},\mathsf{y})$ over its starting values at initializations. This is not surprising, as the objective $\tilde{I}_F(\mathsf{x},\mathsf{y})$ (and its gradients) are undefined for singular Fisher Information matrices (note that in the considered case $\mathrm{rank}(\mathsf{F}_\mathsf{x}) \leq \min\{|\mathsf{x}|,|\mathsf{y}|\} = |\mathsf{y}|$, i.e. $\mathsf{F}_\mathsf{x} \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{x}|}$ is rank-deficient). Further approximations aimed at handling singularities of the gradients (see Section 6.3.1) could lead to slight improvements in $I(\mathsf{x},\mathsf{y})$, though their dynamics was rather inconsistent. In practice, this could often be characterized by non-monotonic changes in $I(\mathsf{x},\mathsf{y})$ with the number of SCG iterations on $\tilde{I}_F(\mathsf{x},\mathsf{y})$. Not surprisingly, after averaging over a number of independent runs, optimization of $\tilde{I}_F(\mathsf{x},\mathsf{y})$ often led to high variances of the mutual information evaluations (Figure 6.1 (*right*)). In contrast, optimization of the local approximation criterion $\tilde{I}(\mathsf{x},\mathsf{y})$ and the variational lower bound $\tilde{I}(\mathsf{x},\mathsf{y})$ typically led to consistent improvements in the exact mutual information. Again, while learning by maximizing the local approximations often resulted in quicker convergence, the variational IM applied to the considered channel usually led to consistently higher values of $I(\mathsf{x},\mathsf{y})$ after a small number of training iterations.

Finally, we note that the imposed constraint on the individual weights $\{w_{ij}\}$ resulted in a hard upper bound on $\|\mathsf{W}\|_F \leq \omega\sqrt{|\mathsf{y}||\mathsf{x}|}$. The comparison of the methods is fair in the sense that the obtained results are not simply influenced by the weight re-scalings. Other types of constraints (e.g. soft constraints on the conditional variances of the stochastic firings) lead to qualitatively similar relations between the learning methods (see Agakov and Barber (2004b)). While the considered problem is low-scale, the observed relations between the learning methods help to understand the effects which optimization of the approximate objective criteria may have on the true mutual information. Specifically, the results confirm the intuitive limitations of the commonly used Fisher approximations of $I(\mathsf{x},\mathsf{y})$.

Hypercubic Constraints and Discrete-Valued Weights

It is interesting to note that by imposing hypercubic constraints on the encoder weights, i.e. constraining each synaptic weight w_{ij} to lie in a symmetric region $[-\omega,\omega]$, and carefully re-scaling the input stimuli, we may often observe the situation when the optimal weight parameters $\mathsf{W} \in [-\omega,\omega]^{|\mathsf{y}|\times|\mathsf{x}|}$ obtained by maximizing any of the three objective criteria $\tilde{I}(\mathsf{x},\mathsf{y})$, $\tilde{I}_F(\mathsf{x},\mathsf{y})$, and $\tilde{I}_L(\mathsf{x},\mathsf{y})$ lie very close to the corners of the hypercube. Not surprisingly, this behavior is strongly influenced by the cube's size ω, dimensionality of the source space $|\mathsf{x}|$, and the moments of the source distribution $\tilde{p}(\mathsf{x})$ (as all these entities influence the effective activation fields of the output units). Here we will not aim at characterizing the relation theoretically, but demonstrate the situation when the effect occurs.

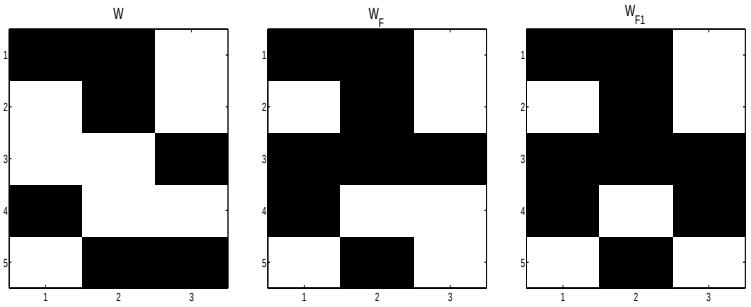


Figure 6.2: Optimal encoder weights $\mathsf{W} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{x}|}$ obtained by maximizing $\tilde{I}(\mathsf{x},\mathsf{y})$, $\tilde{I}_F(\mathsf{x},\mathsf{y})$, and $\tilde{I}_L(\mathsf{x},\mathsf{y})$. Here $|\mathsf{x}| = 3$, $|\mathsf{y}| = 5$, and $\mathsf{M} = 20$. The white squares correspond to the settings $w_{ij} \approx 0.5$. The black squares correspond to the settings $w_{ij} \approx -0.5$. For the illustrated case, $\max_{ij}\{||w_{ij}| - 0.5|\} \sim O(10^{-4})$.

Specifically, we will consider the settings of the previous set of experiments, where $|\mathsf{x}| = 3$, $|\mathsf{y}| = 5$, the weights are constrained as $w_{ij} = \sigma(a_{ij}) - 0.5 \in (-0.5, 0.5)$ for some real-valued parameters a_{ij}, and the sources are sampled as $\mathsf{x} \sim \mathcal{N}_{\mathsf{x}}(\mathbf{0}, 0.7\mathsf{I})$. Optimization of the three objectives with respect to the auxiliary parameters $\mathsf{A} = \{a_{ij}\} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{x}|}$ results in the nearly discrete-valued weights $\mathsf{W} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{x}|}$ shown on Figure 6.2, where $\max_{i,j} ||w_{i,j}| - 0.5| \sim O(10^{-4})$. Figure 6.3 (*left*) shows the changes in the exact mutual information $I(\mathsf{x},\mathsf{y})$ computed at the parameters obtained by maximizing $\tilde{I}(\mathsf{x},\mathsf{y})$, $\tilde{I}_F(\mathsf{x},\mathsf{y})$, and $\tilde{I}_L(\mathsf{x},\mathsf{y})$. Note that the resulting weights have nearly identical norms, i.e. effects which the norm rescalings have on the objective criteria may safely be ignored. Figure 6.3 (*left*) shows typical changes in the exact mutual information $I(\mathsf{x},\mathsf{y})$ computed for the discrete-valued weight parameters $\mathsf{W} \in \{-0.5, 0.5\}^{|\mathsf{y}|\times|\mathsf{x}|}$ around W_I, W_F, and W_L (which are the optimal weights obtained by maximizing $\tilde{I}(\mathsf{x},\mathsf{y})$, $\tilde{I}_F(\mathsf{x},\mathsf{y})$, and $\tilde{I}_L(\mathsf{x},\mathsf{y})$ respectively). As we see from the plot, the variational IM maximizing the bound $\tilde{I}(\mathsf{x},\mathsf{y})$ results in highest values of the exact $I(\mathsf{x},\mathsf{y})$. (Additionally, it turns out that in the considered case the weight $\mathsf{W}_I \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{x}|}$ gives rise to one of the equivalent global optima of the exact $I(\mathsf{x},\mathsf{y})$ computed for the 2^{15} combinations of the discrete weights $\{-0.5, 0.5\}^{|\mathsf{y}|\times|\mathsf{x}|}$).

The result is potentially interesting, as it suggests the existence of encoder models which may be used to transform the combinatorial search over discrete-valued parameters to a continuous optimization problem. It is indeed intuitive that by maximizing the exact $I(\mathsf{x},\mathsf{y})$, the optimal encoder model would encourage spread-out representations in the code space. However, as we showed in Section 6.3.3, for the considered parameterization of the encoder and the considered constraints on the parameters, the weights cannot grow indefinitely. This suggests a curious relation between dimensionality of the sources $|\mathsf{x}|$, the weight range defined by ω, the scalings of the source patterns, and the observed weight saturation effects, which will be interesting to investigate in the future.

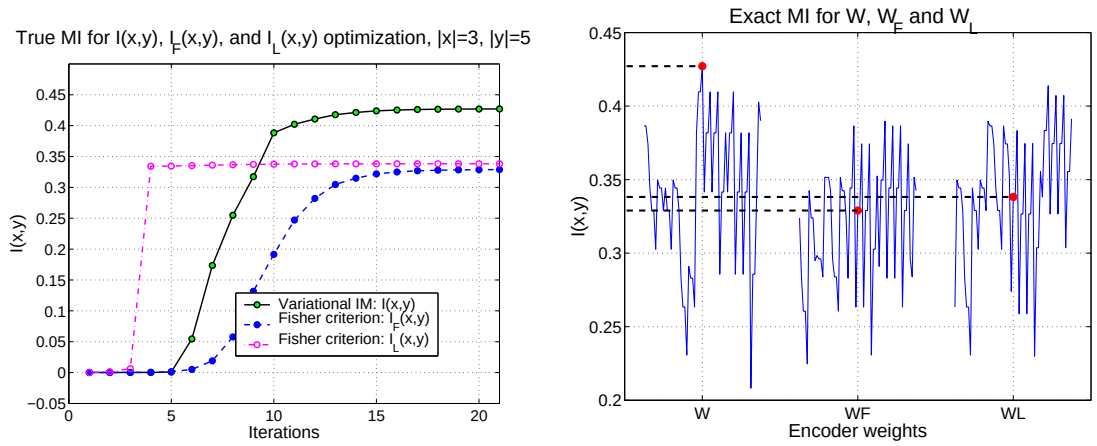
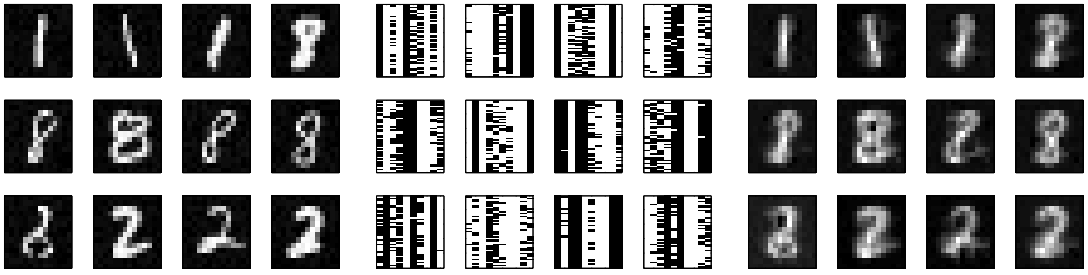


Figure 6.3: Exact mutual information as a function of (approximately) discrete-valued weights. *Left:* typical changes in the exact $I(\mathsf{x},\mathsf{y})$ under maximization of $\tilde{I}(\mathsf{x},\mathsf{y})$, $\tilde{I}_F(\mathsf{x},\mathsf{y})$, and $\tilde{I}_L(\mathsf{x},\mathsf{y})$ for the auxiliary parameters $\{a_{ij}\}$. *Right:* values of the exact mutual information computed at $\mathsf{W} = \{-0.5, 0.5\}^{|\mathsf{y}|\times|\mathsf{x}|}$ around the optimal W_i, W_F, and W_L. The x-axis corresponds to the decimal representations of the binary encodings of the discrete weight vector $\mathsf{w} \in \{-0.5, 0.5\}^{|\mathsf{x}|\times|\mathsf{y}|}$ obtained by flattening $\mathsf{W} \in \mathbb{R}^{|\mathsf{y}|\times\times}$.

Variational IM: Stochastic Representations of the Digit Data

Finally, we have applied the variational IM method to stochastic coding of visual patterns for the case when the number of the input stimuli $|\mathsf{x}|$ significantly exceeds the number of the spiking units $|\mathsf{y}|$. (The problem of coding in this case may be viewed in the context of stochastic compression). Our motivation here was to check whether the variational approach to information maximization could lead to at least vaguely interpretable firing frequencies of small groups of neurons for the considered biologically-inspired channel parameterization; specifically, we were interested in checking whether frequency of stochastic firings of the encoding units could in some sense be representative of different aspects of the visual stimuli. (The stochasticity was an artifact of the constrained channel parameterization, as specified by the encoding distribution (6.1)). We were also interested to see how different firing frequencies could affect reconstructions of input stimuli obtained by applying the variational decoder.

Note that for the considered case the information-theoretic formulation is particularly attractive, as it allows to impose explicit biologically-inspired constraints on parameters of the encoding mappings. Additionally, in the exact formulation learning of encodings is entirely unsupervised, which is particularly attractive from the biological perspective. Our method is an approximation of the exact case, with a specific choice of the variational distribution. In our experiments we used the simplest form of the Gaussian decoder discussed in Section 6.2.2. After numerical optimization for $p(\mathsf{y}|\mathsf{x})$ and $q(\mathsf{x}|\mathsf{y})$ with an explicit constraint on the variance of the conditional firings (enforced by the penalty term $m = 0.2$, see Section 6.3.3), we applied the variational decoder to perform reconstruction of



(a) (b) (c)

Figure 6.4: Stochastic binary encodings of the continuous input stimuli. *(a)*: a subset of the original visual stimuli. *(b)*: 30 samples from the corresponding encoding distribution $p(\mathsf{y}|\mathsf{x})$ for $|\mathsf{y}| = 10$ spiking units (black and white dashes correspond to silent and spiking output units). *(c)*: Reconstructions of the original sources from 50 samples of neural spikes. Note that we imposed soft constraints on the variances of firings.

196-dimensional continuous visual stimuli from 10 spiking neurons (i.e. $|\mathsf{x}| = 196$ and $|\mathsf{y}| = 10$). The training stimuli consisted of 30 instances of digits 1, 2, and 8 (10 of each class). The source variables were reconstructed from 50 stochastic spikes at the mean of the optimal approximate decoder $q(\mathsf{x}|\mathsf{y})$. Note that since $|\mathsf{x}| > |\mathsf{y}|$, the problem could not be efficiently addressed by optimizing the Fisher approximation criterion (6.25). Clearly, the deterministic approaches (see e.g. Bell and Sejnowski (1995) or Shriki et al. (2002)) are not applicable either, since the considered channel is noisy and undercomplete. On the other hand, the variational IM algorithm is applicable and numerically stable.

Figure 6.4 illustrates a subset of the original source signals *(a)*, samples of the corresponding binary responses for units *(b)*, and reconstructions of the source data from the population of binary spikes *(c)*. (For Figure 6.4 *(b)*, the x-axis corresponds to each of the 10 units, and the y-axis corresponds to each of 30 samples from the encoding distribution). We see that while stochastic encodings corresponding to distinct patterns are generally different, there are some visibly distinct input stimuli characterized by very similar firing patterns (see e.g. the rightmost patterns in the first and the second rows), and it is arguably the frequency of firings of a small number of neurons which accounts for differences in the corresponding inputs and reconstructions. For example, the firing patterns of the rightmost stimuli in the first and the second rows are quite similar, apart from the noisy firings of the 2^{nd}, 9^{th} and the 10^{th} units. However, the resulting reconstructions are visibly different (see the sizes of the bottom loops of the digits "8"). The result is consistent with observations that receptive fields of post-synaptic cells in the neocortex are influenced by distinct features of input stimuli, which in turn determines firing frequencies of individual neurons (e.g. Markram et al. (1998)).

Table 6.1: Objective functions for approximate information maximization

1.	*Invertible channels*:	$I(\mathsf{x},\mathsf{y}) = \langle \log \lvert \mathsf{J}_\mathsf{x} \rvert \rangle_{\tilde{p}(\mathsf{x})}$
2.	*Overcomplete deterministic channels*:	$I(\mathsf{x},\mathsf{y}) = \langle \log \lvert \mathsf{J}_\mathsf{x}^T \mathsf{J}_\mathsf{x} \rvert \rangle_{\tilde{p}(\mathsf{x})}$
3.	***Fisher approximation***:	$\tilde{I}_F(\mathsf{x},\mathsf{y}) = \langle \log \lvert \mathsf{F}_\mathsf{x} \rvert \rangle_{\tilde{p}(\mathsf{x})}$
4.	***Local approximation***:	$\tilde{I}_F(\mathsf{x},\mathsf{y}) = \langle \log \exp\{\mathrm{tr}\,\{\mathsf{F}_\mathsf{x}\}\} \rangle_{\tilde{p}(\mathsf{x})}$
5.	***Variational lower bound***:	$\tilde{I}(\mathsf{x},\mathsf{y}) = \langle \log q(\mathsf{x}\lvert\mathsf{y}) \rangle_{p(\mathsf{y}\lvert\mathsf{x})\tilde{p}(\mathsf{x})}$

6.6 Summary

The primary goal of this chapter was to explore applicability of the variational information-maximizing framework in the context of learning *high-dimensional* binary representations of continuous source patterns. We described an application of the variational approach to information maximization for the case when continuous source stimuli were represented by conditionally independent stochastic binary responses. We also showed that for the considered encoding distribution it was possible to derive a local iterative learning rule, which gives rise to a form of weighted variable-rate Hebbian learning (interestingly, the receptive fields of the post-synaptic units were influenced not only by the activations at the pre-synaptic layers, but also (implicitly) by the activations of the neighboring post-synaptic units). This result generalizes the work of Linsker (1997), who derived a local approximation of Bell and Sejnowski's information-maximizing rule for deterministic invertible channels (Bell and Sejnowski (1995)).

Moreover, we have applied the numerical approximations used for deriving the local regularizers of Szummer and Jaakkola (2002), Corduneanu and Jaakkola (2003), and showed that the results may be used to approximate a lower bound on $I(\mathsf{x},\mathsf{y})$. While the resulting approximation typically has better convergence properties than the common Fisher approximation of mutual information (see e.g. Brunel and Nadal (1998)), our empirical results indicate that the variational approach may be more attractive. Additionally, our results indicate that the considered methods addressing approximate information maximization may be viewed as approximations of our variational approach; however, generally they do not preserve a proper bound on the mutual information, and may be less computationally and numerically appealing.

Table 6.1 summarizes effective criteria optimized by Bell and Sejnowski (1995), Shriki et al. (2002), Brunel and Nadal (1998), the local approximation based on the work of Szummer and Jaakkola (2002) and Corduneanu and Jaakkola (2003), and our generic variational approach. Here the symmetric matrix $\mathsf{J}_\mathsf{x} = \{J_{ij}(\mathsf{x})\} \stackrel{\text{def}}{=} \{\partial y_i(\mathsf{x})/\partial x_j\} \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{x}|}$ is the Jacobian of the deterministic invertible mapping $\mathsf{x} \mapsto \mathsf{y}$ (with $|\mathsf{y}| = |\mathsf{x}|$), $\mathsf{F}_\mathsf{x} \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{x}|}$ is the Fisher Information matrix (6.19), and $q(\mathsf{x}|\mathsf{y})$ is an approximate decoder lying in a tractable family. In this chapter we focused primarily on the discussion of the last three approaches (shown in **bold**), since they are applicable for approximate maximization of mutual information in stochastic, rather than deterministic, channels. For this reason we have ignored the comparisons of our method with those of Bell and Sejnowski (1995) and Shriki et al. (2002), which presume noiseless mappings from the code to the data space.

Note that in contrast to the other techniques, the suggested variational method optimizes a proper lower bound independently of the choice of the decoder, dimensionality of the input stimuli, number of post-synaptic neurons, or noise of the stochastic firings. This extends applicability of the variational approach to dimensionality reduction, compression, syndrome decoding, and facilitates applications of the method to undercomplete and overcomplete stochastic coding. Of course, more biologically realistic channels and applications should potentially be considered; our results here mainly serve to illustrate the potential advantages of our variational information-maximizing formulation over the common (Brunel and Nadal (1998), Kang and Sompolinsky (2001)) and less common (Szummer and Jaakkola (2002), Corduneanu and Jaakkola (2003) and Section 6.2.1) approaches to population coding of high-dimensional input stimuli.

Chapter 7

Auxiliary Variational Inference and Variational Mutual Information Maximization

In Chapter 2 we discussed a simple and general variational approach to maximizing the generic lower bound on $I(\mathsf{x},\mathsf{y})$ for a stochastic channel. In the subsequent chapters we considered generalizations and applications of the variational information-maximizing approach (see Chapter 4, Chapter 5, and Chapter 6), and discussed the relation of the variational IM to conditional likelihood training (Chapter 3). Here we change the perspective rapidly and demonstrate that the generic bound on the mutual information may be obtained in the context of *auxiliary variational statistical inference* (Agakov and Barber (2004a)), where we aim to lower-bound a generally intractable normalizing constant of a Markov network[1]. Generally, the approach we describe here defines a completely different family of variational bounds; however, maximization of the generic lower bound on mutual information (2.2) may be seen as its fundamentally important subgoal.

While little work appears to have been done on developing a simple and easily generalizable variational framework for approximate information maximization, variational methods have proved popular and effective for inference and maximum-likelihood learning in intractable graphical models. In this context they are often applied for bounding the likelihoods and the normalizing constants (see e.g. Jaakkola (1997), Jaakkola and Jordan (1998), Barber and Wiegerinck (1998), Bishop et al. (1998), Lawrence (2000), Wainwright et al. (2001), Wainwright et al. (2002), Beal (2003)). The majority of the standard approaches to lower-bounding the normalizing constants are based on non-negativity of the Kullback-Leibler divergence $KL(q(\mathsf{x})\|p(\mathsf{x}))$ between a tractable variational distribution $q(\mathsf{x})$ and the original distribution $p(\mathsf{x})$ (see e.g. Jordan et al. (1998) for a general introduction to variational inference and learning). Here we show that by expressing the bound on the normalizing constant from the Kullback-Leibler divergence $KL(q(\mathsf{x},\mathsf{y})\|p(\mathsf{x},\mathsf{y}))$ in a specifically defined augmented variable space $\{\mathsf{x},\mathsf{y}\}$, we may improve the standard bounds. It turns out that the improvement

[1]We can use a fundamentally similar method for inference and learning in arbitrary distributions, but focus on Markov networks for clarity.

of the proposed *auxiliary variational* method over a convex combination of simple variational bounds is given by a specific form of the generic lower bound on mutual information $I(\mathsf{x}, \mathsf{y})$ derived in Chapter 2 (see expression (2.2)).

In the context of approximate probabilistic inference, we may use the suggested approach for improving on the lower bounds of standard factorized approximations. Indeed, we show that the method described here forms a more powerful class of approximations than any *structured mean field* technique. The existing lower bounds of the variational mixture models (Lawrence et al. (1998), Jaakkola and Jordan (1998)) can be viewed as computationally expensive special cases of our method. A byproduct of our work is an efficient way to calculate a set of mixture coefficients for any set of tractable distributions, which principally improves on the flat combination (Agakov and Barber (2004a)).

7.1 Introduction

Probabilistic graphical models provide a convenient framework for graphical representation of joint probability distributions, and facilitate computation of many quantities of interest required for both inference and learning. Generally, probabilistic treatment of uncertainty offers a consistent and principled framework for inference in complex domains (Chapter 1). However, many distributions used for modeling practical domains are inherently intractable, which motivates the need for accurate and efficient approximations. In this chapter we focus on approximate inference, specifically on computation of lower bounds on normalization constants of undirected graphical models, which can also be used to approximate marginals of a formally intractable distribution. Fundamentally similar methods can be applied for inference and learning in arbitrary distributions; we will focus on inference in undirected models for clarity.

To be explicit, we will consider distributions $p(\mathsf{x})$ of the (Boltzmann) form

$$p(\mathsf{x}) = \exp\{-E(\mathsf{x})\}/Z, \quad Z = \sum_{\mathsf{x}} \exp\{-E(\mathsf{x})\}. \tag{7.1}$$

The complexity of carrying out the summation required to compute Z depends on the graphical structure of $p(\mathsf{x})$. In (poly)trees the normalisation constant can be computed in time linear in the number of variables in the distribution. In a general graph, the time is exponential in the size of the largest clique in the associated junction tree (Lauritzen and Spiegelhalter (1988), Jordan (2005)). Here we are interested in cases where exact computation by the junction tree algorithm is infeasible.

Variational approximations have been extensively used in physics and engineering and more recently applied to approximate inference and learning in intractable graphical models (see e.g. Saul et al. (1996), Jordan et al. (1998); Barber and Wiegerinck (1998); Wainwright et al. (2002)). In this context they were shown to result in relatively simple representations of the induced optimization problems; at the same time, it was shown that they often led to accurate approximations of the generally intractable quantities of interest. Their other

advantage for graphical models is availability of rigorous bounds on the normalizing constant (e.g. Jaakkola and Jordan (1996)), which contrasts with other (e.g. Monte-Carlo) approximations (see e.g. Neal (1993)). Availability of such bounds on the underlying intractable objectives facilitates comparisons of variational approximation techniques, where the tightness of a bound on an unknown quantity is often used as a relative measure of performance of an approximate method.

A popular class of lower bounds on $\log Z$ is based on the non-negativity of the Kullback-Leibler divergence

$$KL(q(\mathsf{x})\|p(\mathsf{x})) = \langle \log q(\mathsf{x}) \rangle_{q(\mathsf{x})} - \langle \log p(\mathsf{x}) \rangle_{q(\mathsf{x})} \geq 0. \tag{7.2}$$

Here $\langle \ldots \rangle_{q(\mathsf{x})}$ denotes an average over $q(\mathsf{x})$, and the bound is saturated if and only if $q(\mathsf{x}) \equiv p(\mathsf{x})$. In the case of the Boltzmann distribution (7.1), non-negativity of the KL divergence (7.2) yields the well-known class of lower bounds

$$\log Z \geq \langle -\log q(\mathsf{x}) \rangle_{q(\mathsf{x})} - \langle E(\mathsf{x}) \rangle_{q(\mathsf{x})}, \tag{7.3}$$

where $q(\mathsf{x})$ is typically restricted to a class of tractable distributions and varied to obtain the tightest bound within the tractable class. Coupled with an upper bound on the normalizing constant (Wainwright et al., 2002), expression (7.3) may be used for bounding expectations in the original model. A further use for this procedure is to provide a lower bound on the marginal likelihood in situations of observed and unobserved variables, which is a natural derivation and extension (Neal and Hinton, 1998) of the EM procedure (Dempster et al. (1977)).

7.1.1 Existing Variational Approximations

The computational tractability of the bound (7.3) depends on the choice of the approximating distribution $q(\mathsf{x})$. Possibly one of the simplest choices is given by the factorized *mean field* model (see Figure 7.1 (a), (b)) with $q_{MF}(\mathsf{x}) = \prod_i q(x_i)$, which discards all the edges from the original graph. The factorized assumption often results in a tractably computable bound (7.3). However, the bound may be inaccurate when the variables in the true distribution are strongly correlated. Moreover, it can be shown that the mean field distribution is log-concave (in the space of functional parameters), which implies a fundamental limitation of the mean field approximation in the case when $p(\mathsf{x})$ is locally multi-modal, since significant mass contributing to the partition function may be missed.

One way to go beyond the factorized assumption for $q(\mathsf{x})$ is to consider a *structured mean field* approximation (Ghahramani and Jordan, 1995; Barber and Wiegerinck, 1998) which retains some of the structure of $p(\mathsf{x})$. In this case it is often assumed that $q(\mathsf{x})$ has a sparse graphical representation (e.g. it is a (poly)tree, see Figure 7.1 (c)), which typically leads to an improvement on the bound at a moderate increase in computational cost. However, we may note that discarded edges in $q(\mathsf{x})$ introduce conditional independencies which may not exist in the original distribution $p(\mathsf{x})$.

Other approaches (Lawrence et al., 1998; Jaakkola and Jordan, 1998) bound the log partition function by considering mixtures of mean field type models

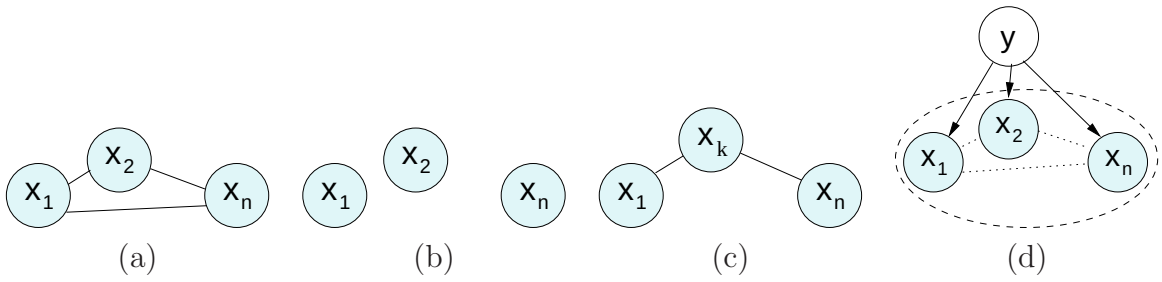


Figure 7.1: Undirected models and their approximations (a schematic plot). (a) A fully connected pairwise Markov network representing the intractable distribution $p(\mathsf{x})$; (b) a standard mean field approximation $q_{MF}(\mathsf{x})$; (c) a structured mean field approximation $q_{SMF}(\mathsf{x})$; (d) a mixture of mean field models. (All the variables x are coupled through the mixture label y. The dotted lines serve to indicate that the marginal $q_{MMF}(\mathsf{x})$ expressed from $q_{MMF}(\mathsf{x}, y)$ is in general fully connected). The transparent node y shows the auxiliary variable. The shaded nodes indicate the variables forming the space $\{\mathsf{x}\}$ of the original model $p(\mathsf{x})$.

(see Figure 7.1 (d)). This formally extends the standard factorized approximations, since the resulting approximating distribution $q(\mathsf{x})$ is generally multi-modal and not factorized in x. However, in general optimization of the bound (7.3) in this case requires minimization of the KL divergence between two fully connected distributions $p(\mathsf{x})$ and $q(\mathsf{x})$, which requires a formally computationally intractable evaluation of the entropy of the mixture $H(\mathsf{x})$. Further approximations suggested by Lawrence et al. (1998), Jaakkola and Jordan (1998), El-Hay and Friedman (2002) circumvent the computational intractability of variational mixture approximations by relaxing (7.3) at the cost of introducing additional variational relaxations. Unfortunately, unless *all* the mixture components $q(\mathsf{x}|\mathsf{y})$ have the same tractable structure, optimization of the resulting bound on $\log Z$ may become numerically unstable, which may complicate generalizations of the existing results to variational mixtures of *arbitrarily structured* tractable experts. Additionally, it is difficult to formally analyze the optimal solution induced by the resulting optimization procedure, and the existing procedure is not easily generalizable to richer families of variational bounds.

7.2 Auxiliary Variational Method

We will now explore the idea of using augmented variable spaces in the context of bounding the normalizing constant, and show that the formulation gives rise to the generic lower bound on mutual information.

Note that the computational problems of the variational mixture approximations arise from the fact that marginalization of the mixture labels y from the joint distribution $q(\mathsf{x}, \mathsf{y})$ results in a fully connected marginal $q(\mathsf{x})$ (see Figure 7.1 (d)). Informally, computation of the bound on $\log Z$ in this case requires minimization of the KL divergence between two fully-connected distributions (see expression (7.2)). It is intuitive that from the computational viewpoint it would

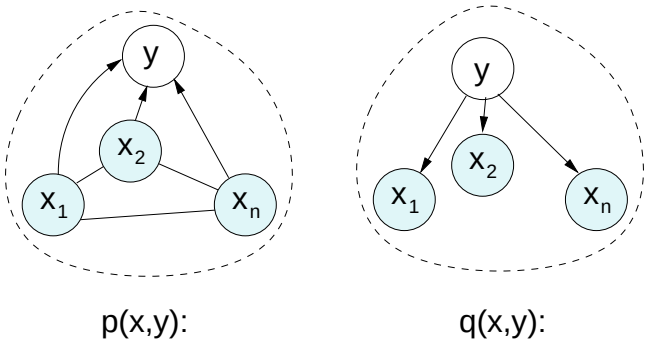


Figure 7.2: An auxiliary mean field approximation. The target distribution $p(\mathsf{x}, y)$ is approximated by $q(\mathsf{x}, y)$, which is structured in the *augmented space*. Note that the mapping $p(y|\mathsf{x})$ was introduced in a way which does not affect the original fully connected pairwise target distribution $p(\mathsf{x})$. The transparent node y shows the auxiliary variable. The shaded nodes indicate the variables forming the space $\{\mathsf{x}\}$ of the original model $p(\mathsf{x})$.

be significantly more beneficial to retain the structural form of the joint distribution and use $q(\mathsf{x}, \mathsf{y})$ as an approximation. The variational distribution would in this case be defined over the augmented variable space $\{\mathsf{x}, \mathsf{y}\}$, while the target $p(\mathsf{x})$ is defined over the space of the original variables $\{\mathsf{x}\}$. In order for the bound (7.3) to be well-defined, we introduce *auxiliary variables* y to the target distribution. This can be readily done in such a way that the marginal $\tilde{p}(\mathsf{x})$ of the joint distribution $p(\mathsf{x}, \mathsf{y})$ has the same graphical structure as the original target $p(\mathsf{x})$ (see Figure 7.2). Then we minimize the KL divergence between $q(\mathsf{x}, \mathsf{y})$ and $p(\mathsf{x}, \mathsf{y})$ in the joint variable spaces. Note that in contrast to standard structured approximations, all the variables x of the marginal $q(\mathsf{x})$ remain coupled, which agrees with the graphical structure of the fully connected Markov network $p(\mathsf{x})$. However, similarly to structured mean field techniques, our *auxiliary variational* method does not require a direct evaluation of the computationally intractable entropy of the mixture $H_q(\mathsf{x})$, as by minimizing the KL divergence in the augmented variable space we would effectively be fitting a structured distribution. Note that the idea of introducing "dummy" variables y is conceptually similar to the auxiliary variational bound on the mutual information discussed in Section 2.3, though the projections and the bounds are defined differently.

Another motivation for introducing the auxiliary variables is the reported success of auxiliary sampling techniques, such as the *Swendsen-Wang* (Swendsen and Wang, 1987), *partial decoupling* (Higdon, 1998) algorithms. It has been shown that by augmenting the original variable space with auxiliary variables and drawing samples from the joint distribution $p(\mathsf{x}, \mathsf{y})$ in the augmented space, one can achieve a significant improvement over standard MCMC approaches. The purpose of the auxiliary variables in this context is to capture (structural) information about clusters of strongly correlated variables. Our hope was that by performing approximations in the augmented space, where the auxiliary variables capture useful regularities about the data, we may improve over standard

variational approaches.

Indeed, we demonstrate that the auxiliary variational technique forms a more powerful class of approximations than any structured mean field approach. Moreover, the method offers an efficient way of calculating a set of mixture coefficients for any choice of tractable approximators (for example, trees with different structures). These coefficients may be used to form a mixture which is principally better than a single best tractable approximator or their flat combination. Finally, we show that our approach provides a computationally and conceptually simple alternative to the existing bounds optimized by the variational mixture methods.

7.2.1 Optimizing the Auxiliary Variational Bound

Let $p(\mathsf{x}, \mathsf{y}) = p(\mathsf{x})p(\mathsf{y}|\mathsf{x})$ define the joint distribution of the original variables x and auxiliary variables y in the augmented $\{\mathsf{x}, \mathsf{y}\}$ space. From the divergence $KL(q(\mathsf{x}, \mathsf{y})\|p(\mathsf{x}, \mathsf{y}))$ in the joint space it is easy to obtain an expression for the lower bound on the log partition function of $p(\mathsf{x}) = \exp\{-E(\mathsf{x})\}/Z$, which is given by

$$\log Z \geq -\langle \log q(\mathsf{x}, \mathsf{y}) \rangle_{q(\mathsf{x},\mathsf{y})} - \langle E(\mathsf{x}) \rangle_{q(\mathsf{x})} + \langle \log p(\mathsf{y}|\mathsf{x}) \rangle_{q(\mathsf{x},\mathsf{y})}. \tag{7.4}$$

Here $p(\mathsf{y}|\mathsf{x}) \stackrel{\text{def}}{=} p(\mathsf{y}|\mathsf{x}; \mathbf{\Psi})$ is an *auxiliary conditional* distribution parameterized by $\mathbf{\Psi}$. (In principle, we do not need to impose parametric constraints on $p(\mathsf{y}|\mathsf{x})$; for sparse discrete models the auxiliary conditional may be defined by the conditional probability table). Equivalently, (7.4) may be written as

$$\begin{aligned} \log Z &\geq -\langle KL(q(\mathsf{x}|\mathsf{y})\|p(\mathsf{x})) \rangle_{q(\mathsf{y})} + \tilde{I}(\mathsf{x}, \mathsf{y}) \\ &= \sum_{\mathsf{y}} q(\mathsf{y}) \left[\langle -E(\mathsf{x}) - \log q(\mathsf{x}|\mathsf{y}) \rangle_{q(\mathsf{x}|\mathsf{y})} \right] + \tilde{I}(\mathsf{x}, \mathsf{y}), \end{aligned} \tag{7.5}$$

where $\tilde{I}(\mathsf{x}, \mathsf{y})$ is defined as

$$\tilde{I}(\mathsf{x}, \mathsf{y}) \stackrel{\text{def}}{=} \sum_{\mathsf{x}} \sum_{\mathsf{y}} q(\mathsf{x}, \mathsf{y}) \log \frac{p(\mathsf{y}|\mathsf{x})}{q(\mathsf{y})} = \langle \log p(\mathsf{y}|\mathsf{x}) \rangle_{q(\mathsf{x},\mathsf{y})} + H(\mathsf{y}). \tag{7.6}$$

Note that up to the notational invariance, (7.6) is exactly the generic lower bound on mutual information $I(\mathsf{x}, \mathsf{y})$ expressed for the encoder model $q(\mathsf{x}, \mathsf{y}) = q(\mathsf{y})q(\mathsf{x}|\mathsf{y})$ (see expression (2.2)). In this context, the mapping to the auxiliary space $p(\mathsf{y}|\mathsf{x})$ may be interpreted as the variational decoder of the IM framework. The leftmost term in the r.h.s. of expression (7.6) defines a convex summation of the standard lower bounds (7.3) on $\log Z$ for the set of $\{\mathsf{x}\}$-variables, which cannot improve on the single best bound in the set. Clearly, (7.5) may improve on the standard bounds only if $\tilde{I}(\mathsf{x}, \mathsf{y}) > 0$. In the case that the auxiliary variables y contain no information about x, i.e. $p(\mathsf{y}|\mathsf{x}) = p(\mathsf{y})$, it is straightforward to show that the method reproduces the standard variational bound which uses the single best approximation $q(\mathsf{x}|\mathsf{y})$. It is intuitive that by considering less trivial auxiliary conditional mappings we may improve on the generic lower bound $\tilde{I}(\mathsf{x}, \mathsf{y})$.

If $p(\mathsf{y}|\mathsf{x}) \equiv q(\mathsf{y}|\mathsf{x})$ then $\tilde{I}(\mathsf{x},\mathsf{y})$ defines the exact mutual information between the original variables x and the auxiliary variables y. However, this specification leads to computational difficulties of evaluating or bounding the intractable entropy of the mixture $H(\mathsf{y})$ (see Section 1.4). By analogy with the variational IM method, we may constrain $p(\mathsf{y}|\mathsf{x})$ so that it lies in a tractable family (we will denote the constrained distribution as $p(\mathsf{y}|\mathsf{x};\boldsymbol{\Psi})$, where $\boldsymbol{\Psi}$ is a set of parameters; structural constraints may be considered similarly). Then a rigorous variational approach to maximizing $\log Z$ would involve maximizing the bound (7.6) with respect to the parameters (or clique potentials) of the auxiliary conditional distribution $p(\mathsf{y}|\mathsf{x},\boldsymbol{\Psi})$ (which corresponds to the variational decoder in the IM terminology), the marginal $q(\mathsf{y})$, and the conditionals $q(\mathsf{x}|\mathsf{y})$. The general iterative optimization algorithm in this case is given as follows:

1. Choose the auxiliary conditional $p(\mathsf{y}|\mathsf{x})$. For the remainder, we choose

$$p(\mathsf{y}|\mathsf{x}) = \exp\{\boldsymbol{\Psi}(\mathsf{y};\mathsf{x})\}/Z_{\mathsf{y}|\mathsf{x}}, \quad Z_{\mathsf{y}|\mathsf{x}} = \sum_{\mathsf{y}} \exp\{\boldsymbol{\Psi}(\mathsf{y};\mathsf{x})\}, \tag{7.7}$$

   though more general distributions may potentially be considered.

2. Initialize $q(\mathsf{x}|\mathsf{y})$, $q(\mathsf{y})$, and parameters $\boldsymbol{\Psi}$ of $p(\mathsf{y}|\mathsf{x})$.

3. For the fixed $q(\mathsf{y},\mathsf{x})$, obtain $\boldsymbol{\Psi}^{new}$ by solving for zeros of

$$\partial \log Z/\partial \boldsymbol{\Psi} = \langle \partial \log p(\mathsf{y}|\mathsf{x})/\partial \boldsymbol{\Psi} \rangle_{q(\mathsf{x},\mathsf{y})}, \tag{7.8}$$

   or performing numerical ascent on $\log Z$ for $\boldsymbol{\Psi}$ (see the bound (7.6)). Clearly, an unconstrained optimization would result in $p(\mathsf{y}|\mathsf{x}) = q(\mathsf{y}|\mathsf{x})$, as this would maximize the generic bound $\tilde{I}(\mathsf{x},\mathsf{y})$.

4. For the fixed $p^{new}(\mathsf{y},\mathsf{x}) \stackrel{\text{def}}{=} p(\mathsf{x})p(\mathsf{y}|\mathsf{x};\boldsymbol{\Psi}^{new})$ and $q(\mathsf{y})$ set

$$q^{new}(\mathsf{x}|\mathsf{y}) \propto p^{new}(\mathsf{y},\mathsf{x}) \tag{7.9}$$

   for all instances y.

5. For the fixed $p^{new}(\mathsf{y},\mathsf{x})$ and $q^{new}(\mathsf{x}|\mathsf{y})$, set

$$q^{new}(\mathsf{y}) \propto \exp\left\{ -\sum_{\mathsf{x}} q^{new}(\mathsf{x}|\mathsf{y}) \log \frac{q^{new}(\mathsf{x}|\mathsf{y})}{p(\mathsf{x})p(\mathsf{y}|\mathsf{x};\boldsymbol{\Psi}^{new})} \right\}. \tag{7.10}$$

6. Iterate steps 3–5 until a termination criterion is met.

Note that in the case of a constrained auxiliary conditional $p(\mathsf{y}|\mathsf{x})$, step 3 is analogous to the M-step of the generalized EM algorithm (Neal and Hinton, 1998), while steps 4 and 5 are analogous to the E-step. An update rule for each term was obtained from (7.6) by computing the corresponding functional derivatives while keeping other terms fixed.

Up to this point, the results are general and applicable (at least, in theory) for arbitrarily parameterized distributions. However, as discussed in Section 1.4, for computational reasons it is fundamentally important to impose constraints on the auxiliary conditional $p(\mathsf{y}|\mathsf{x})$ and the approximate joint $q(\mathsf{x},\mathsf{y})$ (effectively, both distributions define a set of variational parameters). If both distributions are in a tractable family (for example, if each node y_i has a small number of x-parents), the bound (7.6) may be computed and optimized exactly (see the discussion in Section 2.2.1). Another case leading to exact computations is when $p(\mathsf{y}|\mathsf{x}) \sim \mathcal{N}(\mathbf{\Psi}\mathsf{x}, \mathbf{\Sigma})$, which would correspond to Linsker's *as-if Gaussian* bound on mutual information (see Section 2.2.2). Note that fundamentally the variational bound on $\log Z$ is tractable only in cases when $\tilde{I}(\mathsf{x},\mathsf{y})$ may be exactly computed and optimized. Optimization of the remaining terms in the bound (7.6) should not be problematic (at least, this is the case when the variational components $q(\mathsf{x}|\mathsf{y})$ are structured and the underlying distribution $p(\mathsf{x})$ is a pairwise Markov network $p(\mathsf{x}) \propto \exp\{-(\mathsf{x}^T\mathsf{W}\mathsf{x} + \mathsf{x}^T\mathsf{b})\}$ for some $\mathsf{W} \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{x}|}$, $\mathsf{b} \in \mathbb{R}^{|\mathsf{x}|}$).

If one wishes to have a large number of parental variables x influencing y, further approximations may need to be employed. For the special case of distributions of the form (7.7), we can utilize the standard linear upper bound $\log x \leq mx - \log m - 1$. This transforms the objective (7.6) to

$$\log Z \geq 1 + \sum_{\mathsf{y}} q(\mathsf{y}) \langle -E(\mathsf{x}) - \log q(\mathsf{x},\mathsf{y}) \rangle_{q(\mathsf{x}|\mathsf{y})} + \left\langle \mathbf{\Psi}(\mathsf{y};\mathsf{x}) + \mu(\mathsf{x};\mathsf{y}) - e^{\mu(\mathsf{x};\mathsf{y})} Z_{\mathsf{y}|\mathsf{x}} \right\rangle_{q(\mathsf{x},\mathsf{y})} \tag{7.11}$$

where $e^{\mu(\mathsf{x};\mathsf{y})}$ is an additional variational functional of the exponential form (see e.g. Jaakkola and Jordan (1998)). In general, optimization of the bound (7.11) is numerically unstable and computationally expensive. However, by analogy with the general formulation of the IM framework, we may avoid computational difficulties of computing the bound (7.6) by constraining $p(\mathsf{y}|\mathsf{x},\mathbf{\Psi})$ to lie in a tractable family. Generally, such constraints would obviate a recourse to (7.11).

7.2.2 Specific Auxiliary Representations

Effectively, the problem of choosing an appropriate (tractable yet general) mapping to the auxiliary space is strongly related to the problem of choosing an appropriate variational decoder (see e.g. Sections 1.5, 2.2.1). Generally, any kind of tractable constraints on $p(\mathsf{y}|\mathsf{x})$ may be used, such as the Gaussian or a factorized approximation (see e.g. the discussion in Section 2.2.1). Additionally, we note that by analogy with Section 2.3 we may use the auxiliary variational lower bound on mutual information[2] (2.38). Here we will briefly outline other choices of the auxiliary conditional $p(\mathsf{y}|\mathsf{x})$, which in practice often lead to accurate approximations.

[2]This should not be confused with the auxiliary variational lower bound on $\log Z$ (expression (7.5)).

Discrete Spaces with Parametric Auxiliary Distributions

If the auxiliary space is given by a single multinomial variable $y \in \{1, \ldots, M\}$, a natural choice for $p(y|\mathsf{x})$ is to use a *softmax* type representation

$$p(y_k|\mathsf{x}) \propto \exp\left\{f(\mathsf{x}; \mathsf{u}^{(k)})\right\}, \quad \mathsf{U} = \{\mathsf{u}^{(1)}, \ldots, \mathsf{u}^{(M)}\} \in \mathbb{R}^{|\mathsf{x}| \times M}. \tag{7.12}$$

Here $f(\mathsf{x}; \mathsf{U})$ is some differentiable function and $p(y_k|\mathsf{x})$ is the probability of the auxiliary variable y being in state $k = 1, \ldots, M$. Although the expectation $\langle \log p(y_k|\mathsf{x}) \rangle_{q(\mathsf{x}|y_k)}$ is in general intractable, we can make a use of the transformed bound (7.11). A significantly cheaper alternative is to perform optimization of (7.6) by approximating the term at $\log p(y_k|\langle \mathsf{x} \rangle_{q(\mathsf{x}|y_k)})$, which is exact as long as the mapping from the auxiliary to the data space is deterministic, i.e. $q(\mathsf{x}|y) \sim \delta(\mathsf{x} - \langle \mathsf{x}|y \rangle)$. (Note that in the IM terminology this choice of the variational component distribution would corresponds to a noiseless encoding model, which may greatly simplify the computations). Generally, this approximation may be accurate as long as each component $q(\mathsf{x}|y)$ is sharply peaked around the mode.

Additionally, in the case when $|\mathsf{x}|$ is large, $p(\mathsf{y}|\mathsf{x}) = \prod_k p(y_k|\mathsf{x})$, and each factor utilizes a simple (generalized) linear type dependency

$$p(y_k|\mathsf{x}) = p(y_k|a_k), \quad a_k \stackrel{\text{def}}{=} f_k(\mathsf{x}^T \mathsf{u}^{(k)} + b^{(k)}), \tag{7.13}$$

the *Gaussian field* approximation (Barber and Sollich (2000)) can be employed. From the Central Limit Theorem we can assume approximate Gaussianity of the field a_k and approximate the expectation $\langle \log p(y_i|\mathsf{x}^i) \rangle_{q(\mathsf{x}|y_k)}$ by performing 1-D Gaussian integration of the general form $\int_a f(a)p(a)$, where $p(a) \sim \mathcal{N}(\mu_a, \sigma_a^2)$. The means and variances of the fields are readily relatable to the first and second order moments of $q(\mathsf{x}|y_k)$.

In practice, such approximations do not lead to significant deviations and are shown to be both accurate and efficient (Barber and Sollich (2000), Agakov and Barber (2003)). Probably the greatest disadvantage of these relaxations is due to the fact that for any realistic limit of $|\mathsf{x}|$ the bound will no longer be strict. One way to address this problem is to impose additional structural constraints on the conditionals by limiting the number of parental variables for each factor.

Discrete Spaces with Structured Auxiliary Distributions

If $\boldsymbol{\pi}_x(y_i)$ and $\boldsymbol{\pi}_y(y_i)$ correspond to x- and y-parents of the auxiliary variable y_i in the graph of $p(\mathsf{y}|\mathsf{x})$, the generic bound on the mutual information $\tilde{I}(\mathsf{x}, \mathsf{y})$ in the bound (7.6) is expressed as

$$\begin{aligned} \tilde{I}(\mathsf{x}, \mathsf{y}) &= \langle \log p(\mathsf{y}|\mathsf{x}) \rangle_{q(\mathsf{x},\mathsf{y})} + H(\mathsf{y}) \\ &= \sum_{i=1}^{|\mathsf{y}|} \langle \log p(y_i|\boldsymbol{\pi}_x(y_i), \boldsymbol{\pi}_y(y_i)) \rangle_{q(y_i, \boldsymbol{\pi}_y(y_i), \boldsymbol{\pi}_x(y_i))} + H(\mathsf{y}). \end{aligned} \tag{7.14}$$

We can always choose a mapping to the auxiliary space in such a way that computation of the entropic term $H(\mathsf{y})$ is not problematic. Here we will focus

mainly on the discussion of the computational complexity of $\langle \log p(\mathsf{y}|\mathsf{x}) \rangle_{q(\mathsf{x},\mathsf{y})}$. Note that the representational complexity of each conditional $p(y_i|\boldsymbol{\pi}_x(y_i), \boldsymbol{\pi}_y(y_i))$ is $\sim O(s^{|\boldsymbol{\pi}_x(y_i)|+|\boldsymbol{\pi}_y(y_i)|})$, where s is the number of states (for simplicity assumed to be equal for all variables y_i, x_j). Since we are free to choose a form of the distribution $p(\mathsf{y}|\mathsf{x})$, we can limit its parental structure, so that $|\boldsymbol{\pi}_x(y_i)|+|\boldsymbol{\pi}_y(y_i)|$ is low. For discrete variables this allows an exact representation of the conditionals.

In the special case when $q(\mathsf{y},\mathsf{x}) = \prod_{l=1}^{|\mathsf{y}|} q(y_l) \prod_{j=1}^{|\mathsf{x}|} q(x_j|\boldsymbol{\pi}_y(x_j))$ is an irregular sparse bipartite graph (i.e. the number of y-parents $\boldsymbol{\pi}_y(x_j)$ of each variable x_j is low), and the auxiliary conditional $p(\mathsf{y}|\mathsf{x})$ is a sparse structured distribution, one may perform the marginalization (7.14) explicitly, so that

$$\langle \log p(y_i|\boldsymbol{\pi}_x(y_i), \boldsymbol{\pi}_y(y_i)) \rangle_{q(y_i,\boldsymbol{\pi}_y(y_i),\boldsymbol{\pi}_x(y_i))} = \sum_{y_i,\boldsymbol{\pi}_y(y_i)} q(y_i, \boldsymbol{\pi}_y(y_i)) \times \sum_{\boldsymbol{\pi}_x(y_i)} q\left(\boldsymbol{\pi}_x(y_i)|y_i, \boldsymbol{\pi}_y(y_i)\right) \log p(y_i|\boldsymbol{\pi}_x(y_i), \boldsymbol{\pi}_y(y_i)), \tag{7.15}$$

where

$$q\left(\boldsymbol{\pi}_x(y_i)|y_i, \boldsymbol{\pi}_y(y_i)\right) \equiv \sum_{\tilde{\boldsymbol{\pi}}_i} q(\tilde{\boldsymbol{\pi}}_i) \prod_{j\in\boldsymbol{\pi}_x(y_i)} q\left(x_j|\tilde{\boldsymbol{\pi}}_i, y_i, \boldsymbol{\pi}_y(y_i)\right) \tag{7.16}$$

$$= \sum_{\tilde{\boldsymbol{\pi}}_i} q(\tilde{\boldsymbol{\pi}}_i) \prod_{j\in\boldsymbol{\pi}_x(y_i)} q\left(x_j|\boldsymbol{\pi}_y(x_j)\right), \tag{7.17}$$

and $\tilde{\boldsymbol{\pi}}_i \stackrel{\text{def}}{=} \boldsymbol{\pi}_y(\boldsymbol{\pi}_x(y_i))\backslash\{\boldsymbol{\pi}_y(y_i) \cup y_i\}$. Here we extended the definition of the y-parents in the conditional $q(\mathsf{x}|\mathsf{y})$, so that $\boldsymbol{\pi}_y(\boldsymbol{\pi}_x(y_i))$ defines all the y-parents (expressed from $q(\mathsf{x}|\mathsf{y})$) of all the x-parents (expressed from $p(\mathsf{y}|\mathsf{x})$) of the auxiliary variable y_j. The explicit marginalization (7.17) is $\sim O(s^{|\tilde{\boldsymbol{\pi}}_i|})$, and the complexity of explicit computations of the bound (7.15) is $\sim O(s^{|\boldsymbol{\pi}_x(y_i)|+|\boldsymbol{\pi}_y(\boldsymbol{\pi}_x(y_i))|})$. Clearly, it is acceptable as long as both $p(\mathsf{y}|\mathsf{x})$ and $q(\mathsf{x}|\mathsf{y})$ are sparse. For specific sparse structures of the auxiliary conditional $p(\mathsf{y}|\mathsf{x})$ and the variational distribution $q(\mathsf{x},\mathsf{y})$, one may also consider evaluating the bound (7.15) by applying algorithms of the exact inference (e.g. Jensen (1996), Jordan (2005)).

Clearly, if there are no other (e.g. parametric) constraints on the auxiliary conditional $p(y_i|\boldsymbol{\pi}_x(y_i), \boldsymbol{\pi}_y(y_i))$, optimization of the bound (7.5) for $p(\mathsf{y}|\mathsf{x})$ results in the optimal gain

$$\max_{p(\mathsf{y}|\mathsf{x})} \left\{ \tilde{I}(\mathsf{x},\mathsf{y}) \right\} = H(\mathsf{y}) - \sum_{i=1}^{|\mathsf{y}|} H(y_i|\boldsymbol{\pi}_x(y_i), \boldsymbol{\pi}_y(y_i)), \tag{7.18}$$

where $H(y_i|\boldsymbol{\pi}_x(y_i), \boldsymbol{\pi}_y(y_i))$ are the structural relaxations of $H_q(\mathsf{y}|\mathsf{x})$ expressed from the variational distribution $q(\mathsf{x},\mathsf{y})$. Computations of each entropic term $H(y_i|\boldsymbol{\pi}_y(y_i), \boldsymbol{\pi}_x(y_i))$ are of the same order of complexity $\sim O(s^{|\boldsymbol{\pi}_x(y_i)|+|\boldsymbol{\pi}_y(\boldsymbol{\pi}_x(y_i))|})$, and under the sparsity constraints may be performed exactly (see also the discussion in Section 2.2.1).

7.3 Relation to Variational Mixture Models

In Section 7.2 we described optimization of the auxiliary variational bound on the log partition function subject to parametric constraints on the auxiliary conditional distribution. Fundamentally, it is exactly due to the choice of a constrained conditional $p(\mathsf{y}|\mathsf{x})$ (the variational decoder in the IM formulation) that computationally efficient evaluations of the bound (7.5) were possible. The *variational mixture* methods of Jaakkola and Jordan (1998), Bishop et al. (1998) may be viewed as maximizing the same objective criterion (7.5) for the case when there are no utilizable structural or parametric constraints on $p(\mathsf{y}|\mathsf{x})$. In this case, optimization of (7.5) for large-scale models requires arguably less general relaxations of (7.6), such as the one given by (7.11). Our argument here is that by constraining the auxiliary conditional (variational decoder) $p(\mathsf{y}|\mathsf{x})$ to lie in a tractable family of distributions, we may avoid some of the problems of the existing variational mixture formulations.

The existing variational mixture approaches (Jaakkola and Jordan (1998); Bishop et al. (1998); Lawrence et al. (1998); El-Hay and Friedman (2002)) express the mutual information term in the objective criterion (7.5) as

$$\begin{aligned} \tilde{I}(\mathsf{x}, y) &\geq \left\langle \log \frac{\tilde{q}(\mathsf{x}|y)}{q(y)} \right\rangle_{q(\mathsf{x},y)} + \langle \log \lambda(y) \rangle_{q(y)} + 1 - \sum_y \lambda(y) \sum_{\mathsf{x}} \tilde{q}(\mathsf{x}|y) q_{mix}(\mathsf{x}) \\ &\stackrel{\text{def}}{=} \acute{I}(\mathsf{x}, y), \end{aligned} \tag{7.19}$$

which is obtained from (7.6) by applying the upper bound on the logarithmic function

$$\log x \leq \lambda x - \log \lambda - 1 \tag{7.20}$$

(see e.g. Jordan et al. (1998)). Commonly, the bound (7.19) is optimized with respect to the variational functionals $\lambda(y)$, "smoothing" conditionals $\tilde{q}(\mathsf{x}|y)$, and parameters of the mixture distribution $q(y)$ and $q(\mathsf{x}|y)$, where $q_{mix}(\mathsf{x}) \stackrel{\text{def}}{=} \langle q(\mathsf{x}|y) \rangle_{q(y)}$. In this case it is usually presumed that $y \in \{y_1, \ldots, y_{|y|}\}$ is the space of mixing coefficients, so that the marginal $q_{mix}(\mathsf{x})$ may be computed exactly[3]. In order for the computations in (7.19) to be tractable, the smoothing distributions $\tilde{q}(\mathsf{x}|y)$ will need to be factorized (e.g. Jaakkola and Jordan (1998)). While theoretically one could use (7.19) to variationally fit a mixture of trees of different structures (as opposed to the mixture of completely factorized (Lawrence et al. (1998)) or identically structured (El-Hay and Friedman (2002)) models), it may be numerically difficult, as optimization would involve computations of non-factorized ratios of summations of exponentially small terms (see e.g. Bishop et al. (1998)), which follows from the need of computing the high-dimensional convolutions $\langle q_{mix}(\mathsf{x}) \rangle_{\tilde{q}(\mathsf{x}|y)}$.

Moreover, it is not easy to interpret the optimal solutions of (7.19). Effectively, by optimizing (7.19) with respect to the variational parameters $\lambda(y)$, we obtain

$$\max_{\lambda(y)} \acute{I}(\mathsf{x}, y) = \left\langle \log \frac{\tilde{q}(\mathsf{x}|y) q(y)}{\langle \tilde{q}(\mathsf{x}|y) \rangle_{q_{mix}(\mathsf{x})}} \right\rangle_{q(\mathsf{x},y)} + H(y) \leq I(\mathsf{x}, y). \tag{7.21}$$

[3]Richer representations of the y-space may potentially be considered, provided that $q(\mathsf{x}, \mathsf{y})$ is in a tractable family, and the integrals in (7.19) may be computed exactly.

The bound (7.21) is saturated if and only if

$$\tilde{q}(\mathsf{x}|y)q(y)/\langle\tilde{q}(\mathsf{x}|y)\rangle_{q_{mix}(\mathsf{x})} \equiv q(y|\mathsf{x}). \tag{7.22}$$

However, for a general choice of $\tilde{q}(\mathsf{x}|y)$ the term in the l.h.s. of (7.22) does not define a proper distribution in y, which complicates the analysis of effects which relaxations of the structure of the conditional $\tilde{q}(\mathsf{x}|y)$ may have on the bound (*cf* variational information maximization). Probably the key conceptual problems of the existing variational mixture approaches are analytical difficulties of interpreting the optimal solutions, and practical difficulties of generalizing the existing bound (7.19) to richer, non-factorized families of the variational distributions. Some other practical limitations include general numerical instability of the direct computations of the average $\langle q_{mix}(\mathsf{x})\rangle_{\tilde{q}(\mathsf{x}|y)}$ in large-scale models, and the iterative nature of optimization for individual factors $\tilde{q}(x_i|y)$ of the smoothing distribution $\tilde{q}(\mathsf{x}|y) = \prod_{i=1}^{|\mathsf{x}|}\tilde{q}(x_i|y)$ (which may sometimes result in a low convergence speed).

Our method addresses the subgoal of optimizing the generic lower bound $\tilde{I}(\mathsf{x},\mathsf{y})$ by applying the variational IM. Apart from being conceptually and computationally simpler than the existing variational mixture approaches (Jaakkola and Jordan (1998); Bishop et al. (1998); Lawrence et al. (1998); El-Hay and Friedman (2002)), our method is also arguably more general. Specifically, by considering an unconstrained mapping to the auxiliary space and applying the bound on the logarithm (7.20), we can use our formulation (7.5) to arrive at (7.19), but not vice versa (also, note that (7.11) is a formal generalization of (7.19)). Importantly, we are not confined to using the specific relaxations (7.19), and may handle the intractability of computing $I(\mathsf{x},\mathsf{y})$ by imposing constraints on the auxiliary conditional $p(\mathsf{y}|\mathsf{x})$. Generally, the optimal auxiliary conditional distributions $p(\mathsf{y}|\mathsf{x})$ approximate the posteriors $q(\mathsf{y}|\mathsf{x})$ expressed from the variational model $q(\mathsf{y},\mathsf{x})$, which leads to a consistent improvement in the bound on $I(\mathsf{x},\mathsf{y})$. Also, due to the flexibility of choosing the form of the mapping to the auxiliary space, our method suggests extensions of the variational mixture approaches to factorial and structured auxiliary spaces. Specifically, we note that our bound on $\log Z$ may be used in conjunction with the auxiliary variational bound on mutual information (2.38) to improve on simple (e.g. factorized) choices of the auxiliary conditional $p(\mathsf{y}|\mathsf{x})$; it may also be used in situations when the auxiliary space is high-dimensional.

7.4 Demonstrations

Here we demonstrate the results comparing performance of the auxiliary variational framework with the standard factorized approaches. Also, for discrete variable spaces we apply our method to computing reweightings of fixed approximations $q(\mathsf{x}|y)$ of the original model $p(\mathsf{x})$, where each approximating distribution $q(\mathsf{x}|y_i)$, $i = 1,\ldots,|y|$ had a unique structure. We also look at the qualitative effects which the choice of the structure of the auxiliary conditional $p(\mathsf{y}|\mathsf{x})$ may have on the theoretically achievable improvements over a mixture of simple variational approximators.

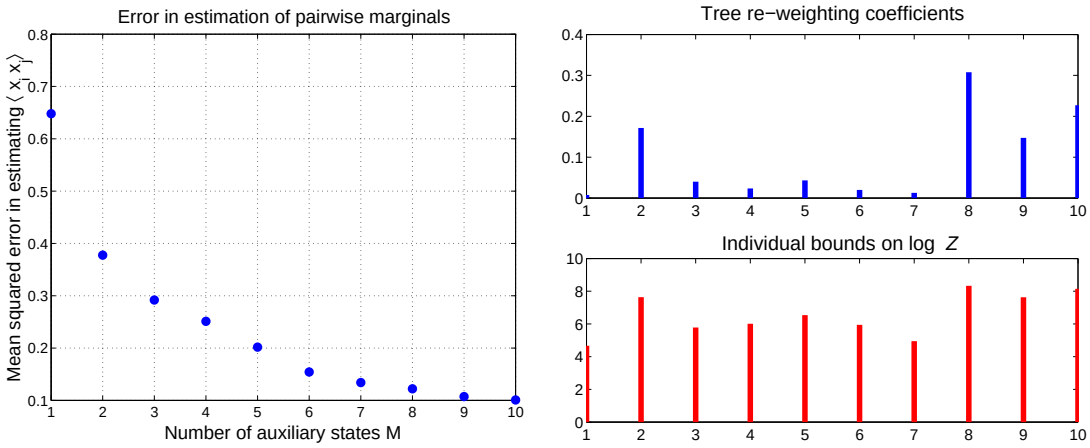


Figure 7.3: An auxiliary variational framework for approximate inference. (a) Systematic changes in the mean squared error for the estimates of the second-order moments with the growth in the number of mixture states M; (b) *Top:* re-weighting coefficients for a set of fixed structured approximators (each $q(\mathsf{x}|y)$ is a uniquely structured spanning tree); *Bottom:* the lower bounds on $\log Z$ obtained by each individual tree. Note that greater mixing coefficients were assigned to the tree-structured distributions which individually had resulted in higher values of lower bounds on $\log Z$.

Inference in Discrete Markov Networks

Here we demonstrate systematic changes in the auxiliary variational estimates of the second-order moments for discrete variable spaces. Throughout the simulations, it was assumed that $p(\mathsf{x})$ was a pairwise Markov network with the energy

$$E(\mathsf{x}) = -(\mathsf{x}^T \mathsf{W} \mathsf{x} + \mathsf{x}^T \mathsf{b}), \quad \mathsf{W} \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{x}|}, \ \mathsf{b} \in \mathbb{R}^{|\mathsf{x}|}, \tag{7.23}$$

and $\mathsf{x} \in \{-1, 1\}^{|\mathsf{x}|}$. In the following experiments it is assumed that $y \in \{1, \ldots, M\}$ is multinomial and $p(y|\mathsf{x})$ has a softmax form (7.13). Initially, we did not impose structural (sparsity) constraints on the auxiliary conditional distribution $p(y|\mathsf{x})$, and approximated the averages in the bound (7.5) at the means (i.e. $\langle \log p(y|\mathsf{x}) \rangle_{q(\mathsf{x}|y)} \approx \log p(y|\langle \mathsf{x} \rangle_{q(\mathsf{x}|y)})$). (Strictly speaking, the approximation is only accurate in the limit of noiseless projections from the auxiliary space, and one way to ensure its validity would be to explicitly constrain $q(x_i|y) \in \{0, 1\}$; we did not impose such constraints, hoping that maximization of the bound on mutual information $I(\mathsf{x}, y)$ in (7.5) would tend to favor low conditional entropies $H(\mathsf{x}|y)$, which would imply sharply peaked distributions). Analogous experiments were repeated for the case of conditionally factorized factorial auxiliary state representations, i.e. $p(\mathsf{y}|\mathsf{x}) = \prod_{y=1}^{|y|} p(y_i|\mathsf{x})$, $|y| \sim O(\log_2 M)$, which led to almost identical results.

To demonstrate the effect of the cardinality of the auxiliary space on the marginals, we generated 100 biases $\mathsf{b} \in \mathbb{I}^{10}$ and matrices $\mathsf{W} \in \mathbb{I}^{10\times 10}$ of the 10-D pairwise Markov network $p(\mathsf{x})$, where $\mathbb{I}$ defines the uniform range $[-1, 1]$. Then by analogy with Lawrence et al. (1998) we computed the squared errors $\epsilon(M)$

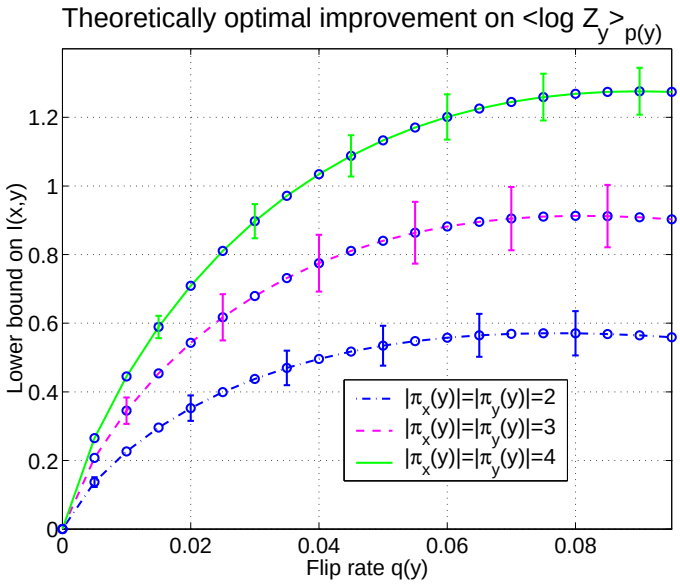


Figure 7.4: Influence of the structure of the auxiliary conditional $p(\mathsf{y}|\mathsf{x})$ and the variational $q(\mathsf{x}|\mathsf{y})$ distributions on the theoretically achievable improvement of the auxiliary variational lower bound on $\log Z$ over a mixture of simple bounds. The *x*-axis correspond to the "flip rate" $q(y)$. The *y*-axis correspond to the bound $\tilde{I}(\mathsf{x}, y)$. The results are computed for $|\mathsf{x}| = 20$, $|\mathsf{y}| = 40$. The curves correspond to the exact values of the bound $\tilde{I}(\mathsf{x}, \mathsf{y})$ for random structures of $q(\mathsf{x}|\mathsf{y})$ and $p(\mathsf{y}|\mathsf{x})$ which satisfy the specified sparsity constraints ($|\boldsymbol{\pi}_y(x)| = 3$, $|\boldsymbol{\pi}_x(y)|$ and $|\boldsymbol{\pi}_y(y)|$ are shown in the legend box). Note that a choice of a richer auxiliary conditional $p(\mathsf{y}|\mathsf{x})$ generally leads to higher values of $\tilde{I}(\mathsf{x}, y)$.

between exact and estimated second moments $\langle x_i x_j \rangle$, averaged for all networks and all $i \neq j$. Note that for $M = 1$ the auxiliary variational representation (7.6) is equivalent to the mean field model. As can be seen from Figure 7.3 (a), we obtain a systematic improvement in the accuracy with an increase in M. The scale of the changes in the accuracy is different from that reported by Lawrence et al. (1998), whose mean field error $\epsilon(1)$ was approximately 0.15, though we observe qualitatively similar improvements. This discrepancy is undoubtedly due to details of the optimization approaches, and does not detract from our conclusion that the auxiliary method conveys a systematic improvement.

Reweighting Structured Approximators

Here we demonstrate how we can easily apply a simple reformulation of our method for computing reweightings of fixed structured approximations $q(\mathsf{x}|y)$. For the weights $\mathsf{W} \in \mathbb{I}^{10\times 10}$ chosen at uniform random, we generated $K = 10$ random spanning trees with the weights $\mathsf{W}^{(m)}$ such that $\mathsf{W}^{(m)}_{ij} = \mathsf{W}_{ij}$ for all i, j and $m = 1, \ldots, K$. Then we re-weighted the trees by recomputing $q(y)$ according

to

$$q(y_k) \propto \exp\left\{\langle -E(\mathsf{x}) - \log q(\mathsf{x}|y_k) + \log p(y_k|\mathsf{x})\rangle_{q(\mathsf{x}|y_k)}\right\}, \quad (7.24)$$

$$= p(y_k) \exp\left\{-KL(q(\mathsf{x}|y_k)\|p(\mathsf{x}|y_k))\right\} \quad (7.25)$$

which follows directly from (7.10). Note that the mixing coefficient $q(y_k)$ of each component $q(\mathsf{x}|y_k)$ in the variational distribution is proportional to the surrogate marginal in the auxiliary space $p(y_k)$, and the discrepancy between the mixture component $q(\mathsf{x}|y_k)$ and the model $p(\mathsf{x}|y_k)$ for a fixed setting of the component's label y_k.

Figure 7.3 (b) illustrates typical re-weightings of fixed structured approximators $q(\mathsf{x}|y)$ and the induced lower bounds on $\log Z$ for the case of the softmax parameterization of $p(y|\mathsf{x})$ and for *fixed* parameters of the auxiliary conditional (selected at uniform random on $\mathbb{I}$). The corresponding bounds in this case were $L_r \approx 8.38$, $L_u \approx 8.87$, $L_b \approx 8.33$, and $L_{av} \approx 9.52$ for the random, uniform, best single, and auxiliary variational weightings respectively. Note that the single-step computation of (7.24) has an acceptable order of computational complexity ($\sim O(|\mathsf{x}|^2)$ if we consider the approximation $\log p(y|\langle\mathsf{x}\rangle_{q(\mathsf{x}|y)})$ and a Markov network (7.23)), and can be easily extended to the case of higher dimensional models.

The results are straight-forward but potentially interesting, as they suggest a quick and simple way to compute mixing coefficient for a set of trained structured approximations $q(\mathsf{x}|y)$. So instead of maximizing the bound for K standard intrinsically tractable approximators, choosing a single best one which leads to the tightest lower bound on $\log Z$, and discarding the remaining $K-1$ variational distributions, we can use all of the approximations to obtain a tighter lower bound on the normalizing constant. Interestingly, we can obtain these results even without learning the optimal mapping $p(\mathsf{y}|\mathsf{x})$ to the auxiliary space, and by performing generally suboptimal approximations of the conditional entropies in $\tilde{I}(\mathsf{x}, y)$ at the mean of $q(\mathsf{x}|y)$. An arguably more principled approach to the problem of re-weighting the variational distributions would involve optimization of the auxiliary variational lower bound on $\log Z$ with respect to $q(y)$ and $p(y|\mathsf{x})$ for a set of given structured components $q(\mathsf{x}|y)$. Formally, a more careful choice of constraints on $p(y|\mathsf{x})$ may need to be considered in order to ensure tractability of computing $\tilde{I}(\mathsf{x}, \mathsf{y})$. Generally, however, it may be practically attractive to apply a simple one-step procedure to produce a tighter bound on $\log Z$ for a given choice of standard variational distributions.

Structured Auxiliary Mappings

It is interesting to see how the choice of structured constraints on both the auxiliary mapping $p(\mathsf{y}|\mathsf{x})$ and the variational conditional $q(\mathsf{x}|\mathsf{y})$ could affect the theoretically optimal improvement over a convex combination of approximations. In order to check this, we assumed that the auxiliary space was binary and high-dimensional, so that $\mathsf{y} \in \{0,1\}^{|\mathsf{y}|}$ (*cf* the variational mixture methods of Jaakkola and Jordan (1998), Bishop et al. (1998)). We also assumed that different dimensions of the auxiliary vector were identically distributed, so that $q(\mathsf{x}, \mathsf{y}) = q(\mathsf{x}|\mathsf{y}) \prod_{i=1}^{|\mathsf{y}|} q_i(y_i)$ and $q_i(y_i = a) \equiv q_j(y_j = a) \stackrel{\text{def}}{=} q(y = a)$ for all

$i, j \in \{1, \ldots, |\mathsf{y}|\}$. The variational conditional $q(\mathsf{x}|\mathsf{y})$ was constrained as

$$q(\mathsf{x}|\mathsf{y}) = \prod_{i=1}^{|\mathsf{x}|} q(x_i|\boldsymbol{\pi}_{\mathsf{y}}(x_i)) = \prod_{i=1}^{|\mathsf{x}|} \delta\left(x_i - \sum_{y_j \in \boldsymbol{\pi}_{\mathsf{y}}(x_i)}^{\oplus} y_j\right), \tag{7.26}$$

where $\boldsymbol{\pi}_{\mathsf{y}}(x_i)$ are the y- parents of variable x_i in the graph for $q(\mathsf{x}|\mathsf{y})$, and $\sum^{\oplus}$ defines the modulo 2 summation. Equivalently, we can define $q(\mathsf{x}|\mathsf{y}) \sim \delta(\mathsf{x} - \mathsf{G}\mathsf{y})$, where $\mathsf{G} \in \{0,1\}^{|\mathsf{x}|\times|\mathsf{y}|}$ is a binary matrix such that $G_{ij} = 1$ if and only if y_j is a parent of x_i in the graph[4] for the variational conditional $q(\mathsf{x}|\mathsf{y})$.

To ensure tractability of the computations, we also imposed structural constraints on the stochastic mapping to the auxiliary space, so that the auxiliary conditional was defined as $p(\mathsf{y}|\mathsf{x}) = \prod_{i=1}^{|\mathsf{y}|} p(y_i|\boldsymbol{\pi}_{\mathsf{x}}(y_i), \boldsymbol{\pi}_{\mathsf{y}}(y_i))$, where $\boldsymbol{\pi}_{\mathsf{x}}(y_i)$ and $\boldsymbol{\pi}_{\mathsf{y}}(y_i)$ are the x- and y-parents of unit y_i under the model $p(\mathsf{y}|\mathsf{x})$, and $|\boldsymbol{\pi}_{\mathsf{x}}(y_i)| \ll |\mathsf{y}|$ (additional care had been taken to ensure that $p(\mathsf{y}|\mathsf{x})$ defined a proper distribution, i.e. its graph was directed acyclic). The variables were defined over discrete high-dimensional domains, so that $\mathsf{x} \in \{0,1\}^{|\mathsf{x}|}$ and $\mathsf{y} \in \{0,1\}^{|\mathsf{y}|}$, and under the imposed sparsity constraints the conditional $p(y_i|\boldsymbol{\pi}_{\mathsf{x}}(y_i), \boldsymbol{\pi}_{\mathsf{y}}(y_i))$ could be immediately obtained by maximizing the exact value of the bound (7.6).

Figure 7.4 shows the improvement $\tilde{I}(\mathsf{x}, \mathsf{y})$ over the convex combination in (7.5) as a function of the flip rate $q(y) \in [0, 0.1]$ for $|\mathsf{x}| = 20$ and $|\mathsf{y}| = 40$. The results are shown for the optimal settings of the auxiliary conditional $p(\mathsf{y}|\mathsf{x})$, obtained by maximizing the bound on $\log Z$ (which immediately leads to the constrained approximations of $q(\mathsf{x}|\mathsf{y})$). Throughout the experiments, it was assumed that $|\boldsymbol{\pi}_{\mathsf{y}}(x_i)| = 3$, and the structure of $p(\mathsf{y}|\mathsf{x})$ varied as shown on Figure 7.4. For each choice of the structural constraints, the results were averaged over 20 runs with random choices of $q(\mathsf{x}|\mathsf{y})$. We see that models with richer structures of the auxiliary conditional typically lead to greater bounds. Note that for the trivial independent setting $p(\mathsf{y}|\mathsf{x}) = p(\mathsf{y})$ (i.e. $|\boldsymbol{\pi}_{\mathsf{x}}(y_i)| = 0$), there would be no theoretical improvements over a standard approximation, so that $\tilde{I}(\mathsf{x}, \mathsf{y}) = 0$. However, the considered choices do lead to consistent improvements in (7.6) with an increase in the complexity of the auxiliary mappings. This also empirically confirms the analytical results of Section 2.2.1, and agrees with the intuition that richer structures of variational distributions tend lead to consistent improvements over simpler approximations.

We have repeated similar experiments for different sizes of the auxiliary space $|\mathsf{y}|$ (assuming that $|\mathsf{x}|$ was fixed). Our initial hope was that by considering high-dimensional factorial representations in the auxiliary space, we could potentially increase the theoretically achievable improvements over mixtures of simple bounds. It is easy to see that if M is the effective number of mixture states then $\tilde{I}(\mathsf{x}, y) \leq I(\mathsf{x}, y) \leq \log M$, with the maximum achieved for flat $q(y)$ and deterministic posteriors $q(y|\mathsf{x})$ (see expression (7.5) and a discussion in Bishop et al.

[4]This formulation is particularly interesting, as it could be shown to be relatable to the settings of a syndrome decoding problem (see e.g. McEliece (1977)), where the marginal $q(y)$ defines the bit error probability (flip rate) of a binary symmetric channel, $\mathsf{x} \in \{0,1\}^{|\mathsf{x}|}$ is the *syndrome* vector, and $\mathsf{y} \in \{0,1\}^{|\mathsf{y}|}$ is the (unknown) noise vector. The goal of decoding in this context would be to perfectly determine the unknown noise y from the syndrome vector x, which for the half-rate codes ($|\mathsf{y}| = 2|\mathsf{x}|$) could theoretically be done for $q(y) \lessapprox 0.11$.

(1998)). By considering the factorial auxiliary state representations, we hoped to massively increase the effective number of mixture states (so that $M \sim O(s^{|\mathsf{y}|})$) and achieve nearly linear improvement in the bound (7.5) with $|\mathsf{y}|$. Unfortunately, our initial intuition has not yet received strong empirical confirmations. While choosing richer structures of the auxiliary conditionals $p(\mathsf{y}|\mathsf{x})$ could indeed be helpful for obtaining consistent improvements in $\tilde{I}(\mathsf{x}, \mathsf{y})$, choices of higher-dimensional auxiliary spaces did not always lead to systematic improvements in the bounds, especially for higher flip rates $q(y)$ and random irregular structures of $p(\mathsf{y}|\mathsf{x})$ and $q(\mathsf{x}|\mathsf{y})$.

This effect may be explained by the sparsity constraints which we needed to impose on the auxiliary conditional. Let us presume that the auxiliary conditional is given by $p(\mathsf{y}|\mathsf{x}) = \prod_{i=1}^{|\mathsf{y}|} p(y_i|\boldsymbol{\pi}_x(y_i), \boldsymbol{\pi}_y(y_i))$, where each factor $p(y_i|\boldsymbol{\pi}_x(y_i), \boldsymbol{\pi}_y(y_i))$. Without loss of generality we may define the exact posterior expressed from the variational model $q(\mathsf{y}, \mathsf{x})$ as $q(\mathsf{y}|\mathsf{x}) = \prod_{i=1}^{|\mathsf{y}|} q(y_i|\tilde{\boldsymbol{\pi}}_x(y_i), \tilde{\boldsymbol{\pi}}_y(y_i))$, where $\tilde{\boldsymbol{\pi}}_x(y_i) \supseteq \boldsymbol{\pi}_x(y_i)$ and $\tilde{\boldsymbol{\pi}}_y(y_i) \supseteq \boldsymbol{\pi}_y(y_i)$. Then the difference between the exact mutual information $I(\mathsf{x}, \mathsf{y})$ and the tightest lower bound (7.6) obtained at the optimal settings of $p(\mathsf{y}|\mathsf{x})$ is given by

$$\sum_{i=1}^{|\mathsf{y}|} I(y_i, \{\tilde{\boldsymbol{\pi}}_x(y_i), \tilde{\boldsymbol{\pi}}_y(y_i)\} \backslash \{\boldsymbol{\pi}_x(y_i), \boldsymbol{\pi}_y(y_i)\} | \{\boldsymbol{\pi}_x(y_i), \boldsymbol{\pi}_y(y_i)\}),$$

which quantifies the overcounting effects which occur when we bound the conditional entropy $H(\mathsf{y}|\mathsf{x})$ by a summation of the marginals. (For example, if both the variational distribution $q(\mathsf{x}, \mathsf{y})$ and the auxiliary conditional $p(\mathsf{y}|\mathsf{x})$ are chains, this difference would be negligible only in situations when $q(y_i|y_{i-1}, x_i) \approx q(y_i|y_{i-1}, \mathsf{x})$, leading to $I(y_i, \mathsf{x} \backslash x_i | y_{i-1}, x_i) \approx 0$. This is effectively analogous to the cases when we may ignore future observations during the inference in state-space models, which clearly imposes significant limitations). One may expect that for general distributions $q(\mathsf{x}, \mathsf{y})$, the growth in $|\mathsf{y}|$ and $|\mathsf{x}|$ leads to increasing overcounting effects, resulting in a greater volume loss. These effects are not accounted for in the current formulation of our method, and will need to be addressed in further extensions of this work. E.g. the overcounting problem may be addressed by cluster variation methods (see e.g. Kikuchi (1951), Yedidia et al. (2000a), Yedidia et al. (2004)), which approximate entropy over a region as a weighted summation of entropies over smaller regions. Generally, however, our motivation here was to retain a rigorous bound on the intractable entropy (leading to a proper lower bound of $\log Z$), and improving our current method in a rigorous and general way is a challenging problem for future research.

7.5 Summary

Here we described an approach which generalizes the standard Kullback-Leibler variational procedures for evaluating the variational lower bounds on the generally intractable normalizing constants of undirected graphical models. Our work here was partially motivated by the success of auxiliary sampling techniques (e.g.

Swendsen and Wang (1987), Higdon (1998), Neal (1993)), which introduce auxiliary variables and draw samples from a specific joint distribution defined in the augmented variable space. The role of the auxiliary variables in this context is to capture (structural) information about clusters of correlated variables, which proves to be effective for decreasing the time gap between subsequent independent samples.

In this chapter we considered an auxiliary variable extension of common variational methods for bounding the log partition function, which could be useful in the context of approximate probabilistic inference. We showed that it is possible to define a tractable variational framework which leads to systematic improvements over the standard theory for any structured approximation. While the method described here is of a potential interest as a general approach for approximate inference, it demonstrates a curious link to our variational information-maximizing framework discussed in Chapter 2. Specifically, it turns out that the improvement of the proposed bound on $\log Z$ over a convex combination of standard bounds is given by a specific form of the generic lower bound on mutual information (see expression (2.2)).

Our auxiliary variational method for inference is related to the family of variational mixture models (see Jaakkola and Jordan (1998), Lawrence et al. (1998)), which can be seen as special and more computationally expensive cases of our approach. Specifically, we can obtain the existing bounds by considering an unconstrained mapping to the auxiliary space and applying a standard relaxation of the logarithmic function (7.20) (but we cannot generally derive our framework from the existing variational mixture approximations). Importantly, we may avoid computational, numerical, and practical problems of applying the existing variational mixture methods by considering tractable constraints on the mappings from the data space $\{\mathsf{x}\}$ to the auxiliary space $\{\mathsf{y}\}$. In this formulation, we may view our variational information-maximizing framework as an integral subgoal of auxiliary variational inference. The method is attractive both computationally and analytically, as the flexibility of the choice of the auxiliary conditional distribution (the variational decoder in the IM terminology) suggests simple generalizations of the generic approach to reacher families of variational decoders (such as chains, polytrees, and mixtures-of-experts, see e.g. Section 2.2.1 and Section 2.3). Additionally, the variational IM algorithm is easy to understand analytically (which becomes particularly attractive when compared with the analysis of the existing variational mixture models, where interpretation of the optimal smoothing factors and their effects on the induced bounds on $I(\mathsf{x}, y)$ at this stage remain, to us, largely unclear). Indeed, the optimization surface of our method is concave in the auxiliary conditional $p(\mathsf{y}|\mathsf{x})$, which for sparse models makes it possible to find the optimal projections analytically (see Section 2.2 and Section 7.2.2).

In the context of probabilistic variational inference, we may use the suggested approach for improving on the lower bounds on $\log Z$ produced by the standard approximations. Indeed, we showed that the method described here forms a more powerful class of approximations than any structured mean field technique. We also showed that a richer choice of structures of the auxiliary conditional

distribution may indeed lead to tighter lower bounds on the log partition function compared with simpler approximations. While these results are potentially interesting (especially when compared with standard variational approaches), the theoretically achievable gains over mixtures of standard approximations are bounded to be logarithmic in the size of the auxiliary space. Our initial hope was that by considering factorial representations of the auxiliary space, we could increase the effective number of the states, which could potentially lead to super-logarithmic improvements. Unfortunately, our current experience suggests that by naively increasing dimensionality of the auxiliary space (and imposing sparsity constraints on $p(\mathsf{x}|\mathsf{y})$), we do not always get significant improvements over the lower-dimensional formulations. This may potentially be explained by the fact that the simple formulation of the method leads to the bounds which ignore the overcounting problem in bounding the conditional entropy $H_q(\mathsf{y}|\mathsf{x})$. Addressing this matter in a systematic way can make the method potentially more attractive for variational inference.

A practical byproduct of our work is an efficient way to calculate a set of mixture coefficients for any set of tractable distributions, which principally improves on the flat combination (Agakov and Barber (2004a)). An extension of this work would be an application to a real-world problem (possibly in a structured context, e.g. by extending and re-formulating the results of Ghahramani and Hinton (1998), Ghahramani and Hinton (2000)).

Chapter 8

Discussion

Finding regularities in observations of the external world is an important task handled by many biological organisms. It is believed that learning about such regularities involves adapting structural and physiological properties of brain synapses, resulting in internal encodings of the external stimuli. Intuitively, meaningful internal representations should in some sense be informative about the environment, in which case biological learning may be viewed as a process of finding informative representations of the observations. Of course, in biology such internal representations would be intrinsically constrained by the neurophysiological properties of biological networks.

Many applications of machine learning are aiming to address a fundamentally similar task, where the key goal is to automatically find meaningful representations of the observations. For example, a system of automated medical diagnostics may be applied for determining hidden diseases which could have given rise to the observed symptoms; images of human faces taken by a robot's camera may be used by a robot to recognize human emotions, etc. The process of finding unknown informative descriptions (such as diseases and emotions) of the observations (such as symptoms and photographs) is generally referred to as the process of *inference*, which is one of the fundamental tasks of machine learning. Another fundamental task is *learning*, which may be viewed as an automated procedure for finding mathematical formalisms which would result in meaningful inferences. In practice learning often corresponds to finding an optimal *model* which provides a description of how the data relates to its representations. Both learning and inference may in principle be very computationally demanding. In this work we aimed to target computational intractability of a class of learning approaches.

Generative vs encoder models

Generally, one may distinguish two different approaches to unsupervised learning. The first class of methods aims to learn a constrained statistical model of the observations. The key idea there is to find a model of a probability distribution which would be likely to generate the data. Constraints on the distributions are introduced by utilizing the prior knowledge about the modeled domain, for example by choosing a specific parameterization and structure for a class of models. In these methods learning typically corresponds to fitting a constrained model to

the observations. Informally, this may be viewed as minimizing a measure of discrepancy between the empirical distribution and the marginal distribution of the observed variables expressed from the model. Inference, i.e. the process of finding unknown informative representations of the observations, can be addressed by applying fundamental probability rules. An example of such class of methods includes maximization of marginal likelihoods in *generative latent variable* models, where for a fixed set of parameters the model specifies how to generate the observations. Usefulness of the extracted hidden variable representations expressed from the model would in this case be quantified by how well the model fits to the set of observations. A different class of methods aims to extract informative representations directly from the set of observations. Rather than learning the density model of the observations, the methods aim to find a mapping (generally, stochastic) from the observations to the latent variable representations by optimizing other measures of informativeness. A popular class of such methods is based on the idea of maximizing the mutual information between the observations and the unknown representations in the *encoder*, or *recognition* models.

There are fundamental differences in how we parameterize and train generative and recognition models. Specifically, while parameterization of a generative model imposes explicit constraints on the distribution of the observations given latent variable representations (which may require knowledge about the process which generates the data), parameterization of an encoder model imposes explicit constraints on the *encoding* distribution. This may be particularly important in situations when constraints on the encoding distribution are partially known from the physical properties of the channels (which may be the case in engineering or neurobiological domains), or when explicit constraints on the encoding distribution form a part of a specific problem formulation (for example, in problems of constrained subspace selection). In some applications it may be easier to parameterize an encoder, rather than a generative model (one example is clustering with a defined family of similarity measures); in other cases, the situation may be completely opposite (for instance, when the family of distributions generating a specific dataset is known).

Generally, a choice of a model (generative *vs* encoder) may be strongly dependent on a specific task. To illustrate this, consider a problem of constructing a system for automated medical diagnostics, where patients' symptoms are associated with diseases. Many medical experts may arguably find it easier to parameterize a stochastic mapping $D \rightarrow S$ from the diseases (D) to the symptoms (S) (which requires encyclopedic knowledge), rather than $S \rightarrow D$ (which requires medical heuristics). Once the model is specified and trained, it may be applied for inferring combinations of hidden diseases from a combination of the observed symptoms. We see that the former specification of the system defines a standard generative model; the latter parameterization corresponds to an encoder model. If we have a good idea about symptoms which could be "generated" by a given disease, the generative model is more easily parameterizable.

Now let us re-formulate the problem slightly. Assume that we are interested in finding clusters of symptoms which could form a new (and unknown) family of diseases, which we may be aiming to name. By specifying a similarity metric

in $\{S\}$ (i.e. defining what it means for the symptoms to be similar), we may find it easier to describe a mapping from a symptom to a disease family, even though we may not necessarily know much about disease families. This would be an example of an encoder model, where the unknown cluster labels (disease families) carry useful information about the observations (symptoms). Effectively, this formulation ($S \rightarrow D$) defines a *discriminative unspervised* framework, where the model is parameterized similarly to a conditionally trained classifier (like in the supervised learning formulation), but the representations D are unknown. Such parameterizations may be particularly useful in situations when little is known about the process which generates the data.

Unfortunately, for both generative and encoder models the problem of finding informative latent variable representations $\{\mathsf{y}\}$ of the observations $\{\mathsf{x}\}$ may become very difficult in the presence of noise, which motivates a need of approximations. In this thesis we considered a class of machine learning methods maximizing the mutual information $I(\mathsf{x}, \mathsf{y})$ in encoder models, and addressed the fundamental computational problems of the exact formulation by applying variational approximations. Our focus on the discussion of variational methods for information maximization was partially motivated by their popularity and effectiveness for inference and likelihood training in graphical models, and the apparent lack of understanding of how the methods could in general be applied in the information-maximizing context (indeed, many of the currently existing approximations of mutual information are either heuristic or too specific). The fundamental advantage of the variational methods over other approximations is availability of rigorous bounds on the intractable quantities, which facilitates comparisons of different variational approximation techniques, and in some cases makes them particularly attractive for learning. For this reason, one of our technical subgoals here was to try to retain rigorous bounds on $I(\mathsf{x}, \mathsf{y})$, whenever possible.

General results

In this thesis we described a family of variational lower bounds on mutual information $I(\mathsf{x}, \mathsf{y})$, which gave rise to a formal and theoretically justified approach to information maximization in noisy channels. While the formulation of the generic bound on the objective criterion is straight-forward, it appears to have attracted little previous attention as a practical tool for approximate maximization of information content. The fundamental idea of the approach is to introduce a *variational decoder* $q(\mathsf{x}|\mathsf{y})$ which is constrained to lie in a tractable family. Effectively, an iterative information-maximizing (IM) algorithm optimizing the generic lower bound $\tilde{I}(\mathsf{x}, \mathsf{y})$ extends the family of the generally intractable Blahut-Arimoto type algorithms (Arimoto (1972), Blahut (1972)), and reduces to them in the special case when the variational decoder is unconstrained. Qualitatively, this is similar to the variational EM algorithm for likelihood maximization (Neal and Hinton (1998)), which reduces to the standard EM (Dempster et al. (1977)) for unconstrained variational posteriors. The generality of the approach allows a flexibility in the choice of variational decoders or specific optimization procedures, which suggests that the method may naturally generalize other techniques for approx-

imate maximization of mutual information. Indeed, in addition to generalizing the conventional Blahut-Arimoto formulation, the IM generalizes the existing *as-if* Gaussian criterion (Linsker (1992)), which may be seen as a specific way of optimizing the variational lower bound on mutual information for a specific choice of linear Gaussian variational decoders. Additionally, factorial relaxations of the generally intractable conditional entropy $H(\mathsf{x}|\mathsf{y})$, which are sometimes used in approximations of $I(\mathsf{x},\mathsf{y})$, may be seen as special cases of the generic lower bound on $I(\mathsf{x},\mathsf{y})$ with structural constraints on the parental structure of the variational decoder distribution.

In our work we also explored general relations of the variational IM algorithm to maximum likelihood learning in generative models and conditional likelihood learning in noiseless and stochastic autoencoders (where by a noiseless autoencoder we mean a self-supervised network with a deterministic encoding distribution $p(\mathsf{y}|\mathsf{x}) \sim \delta$). In contrast to much of the previous work which related the likelihood and the mutual information approaches for relatively simple special cases (e.g. Pearlmutter and Parra (1996), Cardoso (1997), MacKay (1999b)), we aimed at relating the methods for the general variational settings independently of the specific model parameterizations. To make the comparisons possible, we assumed that both the encoder and the generative model led to identical inferences of the latent representations y for all the source patterns x. For this case, we showed that the likelihood of a generative model $\mathsf{y} \to \mathsf{x}$ could be viewed as a relaxation of a specific generic bound on $I(\mathsf{x},\mathsf{y})$, defined for the corresponding model $\mathsf{x} \to \mathsf{y}$ of a noisy channel. Interestingly, while generally it is not easy to relate learning by maximizing the exact likelihood and the exact mutual information for generative and encoder models, we may find specific cases when the fixed points of the generic lower bound on $I(\mathsf{x},\mathsf{y})$ are identical to those of the standard variational Jensen's bound on the likelihood (with identical constraints on the *encoding mapping* of the encoder model and the *variational posterior* of the generative model). For example, this happens for noiseless channels *independently* of invertibility of the mappings $\mathsf{x} \mapsto \mathsf{y}$ (provided that the generative model is characterized by flat priors on latent variables y). This result shows interesting intersections of the variational IM and EM approaches, though in general the methods are quite different.

Another curious result of our study is a link of the simple form of the IM algorithm to conditional likelihood training in stochastic autoencoders. While the relation of the probability of correct *deterministic* reconstructions to the mutual information in the corresponding stochastic channels is well-known (Fano (1961)), its stochastic generalizations appear to have attracted little attention within the machine learning community. (This was probably the reason for a large number of heuristic or very specific approximations of the generally intractable mutual information $I(\mathsf{x},\mathsf{y})$, which to the best of our knowledge are rarely (if at all) compared with the conditional training in stochastic autoencoders). A simple way to maximize the information content which the codes y contain about the sources x would be to maximize the conditional likelihood $p(\tilde{\mathsf{x}}|\mathsf{x})$ in a stochastic autoencoder model $\mathsf{x} \to \mathsf{y} \to \tilde{\mathsf{x}}$ for some choice of the decoding distribution. It is easy to see that in stochastic models the conditional likelihood is generally intractable,

and it is necessary to consider its approximations. As a result of our study we showed that the standard variational Jensen's bounds on the conditional likelihood have the same fixed points as the generic lower bound on $I(\mathsf{x}, \mathsf{y})$. This result is rather disappointing, as it shows that by considering a simple *stochastic* model performing *self-supervised* training we could effectively arrive at the fixed points identical to the ones produced by the simplest formulation of the IM approach. However, our method is easier than standard variational approaches of maximizing the conditional likelihood in stochastic autoencoders, which would effectively correspond to a more expensive way of optimizing the simple generic bound on the mutual information. Generally, this results in the computational and representational efficiency and desirable convergence properties of the IM approaches, which optimize the bound for a significantly lower number of variational parameters (lower by the order of magnitude compared with a naive application of the variational Jensen's bound on $p(\tilde{\mathsf{x}}|\mathsf{x})$).

Another result of the suggested work is the definition of a richer family of tractable *auxiliary variational* lower bounds on $I(\mathsf{x}, \mathsf{y})$, which formally generalizes on the generic lower bounds on $I(\mathsf{x}, \mathsf{y})$. By analogy with the auxiliary sampling techniques (see e.g. Swendsen and Wang (1987), Higdon (1998)), the key idea was to introduce additional variables, which could be used for capturing useful regularities of the source patters and for introducing global dependencies to the decodings. Importantly, projections to the auxiliary space were defined in a way which did not alter properties of the original channel. We demonstrated that the auxiliary variational bounds may indeed be significantly tighter than the simple generic criteria. This result is potentially interesting from the communication-theoretic perspective, as it demonstrates a simple and computationally efficient way to produce tighter bounds on the capacity of a communication channel without assuming that more data is being transmitted across the channel. Generally, this family of variational methods may be used as a simple and tractable approach for improving on simple bounds on mutual information.

Interestingly, the variational formulation of the information maximizing procedure suggests a relation between maximizing the generic lower bound on $I(\mathsf{x}, \mathsf{y})$ and computing an optimal estimate of an intractable posterior $p(\mathsf{x}|\mathsf{y})$ in a generative model, where the codes y are visible and the sources x are hidden. One of the goals of variational inference in graphical models is to approximate moments of the generally intractable posterior $p(\mathsf{x}|\mathsf{y})$ by the moments of a simpler distribution $q(\mathsf{x}|\mathsf{y})$. Standard mean field approaches assume that $q(\mathsf{x}|\mathsf{y}) \propto \prod_{i=1}^{|\mathsf{x}|} q(x_i|\mathsf{y})$, which usually leads to $q(x_i|\mathsf{y})$ approximating any one of a large number of local *modes* of the model-specific posterior $p(x_i|\mathsf{y})$ (where $q(\mathsf{x}|\mathsf{y})$ is assumed to be factorized in x). On the other hand, by imposing parametric constraints on the variational decoder, so that $q(\mathsf{x}|\mathsf{y}, \boldsymbol{\Theta}) = \prod_{i=1}^{|\mathsf{x}|} q(x_i|\mathsf{y}, \boldsymbol{\Theta}_i)$, and optimizing our bound on $I(\mathsf{x}, \mathsf{y})$ for parameters of the decoder, we obtain posterior mean estimates which are good *on average*.

This outlines a fundamental problem of using the variational decoders $q(\mathsf{x}|\mathsf{y}, \boldsymbol{\Theta})$ of the IM formulation for approximate inference, even compared with standard mean field methods. For example, this is the case for the problem of *syndrome decoding* in binary symmetric channels. Specifically, if the exact posterior $p(\mathsf{x}|\mathsf{y})$

is sharply peaked (which is the case for specific choices of the encoder distribution $p(\mathsf{y}|\mathsf{x})$, see e.g. Luby et al. (2001)), a mean field theory may indeed provide an accurate approximation of $p(\mathsf{x}|\mathsf{y})$ for each setting of y. (Generally, however, such accurate approximations may not always be easily found, due to a large number of local modes). In contrast, by learning a factorized variational distribution which is only optimal on average, variational estimates of the exact marginal for each specific codeword y may in general be very poor. One could hope to obtain improvements by considering significantly richer families of variational decoders (e.g. by having no parametric constraints on each factor $q(x_i|\mathsf{y})$). Unfortunately, unless the parental structure of each $q(x_i|\mathsf{y})$ is very sparse, the computational and representational complexity of using such approximations will generally be prohibitively high, while a choice of sparse structures will still lead to the averaging effects. While it may potentially be possible to modify the IM framework to obtain better estimates of the exact posterior means $\langle x_i \rangle_{p(x_i|\mathsf{y})}$, we do not advocate using simple constrained variational decoders obtained by our method as a competitive technique for error correction in binary symmetric channels.

Generally, we stress that a correct way of viewing our variational approach would be to interpret it as a general framework for learning an optimal *encoding* distribution by maximizing a proper bound on $I(\mathsf{x}, \mathsf{y})$. Once the optimal encoder is learned, one may choose any of a number of approximate inference techniques to perform the reconstructions. Using the variational decoder $q(\mathsf{x}|\mathsf{y})$ for inference is only one of such choices (which may be good or poor, depending on the specific application).

Case studies

As a part of our exploration of the information maximizing framework, we considered applications to constrained dimensionality reduction. Specifically, we discussed several ways of applying the framework to information-theoretic clustering, where the encoding distribution was defined either by the exact posterior of the corresponding latent variable model, or by explicitly choosing a specific nonlinear projection into the code space. The former approach could also be used for training generative models, which typically resulted in a more uniform coverage of the data space compared with the conventional training by maximizing the likelihood in mixture models. The latter approach to information-theoretic clustering could be applied to learning parameters of kernel functions (within a specific family), which was beneficial for visualizing the underlying structure of the data. Empirically, the method favorably compared with the Gaussian mixture, feature-space k-means, and non-kernelized information-theoretic clustering.

We also outlined analytical properties of the optimal IM solutions for the case of real-valued encoded representations y. Specifically, we extended the work of Bourlard and Kamp (1988) and Bourlard (2000) to *arbitrary kernelizable* feature mappings in the stochastic information-theoretic context, and showed the nothing could be gained by using nonlinear encoders and linear variational decoders in the context of variational information maximization in noisy Gaussian channels. To handle the intrinsic constraints of linear Gaussian variational decoders applied in the context of nonlinear encodings, we suggested a proper variational

relaxation of the bound on $I(\mathsf{x},\mathsf{y})$. For *nonlinear Gaussian* encoding distributions, this led to kernel PCA (Schoelkopf et al. (1998)) as the optimal solution for encoder's weights. Additionally, in the deterministic limit of non-parametric encoding mappings, our framework led to the Gaussian Process Latent Variable Models (Lawrence (2003)) for a specific choice of the variational decoder.

As an immediate extension of our work, we note that a further study of the bound on $I(\mathsf{x},\mathsf{y})$ for nonlinear Gaussian channels may need to be considered. Our current study indicates that the IM framework may be used to learn non-degenerate kernel parameters, but the obtained visualization and reconstruction results are strongly influenced by the choice of the kernel and the constraints on the feature-to-data decoding mappings. Importantly, we note that if the encoding noise may in principle be reduced to zero, one may consider optimizing simpler objective functions, such as those of Gaussian Process Latent Variable Models (Lawrence (2003)). These models also appear to lead to better visualization results, as they consider noiseless non-parametric encoders and rich families of variational decoders (given in these models by products of Gaussian processes). It is therefore believed that a further study of variational IM in nonlinear Gaussian channels for visualization or dimensionality reduction may be of a practical interest mainly in situations when the noise of the encoding distribution is unavoidable.

An arguably more promising application field to explore would be communication of discrete-valued data over channels with Gaussian noise, where a specific practical application may include code division multiple access in cellular telephony (see e.g. Viterbi (1995)). Additionally, it is interesting to explore the relation of our kernelized information-theoretic clustering approach (Agakov and Barber (2005b), Agakov and Barber (2005c)) to some of the common spectral clustering methods (e.g. Shi and Malik (2000), Yu and Shi (2003)), which were shown to be related to the weighted feature-space k-means (Dhillon et al. (2004)). Some of our preliminary results extending the work of Chapter 5 suggest that by constraining the feature-space coefficients, it may potentially be possible to relate the mentioned spectral clustering methods to a form of the constrained variational information-maximizing procedure (though a direct relation between the methods is not entirely clear at this stage).

As another application of the variational IM approach, we explored applicability of the variational IM framework in the context of learning high-dimensional binary representations of continuous source patterns for a specific biologically inspired parameterization of the encoding distribution (defining a population of point-neuron models). We believe that the application area where the results may be found to be particularly useful is stochastic neural coding (as neurophysiological properties of biological networks suggest a need of explicit constraints on the encoding distribution). For this case, we compared our variational approach with Brunel and Nadal's Fisher approximation of mutual information (Brunel and Nadal (1998)). Moreover, we compared our approach with the recent results of Szummer and Jaakkola (2002) and Corduneanu and Jaakkola (2003), which may be seen as another approximation of a lower bound on $I(\mathsf{x},\mathsf{y})$ formally generalizing on the criteria optimized by a variety of common popula-

tion coding methods (Pouget et al. (1998), Zhang and Sejnowski (1999), Bethge et al. (2002)). Our empirical results indicated that for the considered cases, our variational approach was most preferable (though the results here are rather preliminary, since in order to provide a sensible comparison of our bound with the existing non-bounding approximations, we have constrained our models to be low-dimensional). Additionally, we demonstrated that for the considered encoding distribution it was possible to derive a *local* learning rule, which may potentially be attractive from neuro-biological and computational perspectives (this also generalizes work of Linsker (1997), who had derived local approximations of *infomax* learning for invertible channels). In the future, more biologically realistic channels and applications should be considered. Additionally, for the specific biologically interesting channels, we are planning to conduct a tractability study of other lower bounds on mutual information known from statistical mechanics of supervised learning (Opper and Haussler (1995), Haussler and Opper (1997)) and compare them with our variational approach. Nevertheless, we can say that at this stage our current results illustrate potential advantages of the variational information-maximizing framework over the common (Brunel and Nadal (1998), Kang and Sompolinsky (2001)) and less common (Szummer and Jaakkola (2002), Corduneanu and Jaakkola (2003) and Section 6.2.1) approaches to population coding of high-dimensional input stimuli.

Finally, we considered a seemingly unrelated problem of lower bounding the normalizing constant (partition function) Z of an undirected graphical model (similar methods may be used to bound the likelihood of any probability distribution). Specifically, we introduced an auxiliary variable extension of any structured mean field theory, which could be useful in the context of approximate probabilistic inference. We showed that by considering a projection to the auxiliary variable space, and expressing the Kullback-Leibler divergence between the distributions defined over the augmented variable spaces, it was possible to define a tractable variational framework which leads to systematic improvements over the standard theories. While the auxiliary variational extension of standard approximation theories is of a potential interest as a general approach to approximate inference, it demonstrates a curious link to our variational information-maximizing framework. In particular, it turns out that the improvement of the proposed bound on $\log Z$ over a convex combination of standard bounds is given by a specific form of the generic lower bound on mutual information. The variational IM framework may therefore be seen as addressing an integral subgoal of auxiliary variational inference.

This part of our study suggests interesting directions for further research, including exploration of the effects which a choice of the mapping to the auxiliary space may have on the bounds for high-dimensional auxiliary spaces. Unfortunately, our current results suggest that the sparsity constraints on the structures of auxiliary conditional (variational decoder) distributions may be too restrictive. (For example, in the case when both the variational distribution $q(\mathsf{x}, \mathsf{y})$ and the auxiliary conditional $p(\mathsf{y}|\mathsf{x})$ have chain structures, the theoretically achievable gain produced by our approximation (over a simple theory) would only be high ($\sim O(|\mathsf{y}|)$) in situations when $q(y_i|y_{i-1}, x_i) \approx q(y_i|y_{i-1}, \mathsf{x})$, where y and x

are the auxiliary and latent variables respectively. This is effectively analogous to the cases when we may ignore future observations during inference in state-space models, which may be the case only for a limited number of applications). Nevertheless, we can formally show that by considering richer structures of the auxiliary conditional distributions we formally improve on simple approximations, modeling them as generally suboptimal special cases. Generally, modifying our framework for obtaining even better improvements would be an interesting and challenging direction for future research. We expect, however, that obtaining significantly better improvements over standard theories without sacrificing the rigorous bound may be difficult.

A further study of our auxiliary variational method for inference may also be motivated from a different perspective. Note that even in situations when the exact posterior may be accurately approximated by a factorized uni-modal distribution, finding such good factorized approximations by using the standard techniques may be complicated. While our results suggest that increasing the cardinality of the auxiliary variables, or using richer structures of variational decoders helps to produce better lower bounds on the partition function (by missing less of the probability mass), at this stage it remains largely unclear how helpful the auxiliary variables could be for preventing convergence of each component of the variational distribution to a suboptimal local mode of the exact posterior. A part of our motivation for further exploration of the method is the reported success of the auxiliary sampling techniques, which often result in efficient exploration of the distribution (by allowing potentially large changes in clusters of variables). Motivated by this observation, we are planning to continue studying our framework by empirically applying it to real-world tasks, as well as analyzing the theoretically achievable improvements over simple bounds for other choices of constraints on the auxiliary conditional distribution.

Summary

To summarize, we note that for the purpose of finding informative representations of the data, learning by maximizing the mutual information in encoder models is an alternative to learning probability models of the observations. In some cases, e.g. when specific tasks require explicit parameterization of the encoding distribution, information-maximization in encoder models may be the method of choice. It may be particularly attractive for clustering and constrained dimensionality reduction (for both deterministic and intrinsically noisy encoding mappings). An example of an intrinsically noisy channel is a model of a biological network (for example, a retinal encoder model defining a mapping from photoreceptors to ganglion representations in the retina), where a choice of constraints on encoder parameters is motivated by physiological properties of biological systems.

An important practical complication of maximizing mutual information in stochastic environments is high computational complexity of the exact optimization procedure, which is only tractable in a few simple special cases. In this thesis we suggested a general variational framework for maximizing a family of proper lower bounds on mutual information in stochastic environments. It turned out that in its simplest formulation our framework led to the same fixed points as the

approaches aiming to optimize the variational Jensen's bounds on the likelihoods of stochastic self-supervised models; however, our method is computationally and representationally simpler than naive applications of the variational EM for such models. Additionally, our approach may be easily generalized to define richer types of bounds on mutual information which formally generalize simpler approximations. We believe that this makes our method potentially useful as a general variational framework for approximate information maximization.

We note, however, that the simple formulation of the suggested framework may in some cases be limited. Specifically, our empirical experience and intuition suggest that using parametric or structurally constrained variational decoders for approximating moments of the exact posterior $p(\mathsf{x}|\mathsf{y})$ may in some cases be too restrictive (for example, this is the case for error-correction in binary symmetric channels). This follows from the fact that the IM-optimal constrained variational decoder computes the marginal estimates which are good on average (for all possible codewords $\{\mathsf{y}\}$). Thus, for any *specific* codeword y a constrained IM-optimal decoder may often be inferior to standard approximation theories. In future work it will be interesting to see whether or not using the IM framework for learning the optimal *encoder* may improve on standard inferences in the resulting models. Improving our framework to be competitive for error correction applications may be particularly interesting and challenging, as this is one case where a naive application of our approach is not expected to work well.

Despite the apparent limitations of our framework, we believe that the results presented in this work may be potentially interesting from several perspectives. First of all, in contrast to the majority of the existing approximations of $I(\mathsf{x}, \mathsf{y})$ (see e.g. Brunel and Nadal (1998), Kang and Sompolinsky (2001), Torkkola (2000), Gokcay and Principe (2002), Szummer and Jaakkola (2002), Corduneanu and Jaakkola (2003)), our method optimizes a proper lower bound, rather than a surrogate objective criterion or an approximation of $I(\mathsf{x}, \mathsf{y})$ (which may only be accurate under specific asymptotic assumptions, and weak or even undefined when the assumptions are violated). Secondly, the flexibility of the choice of the variational distribution makes it possible to generalize and improve other bounds on mutual information. For example, we may tractably extend our method to the family of auxiliary variational bounds on $I(\mathsf{x}, \mathsf{y})$, which may be used to improve on any simple generic approach without altering properties of the original channel. Moreover, in contrast to other approaches (e.g. Lawrence et al. (1998), Jaakkola and Jordan (1998), Brunel and Nadal (1998)), our method may be easily generalized to the structured context. Finally, we can demonstrate that by applying the IM framework to optimizing the bounds on $I(\mathsf{x}, \mathsf{y})$ in the augmented space of encoder and variational decoder parameters, we can often obtain simpler optimization procedures than those resulting from expressing the bounds as functions of the encoder alone. This leads to an interesting observation that in some cases the IM framework may be used to derive optimization procedures which only require local computations. This may be particularly attractive from the the neuro-biological, but also from the computational perspectives.

Possibly the most important contribution of this work is a rigorous and general variational framework for maximizing the mutual information in intrinsically

intractable channels. We show that it leads to simple, stable, and easily generalizable optimization procedures, which (despite some intrinsic constraints) outperform and supersede many of the common approximate information-maximizing techniques.

Appendix A

Linsker's Bound: Centering in the Code Space

A brief comment about centering the data in the code space is in order. Consider the *as-if Gaussian* objective

$$I_G(\mathsf{x},\mathsf{y}) \quad \propto \quad \log\left|\langle \mathsf{x}\mathsf{x}^T\rangle - \log\langle \mathsf{x}\mathsf{x}^T\rangle - \langle \mathsf{x}\mathsf{y}^T\rangle\langle \mathsf{y}\mathsf{y}^T\rangle^{-1}\langle \mathsf{y}\mathsf{x}^T\rangle\right| \tag{A.1}$$

with source vectors x and codes y (see Section 2.2.2). Here all the averages are computed over $p(\mathsf{x},\mathsf{y}) = \tilde{p}(\mathsf{x})p(\mathsf{y}|\mathsf{x})$, where $p(\mathsf{y}|\mathsf{x})$ is the encoder and $\tilde{p}(\mathsf{x})$ is the empirical distribution. In Section 2.2.2 we showed that (A.1) is in fact a special case of the generic lower bound on $I(\mathsf{x},\mathsf{y})$

$$\tilde{I}(\mathsf{x},\mathsf{y}) = \langle \log q(\mathsf{x}|\mathsf{y})\rangle_{\tilde{p}(\mathsf{x})p(\mathsf{y}|\mathsf{x})} + const \tag{A.2}$$

for the variational decoder $q(\mathsf{x}|\mathsf{y}) \sim \mathcal{N}_{\mathsf{x}}(\mathsf{U}\mathsf{y},\boldsymbol{\Sigma})$, where we assumed $\langle \mathsf{x}\rangle_{p(\mathsf{x})} = 0$ and $\langle \mathsf{y}\rangle_{p(\mathsf{y})} = 0$. In practice, centering the source vectors $\{\mathsf{x}\}$ should not be problematic, since transformations of the sources are not explicitly affected by parameterization of the encoder $p(\mathsf{y}|\mathsf{x})$. In some cases, $\langle \mathsf{x}\rangle_{p(\mathsf{x})} = 0 \Rightarrow \langle \mathsf{y}\rangle_{p(\mathsf{y})} = 0$ [for example, this is the case for linear Gaussian encoders $p(\mathsf{y}|\mathsf{x}) \sim \mathcal{N}_{\mathsf{y}}(\mathsf{W}\mathsf{x},\sigma^2\mathsf{I}_{|\mathsf{y}|})$]. However, in general the assumption of the centered *encodings* $\{\mathsf{y}\}$ implies additional constraints on the encoder distribution $p(\mathsf{y}|\mathsf{x})$, which may in general be difficult to enforce.

We will now consider non-centered codes (i.e. $\langle \mathsf{y}\rangle_{p(\mathsf{y})} \neq 0$) and a more general form of the linear Gaussian variational decoder, namely $q(\mathsf{x}|\mathsf{y}) \sim \mathcal{N}_{\mathsf{x}}(\mathsf{U}\mathsf{y}+\boldsymbol{\mu},\boldsymbol{\Sigma})$. A straight-forward substitution into (A.2) then results in

$$\tilde{I}(\mathsf{x},\mathsf{y}) = -\frac{1}{2}\left\langle \mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}(\mathsf{x}-\mathsf{U}\mathsf{y}-\boldsymbol{\mu})(\mathsf{x}-\mathsf{U}\mathsf{y}-\boldsymbol{\mu})^T\right\}\right\rangle_{p(\mathsf{x})p(\mathsf{y}|\mathsf{x})} - \frac{1}{2}\log|\boldsymbol{\Sigma}| + const. \tag{A.3}$$

Then by maximizing $\tilde{I}(\mathsf{x},\mathsf{y})$ for $\boldsymbol{\mu} \in \mathbb{R}^{|\mathsf{x}|}$ we obtain $\boldsymbol{\mu} = -\mathsf{U}\langle \mathsf{y}\rangle$. By comparing the mean of the reconstructed sources (computed for the variational distribution $q(\mathsf{x}|\mathsf{y})$) with the empirical mean of the source vectors, we easily see that $\langle \mathsf{x}\rangle_{q(\mathsf{x}|\mathsf{y})p(\mathsf{y})} = \mathsf{U}\langle \mathsf{y}\rangle_{p(\mathsf{y})} + \boldsymbol{\mu} = \langle \mathsf{x}\rangle_{p(\mathsf{x})} = 0$, i.e. the variational decoder does not introduce a bias into the reconstructions. By substituting $\boldsymbol{\mu} = \mathsf{U}\langle \mathsf{y}\rangle \in \mathbb{R}^{|\mathsf{y}|}$ and solving for $\mathsf{U} \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{y}|}$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{x}|}$, we obtain

$$\tilde{I}(\mathsf{x},\mathsf{y}) = -\log|\boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}|, \tag{A.4}$$

where we ignored the irrelevant constants and let $\mathbf{\Sigma}_{xx} = \langle \mathsf{x}\mathsf{x}^T \rangle$, $\mathbf{\Sigma}_{xy} = \mathbf{\Sigma}_{yx}^T = \langle \mathsf{x}\mathsf{y}^T \rangle$, and $\mathbf{\Sigma}_{yy} = \langle \mathsf{y}\mathsf{y}^T \rangle - \langle \mathsf{y} \rangle \langle \mathsf{y} \rangle^T$. This is exactly the general form of Linsker's *as-if Gaussian* criterion (see expression (1.16)).

Appendix B

Analysis of Variational Information Maximization

Here we outline several simple relations of the variational lower bound on mutual information to the standard variational bounds on the likelihood and conditional likelihood. Section B.1 discusses the case of conditional likelihood training in general feed-forward models. A curious result derived there is a simple decoder-specific upper bound on the conditional log-likelihood. Section B.2 discusses a relation of the generic lower bound on $I(\mathsf{x}, \mathsf{y})$ for noiseless or Gaussian channels to the standard bound on the likelihood for flat mixtures. Effectively, it illustrates a sufficient condition for the variational EM and IM algorithms to converge to identical fixed points.

B.1 Variational Information Maximization and Feed-Forward Models

Here we will prove proposition 3.3. By analogy with Section 3.3.1, we will define the feed-forward and the conditional encoder models as

$$\mathcal{M}_C \stackrel{\text{def}}{=} p(\mathsf{y}|\mathsf{x})p(\tilde{\mathsf{x}}|\mathsf{y}), \quad \mathcal{M}_{IC} \stackrel{\text{def}}{=} \tilde{p}(\mathsf{x}, \tilde{\mathsf{x}})p(\mathsf{y}|\mathsf{x}, \tilde{\mathsf{x}}), \tag{B.1}$$

where $p(\mathsf{y}|\mathsf{x}, \tilde{\mathsf{x}})$ is the exact posterior of $\mathcal{M}_C$.

Proposition B.1. *For i.i.d. patterns $\{\mathsf{x}, \tilde{\mathsf{x}}\}$, conditional likelihood learning in the feed-forward model $\mathcal{M}_C$ corresponds to maximization of a* **lower bound** *on the conditional mutual information $I(\tilde{\mathsf{x}}, \mathsf{y}|\mathsf{x})$ in $\mathcal{M}_{IC}$. Up to irrelevant constants, this bound is weaker or as tight as $\hat{I}_C(\tilde{\mathsf{x}}, \mathsf{y}|\mathsf{x}) = \langle \log p(\tilde{\mathsf{x}}|\mathsf{x}, \mathsf{y}) \rangle_{p(\mathsf{y}|\mathsf{x},\tilde{\mathsf{x}})\tilde{p}(\mathsf{x},\tilde{\mathsf{x}})}$.*

Proof. It is easy to see that for i.i.d. data, the average conditional log-likelihood (3.29) may be expressed as

$$\begin{aligned} \mathcal{L}_{\tilde{x}|x} &= \langle \log p(\tilde{\mathsf{x}}|\mathsf{x}) \rangle_{\tilde{p}(\tilde{\mathsf{x}},\mathsf{x})} \\ &= \langle \log p(\mathsf{y}|\mathsf{x}) + \log p(\tilde{\mathsf{x}}|\mathsf{x}, \mathsf{y}) - \log p(\mathsf{y}|\mathsf{x}, \tilde{\mathsf{x}}) \rangle_{\tilde{p}(\mathsf{x},\tilde{\mathsf{x}})}, \end{aligned} \tag{B.2}$$

where $\tilde{p}(\mathsf{x}, \tilde{\mathsf{x}})$ is the empirical distribution, and y is a latent variable in the conditional $p(\tilde{\mathsf{x}}, \mathsf{y}|\mathsf{x})$. By averaging both parts of (B.2) over the exact posterior $p(\mathsf{y}|\mathsf{x}, \tilde{\mathsf{x}})$, we obtain

$$\mathcal{L}_{\tilde{x}|x} = \langle \log p(\tilde{\mathsf{x}}|\mathsf{x}, \mathsf{y}) \rangle_{p(\mathsf{y}|\mathsf{x},\tilde{\mathsf{x}})\tilde{p}(\mathsf{x},\tilde{\mathsf{x}})} - \langle KL(p(\mathsf{y}|\mathsf{x}, \tilde{\mathsf{x}}) \| p(\mathsf{y}|\mathsf{x})) \rangle_{\tilde{p}(\mathsf{x},\tilde{\mathsf{x}})} . \tag{B.3}$$

We will assume that $p(\tilde{\mathsf{x}}, \mathsf{y}|\mathsf{x})$ is in the tractable family, i.e. no variational relaxations of the conditional likelihood are required to perform the computations in (B.3). Note that the Kullback-Leibler divergence in (B.3) cancels if and only if $p(\mathsf{y}|\mathsf{x}, \tilde{\mathsf{x}}) \equiv p(\mathsf{y}|\mathsf{x})$, which is generally not the case for stochastic feed-forward models.

By definition, the conditional mutual information $I(\tilde{\mathsf{x}}, \mathsf{y}|\mathsf{x})$ in the corresponding recognition model $\mathcal{M}_{IC}$ may be expressed as

$$\begin{aligned} I(\tilde{\mathsf{x}}, \mathsf{y}|\mathsf{x}) &= H_{\tilde{p}}(\tilde{\mathsf{x}}|\mathsf{x}) + \langle \log \tilde{p}(\tilde{\mathsf{x}}|\mathsf{y}, \mathsf{x}) \rangle_{\tilde{p}(\mathsf{y}|\mathsf{x},\tilde{\mathsf{x}})\tilde{p}(\mathsf{x},\tilde{\mathsf{x}})} \\ &= H_{\tilde{p}}(\tilde{\mathsf{x}}|\mathsf{x}) + \langle \log \tilde{p}(\tilde{\mathsf{x}}|\mathsf{y}, \mathsf{x}) \rangle_{p(\mathsf{y}|\mathsf{x},\tilde{\mathsf{x}})\tilde{p}(\mathsf{x},\tilde{\mathsf{x}})}, \end{aligned} \tag{B.4}$$

where $\tilde{p}(\mathsf{y}|\mathsf{x}, \tilde{\mathsf{x}}) = p(\mathsf{y}|\mathsf{x}, \tilde{\mathsf{x}})$ by construction (3.30), $H_{\tilde{p}}(\tilde{\mathsf{x}}|\mathsf{x}) \stackrel{\text{def}}{=} -\langle \log \tilde{p}(\tilde{\mathsf{x}}|\mathsf{x}) \rangle_{\tilde{p}(\mathsf{x},\tilde{\mathsf{x}})}$ is the entropy of the empirical distribution (independent of the functional parameters), and $\tilde{p}(\tilde{\mathsf{x}}|\mathsf{y}, \mathsf{x})$ is the Bayes-optimal decoder of $\mathcal{M}_{IC}$ given by

$$\tilde{p}(\tilde{\mathsf{x}}|\mathsf{y}, \mathsf{x}) \propto p(\mathsf{y}|\mathsf{x}, \tilde{\mathsf{x}})\tilde{p}(\mathsf{x}, \tilde{\mathsf{x}}). \tag{B.5}$$

Here $p(\mathsf{y}|\mathsf{x}, \tilde{\mathsf{x}})$ is the exact posterior of the feed-forward model $\mathcal{M}_C$. Again, it is easy to see that due to the mixture form of the Bayesian decoder $\tilde{p}(\tilde{\mathsf{x}}|\mathsf{y}, \mathsf{x})$, the exact evaluations of the conditional mutual information $I(\tilde{\mathsf{x}}, \mathsf{y}|\mathsf{x})$ are in general computationally intractable. However, we may re-define the generic lower bound on the mutual information (2.2) for the conditional case, and obtain a tractable bound on (B.4) as

$$I(\tilde{\mathsf{x}}, \mathsf{y}|\mathsf{x}) \geq \hat{I}_C(\tilde{\mathsf{x}}, \mathsf{y}|\mathsf{x}) \stackrel{\text{def}}{=} H_{\tilde{p}}(\tilde{\mathsf{x}}|\mathsf{x}) + \langle \log p(\tilde{\mathsf{x}}|\mathsf{x}, \mathsf{y}) \rangle_{p(\mathsf{y}|\mathsf{x},\tilde{\mathsf{x}})\tilde{p}(\mathsf{x},\tilde{\mathsf{x}})}, \tag{B.6}$$

which is saturated if and only if $p(\tilde{\mathsf{x}}|\mathsf{x}, \mathsf{y}) \equiv \tilde{p}(\tilde{\mathsf{x}}|\mathsf{x}, \mathsf{y})$. For the considered case of tractable distributions $p(\tilde{\mathsf{x}}, \mathsf{y}|\mathsf{x})$, the bound (B.6) will also be tractable. Moreover, from the non-negativity of the KL divergence we get

$$\hat{I}_C(\tilde{\mathsf{x}}, \mathsf{y}|\mathsf{x}) \geq H_{\tilde{p}}(\tilde{\mathsf{x}}|\mathsf{x}) + \langle \log p(\tilde{\mathsf{x}}|\mathsf{x}, \mathsf{y}) \rangle_{p(\mathsf{y}|\mathsf{x},\tilde{\mathsf{x}})\tilde{p}(\mathsf{x},\tilde{\mathsf{x}})} - \langle KL(p(\mathsf{y}|\mathsf{x}, \tilde{\mathsf{x}}) \| p(\mathsf{y}|\mathsf{x})) \rangle_{\tilde{p}(\mathsf{x},\tilde{\mathsf{x}})} , \tag{B.7}$$

which by transitivity leads to

$$I(\tilde{\mathsf{x}}, \mathsf{y}|\mathsf{x}) \geq \hat{I}_C(\tilde{\mathsf{x}}, \mathsf{y}|\mathsf{x}) \geq H_{\tilde{p}}(\tilde{\mathsf{x}}|\mathsf{x}) + \mathcal{L}_{\tilde{\mathsf{x}}|\mathsf{x}}. \tag{B.8}$$

□

The result shows that the conditional likelihood training in chains $\mathcal{M}_C$ may be viewed as maximization of a specific lower bound on $I(\tilde{\mathsf{x}}, \mathsf{y}|\mathsf{x})$ in the corresponding conditional encoder models $\mathcal{M}_{IC}$. Alternatively (and perhaps more intuitively), we may view the result as *an upper bound* on $\mathcal{L}_{\tilde{\mathsf{x}}|\mathsf{x}}$. From proposition 3.3 we may

immediately obtain an upper bound on the average log-probability of producing the required outputs $\tilde{\mathsf{x}}$ from the encodings y with the chosen model-specific decoder $p(\tilde{\mathsf{x}}|\mathsf{y})$:

$$\mathcal{L}_{\tilde{\mathsf{x}}|\mathsf{x}} = \frac{1}{M}\sum_{i=1}^{M}\log p(\tilde{\mathsf{x}}=\tilde{\mathsf{x}}^{(i)}|\mathsf{x}=\mathsf{x}^{(i)}) \quad \leq \quad \underbrace{\frac{1}{M}\sum_{i=1}^{M}\langle \log p(\tilde{\mathsf{x}}^{(i)}|\mathsf{y})\rangle_{p(\mathsf{y}|\mathsf{x}^{(i)},\tilde{\mathsf{x}}^{(i)})}}_{\hat{I}_C(\tilde{\mathsf{x}},\mathsf{y}|\mathsf{x})+const}, \tag{B.9}$$

where $(\mathsf{x}^{(i)},\tilde{\mathsf{x}}^{(i)}) \in \mathcal{X}_C$ is the i^{th} source-output pair of the training set $\mathcal{X}_C = \{(\mathsf{x}^{(i)},\tilde{\mathsf{x}}^{(i)})|i=1,\ldots,M\}$. For the case when $\tilde{\mathsf{x}}^{(i)} = \mathsf{x}^{(i)}$ for all M training patterns, the feed-forward model $\mathcal{M}_C$ becomes an autoencoder, and the conditional likelihood defines the probability of correct reconstructions of the training set. Note that the bound (B.9) is fundamentally different from Fano's inequality (3.1) in the sense that (B.9) is a functional of the decoder distribution $p(\tilde{\mathsf{x}}|\mathsf{y})$, while Fano's result (3.1) ignores any knowledge about the specific distributions used for source reconstructions. Moreover, whilst Fano's bound on reconstruction error is tractable only for simple channels where it is possible to compute $I(\mathsf{x},\mathsf{y}) = H(\mathsf{y}) - H(\mathsf{y}|\mathsf{x})$ exactly, our result (B.9) does not require computations of entropies of mixture distributions.

B.2 Variational Information Maximization and Flat Mixture Models

Here we compare variational information maximization (IM) and variational expectation maximization (EM) for flat latent variable models (i.e. generative models where the latent variables are uniformly distributed).

Let $\mathcal{M}_L \stackrel{\text{def}}{=} q(\mathsf{y})q(\mathsf{x}|\mathsf{y})$ define a latent variable model, where x is a data pattern, and y is its latent variable representation. (NB: we change the conventional notations for the purpose which will soon become clear). Following the standard variational procedure of maximizing the log-likelihood $\mathcal{L} \stackrel{\text{def}}{=} \langle \log q(\mathsf{x})\rangle_{\tilde{p}(\mathsf{x})}$ in the generative model $\mathcal{M}_L$, it is straight-forward to obtain the standard variational Jensen's bound on $\mathcal{L}$

$$\begin{aligned}
\mathcal{L} &= \left\langle \log \int_{\mathsf{y}} q(\mathsf{x},\mathsf{y})\mathrm{d}\mathsf{y} \right\rangle_{\tilde{p}(\mathsf{x})} \\
&= \left\langle \log \int_{\mathsf{y}} q(\mathsf{x},\mathsf{y})\frac{p(\mathsf{y}|\mathsf{x})}{p(\mathsf{y}|\mathsf{x})}\mathrm{d}\mathsf{y} \right\rangle_{\tilde{p}(\mathsf{x})} \\
&\geq \langle \log q(\mathsf{x}|\mathsf{y})\rangle_{p(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x})} - \langle KL(p(\mathsf{y}|\mathsf{x})\|q(\mathsf{y}))\rangle_{\tilde{p}(\mathsf{x})} \stackrel{\text{def}}{=} \tilde{\mathcal{L}},
\end{aligned} \tag{B.10}$$

where we have applied Jensen's inequality (e.g. Jensen (1906), Hardy et al. (1988)). Here $\tilde{p}(\mathsf{x})$ is the empirical distribution, and $p(\mathsf{y}|\mathsf{x})$ is a variational distribution approximating the true posterior $q(\mathsf{y}|\mathsf{x}) \propto q(\mathsf{y})q(\mathsf{x}|\mathsf{y})$ expressed from the generative model $\mathcal{M}_L$. As usual, in order to simplify computations of (B.10), the variational posterior $p(\mathsf{y}|\mathsf{x})$ needs to be constrained to ensure tractability of

computing the averages. The standard variational extension of the expectation-maximizing algorithm (see e.g. Neal and Hinton (1998)) trains the generative model $\mathcal{M}_L$ by iteratively optimizing (B.10) with respect to the model parameters $q(\mathsf{y})$, $q(\mathsf{x}|\mathsf{y})$, and the variational posterior $p(\mathsf{y}|\mathsf{x})$.

Note that we can equivalently express the bound (B.10) as

$$\tilde{\mathcal{L}} \;=\; \langle \log q(\mathsf{x}|\mathsf{y}) \rangle_{p(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x})} + \langle \log q(\mathsf{y}) \rangle_{p(\mathsf{y}|\mathsf{x})\tilde{p}(\mathsf{x})} + \langle H(p(\mathsf{y}|\mathsf{x})) \rangle_{\tilde{p}(\mathsf{x})}, \quad \text{(B.11)}$$

where $H(p(\mathsf{y}|\mathsf{x})) \stackrel{\text{def}}{=} -\langle \log p(\mathsf{y}|\mathsf{x}) \rangle_{p(\mathsf{y}|\mathsf{x})}$. It is now straight-forward to see that for the case of uniform code distributions ($q(\mathsf{y}) \sim \mathcal{U}_\mathsf{y}$) with deterministic variational posteriors ($p(\mathsf{y}|\mathsf{x}) \sim \delta$), maximization of the bound (B.11) reduces to optimizing

$$\tilde{\mathcal{L}} = \langle \log q(\mathsf{x}|\mathsf{y}) \rangle_{\tilde{p}(\mathsf{x})p(\mathsf{y}|\mathsf{x})} + const. \quad \text{(B.12)}$$

Up to irrelevant constants, the bound on the likelihood for this case is in fact the generic lower bound (2.2) on the mutual information for the noiseless channel model $\mathcal{M}_I(\mathsf{x},\mathsf{y}) \stackrel{\text{def}}{=} \tilde{p}(\mathsf{x})p(\mathsf{y}|\mathsf{x})$, which uses the variational posterior of $\tilde{\mathcal{L}}$ as the *encoder* distribution, i.e.

$$\tilde{I}(\mathsf{x},\mathsf{y}) = H_{\tilde{p}}(\mathsf{x}) + \langle \log q(\mathsf{x}|\mathsf{y}) \rangle_{\tilde{p}(\mathsf{x})p(\mathsf{y}|\mathsf{x})}, \quad \text{where } p(\mathsf{y}|\mathsf{x}) \sim \delta. \quad \text{(B.13)}$$

Here the variational decoder $q(\mathsf{x}|\mathsf{y})$ of the bound $\tilde{I}(\mathsf{x},\mathsf{y})$ is the conditional of the latent variable model $\mathcal{M}_L$. From (B.12) and (B.13), we can see that *for i.i.d. patterns* $\{\mathsf{x}\}$*, maximization of the standard variational lower bound on the likelihood in a* **flat** *latent variable model with a* **deterministic** *variational posterior* $p(\mathsf{y}|\mathsf{x})$ *reduces to maximizing the generic lower bound on the mutual information in the corresponding* **noiseless channel**.

Note that the same result holds if $p(\mathsf{y}|\mathsf{x})$ is a Gaussian with a fixed covariance. Indeed, if $p(\mathsf{y}|\mathsf{x}) \sim \mathcal{N}_\mathsf{y}(\mathsf{f}(\mathsf{x}), \boldsymbol{\Sigma})$, the entropic term in (B.11) is a function of the fixed noise covariance $\boldsymbol{\Sigma}$, i.e. up to irrelevant constants the bound $\tilde{\mathcal{L}}$ defined by (B.11) is identical to the generic lower bound on the mutual information (B.13) for the corresponding Gaussian channel.

The result may be trivially generalized to flat mixtures of latent variable models $\mathcal{M}_{LH}(\mathsf{x},\mathsf{y},z) \stackrel{\text{def}}{=} q(\mathsf{y})q(z)q(\mathsf{x}|\mathsf{y},z)$ with deterministic variational posteriors $p(\mathsf{x}|\mathsf{y},z) \sim \delta$ and flat priors on the latent variables $q(\mathsf{y}) \sim \mathcal{U}_\mathsf{y}$ and mixture coefficients $q(z) = 1/|z|$ (here we assumed that $\mathsf{x} \in \mathbb{R}^{|\mathsf{x}|}$, $\mathsf{y} \in \mathbb{R}^{|\mathsf{y}|}$, and $z \in \{1,\ldots,|z|\}$). By analogy with (B.11)– (B.13), it is easy to see that optimization of the standard variational lower bound on $\mathcal{L}$ for this case reduces to maximizing the generic lower bound on $I(\mathsf{x},\{\mathsf{y},z\})$ in the corresponding **noiseless hybrid channel** $\mathsf{x} \to \{\mathsf{y},z\}$

$$\tilde{I}_H(\mathsf{x},\{\mathsf{y},z\}) = H(\mathsf{x}) + \langle \log q(\mathsf{x}|\mathsf{y},z) \rangle_{\tilde{p}(\mathsf{x})p(\mathsf{y}|\mathsf{x})p(z|\mathsf{x})}, \quad \text{(B.14)}$$

where $p(\mathsf{y}|\mathsf{x}) \sim \delta$, $p(z|\mathsf{x}) \sim \delta$, and $q(\mathsf{x}|\mathsf{y},z)$ is the conditional of the generative mixture of latent variable models $\mathcal{M}_{LH}(\mathsf{x},\mathsf{y},z)$. Similar results hold if $p(z|\mathsf{x}) \sim \delta$ and $p(\mathsf{y}|\mathsf{x}) \sim \mathcal{N}_\mathsf{x}(\mathsf{f}(\mathsf{x}), \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is fixed. In the encoder model formulation, the fixed covariance corresponds to the irreducible independent channel noise.

To summarize, we may note that *while generally the optimization surfaces defined by the standard variational lower bounds on the likelihood and the generic*

variational lower bounds on the mutual information are different, the obtained solutions coincide for flat mixture models $\mathcal{M}_L$ with deterministic variational posteriors, and the corresponding noiseless encoder models $\mathcal{M}_I$ (where the variational decoding distribution corresponds to the conditional of the generative model).

Appendix C

Variational Information Maximization for Isotropic Gaussian Channels

Here we analyze properties of the variational IM algorithm for isotropic Gaussian channels $\mathsf{x} \to \mathsf{y}$. Specifically, we focus on the theoretical analysis of the bounds on $I(\mathsf{x}, \mathsf{y})$ for *nonlinear* Gaussian encoders $p(\mathsf{y}|\mathsf{x}) \sim \mathcal{N}_{\mathsf{y}}(\boldsymbol{\phi}(\mathsf{x}), \sigma^2 \mathsf{I})$. As usual, we will presume intractability of explicit computations in $\mathbb{R}^{|\boldsymbol{\phi}|}$, and derive a kernelized extension of the generic lower bound

$$I(\mathsf{x}, \mathsf{y}) \geq \tilde{I}(\mathsf{x}, \mathsf{y}) \stackrel{\text{def}}{=} H(\mathsf{x}) + \langle \log q(\mathsf{x}|\mathsf{y}) \rangle_{p(\mathsf{x},\mathsf{y})}.$$

As the variational information-maximizing framework offers flexibility in choosing optimal variational decoders $q(\mathsf{x}|\mathsf{y})$, we consider optimizing $\tilde{I}(\mathsf{x}, \mathsf{y})$ for several of such choices. As the first obvious choice, we consider using linear Gaussian decoders, which simplifies computations of the bound and facilitates the analysis of optimal solutions for the encoder parameters. Our motivation here is to check whether there is any gain in using nonlinearities of the codes if the variational decoder distribution remains linear. As expected, we show that linear Gaussian variational decoders may indeed be too restrictive. Specifically, we show that in the case of the isotropic noise of Gaussian channels, nothing is gained by using nonlinear encoders and linear decoders in the context of variational information maximization. This agrees with the more specific result for noiseless one-layer autoencoders with linear output units (Bourlard and Kamp (1988), Bourlard (2000)), but is derived for *stochastic* channels with arguably less specific choices of nonlinearities.

Then we consider variational information maximization for nonlinear variational decoders. Since the choice of nonlinearity may significantly influence the complexity of computing variational lower bounds $\tilde{I}(\mathsf{x}, \mathsf{y})$, relaxations of the generic variational procedure may need to be considered. We show that the generic lower bound $\tilde{I}(\mathsf{x}, \mathsf{y})$ may indeed be formally modified to ensure tractable computations. This naturally relaxes the bound on $I(\mathsf{x}, \mathsf{y})$ to provide an objective function for Kernel PCA. By analogy with a simpler case of discrete nonlinear channels (see Section 5.2.3), the information-theoretic formulation suggests a

proper way for learning kernel parameters of KPCA models; however, the choice of constraints on the kernel matrices will prove to be crucial for avoiding degenerate (asymptotically noiseless) solutions.

Throughout the discussion in this chapter, we will make several references to the optimal bound on $I(\mathsf{x},\mathsf{y})$ for isotropic linear Gaussian decoders $q(\mathsf{x}|\mathsf{y}) \sim \mathcal{N}_{\mathsf{x}}(\mathsf{U}\mathsf{y},\sigma^2\mathsf{I})$. This is a more constrained variational decoder than the correlated linear Gaussian (which gives rise to Linsker's *as-if Gaussian* approximation, see Section 1.4, 4.1.1). However, the solutions obtained for this case will prove to be important for considering nonlinear generalizations of the bound $\tilde{I}(\mathsf{x},\mathsf{y})$, so we will discuss them first. Then in Appendix C.2 we will move to a discussion of nonlinear Gaussian channels.

C.1 Linear Gaussian Channels: Linear Decoders

$p(\mathsf{y}|\mathsf{x}) \sim \mathcal{N}_y(\mathsf{W}\mathsf{x}, s^2\mathsf{I})$, $q(\mathsf{x}|\mathsf{y}) \sim \mathcal{N}_x(\mathsf{U}\mathsf{y}, \sigma^2\mathsf{I})$

Let the encoder and decoder be given by $p(\mathsf{y}|\mathsf{x}) \sim \mathcal{N}_y(\mathsf{W}\mathsf{x}, s^2\mathsf{I})$ and $q(\mathsf{x}|\mathsf{y}) \sim \mathcal{N}_x(\mathsf{U}\mathsf{y}, \sigma^2\mathsf{I})$ respectively. Our goal is to learn optimal settings of $\mathsf{W} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{x}|}$ and $\mathsf{U} \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{y}|}$ (for fixed decoder and encoder variances σ^2 and s^2) by maximizing the simple variational bound (4.1) on the mutual information, which in this case is expressed as

$$\tilde{I}(\mathsf{x},\mathsf{y}) = \frac{1}{\sigma^2}\mathrm{tr}\left\{\mathsf{U}\mathsf{W}\mathsf{S}\right\} - \frac{1}{2\sigma^2}\mathrm{tr}\left\{\mathsf{U}\mathbf{\Sigma}\mathsf{U}^T\right\} - \frac{1}{2\sigma^2}\mathrm{tr}\left\{\mathsf{S}\right\} + c. \tag{C.1}$$

Here $c = -|\mathsf{x}|/2\log(2\pi\sigma^2)$ is a constant which does not affect the optimization surface for the encoder and decoder weights,

$$\mathsf{S} = \langle \mathsf{x}\mathsf{x}^T \rangle = \frac{1}{M}\sum_m \mathsf{x}^{(m)}\left(\mathsf{x}^{(m)}\right)^T \tag{C.2}$$

is the sample covariance of the centered data, and

$$\mathbf{\Sigma} = \mathsf{I}s^2 + \mathsf{W}\mathsf{S}\mathsf{W}^T \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{y}|} \tag{C.3}$$

is the covariance of the marginal distribution of responses $p(\mathsf{y})$. In the following we assume that the weights W and U are non-singular, which in our context means that the dimensionality of the codewords $|\mathsf{y}|$ is sufficiently low. Note that we make no parametric assumptions about the distribution of the sources $p(\mathsf{x})$ or non-singularity of the sample covariance S.

Unsurprisingly, objective (C.1) is closely related to the least squares reconstruction error in a linear autoencoder. What is interesting in this context that it provides a proper bound on the mutual information, independently of the source distribution.

C.1.1 Nature of Optimal Solutions

Unconstrained optimization of (C.1) for the encoder's weights W leads to the extremum condition

$$\mathsf{U}^T\mathsf{S} = \mathsf{U}^T\mathsf{U}\mathsf{W}\mathsf{S}. \tag{C.4}$$

By assuming that y is a compressed representation of the source x (i.e. $|\mathsf{x}| > |\mathsf{y}|$), we obtain $\mathsf{W}\mathsf{S} = (\mathsf{U}^T\mathsf{U})^{-1}\mathsf{U}^T\mathsf{S} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{x}|}$. (It is safe to assume non-singularity; indeed, if $\mathsf{U}^T\mathsf{U}$ is singular, we may ensure non-singularity by reducing dimensionality of the codewords). Then it is clear that

$$\begin{aligned} \mathrm{tr}\left\{\mathsf{U}\mathsf{W}\mathsf{S}\mathsf{W}^T\mathsf{U}^T\right\} &= \mathrm{tr}\left\{\mathsf{U}\mathsf{W}\mathsf{S}\mathsf{W}^T\mathsf{U}^T\right\} = \mathrm{tr}\left\{(\mathsf{U}^T\mathsf{U}\mathsf{W}\mathsf{S})\mathsf{W}^T\right\} \\ &= \mathrm{tr}\left\{\mathsf{U}\mathsf{W}\mathsf{S}\right\} = \mathrm{tr}\left\{\mathsf{U}(\mathsf{U}^T\mathsf{U})^{-1}\mathsf{U}^T\mathsf{S}\right\}, \end{aligned} \tag{C.5}$$

leading to

$$\mathrm{tr}\left\{\mathsf{U}\mathbf{\Sigma}\mathsf{U}^T\right\} = \mathrm{tr}\left\{\mathsf{U}(\mathsf{U}^T\mathsf{U})^{-1}\mathsf{U}^T\mathsf{S}\right\} + s^2\mathrm{tr}\left\{\mathsf{U}\mathsf{U}^T\right\}. \tag{C.6}$$

By substituting expression (C.6) into (C.1), we may express the bound as a function of the decoder alone as

$$\tilde{I}(\mathsf{x},\mathsf{y}) = \frac{1}{2\sigma^2}\mathrm{tr}\left\{\mathsf{U}(\mathsf{U}^T\mathsf{U})^{-1}\mathsf{U}^T\mathsf{S}\right\} - \frac{s^2}{2\sigma^2}\mathrm{tr}\left\{\mathsf{U}\mathsf{U}^T\right\} - \frac{1}{2\sigma^2}\mathrm{tr}\left\{\mathsf{S}\right\}, \tag{C.7}$$

where we have ignored the irrelevant constant c.

Let $\mathsf{U} = \mathsf{V}\mathsf{L}\mathsf{R}^T$ be the singular value decomposition of the decoder weights U (see e.g. Golub and Loan (1996)). By definition, $\mathsf{L} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{y}|}$ is diagonal and invertible, and $\mathsf{V} \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{y}|}$, $\mathsf{R} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{y}|}$ are orthogonal, so that $\mathsf{V}^T\mathsf{V} = \mathsf{R}^T\mathsf{R} = \mathsf{R}\mathsf{R}^T = \mathsf{I}_{|\mathsf{y}|}$. From (C.4) it is clear that $\mathsf{W} = \mathsf{R}\mathsf{L}^{-1}\mathsf{V}^T$, i.e. $\mathsf{W}\mathsf{U} = \mathsf{I}_{|\mathsf{y}|}$. Substitution into (C.7) results in

$$\tilde{I}(\mathsf{x},\mathsf{y}) \propto \mathrm{tr}\left\{\mathsf{V}^T\mathsf{S}\mathsf{V}\right\} - s^2\mathrm{tr}\left\{\mathsf{L}^2\right\} - \mathrm{tr}\left\{\mathsf{S}\right\}, \tag{C.8}$$

which is a function of the decoder's singular vectors and scalings. Optimizing (C.8) for V under the orthonormality constraints on V, we readily obtain the PCA solution (and its rigid rotations). It is clear that in the case of a noiseless channel ($s^2 \to 0$), the solutions are invariant with respect to the scalings L.

For noisy channels (i.e. for $s^2 > 0$), maximization of (C.8) with respect to the scalings of the *decoder* weights results in $\mathsf{L} \to 0$, which leads to the divergence of the Frobenius norm $\|\mathsf{W}\|_F$. Effectively, this corresponds to an approximately noiseless encoder, since distinct source patterns will be mapped to codes which are infinitely spread out in the code space, so that the contribution of the finite channel noise to the bound (C.1) becomes negligible. In order to find the optimal bases of the subspaces spanned by the encoder and decoder weights W and U, we may constrain their singular values. It is straight-forward to see that in this case the right singular vectors of the optimal encoder weights W span the principal subspace of S (see e.g. Bishop (1995)). It is also clear that the obtained solutions are invariant under complimentary rotations of W and U.

This demonstrates the simple result that for the considered variational problem, the optimal lower bound on $I(\mathsf{x},\mathsf{y})$ is provided by PCA. Furthermore, by

comparing this result with Linsker's *as-if* Gaussian bound (see Section 4.1.1), we see that for the considered channel nothing is gained by learning full covariances of linear Gaussian variational decoders. As before, this conclusion is reached without the need for a Gaussian assumption about the source distribution. It is intuitively clear that in order to go beyond the PCA solutions, more complex encoders/decoders are required. In the following section we consider the effects of increasing the complexity of the encoder, while still using a simple linear Gaussian variational decoder defined above.

C.2 Nonlinear Gaussian Channels: Linear Decoders

$p(\mathsf{y}|\mathsf{x}) \sim \mathcal{N}_y(\mathsf{W}\boldsymbol{\phi}(\mathsf{x}), \boldsymbol{\Sigma}_y)$, $q(\mathsf{x}|\mathsf{y}) \sim \mathcal{N}_x(\mathsf{U}\mathsf{y}, \boldsymbol{\Sigma}_x)$

Here we consider the case of a nonlinear Gaussian channel with the encoder $p(\mathsf{y}|\mathsf{x}) \sim \mathcal{N}_y(\mathsf{W}\boldsymbol{\phi}(\mathsf{x}), \boldsymbol{\Sigma}_y)$ and decoder $q(\mathsf{x}|\mathsf{y}) \sim \mathcal{N}_x(\mathsf{U}\mathsf{y}, \boldsymbol{\Sigma}_x)$. Note that for all the data points $\{\mathsf{x}^{(m)}|m = 1, \ldots, M\}$, the set of encodings $\{\mathsf{y}^{(m)}\}$ is given by a noisy linear projection from the (potentially high-dimensional) feature space $\{\boldsymbol{\phi}(\mathsf{x}^{(m)})\}$. In what follows we assume that $|\boldsymbol{\phi}| > M$, i.e. dimensionality of the feature space exceeds the number of training patterns.

It is easy to see that for the considered case the lower bound on the mutual information $I(\mathsf{x}, \mathsf{y})$ is given by

$$\begin{aligned}\tilde{I}(\mathsf{x}, \mathsf{y}) &= -\frac{1}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}_x^{-1}\mathsf{S}\right\} + \mathrm{tr}\left\{\boldsymbol{\Sigma}_x^{-1}\mathsf{U}\mathsf{W}\langle\boldsymbol{\phi}(\mathsf{x})\mathsf{x}^T\rangle\right\} \\ &\quad -\frac{1}{2}\mathrm{tr}\left\{\mathsf{U}^T\boldsymbol{\Sigma}_x^{-1}\mathsf{U}\left(\boldsymbol{\Sigma}_y + \mathsf{W}\langle\boldsymbol{\phi}(\mathsf{x})\boldsymbol{\phi}(\mathsf{x})^T\rangle\mathsf{W}^T\right)\right\}. \qquad \text{(C.9)}\end{aligned}$$

As before, we assume that $\mathsf{S} = \langle\mathsf{x}\mathsf{x}^T\rangle = \sum_m \mathsf{x}^{(m)}(\mathsf{x}^{(m)})^T/M$ is the sample covariance of the zero-mean data, and the averages are computed with respect to the empirical distribution $\tilde{p}(\mathsf{x}) = (1/M)\sum_{m=1}^{M}\delta(\mathsf{x} - \mathsf{x}^{(m)})$. Also, by analogy with Section 5.2.3 we assume that for high-dimensional feature spaces direct evaluation of the averages is implausible. It is therefore desirable to avoid explicit computations in $\{\boldsymbol{\phi}\}$.

C.2.1 Kernelized Representation

Since each row $\tilde{\mathsf{w}}_i^T \in \mathbb{R}^{1\times|\boldsymbol{\phi}|}$ of the weight matrix $\mathsf{W} \in \mathbb{R}^{|\mathsf{y}|\times|\boldsymbol{\phi}|}$ has the same dimensionality as the feature vectors $\boldsymbol{\phi}(\mathsf{x}^{(i)})^T$, it is representable as

$$\tilde{\mathsf{w}}_i = \sum_{m=1}^{M}\alpha_{im}\boldsymbol{\phi}(\mathsf{x}^{(m)}) + \tilde{\mathsf{w}}_i^{\perp}, \qquad \text{(C.10)}$$

where $\tilde{\mathsf{w}}_i^{\perp}$ is orthogonal to the span of $\phi(\mathsf{x}^{(1)}), \ldots, \phi(\mathsf{x}^{(M)})$. Then

$$\mathsf{W} = \mathsf{A}\mathsf{F}^T + \mathsf{W}^{\perp}, \quad \mathsf{F} \stackrel{\text{def}}{=} \left[\boldsymbol{\phi}(\mathsf{x}^{(1)}),\ \ldots,\ \boldsymbol{\phi}(\mathsf{x}^{(M)})\right] \in \mathbb{R}^{|\boldsymbol{\phi}|\times M}, \qquad \text{(C.11)}$$

where $\mathsf{A} = \{\alpha_{ij}\} \in \mathbb{R}^{|\mathsf{y}|\times M}$ is the matrix of coefficients, and the transposed rows of $\mathsf{W}^\perp$ are given by $\tilde{\mathsf{w}}_i^\perp$. In kernel literature F is often referred to as the *design* matrix (e.g. Williams (1998), MacKay (1997), Cristianini and Shawe-Taylor (2000)).

From (C.11), we obtain expressions for the averages

$$\begin{aligned}\mathsf{W}\langle\boldsymbol{\phi}(\mathsf{x})\mathsf{x}^T\rangle &= \mathsf{A}\left\langle\left[\boldsymbol{\phi}(\mathsf{x}^{(1)})^T\boldsymbol{\phi}(\mathsf{x}),\ldots,\boldsymbol{\phi}(\mathsf{x}^{(M)})^T\boldsymbol{\phi}(\mathsf{x})\right]^T\mathsf{x}^T\right\rangle \\ &= \frac{\mathsf{A}}{M}\left[\sum_m \mathsf{x}^{(m)}\boldsymbol{\phi}(\mathsf{x}^{(m)})^T\boldsymbol{\phi}(\mathsf{x}^{(1)}),\ \ldots\right]^T = \frac{\mathsf{A}\mathsf{B}^T}{M} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{x}|}, \end{aligned} \tag{C.12}$$

where we defined

$$\mathsf{B} \stackrel{\text{def}}{=} \sum_{m=1}^{M} \mathsf{x}^{(m)}\mathsf{k}(\mathsf{x}^{(m)})^T \in \mathbb{R}^{|\mathsf{x}|\times M},\ \ \mathsf{k}(\mathsf{x}^{(m)}) \stackrel{\text{def}}{=} \mathsf{F}^T\boldsymbol{\phi}(\mathsf{x}^{(m)}) \in \mathbb{R}^M \tag{C.13}$$

and used the fact that from the orthogonality assumption (C.11) we get $\mathsf{W}^\perp\mathsf{F} = \mathbf{0} \in \mathbb{R}^{|\mathsf{x}|\times M}$. Here $\mathsf{k}(\mathsf{x}^{(m)})$ corresponds to the m^{th} column of the Gram matrix $\mathsf{K} \stackrel{\text{def}}{=} \{K_{ij}\} \stackrel{\text{def}}{=} \{\boldsymbol{\phi}(\mathsf{x}^{(i)})^T\boldsymbol{\phi}(\mathsf{x}^{(j)})\} \in \mathbb{R}^{M\times M}$. Clearly, for a fixed $\mathsf{K}\mathbb{R}^{M\times M}$, the computed expectation $\mathsf{W}\langle\boldsymbol{\phi}(\mathsf{x})\mathsf{x}^T\rangle$ is a function of the coefficients $\mathsf{A} \in \mathbb{R}^{|\mathsf{y}|\times|M|}$, which does not require explicit computations in the high-dimensional feature space.

Analogously, we can express

$$\begin{aligned}\mathsf{W}\langle\boldsymbol{\phi}(\mathsf{x})\boldsymbol{\phi}(\mathsf{x})^T\rangle\mathsf{W}^T &= \left(\mathsf{A}\mathsf{F}^T + \mathsf{W}^\perp\right)\left\langle\boldsymbol{\phi}(\mathsf{x})\boldsymbol{\phi}(\mathsf{x})^T\right\rangle\left(\mathsf{A}\mathsf{F}^T + \mathsf{W}^\perp\right)^T \\ &= \frac{1}{M}\mathsf{A}\sum_{m=1}^{M}\mathsf{k}(\mathsf{x}^{(m)})\mathsf{k}(\mathsf{x}^{(m)})^T\mathsf{A}^T, \end{aligned} \tag{C.14}$$

where we have used the fact that $\mathsf{W}^\perp\mathsf{F} = \mathbf{0}$, as follows from construction (C.11). Again, for the fixed Gram matrix the term is a quadratic function of coefficients A alone.

By substitution, we may re-express the bound (C.9) as

$$\begin{aligned}\tilde{I}(\mathsf{x},\mathsf{y}) &= \frac{1}{M}\mathrm{tr}\left\{\boldsymbol{\Sigma}_x^{-1}\mathsf{U}\mathsf{A}\mathsf{B}^T\right\} - \frac{1}{2}\mathrm{tr}\left\{\mathsf{U}^T\boldsymbol{\Sigma}_x^{-1}\mathsf{U}\boldsymbol{\Sigma}_y\right\} \\ &\quad -\frac{1}{2M}\mathrm{tr}\left\{\mathsf{U}^T\boldsymbol{\Sigma}_x^{-1}\mathsf{U}\mathsf{A}\mathsf{K}^2\mathsf{A}^T\right\} - \frac{1}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}_x^{-1}\mathsf{S}\right\} + c, \end{aligned} \tag{C.15}$$

where $c = -|\mathsf{x}|/2\log(2\pi\sigma^2)$ does not affect the optimization surface and be ignored in the rest of the discussion. In the simplest case when $\boldsymbol{\phi}(\mathsf{x}) \equiv \mathsf{x} \in \mathbb{R}^{|\mathsf{x}|}$, we obtain $\mathsf{K}^2 \propto \mathsf{X}^T\mathsf{S}\mathsf{X} \in \mathbb{R}^{M\times M}$, $\mathsf{B} \propto \mathsf{S}\mathsf{X} \in \mathbb{R}^{|\mathsf{x}|\times M}$ where $\mathsf{X} \stackrel{\text{def}}{=} [\mathsf{x}^{(1)},\ \ldots,\ \mathsf{x}^{(M)}] \in \mathbb{R}^{|\mathsf{x}|\times M}$ contains the training data. As expected, this transforms the bound (C.15) to the corresponding expression (C.7) for the linear Gaussian channel, thus resulting in PCA on the sample covariance S for both the encoder and decoder weights U, W^T as the optimal choice.

The objective (C.15) may be used to learn optimal decoder weights $\mathsf{U} \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{y}|}$ and optimal coordinates $\mathsf{A} \in \mathbb{R}^{|\mathsf{y}|\times|M|}$ in the space spanned by the feature vectors $\{\boldsymbol{\phi}(\mathsf{x}^{(i)})|i \in [1,M]\cap\mathbb{N}\}$. Moreover, we may use the so-called kernel trick (e.g.

Vapnik (1998), Cristianini and Shawe-Taylor (2000), Scholkopf and Smola (2002)) and compute the entries of the Gram matrix $\mathsf{K} = \{K_{ij}\} \in \mathbb{R}^{M\times M}$ as $K_{ij} = \boldsymbol{\phi}(\mathsf{x}^{(i)})^T\boldsymbol{\phi}(\mathsf{x}^{(j)}) = \mathcal{K}(\mathsf{x}^{(i)}, \mathsf{x}^{(j)}; \boldsymbol{\Theta})$, where $\mathcal{K}_{\boldsymbol{\Theta}} : |\mathsf{x}| \times |\mathsf{x}| \to \mathbb{R}$ defines a symmetric positive-definite *kernel* function. In principle, we may use the objective $\tilde{I}(\mathsf{x}, \mathsf{y})$ to learn the optimal parameters[1] of the kernel function. However, as we showed in Agakov and Barber (2004c) and will discuss later, such learning may be strongly influenced by the choice of constraints on channel parameters.

C.2.2 Nature of optimal solutions

In the following we assume for simplicity that $\boldsymbol{\Sigma}_y = s^2\mathsf{I}$ and $\boldsymbol{\Sigma}_x = \sigma^2\mathsf{I}$. We also assume that $|\mathsf{y}| \leq |\mathsf{x}| \leq |\boldsymbol{\phi}|$ and $|\mathsf{x}| \leq M$, so that y is a compressed representation of $\boldsymbol{\phi}(\mathsf{x})$, and the number of training points is sufficient to ensure invertibility of the sample covariance.

C.2.2.1 Optimal Decoder

Similarly to Appendix C.2.2, we will express the bound on $I(\mathsf{x}, \mathsf{y})$ as a function of decoder parameters alone and analyze the optimal solutions. In order to do this, we will first find an analytical expression for optimal encoder coefficients, re-express the bound, and optimize it for the variational decoder.

Optimization of $\tilde{I}(\mathsf{x}, \mathsf{y})$ for the matrix of coefficients $\mathsf{A} \in \mathbb{R}^{|\mathsf{y}|\times|M|}$ leads to the fixed point condition

$$\partial\tilde{I}(\mathsf{x}, \mathsf{y})/\partial\mathsf{A} = 0 \quad \Rightarrow \mathsf{U}^T\boldsymbol{\Sigma}_x^{-1}\mathsf{B} = \mathsf{U}^T\boldsymbol{\Sigma}_x^{-1}\mathsf{U}\mathsf{A}\mathsf{K}^2 \in \mathbb{R}^{|\mathsf{y}|\times|M|}, \tag{C.16}$$

which for the considered isotropic case $\boldsymbol{\Sigma}_x = \sigma^2\mathsf{I}$ leads to

$$\mathsf{A}\mathsf{K}^2 = (\mathsf{U}^T\mathsf{U})^{-1}\mathsf{U}^T\mathsf{B} \in \mathbb{R}^{|\mathsf{y}|\times|M|}. \tag{C.17}$$

Assuming that the Gram matrix K is non-singular, we obtain $\mathsf{A}\mathsf{K} = (\mathsf{U}^T\mathsf{U})^{-1}\mathsf{U}^T\mathsf{B}\mathsf{K}^{-1}$. Then a substitution into (C.15) leads to

$$\begin{aligned}
\tilde{I}(\mathsf{x}, \mathsf{y}) &= \frac{1}{2M\sigma^2}\mathrm{tr}\left\{\mathsf{U}\mathsf{A}\mathsf{K}^2\mathsf{A}^T\mathsf{U}^T\right\} - \frac{s^2}{2\sigma^2}\mathrm{tr}\left\{\mathsf{U}\mathsf{U}^T\right\} - \frac{1}{2\sigma^2}\mathrm{tr}\left\{\mathsf{S}\right\} \\
&= \frac{1}{2M\sigma^2}\mathrm{tr}\left\{\mathsf{U}(\mathsf{U}^T\mathsf{U})^{-1}\mathsf{U}^T\mathsf{B}\mathsf{K}^{-2}\mathsf{B}^T\mathsf{U}(\mathsf{U}^T\mathsf{U})^{-1}\mathsf{U}^T\right\} - \frac{s^2}{2\sigma^2}\mathrm{tr}\left\{\mathsf{U}\mathsf{U}^T\right\} - \frac{1}{2\sigma^2}\mathrm{tr}\left\{\mathsf{S}\right\} \\
&= \frac{1}{2M\sigma^2}\mathrm{tr}\left\{\mathsf{U}(\mathsf{U}^T\mathsf{U})^{-1}\mathsf{U}^T\mathsf{B}\mathsf{K}^{-2}\mathsf{B}^T\right\} - \frac{s^2}{2\sigma^2}\mathrm{tr}\left\{\mathsf{U}\mathsf{U}^T\right\} - \frac{1}{2\sigma^2}\mathrm{tr}\left\{\mathsf{S}\right\}, \qquad \text{(C.18)}
\end{aligned}$$

where we ignored the terms independent of the decoder parameters $\mathsf{U} \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{y}|}$. Finally, by noticing that $\mathsf{B} = \mathsf{X}\mathsf{K} \in \mathbb{R}^{|\mathsf{x}|\times M}$ and

$$\mathsf{B}\mathsf{K}^{-2}\mathsf{B}^T = \mathsf{X}\mathsf{X}^T = M\mathsf{S}, \tag{C.19}$$

we may transform the bound (C.18) to

$$\tilde{I}(\mathsf{x}, \mathsf{y}) = \frac{1}{2\sigma^2}\mathrm{tr}\left\{\mathsf{U}(\mathsf{U}^T\mathsf{U})^{-1}\mathsf{U}^T\mathsf{X}\mathsf{X}^T\right\} - \frac{s^2}{2\sigma^2}\mathrm{tr}\left\{\mathsf{U}\mathsf{U}^T\right\} - \frac{1}{2\sigma^2}\mathrm{tr}\left\{\mathsf{S}\right\}, \tag{C.20}$$

[1] It is also possible to learn K directly by constraining it to satisfy properties of inner products; this may be useful, for example, when the source alphabet is exhausted by M training patterns.

which is exactly the objective of the linear Gaussian channel (C.7).

Note that optimally $\|\mathsf{U}\|_F \to 0$, which (for a fixed $\mathcal{K}_{\boldsymbol{\Theta}}$) leads to the diverging encoder's coefficients $\mathsf{A} \in \mathbb{R}^{|\mathsf{y}|\times|M|}$, resulting in the divergent weights $\|\mathsf{W}\|_F \to \infty$. As before, this may be easily explained from the information-theoretic perspective, as the resulting channel is approximately noiseless, and therefore it is characterized by a low value of the the conditional entropy $H(\mathsf{y}|\mathsf{x})$ and a high value of the mutual information. Similarly to the case of a linear channel, the convergence of $\|\mathsf{W}\|_F$ for $s^2 > 0$ may be insured by imposing norm constraints on $\mathsf{U} \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{y}|}$. Then, by the exact analogy with (C.7), we can see that under the specific assumption of isotropic Gaussian noise and a non-singular kernel matrix, optimal weights $\mathsf{U} \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{y}|}$ of the linear Gaussian decoder correspond to principal components of the sample covariance S (and their rotations). Fundamentally, we may note that the considered linear Gaussian variational decoder restricts the power of the approach, as for the optimal settings of encoder and decoder parameters we cannot improve on the PCA bounds (C.7).

C.2.2.2 Optimal Encoder

For completeness, we may express the optimal encoder weights. From (C.16) it is clear that optimal solutions for the encoder are given by

$$\mathsf{WF} = \mathsf{AK} \propto \mathsf{U}^{+}\mathsf{X} \in \mathbb{R}^{|\mathsf{y}|\times M}, \tag{C.21}$$

where $\mathsf{U}^{+} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{x}|}$ denotes the pseudo-inverse, and left singular values of $\mathsf{U} \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{y}|}$ correspond to principal eigenvectors of $\mathsf{S} \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{x}|}$. In the case when $\boldsymbol{\phi}(\mathsf{x}) \equiv \mathsf{x} \in \mathbb{R}^{|\mathsf{x}|}$ and the sample covariance $\mathsf{X}\mathsf{X}^T/M$ is non-singular, condition (C.21) results in $\mathsf{W} = \mathsf{U}^{+} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{x}|}$, which is the PCA solution of the linear Gaussian channel. However, for general nonlinear mappings, optimal encoder weights $\mathsf{W}^T \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{y}|}$ do not necessarily give rise to the nonlinear PCA solution.

Finally, note that if the channel noise is isotropic and there are no constraints preventing the weights from taking optimal solutions according to (C.17) and (C.21), then the bound is given by the summation of $|\mathsf{y}|$ principal eigenvalues of the sample covariance S. It is also important to note that for both of the considered channels (with linear and kernelized Gaussian encoders), the bound on the mutual information is maximized when the projection weights are unconstrained, since this situation leads to insignificant channel noise contributions. By imposing norm constraints on the weights of a linear Gaussian decoder, both channels are shown to result in the same optimization surface for U independently of the choice of nonlinearity. Hence, we reach an important conclusion: *for isotropic channel noise, if the decoder is linear, nothing is gained by using a nonlinear encoder* in the proposed variational settings with the considered norm constraints. This agrees with the related result of Bourlard and Kamp (1988) and Bourlard (2000) for noiseless autoencoders, but is derived in the context of variational information maximization for a stochastic channel $\mathsf{x} \to \mathsf{y}$. Moreover, in our derivation we did not have to assume approximate linearity of $\phi(x)$ around $x = 0$ (*cf* Bourlard (2000)).

These results are somewhat disappointing. In order to improve the power of the method, we need to consider both nonlinear encoders and decoders. However,

from (2.2) it is clear that in the stochastic context, the naive approach of using a nonlinear decoder will typically result in intractable averages over y in the expression for the variational bound $\tilde{I}(\mathsf{x},\mathsf{y})$. In order to avoid this computational difficulty, we derive a modified bound on the mutual information by considering further relaxations of the generic bound and performing decoding in the feature space.

C.3 Nonlinear Gaussian Channels: Nonlinear Decoders and KPCA

Since projections to the feature space are deterministic, one may derive an alternative form of the bound on the mutual information. For any choice of a functional nonlinearity $\boldsymbol{\phi} : \mathsf{x} \mapsto \mathsf{f}$, the codes y are as predictable from the feature vectors $\mathsf{f} \stackrel{\text{def}}{=} \boldsymbol{\phi}(\mathsf{x}) \in \mathbb{R}^{|\boldsymbol{\phi}|}$, as they are from the source variables x themselves. This may be used to modify the bound on $I(\mathsf{x},\mathsf{y})$ in such a way that the reconstruction is performed in the feature space, which results in a simple nonlinear generalization of the PCA solutions for optimal parameters.

The nonlinear Gaussian channel discussed in Section C.2 may be represented by the Markov chain $\mathsf{x} \to \mathsf{f} \to \mathsf{y}$, where $\mathsf{f} \in \mathbb{R}^{|\boldsymbol{\phi}|}$ and $p(\mathsf{f}|\mathsf{x}) \sim \boldsymbol{\delta}(\mathsf{f} - \boldsymbol{\phi}(\mathsf{x}))$, $p(\mathsf{y}|\mathsf{f}) \sim \mathcal{N}_{\mathsf{y}}(\mathsf{W}\mathsf{f}, \boldsymbol{\Sigma}_y)$. Indeed, by marginalizing the feature variables f it is clear that the encoder is given[2] by $p(\mathsf{y}|\mathsf{x}) = \int_{\mathsf{f}} \boldsymbol{\delta}(\mathsf{f}-\boldsymbol{\phi}(\mathsf{x}))\mathcal{N}_y(\mathsf{W}\mathsf{f}, \boldsymbol{\Sigma}_y) = \mathcal{N}_y(\mathsf{W}\boldsymbol{\phi}(\mathsf{x}), \boldsymbol{\Sigma}_y)$.

Proposition C.1. *Let* $\mathsf{s} \to \mathsf{t} \to \mathsf{r}$ *define a Markov chain, such that* $p(\mathsf{t}|\mathsf{s}) = \delta(\mathsf{t} - \mathsf{f}(\mathsf{s}))$*, and* $p(\mathsf{r}|\mathsf{t})$ *is a continuous differentiable density function satisfying* $\forall \mathsf{r}.\forall \mathsf{t}.p(\mathsf{r}|\mathsf{t}) \neq 0$*. Then* $I(\mathsf{s},\mathsf{r}) = I(\mathsf{t},\mathsf{r})$*.*

Proof. From basic properties of the mutual information (see e.g. Cover and Thomas (1991)) it is easy to see that

$$\begin{aligned} I(\mathsf{s},\mathsf{t};\mathsf{r}) &= H(\mathsf{r}) - H(\mathsf{r}|\mathsf{t},\mathsf{s}) = I(\mathsf{t},\mathsf{r}), && \text{(C.22)} \\ I(\mathsf{s},\mathsf{t};\mathsf{r}) &= H(\mathsf{s}) + H(\mathsf{t}|\mathsf{s}) - H(\mathsf{s}|\mathsf{r}) - H(\mathsf{t}|\mathsf{s},\mathsf{r}) \\ &= I(\mathsf{s},\mathsf{r}) + H(\mathsf{t}|\mathsf{s}) - H(\mathsf{t}|\mathsf{s},\mathsf{r}). && \text{(C.23)} \end{aligned}$$

Utilizing the chain structure and the deterministic mapping $p(\mathsf{t}|\mathsf{s})$, we obtain

$$p(\mathsf{t}|\mathsf{s},\mathsf{r}) = \frac{\boldsymbol{\delta}(\mathsf{t} - \mathsf{f}(\mathsf{s}))p(\mathsf{r}|\mathsf{t})}{\int_{\mathsf{t}} \boldsymbol{\delta}(\mathsf{t} - \mathsf{f}(\mathsf{s}))p(\mathsf{r}|\mathsf{t})} = \frac{\boldsymbol{\delta}(\mathsf{t} - \mathsf{f}(\mathsf{s}))p(\mathsf{r}|\mathsf{f}(\mathsf{s}))}{p(\mathsf{r}|\mathsf{f}(\mathsf{s}))}, \qquad \text{(C.24)}$$

i.e. $p(\mathsf{t}|\mathsf{s},\mathsf{r}) = p(\mathsf{t}|\mathsf{s})$. Here we used $f(\mathsf{x})\delta(\mathsf{x}-\mathsf{a}) = \lim_{\mathsf{e}\to 0} [f(\mathsf{a}-\mathsf{e}) + f(\mathsf{a}+\mathsf{e})]\, \delta(\mathsf{x}-\mathsf{a})/2$ (see e.g. Korn and Korn (1968)). Then $H(\mathsf{t}|\mathsf{s},\mathsf{r}) = H(\mathsf{t}|\mathsf{s})$, and from (C.22), (C.23) we obtain $I(\mathsf{s},\mathsf{r}) = I(\mathsf{t},\mathsf{r})$. □

[2] We assume Cartesian coordinates, i.e. $\boldsymbol{\delta}(\mathsf{x} - \mathsf{a}) = \prod_i \delta(x^{(i)} - a_i)$, see e.g. Korn and Korn (1968).

From proposition C.1, the mutual information $I(\mathsf{x},\mathsf{y})$ may be bounded as

$$I(\mathsf{x},\mathsf{y}) = I(\mathsf{f},\mathsf{y}) \geq \tilde{I}(\mathsf{f},\mathsf{y}), \text{ where } \tilde{I}(\mathsf{f},\mathsf{y}) \stackrel{\text{def}}{=} \langle \log q(\mathsf{f}|\mathsf{y}) \rangle_{\tilde{p}(\mathsf{x})p(\mathsf{f}|\mathsf{x})p(\mathsf{y}|\mathsf{f})} + H(\mathsf{f}). \quad \text{(C.25)}$$

We make the simple assumption that the feature decoder is Gaussian, $q(\mathsf{f}|\mathsf{y}) \sim \mathcal{N}_{\mathsf{f}}(\mathsf{U}\mathsf{y}, \boldsymbol{\Sigma}_f)$, which leads to

$$\tilde{I}(\mathsf{x},\mathsf{y}) = -\frac{1}{2}\mathrm{tr}\left\{\mathsf{U}^T\boldsymbol{\Sigma}_f^{-1}\mathsf{U}\left(\boldsymbol{\Sigma}_y + \mathsf{W}\mathsf{S}_F\mathsf{W}^T\right)\right\} + \mathrm{tr}\left\{\boldsymbol{\Sigma}_f^{-1}\mathsf{U}\mathsf{W}\mathsf{S}_F\right\} + H(\mathsf{f}) + c \quad \text{(C.26)}$$

where $\mathsf{S}_F \stackrel{\text{def}}{=} \langle \mathsf{f}\mathsf{f}^T \rangle_{p(\mathsf{f})} \in \mathbb{R}^{|\mathsf{f}|\times|\mathsf{f}|}$ corresponds to the matrix of second-order moments in the feature space. By centering the data in the feature space (see e.g. Schoelkopf et al. (1998)), we may also express the covariance of the feature vectors; we will omit this discussion, as it proves to be secondary for the consequent analysis.

The matrix of the second order moments is easy to compute from the training set as

$$\mathsf{S}_F = \frac{1}{M}\sum_{i=1}^{M} \boldsymbol{\phi}(\mathsf{x}^{(i)})\boldsymbol{\phi}(\mathsf{x}^{(i)})^T, \quad \text{(C.27)}$$

where $\mathsf{f} = \boldsymbol{\phi}(\mathsf{x})$. As expected, in the special case of linear mappings $\boldsymbol{\phi}(\mathsf{x}) \equiv \mathsf{x}$ we get $\mathsf{S}_F = \langle \mathsf{x}\mathsf{x}^T \rangle \equiv \mathsf{S}$, which transforms the objective (C.26) to the simpler bound (C.1).

C.3.1 Constraints on the Feature Mappings $p(\mathsf{f}|\mathsf{x})$

Evaluation of the bound (C.26) is complicated by the need of computing the entropic term $H(\mathsf{f})$. Despite the fact that the mapping to the feature space is deterministic, generally we do not know explicit feature space representations of the training patterns, i.e. numeric approximations due to Brunel and Nadal (1998), Shriki et al. (2002), Corduneanu and Jaakkola (2003) are not directly applicable (moreover, such approximations will not generally retain a proper bound on $I(\mathsf{x},\mathsf{y})$). To simplify the problem of computing $H(\mathsf{f})$, we may note that if $\boldsymbol{\phi} : \mathsf{x} \to \mathsf{f}$ is deterministic and $\tilde{p}(\mathsf{x}) = \sum_{i=1}^{M} \boldsymbol{\delta}(\mathsf{x} - \mathsf{x}^{(i)})/M$ then $\boldsymbol{\phi}(\mathsf{x}^{(i)})$ corresponds to a re-labeled source pattern. If no two distinct points $\mathsf{x}^{(i)}$ and $\mathsf{x}^{(j)}$ in the data space are mapped to the same point in the feature space, then $H(\mathsf{f}) = H(\mathsf{x}) = \log M/\delta(0) = const$, i.e. it may be ignored during the optimization. One way to ensure that this one-to-one condition is satisfied is by imposing inequality constraints on the kernel function, so that

$$\forall i, j \in \{1, \ldots, M\}.\ \mathcal{K}_{i,i} - \mathcal{K}_{i,j} > \epsilon > 0. \quad \text{(C.28)}$$

Clearly, this is a sufficient (but not necessary) condition for the distinct sources $\mathsf{x}^{(i)}$ and $\mathsf{x}^{(j)}$ to have different images in the feature space.

By enforcing the constraints, one may simply ignore the (constant) entropic term $H(\mathsf{f})$ and optimize the objective (C.26) by analogy with the simple bound (C.1) for isotropic linear Gaussians. In conjunction with the set of constraints

(C.28), the objective may be further optimized with respect to parameters of the kernel function. Note that (C.26) remains a proper bound on $I(\mathsf{x},\mathsf{y})$ for the considered constrained nonlinear Gaussian channel.

For the considered choice of the variational feature-space decoder $q(\mathsf{f}|\mathsf{y}) \sim \mathcal{N}_f(\mathsf{U}\mathsf{y}, \sigma_f^2\mathsf{I})$, we may express the corresponding stochastic mapping from the code to the data space as $q(\mathsf{x}|\mathsf{y}) = \int_{\mathsf{f}} p(\mathsf{x}|\mathsf{f})\mathcal{N}_f(\mathsf{U}\mathsf{y}, \sigma_f^2\mathsf{I})$. Unfortunately, these variational posteriors may in principle be difficult to compute. This is a possible limitation of using the variational decoders for reconstruction, since for a general feature mapping $\boldsymbol{\phi}(\mathsf{x})$ it may be difficult to reconstruct a source vector x from the corresponding feature vector f. However, as we mentioned in Chapter 2.1.3, there could be multiple ways of reconstructing source vectors from their noisy representations. In fact, one may view optimization of the bound (C.26) as a way of learning optimal encoders. Once the encoder $p(\mathsf{y}|\mathsf{x})$ is learned, it may be used by any method which can perform exact or approximate inference. For example, if the empirical distribution is available, and if it exhausts all the possible training instances, one may retrieve the transmitted source x from the noisy encoding y by reconstructing with the exact posterior $p(\mathsf{x}|\mathsf{y}) \propto \sum_{i=1}^{M} \delta(\mathsf{x} - \mathsf{x}^{(i)})p(\mathsf{y}|\mathsf{x})$.

C.3.2 Variational Bounds on $H(\mathsf{f})$

As we have discussed above, one way to avoid computation and optimization of the generally intractable entropy of the feature variables $H(\mathsf{f})$ is to impose constraints on the kernel function. This ensures that $H(\mathsf{f}) = H(\mathsf{x}) = const$, which may significantly simplify optimization of (C.26). Moreover, constraints on the off-diagonal elements of the kernel matrices may help to avoid possible singularities, as they will intuitively favor more uniform eigenspectra.

An alternative way to address intractability of $H(\mathsf{f})$ is by relaxing the bound on $I(\mathsf{x},\mathsf{y})$. Clearly, from the joint rule for entropies we get

$$H(\mathsf{f}) = H(\mathsf{f}|\mathsf{x}) + H(\mathsf{x}) - H(\mathsf{x}|\mathsf{f}) \geq H(\mathsf{x}) + \langle \log q(\mathsf{x}|\mathsf{f}) \rangle_{p(\mathsf{x},\mathsf{f})} + c, \tag{C.29}$$

where we used the fact that the mapping $\mathsf{x} \mapsto \mathsf{f}$ is deterministic (but of course not in general one-to-one). By substituting (C.29) into the bound (C.25), we obtain

$$\tilde{I}(\mathsf{x},\mathsf{y}) \geq \langle \log q(\mathsf{f}|\mathsf{y}) \rangle_{\tilde{p}(\mathsf{x})p(\mathsf{f}|\mathsf{x})p(\mathsf{y}|\mathsf{f})} + \langle \log q(\mathsf{x}|\mathsf{f}) \rangle_{p(\mathsf{x},\mathsf{f})} + c \tag{C.30}$$

where c are irrelevant constants. As before, we will be optimizing (C.30) with respect to parameters of the encoder $p(\mathsf{y}|\mathsf{f})$, the feature decoder $q(\mathsf{f}|\mathsf{y})$, and the data decoder $q(\mathsf{x}|\mathsf{f})$.

One tractable way to constrain $q(\mathsf{x}|\mathsf{f})$ is again to use a linear Gaussian $q(\mathsf{x}|\mathsf{f}) \sim \mathcal{N}_{\mathsf{x}}(\mathsf{V}\mathsf{f}, \boldsymbol{\Sigma}_{x|f})$. This parameterization results in a simple linear Gaussian decoder $q(\mathsf{x}|\mathsf{y})$ with a structured covariance matrix, which may be easily used for variational decoding. Other tractable parameterizations may also be possible. Indeed, since integration over x and f reduces to evaluations of the empirical averages, computing the average $\langle \log q(\mathsf{x}|\mathsf{f}) \rangle_{p(\mathsf{x},\mathsf{f})}$ is easy for any data decoder $q(\mathsf{x}|\mathsf{f})$ (provided that it is kernelized, so that there are no explicit computations in the feature space). Fundamentally, however, the choice of the data decoder $q(\mathsf{x}|\mathsf{f})$ does not affect the nature of optimal solutions for parameters of the encoder and feature decoder, provided that at each iteration of the IM the kernel function $\mathcal{K}_{\boldsymbol{\Theta}}$ is fixed.

C.3.3 Kernelized Representation

Here we show that for a fixed kernel function $\mathcal{K}_{\boldsymbol{\Theta}}$, optimal parameters of the encoder $p(\mathsf{y}|\mathsf{f}) \sim \mathcal{N}_{\mathsf{y}}(\mathsf{W}\mathsf{f}, \boldsymbol{\Sigma}_y)$ and feature decoder $q(\mathsf{f}|\mathsf{y}) \sim \mathcal{N}_{\mathsf{f}}(\mathsf{U}\mathsf{y}, \boldsymbol{\Sigma}_f)$ result in a simple nonlinear generalization of the PCA solutions, which justifies nonlinear PCA as a lower bound on mutual information $I(\mathsf{x}, \mathsf{y})$ between the sources and the codes[3].

In what follows we assume that $\boldsymbol{\Sigma}_y = s^2\mathsf{I} \in \mathbb{R}^{|\mathsf{y}|}$, $\boldsymbol{\Sigma}_f = \sigma_f^2\mathsf{I} \in \mathbb{R}^{|\boldsymbol{\phi}|}$, and $|\mathsf{y}| < M < |\boldsymbol{\phi}|$. By analogy with Section C.2 we notice that rows of W and columns of U have dimension $|\boldsymbol{\phi}|$. Then they may be represented in the basis defined by the span of $\{\boldsymbol{\phi}(\mathsf{x}^{(m)})|m = 1, \ldots, M\}$ and its orthogonal compliment as

$$\mathsf{W} = \mathsf{A}\mathsf{F}^T + \mathsf{W}^\perp \in \mathbb{R}^{|\mathsf{y}|\times|\boldsymbol{\phi}|}, \tag{C.31}$$

$$\mathsf{U} = \mathsf{F}\mathsf{C} + \mathsf{U}^\perp \in \mathbb{R}^{|\boldsymbol{\phi}|\times|\mathsf{y}|}. \tag{C.32}$$

Here $\mathsf{F} \in \mathbb{R}^{|\boldsymbol{\phi}|\times M}$ is the design matrix, $\mathsf{U}^\perp$, $\mathsf{W}^\perp$ are orthogonal to F, and $\mathsf{A} \in \mathbb{R}^{|\mathsf{y}|\times M}$, $\mathsf{C} \in \mathbb{R}^{M\times|\mathsf{y}|}$ are matrices of coefficients to be learned. A substitution into expression (C.30) results in

$$\begin{aligned}\tilde{I}(\mathsf{x}, \mathsf{y}) \propto\ & 2\mathrm{tr}\left\{\mathsf{C}\mathsf{A}\mathsf{K}^2\right\} - s^2 M\mathrm{tr}\left\{\mathsf{C}^T\mathsf{K}\mathsf{C}\right\} - \mathrm{tr}\left\{\mathsf{A}\mathsf{K}^2\mathsf{A}^T\left[\mathsf{C}^T\mathsf{K}\mathsf{C} + (\mathsf{U}^\perp)^T\mathsf{U}^\perp\right]\right\} \\ & - Ms^2\mathrm{tr}\left\{(\mathsf{U}^\perp)^T\mathsf{U}^\perp\right\} - \mathrm{tr}\left\{\mathsf{K}\right\} + \frac{1}{2M\sigma_f^2}\langle\log q(\mathsf{x}|\mathsf{f})\rangle_{p(\mathsf{x},\mathsf{f})},\end{aligned} \tag{C.33}$$

where we have used the orthogonality conditions $\mathsf{W}^\perp\mathsf{F} = \mathbf{0} \in \mathbb{R}^{|\mathsf{y}|\times M}$ and $\mathsf{F}^T\mathsf{U}^\perp \in \mathbb{R}^{M\times|\mathsf{y}|}$. It is clear that unconstrained optimization of the objective (C.33) for the complimentary basis $\mathsf{U}^\perp$ leads to $\mathsf{U}^\perp = \mathbf{0} \in \mathbb{R}^{|\boldsymbol{\phi}|\times|\mathsf{y}|}$.

The objective (C.33) may be readily used for learning coefficients A, C and parameters of $q(\mathsf{x}|\mathsf{f})$. In the considered case it may also be applied to learning parameters $\boldsymbol{\Theta}$ of the kernel function $\mathcal{K}_{\boldsymbol{\Theta}}(\mathsf{x}^{(i)}, \mathsf{x}^{(j)}; \boldsymbol{\Theta})$ which gives rise to the Gram matrix K.

C.3.4 Nature of Optimal Solutions

Optimization of the bound (C.33) for the coefficients A leads to

$$\partial\tilde{I}(\mathsf{x}, \mathsf{y})/\partial\mathsf{A} \propto \mathsf{C}^T\mathsf{K}^2 - \mathsf{C}^T\mathsf{K}\mathsf{C}\mathsf{A}\mathsf{K}^2, \tag{C.34}$$

resulting in

$$\mathsf{A} = (\mathsf{C}^T\mathsf{K}\mathsf{C})^{-1}\mathsf{C}^T \tag{C.35}$$

for the case when the Gram matrix $\mathsf{K} \in \mathbb{R}^{M\times M}$ is non-singular. This transforms the objective (C.33) to

$$\tilde{I}(\mathsf{x}, \mathsf{y}) = \mathrm{tr}\left\{\mathsf{K}^2\mathsf{C}(\mathsf{C}^T\mathsf{K}\mathsf{C})^{-1}\mathsf{C}^T\right\} - s^2 M\mathrm{tr}\left\{\mathsf{C}^T\mathsf{K}\mathsf{C}\right\} + const. \tag{C.36}$$

[3]For clarity, we assume that the feature vectors $\mathsf{f} \in \mathbb{R}^{|\boldsymbol{\phi}|}$ are centered (see e.g. Schoelkopf et al. (1998)); otherwise we would consider $p(\mathsf{y}|\mathsf{f}) \sim \mathcal{N}_{\mathsf{y}}(\mathsf{W}(\mathsf{f}-\langle\mathsf{f}\rangle), \boldsymbol{\Sigma}_y), q(\mathsf{f}|\mathsf{y}) \sim \mathcal{N}_{\mathsf{f}}(\mathsf{U}\mathsf{y}+\langle\mathsf{f}\rangle, \boldsymbol{\Sigma}_f)$. Note that the explicit computations in the feature space will not be required, as evaluation of $\langle\log q(\mathsf{f}|\mathsf{y})\rangle$ will only involve computations of scalar products of the feature vectors.

Here we ignored the terms independent of the coefficients A and C, and used parameterization (C.31 – C.32) with the optimal settings $\mathsf{U}^{\perp} = \mathbf{0}$. Moreover, for the fixed kernel matrix K we obtain

$$\begin{aligned} \tilde{I}(\mathsf{x},\mathsf{y}) &= M\mathrm{tr}\left\{\mathsf{F}^T\mathsf{S}_F\mathsf{F}\mathsf{C}(\mathsf{C}^T\mathsf{F}^T\mathsf{F}\mathsf{C})^{-1}\mathsf{C}^T\right\} - s^2 M\mathrm{tr}\left\{\mathsf{C}^T\mathsf{F}^T\mathsf{F}\mathsf{C}\right\} + const \\ &\propto \mathrm{tr}\left\{\mathsf{U}(\mathsf{U}^T\mathsf{U})^{-1}\mathsf{U}^T\mathsf{S}_F\right\} - s^2\mathrm{tr}\left\{\mathsf{U}\mathsf{U}^T\right\} + const. \end{aligned} \tag{C.37}$$

By analogy with Section C.2, maximization of (C.37) gives rise to the nonlinear PCA solution (and rotations) for the left singular vectors of U and W^T. (As before, we have ignored $\mathsf{W}^{\perp}$ and $\mathsf{U}^{\perp}$ in the definitions (C.31), (C.32) since they should optimally cancel).

Just as in the linear case, in order to prevent divergence of $\|\mathsf{W}\|_F$ for $s^2 \neq 0$, it is useful to constrain the singular values of $\mathsf{U} \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{y}|}$. In the special case when $\mathsf{U}^T\mathsf{U} = \mathsf{C}^T\mathsf{K}\mathsf{C} = \mathsf{I}_{|\mathsf{y}|}$, expressions (C.32) and (C.37) lead to the optimal settings

$$\mathsf{K}^2\mathsf{C}\mathsf{R} = \mathsf{K}\mathsf{C}\mathsf{R}\boldsymbol{\lambda}_{S_F} \in \mathbb{R}^{M\times|\mathsf{y}|}. \tag{C.38}$$

Here $\boldsymbol{\lambda}_{S_F} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{y}|}$ is a diagonal matrix of $|\mathsf{y}|$ eigenvalues of S_F (and K), and $\mathsf{R} \in \mathbb{R}^{|\mathsf{y}|\times|\mathsf{y}|}$ is a rotation matrix. From (C.35) and (C.38) it is clear that optimal C and A^T correspond to rotations of principal eigenvectors of the Gram matrix K, which is the kernel PCA solution. Hence, *for a fixed kernel function $\mathcal{K}_{\boldsymbol{\Theta}}$ and data decoder $q(\mathsf{x}|\mathsf{f})$ of a nonlinear Gaussian channel, the variational lower bound $\tilde{I}(\mathsf{x},\mathsf{y})$ is maximized by the kernel PCA solutions for encoder and feature decoder weights.*

C.3.5 Optimal Kernel Functions

The bound (C.30) enables us, in a principled way, to choose between different parameters of the kernel function, or to choose between competing kernels. Indeed, for any kernelized representation of the data decoder $q(\mathsf{x}|\mathsf{f})$, optimal parameters $\boldsymbol{\Theta}$ of the kernel function $\mathcal{K}_{\boldsymbol{\Theta}} : |\mathsf{x}| \times |\mathsf{x}| \to \mathbb{R}$ may be obtained by maximizing the general objective (C.30). Another alternative which we have discussed is optimization of (C.26) subject to the distance constraints on the Gram matrix (C.28). The optimization procedure may be viewed as a special case of the IM algorithm, which for this case may be formulated as follows:

1. For the fixed $\mathcal{K}_{\boldsymbol{\Theta}}$, optimize the bound (C.26) with respect to U, W (or the dual parameters C, A), and parameters of the data decoder $q(\mathsf{x}|\mathsf{f})$.

2. For the fixed C, A, and $q(\mathsf{x}|\mathsf{f})$, optimize the bound (C.26) with respect to the kernel parameters $\boldsymbol{\Theta}$ of $\mathcal{K}_{\boldsymbol{\Theta}}$.

Depending on parameterization of the data decoder $q(\mathsf{x}|\mathsf{f})$ or the choice of constraints (C.28), this procedure generally results in non-trivial settings of the kernel parameters $\boldsymbol{\Theta}$.

C.3.5.1 Learning Kernel Functions for KPCA Channels

Before discussing a general way of optimizing the bound (C.26) for kernel parameters $\mathbf{\Theta}$, we will mention a simple pitfall of a careless interpretation of projections into the feature space. Effectively, this case corresponds to learning optimal kernels for KPCA channels by maximizing $I(\mathsf{f},\mathsf{y})$.

As discussed in Section C.3.1, by presuming that no two data patterns $\mathsf{x}^{(i)}$, $\mathsf{x}^{(j)}$ have identical feature space representations $\forall i \neq j$, one could ignore the entropic term $H(\mathsf{f})$ and optimize the simpler bound (C.37) instead. In our variational formulation, this would correspond to a simple special case when the contribution of the data decoder $\langle \log q(\mathsf{x}|\mathsf{f}) \rangle_{p(\mathsf{x},\mathsf{f})}$ to the bound (C.26) is independent of parameters of the kernel function $\mathcal{K}_{\mathbf{\Theta}}$; for example, this is the case for the trivial feature-independent setting of the variational data decoder $q(\mathsf{x}|\mathsf{f}) \equiv q(\mathsf{x})$. In our framework, this special case would effectively be identical to the unconstrained learning of optimal kernel transformations for nonlinear PCA channels.

For this case, optimization of the objective (C.37) for $\mathbf{\Theta}$ would reduce to maximizing

$$\tilde{I}(\mathsf{x},\mathsf{y}) = \langle \log q(\mathsf{f}|\mathsf{y}) \rangle_{p(\mathsf{f},\mathsf{y})} = -\frac{1}{2} \log \left(\sigma_y^2 + \operatorname{tr}\{\mathsf{K}\} - \operatorname{tr}\left\{ \mathsf{C}^T \mathsf{K} \mathsf{C} \right\} \right), \tag{C.39}$$

where the matrix of the decoder coefficients $\mathsf{C} \in \mathbb{R}^{M \times |\mathsf{y}|}$ performs PCA on $\mathsf{K} \in \mathbb{R}^{M \times M}$ (see expressions (C.37), (C.38) and the discussion in section C.3.2). Equivalently, it may be written in terms of the eigenvalues of the Gram matrix as

$$\tilde{I}(\mathsf{x},\mathsf{y}) = -\frac{1}{2} \log \left(\sigma_y^2 + \sum_{i=|\mathsf{y}|+1}^{M} \lambda_i(\mathsf{K}(\mathbf{\Theta})) \right), \tag{C.40}$$

where $\lambda_i(\mathsf{K}(\mathbf{\Theta}))$ is the i^{th} principal component of the Gram matrix $\mathsf{K} \in \mathbb{R}^{M \times M}$ corresponding to the kernel function $\mathcal{K}_{\mathbf{\Theta}}$ (everywhere in our discussion, we implied that $\mathsf{K} = \mathsf{K}(\mathbf{\Theta})$ is a function of the kernel parameters $\mathbf{\Theta}$).

Clearly, an alternative formulation of the optimization problem for $\mathbf{\Theta}$ in this case is given by

$$\mathbf{\Theta}^{\star} = \arg\max_{\mathbf{\Theta}} \sum_{i=1}^{|\mathsf{y}|} \lambda_i(\mathsf{K}(\mathbf{\Theta})), \tag{C.41}$$

where $\mathbf{\Theta}^{\star}$ indicates the optimal kernel parameter settings. It is easy to see that if the achievable ranks and norms of $\mathsf{K}(\mathbf{\Theta})$ are unconstrained, nothing would prevent the method optimizing (C.41) from generating degenerate Gram matrices (Agakov and Barber (2004c)). In fact, there are at least two possible sources of degeneracy of K: due to the norm divergence and due to K's singularity. The first case is somewhat analogous to what has been described for the simple linear case (see Section C.1). It arises, for example, when $\mathbf{\Theta} > 0$ is a scaling factor of some fixed positive semi-definite function. In this case it is intuitive that the optimal channel will be characterized by the divergent norm $\|\mathsf{K}\|_F$, resulting in the diminishing noise effects and the optimality of the divergent settings of $\mathbf{\Theta} \to \infty$.

Effects of near-singular Gram matrices on parameters of kernel functions may in general be less trivial. From (C.41) it is clear that the worst kernel has a flat spectrum (which is the case for $\mathsf{K} = c\mathsf{I}$), while the optimal kernel function results in the Gram matrix K with the eigenspectrum concentrated at $|\mathsf{y}|$ principal components. Intuitively, by allowing changes in eigenspectra of Gram matrices, we effectively choose an M-dimensional subspace of the feature space which could be well modeled by $|\mathsf{y}|$-KPCA. For example, for trace-constrained matrices, we expect to reach the maximum of the objective (C.41) by choosing the parameters $\boldsymbol{\Theta}$ in such a way that $rank(\mathsf{K}(\boldsymbol{\Theta})) \leq |\mathsf{y}|$. A degenerate solution which satisfies the optimality condition is when K is approximately rank-1, i.e. $\mathrm{tr}\{\mathsf{K}\} \approx \lambda_1(\mathcal{K}_{\boldsymbol{\Theta}})$ and $K_{ij} \approx K_{ii}$ for all $i, j \in \{1, \ldots, M\}$.

It may be noticed that the rank degeneracy is largely an artifact of the trivial data decoder $q(\mathsf{x}|\mathsf{f})$. Indeed, for noiseless channels and rank-1 kernels, it is possible to achieve a perfect reconstruction of *feature* vectors f from their encoded representations y. However, the objective (C.39) does not explicitly favour a good reconstruction of the source vectors x from features f. As a result, all the source vectors may potentially be mapped to a single vector in the feature space, which would lead to a degenerate optimum of the bound. For many interesting kernels, this solution would be characterized by degenerate values of $\boldsymbol{\Theta}$. For example, for trace-constrained radial basis and mixture kernels, the objective (C.39) would be a monotonic function of $\boldsymbol{\Theta}$, which leads to trivial settings of the kernel parameters (Agakov and Barber (2004c)).

C.3.5.2 Learning Kernel Functions by Variational Information Maximization

As we mentioned earlier, one simple way to handle a possible degeneracy of the kernel parameters is by constraining the off-diagonal values of the Gram matrix according to (C.28). It is easy to see that this constraint helps to avoid rank-deficiency of the trivial projections to the feature space. Moreover, numerically it ensures that $H(\mathsf{f}) = const$, i.e. the entropy of the features may be safely ignored during the optimization of the general objective (C.25). Of course, the choice of the threshold ϵ may in practice require similar kinds of heuristics as a choice of kernel parameters themselves. Nevertheless, the thresholds may be easier to interpret geometrically, as they define minimal angles between distinct vectors in the feature space.

Alternatively, more complex *data decoders* $q(\mathsf{x}|\mathsf{f}, \boldsymbol{\Theta}_q)$ may be considered, though by analogy with Section C.2, a proper care should be taken for $q(\mathsf{x}|\mathsf{f}, \boldsymbol{\Theta}_q)$ to be kernelized and $\boldsymbol{\Theta}_q$ to be properly constrained. For both of these cases, a formal analysis of the optimal solutions for the kernel parameters $\boldsymbol{\Theta}$ depends on the constraints and the specifics of $\mathcal{K}_{\boldsymbol{\Theta}}$, and is difficult in general.

Bibliography

Agakov, F. V. and Barber, D. (2003). Approximate Learning in Temporal Hidden Hopfield Models. In 13^{th} *International Conference on Artificial Neural Networks*. Springer.

Agakov, F. V. and Barber, D. (2004a). An Auxiliary Variational Method. In *International Conference on Neural Information Processing*. Springer.

Agakov, F. V. and Barber, D. (2004b). Variational Information Maximization for Neural Coding. In *International Conference on Neural Information Processing*. Springer.

Agakov, F. V. and Barber, D. (2004c). Variational Information Maximization in Gaussian Channels. Technical Report EDI-INF-RR-0206, School of Informatics, University of Edinburgh.

Agakov, F. V. and Barber, D. (2005a). Auxiliary Variational Information Maximization for Dimensionality Reduction. In *PASCAL: Subspace, Latent Structure and Feature Selection techniques: Statistical and Optimisation perspectives Workshop*. Available at `http://homepages.inf.ed.ac.uk/felixa/Papers/aux_pasc2.pdf`.

Agakov, F. V. and Barber, D. (2005b). Kernelized Infomax Clustering. In *Neural Information Processing Systems*. MIT Press.

Agakov, F. V. and Barber, D. (2005c). Nonlinear Encoder Models for Information-Theoretic Clustering. In *PASCAL: Statistics and Optimization of Clustering Workshop*.

Aizerman, M., Braverman, E., and Rozonoer, L. (1964). Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25.

Arfken, G. (1985). *Mathematical Methods for Physicists*. Academic Press.

Arimoto, S. (1972). An Algorithm for computing the capacity of arbitrary discrete memoryless channels. *IT*, 18.

Atick, J. J. (1992). Entropy minimization: A design principle for sensory perception? *International Journal of Neural Systems*, 3:81–90.

Bach, F. R. and Jordan, M. I. (2003). Learning spectral clustering. In *Proceedings of Neural Information Processing Systems*, volume 16.

Baldi, P. and Hornik, K. (1989). Neural networks and principal components analysis: Learning from examples without local minima. *Neural Networks*, 2.

Barber, D. and Agakov, F. V. (2003). The IM Algorithm: A Variational Approach to Information Maximization. In *NIPS*. MIT Press.

Barber, D. and Sollich, P. (2000). Gaussian Fields for Approximate Inference. In Solla, S. A., Leen, T., and Müller, K.-R., editors, *Advances in Neural Information Processing Systems 12*, pages 393–399. MIT Press, Cambridge, MA.

Barber, D. and Wiegerinck, W. (1998). Tractable variational structures for approximating graphical models. *Neural Information Processing*.

Barber, D. and Williams, C. K. I. (1997). Gaussian processes for bayesian classification via hybrid monte carlo. In Mozer, M. C., Jordan, M. I., and Petsche, T., editors, *Advances in Neural Information Processing Systems*, volume 9. The MIT Press.

Barlow, H. (1989). Unsupervised Learning. *Neural Computation*, 1:295–311.

Bartholomew, D. (1987). *Latent Variable Models and Factor Analysis*. Charles Griffin and Co. Ltd., London.

Beal, M. J. (2003). *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, University College London.

Becker, S. (1992). *An Information-theoretic unsupervised learning algorithm for neural networks*. PhD thesis, University of Toronto.

Becker, S. and Hinton, G. (1992). A self-organized neural network that discovers surfaces in random-dot stereograms. *Nature*, 355:161–163.

Bell, A. J. and Sejnowski, T. J. (1994). A non-linear information maximization algorithm that performs blind separation. In *Advances in Neural Information Processing Systems*, volume 7. MIT Press.

Bell, A. J. and Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159.

Berger, A., Pietra, S. D., and Pietra, V. D. (1996). A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1).

Bertsekas, D. P. (1996). *Constrained Optimization and Lagrange Multiplier Methods*. Athena Scientific.

Bertsekas, D. P. (1999). *Nonlinear Programming, 2nd ed.* Athena Scientific.

Bethge, M., Rotermund, D., and Pawelzik, K. (2002). Optimal Short-Term Population Coding: When Fisher Information Fails. *Neural Computation*, 14:2317–2351.

Bishop, C. (1994). Mixture Density Networks. Technical Report NCRG 4288, Aston University.

Bishop, C. (1995). *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford.

Bishop, C. M., Lawrence, N., Jaakkola, T., and Jordan, M. I. (1998). Approximating posterior distributions in belief networks using mixtures. In Jordan, M. I., Kearns, M. J., and Solla, S. A., editors, *Advances in Neural Information Processing Systems*, volume 10. The MIT Press.

Bishop, Svensen and Williams (1998a). Developments of the Generative Topographic Mapping. *Neurocomputing*, 21.

Bishop, Svensen and Williams (1998b). GTM: The Generative Topographic Mapping. *Neural Computation*, 10(1):215–234.

Blahut, R. (1972). Computation of channel capacity and rate-distortion functions. *IT*, 18.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*. ACM Press.

Bourlard, H. (2000). Auto-association by multilayer perceptrons and singular value decomposition. Technical Report IDIAP RR 00-16, IDIAP.

Bourlard, H. and Kamp, Y. (1988). Auto-association by multilayer perceptrons and singular value decomposition. *Biological Cybernetics*, 59.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Brunel, N. and Nadal, J.-P. (1998). Mutual Information, Fisher Information and Population Coding. *Neural Computation*, 10:1731–1757.

Burshtein, D. and Miller, G. (2002). Bounds on the performance of belief propagation decoding. *IEEE Transactions on Information Theory*, 48(1):112–122.

Cardoso, J. F. (1997). Infomax and maximum likelihood for blind source separation. In *IEEE Signal Processing Letters*, volume 4.

Chechik, G., Globerson, A., Tishby, N., and Weiss, Y. (2003). Information bottleneck for gaussian variables. In *Advances in Neural Information Processing Systems*, volume 16. The MIT Press.

Chechik, G. and Tishby, N. (2002). Extracting relevant structures with side information. In Becker, S., Thrun, S., and Obermayer, K., editors, *Advances in Neural Information Processing Systems*, volume 15. The MIT Press.

Collins, M., Schapire, R. E., and Singer, Y. (2002). Logistic regression, AdaBoost and Bregman distances. *Machine Learning*, 48.

Corduneanu, A. and Jaakkola, T. (2003). On Information Regularization. In *Uncertainty in Artificial Intelligence: Proceedings of the Nineteenth Conference (UAI-2003)*. Morgan Kaufmann Publishers.

Cottrell, G., Munro, P., and Zipser, D. (1987). Learning internal representations from gray-scale images: An example of extensional programming. In *Proc. 9th Ann. Conf.of the Cognitive Science Society*.

Courant, R. and Hilbert, D. (1953). *Methods of Mathematical Physics*. Interscience.

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. Wiley, New York.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. J. (1999). *Probabilistic Networks and Expert Systems (Information Science and Statistics S.)*. Springer-Verlag.

Cramer, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press.

Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press.

Dahlquist, G. (2003). *Numerical Methods*. Dover Publications.

Darroch, J. N. and Ratcliff, D. (1972). Generalized Iterative Scaling for Log-Linear Models. *The Annals of Mathematical Statistics*, 43(1).

Dayan, P. (2001). Unsupervised Learning. In Wilson, R. A. and Keil, F., editors, *The MIT Encyclopedia of the Cognitive Sciences*. MIT Press.

Dayan, P., Hinton, G., Neal, R., and Zemel, R. (1995). The helmholtz machine. *Neural Computation*, 7.

Dempster, A. P., Laird, M., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1).

Dennis, J. E. and Schnabel, R. B. (1996). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Dover.

Dhillon, I. S. and Guan, Y. (2003). Information Theoretic Clustering of Sparse Co-Occurrence Data. In *Proceedings of the Third IEEE International Conference on Data Mining*.

Dhillon, I. S., Guan, Y., and Kogan, J. (2002). Iterative clustering of high dimensional text data augmented by local search. In *Proceedings of the Second IEEE International Conference on Data Mining*.

Dhillon, I. S., Guan, Y., and Kulis, B. (2004). Kernel k-means, Spectral Clustering and Normalized Cuts. In *Proceedings of The Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD)*.

Diamantaras, K. I. and Kung, S. Y. (1996). *Principal Component Neural Networks: Theory and Applications*. John Wiley, New York.

Doi, E. and Lewicki, M. (2004). Sparse coding of natural images using an overcomplete set of limited capacity units. In *NIPS*.

Eckmiller, R., Neumann, D., and Baruth, O. (2005). Tunable retina encoders for retina implants: why and how. *J. Neural Eng.*, 2:91–104.

El-Hay, T. and Friedman, N. (2002). Incorporating Expressive Graphical Models in Variational Approximations: Chain-Graphs and Hidden Variables. In *UAI: Proceedings of the 18th Conference*.

Everitt, B. S. (1984). *An Introduction to Latent Variable Models*. Chapman and Hall.

Ewig, G. M. (1985). *Calculus of Variations with Applications*. Dover.

Fano, R. M. (1961). *Transmission of Information: A Statistical theory of Communications*. Wiley: New York.

Feller, W. (1971). *An Introduction to Probability Theory and Its Applications, Vol.2,* 2^{nd} *edition*. John Wiley.

Field, D. J. (1994). What is the goal of sensory coding? *Neural Computation*, 6:559–601.

Fisher, J. W., Darrell, T., Freeman, W. T., and Viola, P. (2000). Learning joint statistical models for audio-visual fusion and segregation. In *Proc. Conf. Advances in Neural Information Processing Systems, NIPS*, volume 13. MIT Press.

Fisher, J. W. and Principe, J. C. (1998). A methodology for information theoretic feature extraction. In *Proc. of the IEEE International Joint Conference on Neural Networks*.

Fisher, R. A. (1922). On the Mathematical Foundations of the Theoretical Statistics. *Philosophical Transactions of the Royal Society, A*, 222:309–368.

Fisher, R. A. (1925). Theory of statistical estimation. In *Proceedings of the Cambridge Philosophical Society*, volume 22. Cambridge University Press.

Fisher, R. A. (1950). *Contributions to Mathematical Statistics*. John Wiley and Sons, New York, London.

Fox, C. (1987). *An Introduction to the Calculus of Variations*. Dover.

Friedman, N., Mosenzon, O., Slonim, N., and Tishby, N. (2001). Multivariate information bottleneck. In *Uncertainty in Artificial Intelligence: Proceedings of the Seventeenth Conference (UAI-2001)*, pages 152–161, San Francisco, CA. Morgan Kaufmann Publishers.

Galeev, E. M. and Tihomirov, V. M. (2000). *Optimization*. Editorial URSS, Moscow.

Gallager, R. G. (1963). *Low Density Parity Check Codes*. MIT Press, Cambridge, Massachusetts.

Gamerman, D. (1997). *Markov Chain Monte Carlo: Stochastic Simulation of Bayesian Inference (Chapman and Hall Texts in Statistical Science)*. Taylor and Francis.

Gelfand, I. M. and Fomin, S. V. (1963). *Calculus of Variations*. Prentice-Hall.

Ghahramani, Z. and Hinton, G. E. (1996). The EM Algorithm for Mixtures of Factor Analyzers. Technical Report CRG-TR-96-1, University of Toronto.

Ghahramani, Z. and Hinton, G. E. (1998). Switching state-space models. Technical report, Dept. of Computer Science, University of Toronto, 6 King's College Road, Toronto M5S 3H5, Canada.

Ghahramani, Z. and Hinton, G. E. (2000). Variational learning for switching state-space models. *Neural Computation*, 12(4):831–864.

Ghahramani, Z. and Jordan, M. (1995). Factorial hidden Markov models. In Touretzky, D. S., Mozer, M. C., and Hasselmo, M. E., editors, *Proc. Conf. Advances in Neural Information Processing Systems, NIPS*, volume 8, pages 472–478. MIT Press.

Gibbs, M. N. and MacKay, D. J. C. (2000). Variational gaussian process classifiers. *IEEE-NN*, 11(6).

Glover, I. and Grant, P. (2003). *Digital Communications*. Prentice Hall.

Gokcay, E. and Principe, J. C. (2002). Information theoretic clustering. In *IEEE Trans. Pattern Analysis and Machine Intelligence*, volume 24, pages 158–171.

Golub, G. H. and Loan, C. F. V. (1996). *Matrix Computations (3rd ed.)*. Johns Hopkins University Press.

Hardy, G. H., Littlewood, J. E., and Polya, G. (1988). *Inequalities, 2nd ed.* Cambridge University Press.

Hartigan, J. A. and Wong, M. A. (1979). A K-means Clustering Algorithm. *Applied Statistics*, 28:100–108.

Haussler, D. and Opper, M. (1997). Mutual Information, Metric Entropy, and Cumulative Relative Entropy Risk. *The Annals of Statistics*, 25(6).

Hertz, J., Krogh, A., and Palmer, R. G. (1991). *Introduction to the Theory of Neural Computation*. MA: Addison-Wesley Publishing Company.

Heskes, T. (2002). Stable fixed points of loopy belief propagation are minima of the Bethe free energy. In *NIPS*.

Higdon, D. M. (1998). Auxiliary variable methods for Markov chain Monte Carlo with applications. *Journal of the American Statistical Association*, 93(442):585–595.

Hinton, G. (1989). Connectionist learning procedures. *Artificial Intelligence*, 40.

Hinton, G. E. and Sejnowski, T. J., editors (1999). *Unsupervised Learning*. MIT Press.

Hoch, T., Wenning, G., and Obermayer, K. (2003). Optimal noise-aided signal transmission through populations of neurons. *Physical Review E*, 68.

Ihler, A. T., III, J. W. F., and Willsky, A. S. (2005). Loopy Belief Propagation: Convergence and Effects of Message Errors. *JMLR*, 6.

Jaakkola, T. (1997). *Variational Methods for Inference and Estimation in Graphical Models*. PhD thesis, Department of Cognitive Sciences, MIT.

Jaakkola, T. S. and Jordan, M. I. (1996). Computing upper and lower bounds on likelihoods in intractable networks. Technical Report AIM-1571, MIT.

Jaakkola, T. S. and Jordan, M. I. (1998). Improving the Mean Field Approximation via the Use of Mixture Distributions. In Jordan, M. I., editor, *Learning in Graphical Models*. Kluwer Academic Publishers.

Jacobs, R. A., I.Jordan, M., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive Mixtures of Local Experts. *Neural Computation*, 3.

Jaynes, E. T. (1996). *Probability Theory: The Logic of Science*. Washington University, St. Louis. Fragmentary edition.

Jensen, F. (1996). *An introduction to Bayesian networks*. UCL Press, London.

Jensen, J. L. W. V. (1906). Sur les fonctions convexes et les inegalites entre les valeurs moyennes. *Acta Math*, 30.

Jenssen, R., Hild, K. E., Erdogmuz, D., Principe, J. C., and Eltoft, T. (2003). Clustering using Renyi's Entropy. In *International Joint Conference on Artificial Neural Networks*.

Johnson, D. (2003). Cramer-Rao Bound. The Connexions Project: m11266.

Jordan, M. (2005). Introduction to graphical models. Unfinished.

Jordan, M., Ghahramani, Z., Jaakkola, T., and Saul, L. (1998). An Introduction to Variational Methods for Graphical Models. In Jordan, M. I., editor, *Learning in Graphical Models*, chapter 1. Kluwer Academic Publishers.

Jordan, M. I., editor (1998). *Learning in Graphical Models*. Kluwer Academic Publishers.

Kang, K. and Sompolinsky, H. (2001). Mutual information of population codes and distance measures in probability space. *Physical Review Letters*, 86(21).

Kannan, R., Vempala, S., and Vetta, A. (2000). On clusterings – good, bad, and spectral. In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, volume 41.

Kikuchi, R. (1951). A theory of cooperative phenomena. *Phys. Rev.*, 81(6).

Korn, G. A. and Korn, T. M. (1968). *Mathematical Handbook for Scientists and Engineers: Definitions, Theorems, and Formulas for Reference and Review*. McGraw-Hill, New York.

Kullback, S. (1959). *Information Theory and Statistics*. New York: Dover Publications.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22.

Landau, L. D. and Lifshitz, E. M. (1996). *Statistical Physics: Course of Theoretical Physics*. Butterworth-Heinemann.

Lauritzen, S. L. and Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their applications to expert systems. *Journal of Royal Statistical Society B*, 50(2):157 – 224.

Lawrence, N. (2000). *Variational Inference in Probabilistic Models*. PhD thesis, Computer Laboratory, University of Cambridge.

Lawrence, N. D. (2003). Gaussian Process Latent Variable Models for Visualization of High Dimensional Data. In *NIPS*.

Lawrence, N. D., Bishop, C. M., and Jordan, M. I. (1998). Mixture Representations for Inference and Learning in Boltzmann Machines. In *UAI: Proceedings of the 14th Conference*.

LeCun, Y. and Cortes, C. (1998). The MNIST Database. Available at `http://yann.lecun.com/exdb/mnist/`.

Lee, B. B., Kremer, J., and Yeh, T. (1998). Receptive fields of primate retinal ganglion cells studied with a novel technique. *Visual Neuroscience*, 15.

Linsker, R. (1988). Towards an Organizing Principle for a Layered Perceptual Network. In *Advances in Neural Information Processing Systems*. American Institute of Physics.

Linsker, R. (1989a). An Application of the Principle of Maximum Information Preservation to Linear Systems. In Touretzky, D., editor, *Advances in Neural Information Processing Systems*, volume 1. Morgan-Kaufmann.

Linsker, R. (1989b). How to generate ordered maps by maximizing the mutual information between input and output signals. *Neural Computation*, 1.

Linsker, R. (1992). Deriving Receptive Fields Using an Optimal Encoding Criterion. In Steven Hanson, Jack Cowan, L. G. e., editor, *Advances in Neural Information Processing Systems*, volume 5. Morgan-Kaufmann.

Linsker, R. (1997). A local learning rule that enables information maximization for arbitrary input distributions. *Neural Computation*, 9.

Luby, M. G., Mitzenmacher, M., Shokrollahi, M. A., and Spielman, D. A. (2001). Improved low-density parity-check codes using irregular graphs and belief propagation. *IEEE Trans. Info Theory*, 47(2).

Luenberger, D. (1973). *Introduction to Linear and Nonlinear Programming*. Addison-Wesley.

Luenberger, D. G. (1998). *Optimization by Vector Space Methods*. Wiley Professional Paperbacks.

MacKay, D. (1998). Introduction to Monte Carlo methods. In Jordan, M., editor, *Learning in Graphical Models*, chapter 1. MIT Press.

MacKay, D. (1999a). Good error correcting codes based on very sparse matrices. *IEEE Trans. Info. Theory*, 45(2).

MacKay, D. (1999b). Maximum likelihood and covariant algorithms for independent components analysis. Technical report, University of Cambridge.

MacKay, D. and Neal, R. (1999). Good Error-Correcting Codes based on Very Sparse Matrices. In *IEEE-IT*.

MacKay, D. J. C. (1997). Gaussian processes - a replacement for supervised neural networks? Lecture notes for a tutorial at NIPS.

MacKay, D. J. C. (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge University Press.

Magnus, J. R. and Neudecker, H. (1999). *Matrix differential calculus with applications in statistics and econometrics*. Wiley, New York, second edition.

Markram, H., Wang, Y., and Tsodyksf, M. (1998). Differential signalling via the same axon of neocortical pyramidal neurons. *Neurobiology*, 95(9).

McEliece, R. J. (1977). *The Theory of Information and Coding*. Addison-Wesley.

Mercer, J. (1909). Functions of positive and negative type and their connection with the theory of integral equations. *Philos. Trans. Roy. Soc.*, 209.

Mika, S., Schoelkopf, B., Smola, A., Muller, K.-R., Scholz, M., and Ratsch, G. (1999). Kernel PCA and De-Noising in Feature Spaces. *Advances in Neural Information Processing Systems*, 11.

Minka, T. (2000). Old and new matrix algebra useful for statistics. Technical report, MIT, Available at `ftp://vismod.www.media.mit.edu/pub/tpminka/papers/minka-matrix.ps.gz`.

Minka, T. (2003). A comparison of numerical optimizers for logistic regression. Technical report, CMU.

Mooij, J. M. and Kappen, H. J. (2005). Sufficient Conditions for Convergence of Loopy Belief Propagation. In *UAI: Proceedings of the 21st Conference*.

Morris, R. (1999). Auxiliary variables for markov random fields with higher order interactions. In Hancock, E. R. and Pelillo, M., editors, *Energy Minimization Methods in Computer Vision and Pattern Recognition*, volume 1654 of *Lecture Notes in Computer Science*. Springer.

Murphy, K. P., Weiss, Y., and Jordan, M. I. (1999). Loopy belief propagation for approximate inference: An empirical study. In *Uncertainty in Artificial Intelligence*, volume 9.

Nadal, J.-P., Brunel, N., and Parga, N. (1998). Nonlinear feedforward networks with stochastic outputs: infomax implies redundancy reduction. *Network: Computation in Neural Systems*, 9(2):207–217.

Nadal, J.-P. and Parga, N. (1994). Nonlinear neurons in the low-noise limit: a factorial code maximizes information transfer. *NETWORK*, 5:565 – 581.

Neal, R. (1993). Probabilistic inference using Markov Chain Monte Carlo methods. CRG-TR-93-1, Dept. of Computer Science, University of Toronto.

Neal, R. M. (1992). Connectionist learning of belief networks. *Artificial Intelligence*, 56:71 – 113.

Neal, R. M. and Hinton, G. E. (1998). A View of the EM Algorithm That Justifies Incremental, Sparse, and Other Variants. In Jordan, M., editor, *Learning in Graphical Models*, chapter 1. MIT Press.

Neuneier, R., Hergert, F., Finnoff, W., and Ormoneit, D. (1994). Estimation of Conditional Densities: A Comparison of Neural Network Approaches. In *ICANN*.

Ng, A. Y., Jordan, M., and Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. In *Proceedings of Neural Information Processing Systems*, volume 14.

Oja, E. (1989). Neural networks, principal components, and subspaces. *International Journal on Neural Systems*, 1.

Opper, M. and Haussler, D. (1995). Bounds for Predictive Errors in the Statistical Mechanics of Supervised Learning. *Physical Review Letters*, 75(20).

Papoulis, A. (1984). *Probability, Random Variables, and Stochastic Processes, 2nd ed.* McGraw-Hill.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufman, San Mateo.

Pearlmutter, B. A. and Parra, L. C. (1996). A context-sensitive generalization of ica. In *Proceedings of the International Conference on Neural Information Processing*.

Penney, R. W. and Sherrington, D. (1993). The weight-space of the binary perceptron. *J. Phys. A: Math. Gen.*, 26:6173–6185.

Perona, P. and Freeman, W. T. (1998). A factorization approach to grouping. In Burkardt, H. and Neumann, B., editors, *Proceedings of European Conference on Computer Vision*.

Pinsker, M. S. (1964). *Information and information stability of random variables and processes*. San Francisco: Holden-Day.

Pouget, A., Zhang, K. C., Deneve, S., and Latham, P. E. (1998). Statistically efficient estimation using population code. *Neural Computation*, 10:373–401.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992). *Numerical Recipes in C*. Cambridge University Press, Second edition.

Pretzel, O. (1996). *Error-Correcting Codes and Finite Fields*. Clarendon Press, Oxford.

Principe, J., Fisher, J., and Xu, D. (2000). Information theoretic learning. In Haykin, S., editor, *Unsupervised Adaptive Filtering*. Wiley.

Principe, J., Xu, D., and Fisher, J. (1998). Pose estimation in sar using an information-theoretic criterion. In *Proceedings of SPIE98*.

Redlich, A. N. (1993). Redundancy reduction as a strategy for unsupervised learning. *Neural Computation*, 5:289–304.

Riley, K. F., Hobson, M. P., and Bence, S. J. (2002). *Mathematical Methods for Physics and Engineering: A Comprehensive Guide*. Cambridge U Press.

Rosen-Zvi, M. and Kanter, I. (2001). Training a perceptron in a discrete weight space. *Physical Review E*, 64.

Roweis, S. and Ghahramani, Z. (1999). A unifying review of linear gaussian models. *Neural Computation*, 11(2).

Rubinov, A. and Yang, X. (2003). *Lagrange-type Functions in Constrained Non-convex Optimization (Applied Optimization S.)*. Kluwer Academic Publishers.

Saad, D. and Opper, M. (2001). *Advanced Mean Field Methods Theory and Practice*. MIT Press.

Samengo, I. and Treves, A. (2001). Representational capacity of a set of independent neurons. *Physical Review E*, 63.

Saul, L., Jaakkola, T., and Jordan, M. (1996). Mean field theory for sigmoid belief networks. *Journal of Artificial Intelligence Research*, 4.

Saul, L. and Jordan, M. (1998). A mean field learning algorithm for unsupervised neural networks. In Jordan, M. I., editor, *Learning in Graphical Models*, chapter 3. MIT Press.

Schoelkopf, B., Smola, A., and Mueller, K. (1998). Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation*, 10.

Scholkopf, B. and Smola, A. (2002). *Learning with Kernels*. MIT Press.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell Systems Technical Journal*, 27.

Shi, J. and Malik, J. (2000). Normalized Cuts and Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905.

Shriki, O., Sompolinsky, H., and Lee, D. D. (2002). An information maximization approach to overcomplete and recurrent representations. In *NIPS*, pages 612 – 618. MIT Press.

Slonim, N. (2002). *The Information Bottleneck: Theory and Applications N. Slonim*. PhD thesis, The Hebrew University.

Slonim, N., Friedman, N., and Tishby, N. (2001). Agglomerative multivariate information bottleneck. In *Advances in Neural Information Processing Systems (NIPS)*.

Slonim, N. and Weiss, Y. (2002). Maximum Likelihood and the Information Bottleneck. In *NIPS*. MIT Press.

Smith, D. R. (1998). *Variational Methods in Optimization*. Dover.

Smola, A. J. (1998). *Learning with Kernels*. PhD thesis, Technische Universitat Berlin.

Still, S. and Bialek, W. (2004). How many clusters? an information theoretic perspective. *Neural Computation*, 16(12).

Still, S., Bialek, W., and Bottou, L. (2004). Geometric Clustering using the Information Bottleneck method. In *NIPS*. MIT Press.

Stocks, N. G. and Mannella, R. (2001). Generic noise-enhanced coding in neuronal arrays. *Physical Review E*, 64.

Swendsen, R. and Wang, J.-S. (1987). Nonuniversal critical dynamics in monte carlo simulations. *Physical Review Letters*, 58:86–88.

Szummer, M. and Jaakkola, T. (2002). Information regularization with partially labeled data. In *Advances in Neural Information Processing Systems (NIPS)*.

Tipping, M. E. and Bishop, C. M. (1999a). Mixtures of Probabilistic Principal Component Analyzers. *Neural Computation*, 11(2):443–482.

Tipping, M. E. and Bishop, C. M. (1999b). Probabilistic principal components analysis. *J. Roy. Statistical Society B*, 61(3):611–622.

Tishby, N., Pereira, F. C., and Bialek, W. (1999). The information bottleneck method. In *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377.

Torkkola, K. (2000). Visualizing class structure in data using mutual information. In *NNSP*.

Torkkola, K. (2001). Learning discriminative feature transforms to low dimensions in low dimensions. In *NIPS*.

Torkkola, K. and Campbell, W. M. (2000). Mutual Information in Learning Feature Transformations. *Proc. 17th International Conf. on Machine Learning*.

Vapnik, V. (1998). *Statistical Learning Theory*. Wiley.

Viterbi, A. J. (1995). *CDMA, Principles of Spread Spectrum*. Addison-Wesley.

von Mises, R. (1964). *Mathematical Theory of Probability and Statistics*. Academic Press, New York.

Wainwright, M. (2002). *Stochastic processes on graphs with cycles: geometric and variational approaches*. PhD thesis, Electrical Engineering and Computer Science, MIT.

Wainwright, M., Jaakkola, T., and Willsky, A. (2001). Tree-based reparameterization framework for approximate estimation on graphs with cycles. In *IEEE Trans. on Information Theory*.

Wainwright, M., Jaakkola, T., and Willsky, A. (2002). A new class of upper bounds on the log partition function. In *Uncertainty in Artificial Intelligence*.

Wang, Y., Jiar, Y., Hu, C., and Turk, M. (2004). Face recognition based on kernel radial basis function networks. In *Asian Conference on Computer Vision*.

Watanabe, M. and Rodieck, R. W. (1989). Parasol and midget ganglion cells of the primate retina. *J. Comp. Neurol.*, 289:333–353.

Weinstock, R. (1974). *Calculus of Variations, With Applications to Physics and Engineering: With Applications to Physics and Engineering*. Dover.

Weiss, Y. (1999). Segmentation using eigenvectors: A unifying view. In *International Conference on Computer Vision*.

Weiss, Y. and Freeman, W. T. (2001). Correctness of belief propagation in Gaussian graphical models of arbitrary topology. *Neural Computation*, 13(10).

Williams, C. K. I. (1997). Regression with Gaussian Processes . In S. W. Ellacott, J. C. M. and Anderson, I. J., editors, *Mathematics of Neural Networks: Models, Algorithms and Applications*. Kluwer.

Williams, C. K. I. (1998). Prediction with Gaussian Processes: From Linear Regression to Linear Prediction and Beyond. In Jordan, M. I., editor, *Learning in Graphical Models*. Kluwer Academic Publishers.

Williams, C. K. I. and Barber, D. (1998). Bayesian Classification with Gaussian Processes. *IEEE Trans Pattern Analysis and Machine Intelligence*, 20(12).

Williams, C. K. I. and Rasmussen, C. E. (1996). Gaussian processes for regression. In Touretzky, D. S., Mozer, M. C., and Hasselmo, M. E., editors, *Advances in Neural Information Processing Systems*, volume 8. The MIT Press.

Yedidia, J., Freeman, W., and Weiss, Y. (2000a). Bethe Free Energy, Kikuchi Approximations, and Belief Propagation Algorithms. Technical report, MERL.

Yedidia, J. S., Freeman, W. T., and Weiss, Y. (2000b). Generalized belief propagation. In *NIPS*, pages 689–695.

Yedidia, J. S., Freeman, W. T., and Weiss, Y. (2004). Constructing Free Energy Approximations and Generalized Belief Propagation Algorithms. Technical Report TR-2004-040, MERL.

Yu, S. X. and Shi, J. (2003). Multiclass spectral clustering. In *International Conference on Computer Vision*.

Yuille, A. L. (2002). CCCP algorithms to minimize the Bethe and Kikuchi free energies: Convergent alternatives to belief propagation. *Neural Computation*, 14(7).

Zemel, R. (1993). *A Minimum Description Length Framework for Unsupervised Learning*. PhD thesis, University of Toronto.

Zemel, R. and Hinton, G. (1994). Autoencoders, minimum description length and helmholtz free energy. In *NIPS*.

Zemel, R. and Hinton, G. (1995). Developing population codes by minimizing description length. *Neural Computation*, 7(3).

Zha, H., Ding, C., Gu, M., He, X., and Simon, H. (2001). Spectral relaxation for k-means clustering. In *Proceedings of Neural Information Processing Systems*, volume 14.

Zhang, K. and Sejnowski, T. J. (1999). Neuronal tuning: To sharpen or broaden? *Neural Computation*, 11:75–84.

Zhang, R. and Rudnicky, A. I. (2002). A large scale clustering scheme for kernel k-means. In *Proceedings of International Conference on Pattern Recognition*.