PRINCIPAL MANIFOLD BASED CORRELATION ANALYSIS applied to GENE REGULATION ANALYSIS OF GLIOBLASTOMA MULTIFORME

Diplomarbeit
von

Patrick Michl

Betreuer: Prof. Dr. Willi Jäger

Universität Heidelberg
Fakultät für Mathematik
September 2017

Acknowledgements

I would like to thank my professor, Dr. Willi Jäger, for the great opportunity to write this thesis. All of his support and advice have been greatly appreciated. Furthermore, I want to thank Dr. Rainer König for leading me into the subject. I owe my interest in the intriguing area of living cells to his teaching and influence. Also I want to thank Rebecca, my parents and wonderful and great friends to whom I owe that I came this far.

Abstract

Gene regulation analysis is a challenging task, which requires the consideration of intricate dependency structures. These structures, however, frequently are only selectively understood in terms of parametric relationships, which also impedes the derivation of meaningful correlation measures. The present thesis addresses this issue by introducing a generalized correlation measure, which is based on principal manifolds. This is motivated by recent advances within the approximation of principal manifolds by deep structured Energy Base Models. Finally the application of this approach is demonstrated for gene regulation analysis of cDNA microarray data of Glioblastoma Multiforme.

Contents

List of Figures

List of Algorithms

Introduction

Various diseases, like cancer, are characterized by modifications of gene expression that target specific cellular functions to support their growth or spreading. Accordingly identifying such modifications can be of tremendous use to understand the respective mechanisms of the diseases and eventually to reveal approaches to counteract them. Fortunately, great advances in high-throughput screening technologies increasingly support the identification of gene interactions by statistical association measures. This task can be quite challenging due to intricate systematic and random variations, that for finite sample sizes of gene expression profiles are only hardly distinguishable. It is therefore required to incorporate structural beliefs. Nevertheless in the domain of gene expression structural assumptions usually are subject to uncertainties and therefore in many cases have to be treated as weak constraints, rather than proven truths. As the assumption of an underlying structure, may have severe consequences for the tractability of statistical inference, such assumptions frequently are incorporated to simplify the statistical model. In the last several decades, the bar in computational statistics was continuously raised. Furthermore this development was underpinned by increased computational capabilities. Thereby amongst different modelling approaches, that benefited from this circumstance, a substantial part can be traced back to the incorporation of a differential structure within the statistical model. This structural assumption is biologically justified by the manifold hypothesis, which assumes, that the gene expression profiles of identical cell types are to be found scattered about a low dimensional smooth submanifold, which covers large parts of the embedding space.

Previous Works

The statistical framework to address the manifold hypothesis was introduced in the late 80s by (Hastie et al. 1989) and generalizes principal component analysis to smooth principal manifolds. Since the approximation of non-linear principal manifolds, however, is accompanied by high computational efforts, it nearly took another twenty years for the development of an applicable method (Gorban et al. 2008). An even more acknowledged breakthrough in this direction, was the utilization of deep structured Energy Based Models (Salakhudinov et al. 2009). These provide many advantages in the approximation, foremost by the consideration of unobserved causal structures, which in particular motivated an application in gene expression analysis (Angermüller et al. 2016, Chen et al. 2016, Syafiandini et al. 2016). Nevertheless, in order to finally utilize principal manifolds for the quantification of gene interactions, there are several obstacles that have to be overcome.

Contribution

A severe gap in the literature about principal manifolds regards the unavailability of a tractable association measure that considers its smooth manifold structure: Usually in such a case one would tend to use the mutual information, which - with respect to a density estimate - measures the expected deviation between the joint distribution and the product of marginal distributions over the realizations. In Energy Based Models, however, due to the intractability of the partition function, the probability of a realization can usually not be calculated directly. In order to close this gap a new association measure is introduced, by the $\mathcal{M}$-Correlation, which is shown to generalize the Pearson Correlation to principal manifolds. By its definition the application of the$\mathcal{M}$-Correlation in turn requires an approximation of the underlying principal manifold $\mathcal{M}$. Therefore the methods required for such an approximation by an Energy Based Model are comprehensively reconstituted. Thereby the focus in this thesis is put on the assumptions that are required for convergence and rapid mixing. Since the thesis chronologically started with the issue to model gene expression data, it is not surprising that these assumptions afterwards turn out to be suitable in this

very domain. Therefore the use of an $\mathcal{M}$-Correlation based analysis is empirically demonstrated in the gene regulation analysis of Glioblastoma Multiforme.

Overview of the Thesis

The first chapter introduces the concepts that are required to generalize the Pearson correlation to smooth submanifolds of its embedding space. Thereby an approximation of the Pearson correlation is derived, which in a first step is extended to linear principal manifolds (L-Correlation) and finally to smooth principal manifolds ($\mathcal{M}$-Correlation). As the evaluation of the $\mathcal{M}$-Correlation requires an estimation of $\mathcal{M}$, the second chapter introduces Energy Based Models (EBM) as a statistical framework. Thereby it is shown that for some "mild" analytical assumptions a local ML-estimation can be derived by a respective gradient descent. Afterwards the third chapter is concerned with a tractable approximation of this estimation, as well as its application to a respective gradient descent. Thereby in a first step the use of Monte Carlo (MC) integration yield the Mini-batch stochastic gradient descent (MBGD), whose convergence is assured by the incorporation of a supplementary statistical assumption. Afterwards the MC approximation by itself is approximated by a Markov Chain Monte Carlo (MCMC) approximation, which applied to the MBGD algorithm then provides the Contrastive Divergence (CD) algorithm. In order to assure the convergence of the CD algorithm, some further analytical assumptions regarding the function space are introduced. However, to also support fast convergence, a sparsity assumption is introduced, that applied to specifically structured EBMs, given by Deep Boltzmann Machines (DBM), is shown to assure rapid mixing of the Markov Chain. In the fourth chapter, the assumptions previously introduced, are evluated with respect to their applicability to gene expression data. Finally as all conditions are met, the $\mathcal{M}$-Correlation is empirically studied by comparison to other correlation measures with respect to their application on Glioblastoma Multiforme.

Chapter 1

Principal Manifold based Correlation Analysis

Overview

This chapter is intended to introduce the necessary concepts that are required to generalize the Pearson correlation to smooth submanifolds of its embedding space. In this purpose the first section reconstitutes elementary properties of the Pearson correlation to derive a representation with respect to the a regression line. The second section introduces principal components to derive this respective line, and thereupon to generalize it to linear subspaces. Thereby the theory is derived for generic elliptical distributions and the principal components are introduced as an orthogonal basis of the embedding space, that decorrelates elliptically distributed random vectors. In the subsequent section the spaces, that are spanned by principal components, are used to identify the tangent spaces of principal manifolds. As principal manifolds, however, are not assured to exist for arbitrary underlying densities, a class of smooth manifold based densities is introduced, that closes this gap. Finally with the $\mathcal{M}$-correlation an association measure is defined, which is shown to generalize the Pearson correlation to densities of this respective class.

1.1 Correlation and Regression Dilution

A fundamental issue, that accompanies the analysis of multivariate data, concerns the quantification of statistically dependency structures by association measures. Many approaches in this direction can be traced back to the late 19^{th} century, where the issue was closely related to the task, to extract laws of nature from two dimensional scatter plots. This in particular applies to the widespread Pearson correlation coefficient.

Definition (Pearson Correlation)**.** *Let $X\colon \Omega \to \mathbb{R}$ and $Y\colon \Omega \to \mathbb{R}$ be random variables with finite variances σ_X^2 and σ_Y^2. Then the Pearson correlation $\rho_{X,Y}$ is defined by:*

$$\rho_{X,Y} \coloneqq \frac{\mathrm{Cov}(X,\, Y)}{\sigma_X \sigma_Y} \tag{1.1}$$

Due to its popularity and simplicity the Pearson correlation has been generalized to a variety of different domains of application, including generic monotonous relationships, relationships between sets of random variables and asymmetric relationships (Zheng et al. 2010). In the purpose to provide a generalization to smooth curves and submanifolds, that allow an incorporation of structural assumptions, some elementary considerations have to be taken into account, that allow a separation between the pairwise quantification of dependencies and their global modelling. Pearson´s original motivation, was the regression of a straight line, that minimizes the averaged Euclidean distance to points, that are scattered about it (Pearson 1901, p561). Thereby his investigations were preceded by the observation, that for a measurement series the assumed "direction of causality" influences the estimate of the slope of the regression line. Thereby the direction of causality is implicated by the choice of an error model, that assumes one random variable to be error free and the other to account for the whole observed error. Pearson empirically observed, that for $n \in \mathbb{N}$ points, given by i.i.d. realizations $\boldsymbol{x} \in \mathbb{R}^n$ of X and $\boldsymbol{y} \in \mathbb{R}^n$ of Y, the **least squares** regression line of $\boldsymbol{y}$ on $\boldsymbol{x}$ only equals the regression line of $\boldsymbol{x}$ on $\boldsymbol{y}$, if all points perfectly fit on a straight line. In all other cases, however, the slopes of the respective regression lines turned out, not to be reciprocal and their product was found within the interval $[0,\, 1)$. This observation was decisive for Pearson´s definition of the correlation coefficient. Thereby $\rho_{X,Y}$ is estimated by its empirical counterpart $\rho_{x,y}$, that replaces variances by sample variances and the covariance by the sample variance.

Lemma 1.1. *Let $X\colon \Omega \to \mathbb{R}$ and $Y\colon \Omega \to \mathbb{R}$ be random variables with $n \in \mathbb{N}$ i.i.d. realizations $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{y} \in \mathbb{R}^n$. Furthermore let $\beta_x \in \mathbb{R}$ denote the slope of the linear regression of $\boldsymbol{y}$ on $\boldsymbol{x}$ and $\beta_y \in \mathbb{R}$ the slope of the linear regression of $\boldsymbol{x}$ on $\boldsymbol{y}$. Then:*

$$\rho_{x,y}^2 = \beta_x \beta_y \tag{1.2}$$

Proof of Lemma 1.1. The following proof is based on (Kenney et al. 1962). The least squares regression of $\boldsymbol{y}$ on $\boldsymbol{x}$ implicates, that for regression coefficients $\alpha_x, \beta_x \in \mathbb{R}$ and a normal distributed random error$\varepsilon \coloneqq Y - (\beta_x X + \alpha_x)$ the log-likelihood of the realizations is maximized, if and only if the ℓ^2-norm of the realizations of ε is minimized, such that:

$$SSE_y(\alpha_x,\, \beta_x) \coloneqq \sum_{i=1}^{n} (y_i - (\beta_x x_i + \alpha_x))^2 \to \min \tag{1.3}$$

Since SSE_y is a quadratic function of α_x and β_x and therefore convex, it has a unique global minimum at:

$$\frac{\partial}{\partial \alpha_x} SSE_y = 2 \sum_{i=1}^{n} (y_i - (\beta_x x_i + \alpha_x))(-x_i) = 0 \tag{1.4}$$

$$\frac{\partial}{\partial \beta_x} SSE_y = 2 \sum_{i=1}^{n} (y_i - (\beta_x x_i + \alpha_x))(-1) = 0 \tag{1.5}$$

By equating the coefficients, equations 1.4 and 1.5 can be rewritten as a system of linear equations of α_x and β_x:

$$\alpha_x n + \beta_x \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \tag{1.6}$$

$$\alpha_x \sum_{i=1}^{n} x_i + \beta_x \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \tag{1.7}$$

Consequently in matrix notation the vector $(\alpha_x,\, \beta_x)^{\mathrm{T}}$ is determined by:

$$\begin{pmatrix} \alpha_x \\ \beta_x \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{pmatrix} \tag{1.8}$$

Let $\overline{x}$, $\overline{y}$ respectively denote the *sample means*. Then by calculating the matrix inverse, the slope β_x equates to:

$$\beta_x = \left(\sum_{i=1}^{n} x_i y_i - n\overline{xy}\right)\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right)^{-1} \tag{1.9}$$

Thereupon by substituting the *sample variance*:

$$\begin{aligned}\sigma_x^2 &\coloneqq \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \\ &= \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right)\end{aligned} \tag{1.10}$$

And the *sample covariance*:

$$\begin{aligned}\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{y}) &\coloneqq \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \\ &= \frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i - n\overline{xy}\right)\end{aligned} \tag{1.11}$$

It follows from equation 1.9, that:

$$\beta_x = \frac{\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{y})}{\sigma_x^2} \tag{1.12}$$

Conversely the slope β_y of the linear regression of $\boldsymbol{x}$ on $\boldsymbol{y}$ mutatis mutandis equates to:

$$\beta_y = \frac{\mathrm{Cov}(\boldsymbol{y}, \boldsymbol{x})}{\sigma_y^2} \tag{1.13}$$

By the symmetry of Cov it the follows, from equations 1.12 and 1.13 that:

$$\beta_x \beta_y = \frac{\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{y})^2}{\sigma_x^2 \sigma_y^2} = \rho_{xy}^2 \tag{1.14}$$

□

Lemma 1.1 shows, that $\rho_{x,y}$ may be regarded as the geometric mean of the regression slopes β_x and β_y, where β_x and β_y respectively describe the causal relationships $X \to Y$ and $Y \to X$. Thereby X and Y respectively are treated as error free regressor

variables to predict the corresponding response variable, that captures the overall error. The mutual linear relationship $X \leftrightarrow Y$ is then described by a regression line, that equally treats errors in both variables. As an immediate consequence of this symmetry it follows, that this **total least squares** regression line is unique, and its slope $\beta_x^\star$, that describes $\boldsymbol{y}$ by $\boldsymbol{x}$ is reciprocal to the slope $\beta_y^\star$, that describes $\boldsymbol{x}$ by $\boldsymbol{y}$ such that $\beta_x^\star \beta_y^\star = 1$. In this sense β_x and β_y may be regarded as biased estimations of $\beta_x^\star$ and $\beta_y^\star$. Thereby the bias generally is known as "regression dilution" or "regression attenuation". For the case that both errors are independent and normal distributed, this bias can be corrected by a prefactor, that incorporates the error of the respective regressor variable. An application of this correction to lemma 1.1 then shows, that $\rho_{X,Y}$ has a consistent estimations by the sample variances of $\boldsymbol{x}$ and $\boldsymbol{y}$ and the variances of their respective errors ε_X and ε_Y.

Proposition 1.1. *Let $X\colon \Omega \to \mathbb{R}$ and $Y\colon \Omega \to \mathbb{R}$ be random variables with $n \in \mathbb{N}$ i.i.d. realizations $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{y} \in \mathbb{R}^n$ and random errors $\varepsilon_X \sim \mathcal{N}(0, \eta_X^2)$ and $\varepsilon_Y \sim \mathcal{N}(0, \eta_Y^2)$. Then:*

$$\rho_{x,y}^2 \xrightarrow{P} \left(1 - \frac{\eta_X^2}{\sigma_x^2}\right)\left(1 - \frac{\eta_Y^2}{\sigma_y^2}\right), \text{ for } n \to \infty \tag{1.15}$$

Proof of Proposition 1.1. Let $\beta_x \in \mathbb{R}$ be the slope of the ordinary least squares (OLS) regression line of $\boldsymbol{y}$ on $\boldsymbol{x}$, where $\boldsymbol{x}$ is assumed to realize X with a normal distributed random error $\varepsilon_X \sim \mathcal{N}(0, \eta_X^2)$. Then X decomposes into (i) an unobserved error free regressor variable $X^\star$ and (ii) the random error ε_X, such that:

$$X \sim X^\star + \varepsilon_X \tag{1.16}$$

With respect to this decomposition, the slope $\beta_x^\star$ of the total least squares (TLS) regression line, that also considers ε_X, then is identified by the slope of the OLS regression of $\boldsymbol{y}$ on $\boldsymbol{x}^\star$, where $\boldsymbol{x}^\star$ realizes $X^\star$. Thereupon let $\sigma_{x^*}^2$ be the empirical variance of $\boldsymbol{x}^\star$, then according to (Snedecor et al. 1967) it follows, that:

$$\beta_x \xrightarrow{P} \frac{\sigma_{x^*}^2}{\sigma_{x^*}^2 + \eta_X^2}\beta_x^\star, \text{ for } n \to \infty \tag{1.17}$$

Since furthermore ε_X by definition is statistically independent from $X^\star$, it can be

concluded, that:

$$\begin{aligned}\sigma_x^2 &= \mathrm{Var}(X^\star + \varepsilon_X) \\ &= \mathrm{Var}(X^\star) + \mathrm{Var}(\varepsilon_X) = \sigma_{x^*}^2 + \eta_X^2 \end{aligned} \tag{1.18}$$

Such that:

$$\frac{\sigma_{x^*}^2}{\sigma_{x^*}^2 + \eta_X^2} \overset{1.18}{=} \frac{\sigma_x^2 - \eta_X^2}{\sigma_x^2} = 1 - \frac{\eta_X^2}{\sigma_x^2} \tag{1.19}$$

And therefore by equation 1.17 that:

$$\beta_x \xrightarrow{P} \left(1 - \frac{\eta_X^2}{\sigma_x^2}\right) \beta_x^\star, \text{ for } n \to \infty \tag{1.20}$$

Conversely let now $\beta_y \in \mathbb{R}$ be the the slope of the OLS regression of $\boldsymbol{x}$ on $\boldsymbol{y}$, where $\boldsymbol{y}$ is assumed to realize Y with a random error $\varepsilon_X \sim \mathcal{N}(0,\, \eta_Y^2)$. Then also the corrected slope $\beta_y^\star$ mutatis mutandis satisfies the relation given by equation 1.20 and by the representation of $\rho_{x,y}$, as given by lemma 1.1, it then can be concluded, that:

$$\rho_{x,y}^2 \overset{1.1}{=} \beta_x \beta_x \xrightarrow{P} \left(1 - \frac{\eta_X^2}{\sigma_x^2}\right)\left(1 - \frac{\eta_Y^2}{\sigma_y^2}\right) \beta_x^\star \beta_y^\star, \text{ for } n \to \infty \tag{1.21}$$

The proposition then follows by the uniqueness of the total least squares regression line for known variances η_X^2 and η_Y^2, such that:

$$\beta_y^\star = \frac{1}{\beta_x^\star}$$

□

1.2 Linear Principal Manifold based Correlation

Within the same publication, in which Pearson introduced the correlation coefficient, he also developed a structured approach that determines the straight line, that minimizes the Euclidean distance (Pearson 1901, p563). His method, which later received attribution as the method of **Principal Component Analysis** (**PCA**), however, even went further and allowed a canonical generalization of the problem in the following sense: For $d \in \mathbb{N}$ let $\boldsymbol{X} \colon \Omega \to \mathbb{R}^d$ be a multivariate random vector and for

$n \in \mathbb{N}$ let $\boldsymbol{x} \in \mathbb{R}^{n \times d}$ be an i.i.d. realization of $\boldsymbol{X}$. Then for any given $k \in \mathbb{N}$ with $k \leq d$ the goal is, to determine an affine linear subspace $L \subseteq \mathbb{R}^d$ of dimension k, that minimizes the summed Euclidean distance to $\boldsymbol{x}$. In order to solve this problem, the fundamental idea of Pearson was, to transfer the principal axis theorem from ellipsoids to multivariate Gaussian distributed random vectors. Thereupon, however, the method also can be formulated with respect to generic elliptical distributions.

Definition (Elliptical Distribution)**.** *For $d \in \mathbb{N}$ let $\boldsymbol{X} \colon \Omega \to \mathbb{R}^d$ be a random vector. Then $\boldsymbol{X}$ is elliptically distributed, iff there exists a random vector $\boldsymbol{S} \colon \Omega \to \mathbb{R}^k$ with $k \leq d$, which distribution is invariant to rotations, a matrix $A \in \mathbb{R}^{d \times k}$ of rank k and a vector $\boldsymbol{b} \in \mathbb{R}^d$, such that:*

$$\boldsymbol{X} \sim A\boldsymbol{S} + \boldsymbol{b} \tag{1.22}$$

Consequently a random vector $\boldsymbol{X}$ is elliptically distributed, if it can be represented by an affine transformation of a radial symmetric distributed random vector $\boldsymbol{S}$. The decisive property, that underpins the choice of elliptical distributions, lies within their coincidence of linear and statistical dependencies, which allows to decompose $\boldsymbol{X}$ in statistically independent components by a linear decomposition. This property allows, to substantiate the multidimensional "linear fitting problem" with respect to an orthogonal projection.

Proposition 1.2. *For $d \in \mathbb{N}$ let $\boldsymbol{X} \colon \Omega \to \mathbb{R}^d$ be an elliptically distributed random vector, $L \subseteq \mathbb{R}^d$ an affine linear subspace of $\mathbb{R}^d$ and π_L the orthogonal projection of $\mathbb{R}^d$ onto L. Then the following statements are equivalent:*

(i) *L minimizes the Euclidean distance to $\boldsymbol{X}$*

(ii) *$\mathbb{E}(\boldsymbol{X}) \in L$ and L maximizes the variance* $\mathrm{Var}(\pi_L(\boldsymbol{X}))$

Proof. Let $\boldsymbol{Y}_L \coloneqq \boldsymbol{X} - \pi_L(\boldsymbol{X})$, then the Euclidean distance between $\boldsymbol{X}$ and L can be written:

$$d(\boldsymbol{X},\, \pi_L(\boldsymbol{X}))^2 = \mathbb{E}(\|(\boldsymbol{X} - \pi_L(\boldsymbol{X}))\|_2^2) = \mathbb{E}(\boldsymbol{Y}_L^2) \tag{1.23}$$

This representation can furthermore be decomposed by using the algebraic formula for the variance:

$$\mathbb{E}(\boldsymbol{Y}_L^2) = \mathrm{Var}(\boldsymbol{Y}_L) + \mathbb{E}(\boldsymbol{Y}_L)^2 \tag{1.24}$$

Let now be $\boldsymbol{Y}_L^{\perp} \coloneqq \pi_L(\boldsymbol{X})$, then $\boldsymbol{X} = \boldsymbol{Y}_L + \boldsymbol{Y}_L^{\perp}$ and $\boldsymbol{Y}_L$ and $\boldsymbol{Y}_L^{\perp}$ are uncorrelated, such that:

$$\mathrm{Var}(\boldsymbol{X}) = \mathrm{Var}(\boldsymbol{Y}_L + \boldsymbol{Y}_L^{\perp}) = \mathrm{Var}(\boldsymbol{Y}_L) + \mathrm{Var}(\boldsymbol{Y}_L^{\perp}) \tag{1.25}$$

From equations 1.23, 1.24 and 1.25 it follows, that:

$$d(\boldsymbol{X}, \pi_L(\boldsymbol{X}))^2 = \mathrm{Var}(\boldsymbol{X}) - \mathrm{Var}(\boldsymbol{Y}_L^{\perp}) + \mathbb{E}(\boldsymbol{Y}_L)^2 \tag{1.26}$$

Consequentially the Euclidean distance is minimized, if and only if the right side of equation 1.26 is minimized. The first term $\mathrm{Var}(\boldsymbol{X})$, however, does not depend on L and since $\boldsymbol{X}$ is elliptically distributed, the linear independence of $\boldsymbol{Y}_L^{\perp}$ and $\boldsymbol{Y}_L$ is sufficient for statistically independence. It follows, that the Euclidean distance is minimized, if and only if: (1) The term $\mathbb{E}(\boldsymbol{Y}_L)^2$ is minimized and (2) the term $\mathrm{Var}(\boldsymbol{Y}_L^{\perp})$ is maximized. Concerning (1) it follows, that:

$$\mathbb{E}(\boldsymbol{Y}_L)^2 = \mathbb{E}(\boldsymbol{X} - \pi_L(\boldsymbol{X}))^2 = (\mathbb{E}(\boldsymbol{X}) - \pi_L(\mathbb{E}(\boldsymbol{X})))^2$$

Therefore the term $\mathbb{E}(\boldsymbol{Y}_L)^2$ is minimized, if and only if $\pi_L(\mathbb{E}(\boldsymbol{X})) = \mathbb{E}(\boldsymbol{X})$, which in turn means that $\mathbb{E}(\boldsymbol{X}) \in L$. Concerning (2), the proposition immediately follows by the definition of $\boldsymbol{Y}_L^{\perp}$. □

Let now be $k \leq d$. In order to derive an affine linear subspace $L \subseteq \mathbb{R}^d$ that minimizes the Euclidean distance to $\boldsymbol{X}$, proposition ... states, that is suffices to provide an L which (1) is centred in $\boldsymbol{X}$, such that $\mathbb{E}(\boldsymbol{X}) \in L$, and (2) maximizes the variance of the projection. In order to maximize $\mathrm{Var}(\pi_L(\boldsymbol{X}))$, however, it is beneficial to give a further representation.

Lemma 1.2. *For $d \in \mathbb{N}$ let $\boldsymbol{X} \colon \Omega \to \mathbb{R}^d$ be an elliptically distributed random vector, $L \subseteq \mathbb{R}^d$ an affine linear subspace of $\mathbb{R}^d$, which for an $k \leq d$, a vector $\boldsymbol{v} \in \mathbb{R}^d$ and an orthonormal basis $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k \in \mathbb{R}^d$ is given by:*

$$L = \boldsymbol{v} + \bigoplus_{i=1}^{k} \mathbb{R}\boldsymbol{u}_i$$

Let further be $\pi_L \colon \mathbb{R}^d \to L$ the orthogonal projection of $\mathbb{R}^d$ onto L. Then the variance

of the projection is given by:

$$\mathrm{Var}(\pi_L(\boldsymbol{X})) = \sum_{i=1}^{k} \boldsymbol{u}_i^T \mathrm{Cov}(\boldsymbol{X}) \boldsymbol{u}_i$$

Proof. Let $L' \coloneqq L + \mathbb{E}(\boldsymbol{X}) - \boldsymbol{v}$, then the orthogonal projection $\pi_{L'}(\boldsymbol{X})$ decomposes into individual orthogonal projections to the respective basis vectors, such that:

$$\pi_{L'}(\boldsymbol{X}) = \mathbb{E}(\boldsymbol{X}) + \sum_{i=1}^{k} \langle \boldsymbol{X} - \mathbb{E}(\boldsymbol{X}),\, \boldsymbol{u}_i \rangle \, \boldsymbol{u}_i \tag{1.27}$$

Let $\hat{X}_i \coloneqq \langle \boldsymbol{X},\, \boldsymbol{u}_i \rangle \, \boldsymbol{u}_i$, for $i \in \{1, \ldots, k\}$. The total variance of this projection is then given by:

$$\mathrm{Var}(\pi_{L'}(\boldsymbol{X})) \overset{1.27}{=} \mathrm{Var}\left(\mathbb{E}(\boldsymbol{X}) + \sum_{i=1}^{k} \hat{X}_i - \sum_{i=1}^{k} \mathbb{E}(\hat{X}_i) \right) = \mathrm{Var}\left(\sum_{i=1}^{k} \hat{X}_i \right) \tag{1.28}$$

Since the random variables $\hat{X}_i$ by definition are uncorrelated, the algebraic formula for the variance can be used to decompose the variance:

$$\mathrm{Var}\left(\sum_{i=1}^{k} \hat{X}_i \right) = \sum_{i=1}^{k} \mathrm{Var}(\hat{X}_i) \tag{1.29}$$

By equating the term $\mathrm{Var}(\hat{X}_i)$, for $i \in \{1, \ldots, k\}$ it follows, that:

$$\mathrm{Var}(\hat{X}_i) = \mathrm{Var}\left(\langle \boldsymbol{X},\, \boldsymbol{u}_i \rangle \, \boldsymbol{u}_i \right) = \mathrm{Var}\left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{u}_i \right) \boldsymbol{u}_i^2 = \mathrm{Var}\left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{u}_i \right) \tag{1.30}$$

And furthermore by introducing the covariance matrix $\mathrm{Cov}(\boldsymbol{X})$:

$$\mathrm{Var}\left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{u}_i \right) \overset{\mathrm{def}}{=} \mathbb{E}\left((\boldsymbol{X}^{\mathrm{T}} \boldsymbol{u}_i)^{\mathrm{T}} (\boldsymbol{X}^{\mathrm{T}} \boldsymbol{u}_i) \right) = \boldsymbol{u}_i^{\mathrm{T}} \mathbb{E}\left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{X} \right) \boldsymbol{u}_i \overset{\mathrm{def}}{=} \boldsymbol{u}_i^T \mathrm{Cov}(\boldsymbol{X}) \boldsymbol{u}_i \tag{1.31}$$

Summarized the equations 1.28, 1.29, 1.30 and 1.31 provide a representation for the variance of the projection to L':

$$\mathrm{Var}(\pi_{L'}(\boldsymbol{X})) = \sum_{i=1}^{k} \boldsymbol{u}_i^T \mathrm{Cov}(\boldsymbol{X}) \boldsymbol{u}_i$$

Finally the total variance of the projection is invariant under translations of L, such that:

$$\begin{aligned} \mathrm{Var}(\pi_L(\boldsymbol{X})) = \mathrm{Var}(\pi_L(\boldsymbol{X}) + \mathbb{E}(\boldsymbol{X}) - \boldsymbol{v}) = \mathrm{Var}(\pi_{L'}(\boldsymbol{X})) \\ \overset{1.29}{=} \sum_{i=1}^{k} \boldsymbol{u}_i^T \mathrm{Cov}(\boldsymbol{X}) \boldsymbol{u}_i \end{aligned} \tag{1.32}$$

□

Lemma 1.2, shows, that for elliptically distributed random vectors $\mathbf{X}$ the best fitting linear subspaces are completely determined by the expectation $\mathbb{E}(\boldsymbol{X})$ and the covariance matrix $\mathrm{Cov}(\boldsymbol{X})$. On this point it is important to notice, that the covariance matrix is symmetric, which allows its diagonalization with regard to real valued Eigenvalues.

Lemma 1.3. *For $d \in \mathbb{N}$ let $\boldsymbol{X} \colon \Omega \to \mathbb{R}^d$ be an elliptically distributed random vector, $L \subseteq \mathbb{R}^d$ an affine linear subspace of $\mathbb{R}^d$, which for an $k \leq d$, a vector $\boldsymbol{v} \in \mathbb{R}^d$ and an orthonormal basis $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k \in \mathbb{R}^d$ is given by:*

$$L = \boldsymbol{v} + \bigoplus_{i=1}^{k} \mathbb{R}\boldsymbol{u}_i$$

Let further be $\pi_L \colon \mathbb{R}^d \to L$ the orthogonal projection of $\mathbb{R}^d$ onto L, as well as $\lambda_1, \ldots, \lambda_d \in \mathbb{R}$ the eigenvalues of $\mathrm{Cov}(\boldsymbol{X})$. Then there exist numbers $a_1, \ldots, a_d \in [0, 1]$ with $\sum_{i=1}^{d} a_i = k$, such that:

$$\mathrm{Var}(\pi_L(\boldsymbol{X})) = \sum_{i=1}^{d} \lambda_i a_i$$

Proof. From lemma 1.2 it follows, that:

$$\mathrm{Var}(\pi_L(\boldsymbol{X})) = \sum_{j=1}^{k} \boldsymbol{u}_j^T \mathrm{Cov}(\boldsymbol{X}) \boldsymbol{u}_j$$

Since the covariance matrix $\mathrm{Cov}(\boldsymbol{X})$ is a symmetric matrix, there exists an orthonormal basis transformation matrix $S \in \mathbb{R}^{d \times d}$ and a diagonal matrix $D \in \mathbb{R}^{d \times d}$, such

that $\mathrm{Cov}(\boldsymbol{X}) = S^T DS$. Then the variance $\mathrm{Var}(\pi_L(\boldsymbol{X}))$ has a decomposition, given by:

$$\mathrm{Var}(\pi_L(\boldsymbol{X})) = \sum_{i=1}^{k} \boldsymbol{u}_j{}^{\mathrm{T}} S^{\mathrm{T}} DS\boldsymbol{u}_j = \sum_{i=1}^{k} (S\boldsymbol{u}_j)^{\mathrm{T}} DS\boldsymbol{u}_j$$

For $j \in \{1, \ldots, k\}$ let now $\boldsymbol{c}_j \coloneqq S\boldsymbol{u}_j$ and for $i \in \{1, \ldots, n\}$ let the number $a_i \in \mathbb{R}$ be defined by:

$$a_i \coloneqq \sum_{j=1}^{k} (\boldsymbol{c}_{ji})^2$$

Then according to Lemma 1.2 the variance $\mathrm{Var}(\pi_L(\boldsymbol{X}))$ can be decomposed:

$$\begin{aligned}\sum_{i=1}^{k} \boldsymbol{c}_j{}^T D\boldsymbol{c}_j &= \sum_{j=1}^{k}\sum_{i=1}^{d} \boldsymbol{c}_{ji}\lambda_i \boldsymbol{c}_{ji} \\ &= \sum_{i=1}^{d} \lambda_i \sum_{j=1}^{k} (\boldsymbol{c}_{ji})^2 = \sum_{i=1}^{d} \lambda_i a_i\end{aligned}$$

Furthermore since $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k$ is an orthonormal basis and S an orthonormal matrix it follows that also $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_k$ is an orthonormal basis. Consequentially for $i \in \{1, \ldots, d\}$ it holds, that:

$$a_i = \sum_{j=1}^{k} (\boldsymbol{c}_{ji})^2 \leq \sum_{j=1}^{k} \|\boldsymbol{c}_{ji}\|_2 \leq 1$$

And furthermore by its definition it follows, that $a_i \geq 0$, such that $a_i \in [0, 1]$. Besides this the sum over all a_i equates to:

$$\sum_{i=1}^{d} a_i = \sum_{i=1}^{d}\sum_{j=1}^{k} (\boldsymbol{c}_{ji})^2 = \sum_{j=1}^{k} \boldsymbol{c}_j^T \boldsymbol{c}_j = k$$

□

With reference to the principal axis transformation, the eigenvectors of the covariance matrix are then termed **principal components** and affine linear subspaces of the embedding space as **linear principal manifolds**.

Definition (Linear Principal Manifold)**.** *For $d \in \mathbb{N}$ let $\boldsymbol{X}\colon \Omega \to \mathbb{R}^d$ be a random vector. Then a vector $\boldsymbol{c} \in \mathbb{R}^d$ with $\boldsymbol{c} \neq 0$ is a principal component for $\boldsymbol{X}$, iff there exists an $\lambda \in \mathbb{R}$, such that :*

$$\mathrm{Cov}(\boldsymbol{X}) \cdot \boldsymbol{c} = \lambda \boldsymbol{c} \tag{1.33}$$

Furthermore let $L \subseteq \mathbb{R}^d$ be an affine linear subspace of $\mathbb{R}^d$ with dimension $k \leq d$. Then L is a linear k-principal manifold for $\boldsymbol{X}$, if there exists a set set $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_k$ of linear independent principal components for $\boldsymbol{X}$, such that:

$$L = \mathbb{E}(\boldsymbol{X}) + \bigoplus_{i=1}^{k} \mathbb{R}\boldsymbol{c}_i$$

Then L is termed maximal, iff the sum of the Eigenvalues $\lambda_1, \ldots, \lambda_k$, that correspond to the principal components $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_k$ is maximal.

Proposition 1.3. *For $d \in \mathbb{N}$ let $\boldsymbol{X}\colon \Omega \to \mathbb{R}^d$ be an elliptically distributed random vector and $L \subseteq \mathbb{R}^d$ an affine linear subspace. Then the following statements are equivalent:*

(i) *L minimizes the Euclidean distance to $\boldsymbol{X}$*

(ii) *L is a maximal linear principal manifold for $\boldsymbol{X}$*

Proof. "$\Longrightarrow$" Let $\pi_L\colon \mathbb{R}^d \hookrightarrow L$ denote the orthogonal projection of $\mathbb{R}^d$ onto L. Then according to proposition 1.2 L minimizes the averaged Euclidean distance to $\boldsymbol{X}$, if and only if (i) $\mathbb{E}(\boldsymbol{X}) \in L$ and (ii) L maximizes the variance $\mathrm{Var}(\pi_L(\boldsymbol{X}))$. In particular (i) is satisfied, if and only if an orthonormal basis $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k \in \mathbb{R}^d$ can be chosen, such that:

$$L = \mathbb{E}(\boldsymbol{X}) + \bigoplus_{i=1}^{k} \mathbb{R}\boldsymbol{u}_i$$

Then according to Lemma 1.3 there exist numbers $a_1, \ldots, a_d \in [0, 1]$ with $\sum_{i=1}^{d} a_i = k$, such that:

$$\mathrm{Var}(\pi_L(\boldsymbol{X})) = \sum_{i=1}^{d} \lambda_i a_i$$

Thereupon (ii) is satisfied, if and only if the numbers a_i maximize this sum. Since the covariance matrix $\mathrm{Cov}(\boldsymbol{X})$ is positive semi-definite, the eigenvalues λ_i are not negative such that the sum is maximized for:

$$a_i = \begin{cases} 1 & \text{for } i \in \{1, \dots k\} \\ 0 & \text{else} \end{cases}$$

Such that:

$$\begin{aligned} \sum_{j=1}^{k} \boldsymbol{u}_j^T \mathrm{Cov}(\boldsymbol{X}) \boldsymbol{u}_j &= \sum_{i=1}^{d} \lambda_i a_i \\ &= \sum_{i=1}^{k} \lambda_i = \sum_{j=1}^{k} \boldsymbol{c}_j^T \mathrm{Cov}(\boldsymbol{X}) \boldsymbol{c}_j \end{aligned}$$

Accordingly the choice $\boldsymbol{u}_j = \boldsymbol{c}_j$ for $i \in \{1, \dots, k\}$ maximizes $\mathrm{Var}(\pi_L(\boldsymbol{X}))$ and L has a representation, given by:

$$L = \mathbb{E}(\boldsymbol{X}) + \bigoplus_{i=1}^{k} \mathbb{R}\boldsymbol{c}_i$$

"$\Longleftarrow$" Let L have a representation as given by (ii), then (1) $\mathbb{E}(\boldsymbol{X}) \in L$ and (2) the variance $\mathrm{Var}(\pi_L(\boldsymbol{X}))$ is maximized. According to Proposition 1.2 it follows, that L minimizes the Euclidean distance to $\boldsymbol{X}$. □

Definition (L-Correlation)**.** *For $d \in \mathbb{N}$ let $\boldsymbol{X} \colon \Omega \to \mathbb{R}^d$ be a random vector and L a maximal linear principal manifold for $\boldsymbol{X}$. Then for any $i, j \in \{1, \dots, d\}$ let the L-Correlation between X_i and X_j be defined by:*

$$\rho^2_{X_i, X_j | L} \coloneqq R_i R_j \tag{1.34}$$

*where with the orthogonal projection $\pi_L \colon \mathbb{R}^d \to L$ for any $i \in \{1, \dots, d\}$ the **reliability** of X_i with respect to L is given by:*

$$R_i \coloneqq 1 - \frac{\mathrm{Var}_i\left(\boldsymbol{X} - \pi_L(\boldsymbol{X})\right)}{\mathrm{Var}_i(\boldsymbol{X})} \tag{1.35}$$

Proposition 1.4. *For an elliptically distributed random vector the L-Correlation generalizes the Pearson Correlation to maximal linear principal manifolds.*

Proof. Let $\boldsymbol{X} \colon \Omega \to \mathbb{R}^2$ be an elliptically distributed random vector and L a maximal

linear 1-principal manifold for $\boldsymbol{X}$. Then for $i \in \{1, 2\}$ the random error of the variable X_i has a variance:

$$\eta_{X_i}^2 = \mathrm{Var}_{X_i}\left(\boldsymbol{X} - \pi_L(\boldsymbol{X})\right)$$

Such that by the definition of the reliability it follows, that:

$$R_i \overset{1.35}{=} 1 - \frac{\eta_{X_i}^2}{\sigma_{X_i}^2}$$

Consequently:

$$\rho_{X_i,X_j|L}^2 = \left(1 - \frac{\eta_{X_i}^2}{\sigma_{X_i}^2}\right)\left(1 - \frac{\eta_{X_j}^2}{\sigma_{X_j}^2}\right)$$

With $n \in \mathbb{N}$ i.i.d. realizations $\boldsymbol{x} \in \mathbb{R}^{n\times 2}$ of $\boldsymbol{X}$ an empirical L-Correlation $\rho_{x_i,x_j|L}^2$ is then given by replacing the variances by the sample variances. Then by proposition 1.1 it follows, that:

$$\rho_{x_i,x_j}^2 \xrightarrow{P} \rho_{x_i,x_j|L}^2, \text{ for } n \to \infty$$

□

1.3 Principal Manifold based Correlation

Linear principal manifolds allow the projection of a random vector $\boldsymbol{X}\colon \Omega \to \mathbb{R}^d$ onto a linear subspace $L \subseteq \mathbb{R}^d$, which maximally preserves the linear dependency structure of $\boldsymbol{X}$ in terms of its covariances. Thereby for the orthogonal projection $\pi_L\colon \mathbb{R}^d \hookrightarrow L$, the variance on L, given by $\mathrm{Var}(\pi_L(\boldsymbol{X}))$, is referred as the **explained variance** and the orthogonal deviation $\mathrm{Var}(\boldsymbol{X} - \pi_L(\boldsymbol{X}))$ as the **unexplained variance**. Thereupon by the assumption, that $\boldsymbol{X}$ is elliptically distributed, it can be concluded, that linear independence coincides with statistically independence, that that $\pi_L(\boldsymbol{X})$ and $\boldsymbol{X} - \pi_L(\boldsymbol{X})$ are statistically independent and therefore allow the following decomposition:

$$\underbrace{\mathrm{Var}(\boldsymbol{X})}_{\text{total variance}} = \underbrace{\mathrm{Var}(\pi_L(\boldsymbol{X}))}_{\text{explained variance}} + \underbrace{\mathrm{Var}(\boldsymbol{X} - \pi_L(\boldsymbol{X}))}_{\text{unexplained variance}}$$

This decomposition, as shown by theorem 1.4, is of fundamental importance for the correlation over linear Principal Manifolds, since it determines the reliabilities of the

respective random variable X by the ratio:

$$R = 1 - \frac{\text{explained variance}}{\text{total variance}}$$

On this point of the discussion it´s just a small step to generalize the principal

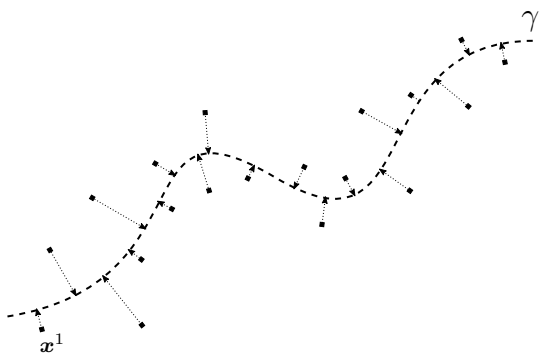

Figure 1.1: Principal Curve for a 2-dimensional realization

components, by a smooth curves $\gamma\colon [a, b] \to \mathbb{R}^d$ (figure 1.1). This is particular appropriate, if the assumption of an elliptically distribution can only hardly be justified, like for **observed dynamical systems**. Thereby the evolution function generates a smooth submanifold $\mathcal{M} \subseteq \mathbb{R}^d$ within the observation space $\mathbb{R}^d$, and an "error free" observation can be identified by a random vector $\boldsymbol{X}^\star$, with outcomes on $\mathcal{M}$. Additionally, however, the observation function may be regarded to be subjected to a measurement error $\boldsymbol{\varepsilon}$. By the assumption, that $\boldsymbol{\varepsilon}$ has an elliptical distribution, then the distribution of the observable random vector $\boldsymbol{X}$ is represented by a elliptical $\mathcal{M}-$distribution.

Definition ($\mathcal{M}$-Distribution)**.** *For $d \in \mathbb{N}$ let $\boldsymbol{X}^\star\colon \Omega \to \mathbb{R}^d$ be a random vector and $\mathcal{M} \subseteq \mathbb{R}^d$ a smooth $k-$submanifold of $\mathbb{R}^d$ with $k \leq d$. Then $\boldsymbol{X}^\star$ is $\mathcal{M}-$distributed, iff for the probability density P, which is induced by $\boldsymbol{X}^\star$, it holds, that:*

$$P(\boldsymbol{X}^\star = \boldsymbol{x}) > 0 \Leftrightarrow \boldsymbol{x} \in \mathcal{M} \tag{1.36}$$

Thereupon a random vector $\boldsymbol{X}\colon \Omega \to \mathbb{R}^d$ is elliptically $\mathcal{M}$-distributed, iff there exists an $\mathcal{M}$-distributed random vector $\boldsymbol{X}^\star\colon \Omega \to \mathbb{R}^d$ and an elliptically distributed random

error $\boldsymbol{\varepsilon}\colon \Omega \to \mathbb{R}^d$*, such that:*

$$\boldsymbol{X} \sim \boldsymbol{X}^\star + \boldsymbol{\varepsilon} \tag{1.37}$$

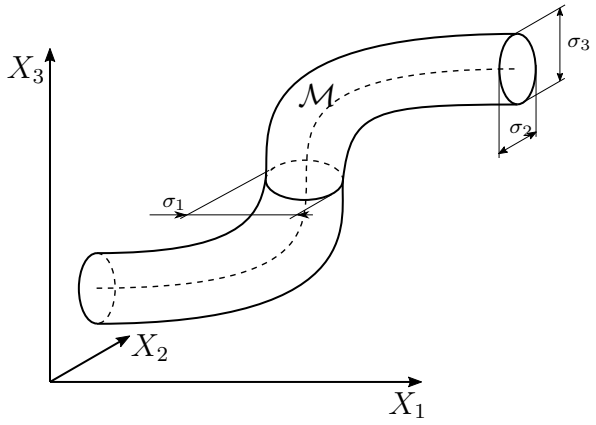

Figure 1.2: Elliptical $\mathcal{M}$-distribution in 3 dimensions

The assumption, that the observed random vector $\boldsymbol{X}$, is elliptically $\mathcal{M}$-distributed, is very general, but allows an estimation of $\mathcal{M}$ by minimizing the averaged Euclidean distance to $\boldsymbol{X}$. Thereby the tangent spaces $T_x\mathcal{M}$ have a basis, given by k principal components of local infinitesimal covariances, such that the remaining $d-k$ principal components describe the normal space $N_x\mathcal{M}$, which is orthogonal to the tangent space $T_x\mathcal{M}$. Since $T_x\mathcal{M}$ and $N_x\mathcal{M}$ are equipped with an induced Riemannian metric, which is simply given by the standard scalar product, there exists a minimal orthogonal projection $\pi_{\mathcal{M}}\colon \mathbb{R}^d \hookrightarrow \mathcal{M}$, that maps any realization $\boldsymbol{x}$ of $\boldsymbol{X}$ to a closest point on $\mathcal{M}$. Then proposition 1.2 motivates properties for $\mathcal{M}$ to minimize the averaged Euclidean distance to realizations of $\boldsymbol{X}$. This provides the definition of smooth k-principal manifolds (Hastie et al. 1989, p513).

Definition (Principal Manifold)**.** *For $d \in \mathbb{N}$ let $\boldsymbol{X}\colon \Omega \to \mathbb{R}^d$ be a random vector, $\mathcal{M} \subseteq \mathbb{R}^d$ a (smooth) k-submanifold of $\mathbb{R}^d$ with $k \leq d$ and $\pi_{\mathcal{M}}\colon \mathbb{R}^d \hookrightarrow \mathcal{M}$ a minimal orthogonal projection onto $\mathcal{M}$. Then $\mathcal{M}$ is a (smooth) k-principal manifold for $\boldsymbol{X}$, iff $\forall \boldsymbol{x} \in \mathcal{M}$ it holds, that:*

$$\mathbb{E}(\boldsymbol{X} \in \pi_{\mathcal{M}}^{-1}(\boldsymbol{x})) = \boldsymbol{x} \tag{1.38}$$

Furthermore $\mathcal{M}$ is termed maximal, iff $\mathcal{M}$ maximizes the explained variance $\mathrm{Var}(\pi_{\mathcal{M}}(\boldsymbol{X}))$.

By extending the local properties of the tangent spaces to the underlying manifold, by propositions 1.2 and 1.3 it can be concluded, that maximal principal manifolds minimize the Euclidean distance to $\boldsymbol{X}$. Intuitively this can be understood as follows: The principal manifold property assures, that:

$$\mathrm{Var}(\boldsymbol{X}) = \mathrm{Var}(\pi_{\mathcal{M}}(\boldsymbol{X})) + \mathrm{Var}(\boldsymbol{X} - \pi_{\mathcal{M}}(\boldsymbol{X}))$$

Consequently the choice of $\mathcal{M}$ maximizes $\mathrm{Var}(\pi_{\mathcal{M}}(\boldsymbol{X}))$ if and only if it minimizes $\mathrm{Var}(\boldsymbol{X} - \pi_{\mathcal{M}}(\boldsymbol{X}))$, which equals the variance of the error and therefore the Euclidean distance. At closer inspection, however, it turns out, that in difference to linear principal manifolds, the maximization problem is ill-defined for arbitrary smooth principal manifolds, since for any finite number of realizations trivial solutions can be found by smooth principal manifolds, that interpolate the realizations and therefore provide a perfect explanation. In order to close this gap, further structural structural assumptions have to be incorporated, ether by a parametric family $\{\boldsymbol{f}_\theta\}_{\theta\in\Theta}$ that restricts the possible solutions - or by a regularization, as given in the elastic map algorithm that penalizes long distances and strong curvature (Gorban et al. 2008). Due to the complexity of this topic, however, it is left to the second chapter, where Energy based models are used to overcome this deficiency. In the following the generalization of the correlation to smooth principal manifolds for convenience is defined with respect to a principal manifold $\mathcal{M}$, which is maximal "with respect to appropriate restrictions".

Definition ($\mathcal{M}$-Correlation)**.** *For $d \in \mathbb{N}$ let $\boldsymbol{X}\colon \Omega \to \mathbb{R}^d$ be a random vector, $\mathcal{M}$ a smooth principal manifold for $\boldsymbol{X}$, which is maximal "with respect to appropriate restrictions" and $\pi_{\mathcal{M}}\colon \mathbb{R}^d \to \mathcal{M}$ a minimal orthogonal projection. Then for any $i, j \in \{1, \dots, d\}$ let the $\mathcal{M}$-Correlation between X_i and X_j be defined by:*

$$\rho^2_{X_i,X_j|\mathcal{M}} := R_i R_j \int_{\mathcal{M}} S_{i,j}(\boldsymbol{x}) S_{j,i}(\boldsymbol{x}) \,\mathrm{d}P_{\mathcal{M}} \tag{1.39}$$

*where for $i \in \{1, \dots, d\}$ the **reliability** of X_i with respect to $\mathcal{M}$ is given by:*

$$R_i := 1 - \frac{\mathrm{Var}_i\left(\boldsymbol{X} - \pi_{\mathcal{M}}(\boldsymbol{X})\right)}{\mathrm{Var}_i(\boldsymbol{X})} \tag{1.40}$$

and for $i, j \in \{1, \ldots, d\}$ *the* ***local sensitivity*** *of* X_i *with respect to* X_j *by:*

$$S_{i,j}(\boldsymbol{x}) \coloneqq \frac{\partial}{\partial x_j}(\boldsymbol{x} - \pi_{\mathcal{M}}(\boldsymbol{x}))\bigg|_i \tag{1.41}$$

Proposition 1.5. *For an elliptical* $\mathcal{M}-$*distributed random vector* $\boldsymbol{X}$ *the* $\mathcal{M}$*-Correlation generalizes the L-Correlation to smooth principal manifolds.*

Proof. Let L be a maximal linear principal manifold for $\boldsymbol{X} \colon \Omega \to \mathbb{R}^d$, and $\boldsymbol{x}$ a realization of $\boldsymbol{X}$ then there exists an $\beta \in \mathbb{R}$ with:

$$S_{i,j}(\boldsymbol{x}) = \frac{\partial}{\partial x_j}(\boldsymbol{x} - \pi_L(\boldsymbol{x}))\bigg|_i \equiv \beta$$

Furthermore for $c \neq 0$:

$$S_{j,i}(\boldsymbol{x}) = \frac{\partial}{\partial x_i}(\boldsymbol{x} - \pi_L(\boldsymbol{x}))\bigg|_j \equiv \frac{1}{\beta}$$

Such that $S_{i,j}(\boldsymbol{x})S_{j,i}(\boldsymbol{x}) = 1$. Consequently for $\mathcal{M} = L$ it follows, that:

$$\begin{aligned}
\rho^2_{X_i,X_j|\mathcal{M}} &= R_i R_j \int_{\mathcal{M}} S_{i,j}(\boldsymbol{x}) S_{j,i}(\boldsymbol{x}) \,\mathrm{d}P_{\mathcal{M}} \\
&= R_i R_j \int_{\mathcal{M}} \mathrm{d}P_{\mathcal{M}} \\
&= R_i R_j = \rho^2_{X_i,X_j|L}
\end{aligned}$$

□

Chapter 2

Estimation of Principal Manifolds by Energy Based Models

Overview

The evaluation of the generalized Pearson correlation requires the estimation of a principal manifold. This chapter is intended to introduce a statistical framework for this estimation, which allows the incorporation of meaningful structural assumptions. In this purpose, in the first section, random vectors with respect to their dependency structure are equipped with a topology, which provides the notation of Markov random field. Subsequently by the Hammersley-Clifford Theorem is shown, that this structural assumption allows a clique factorization of the joint probability distribution. Thereupon in the second section the positivity assumption of $\mathcal{M}$-distributions over their respective manifold $\mathcal{M}$, is incorporated within clique factorized distributions to define a statistical model by an Energy Based Model. Afterwards it is shown, that the resulting model has a canonical representation by a log-linear parametrisation, which allows an explicit representation of the log-likelihood gradient, which finally motivates a local maximum likelihood estimation of a principal manifold by a steepest gradient descent in the negative log-likelihood gradient. The chapter closes by the proof of its convergence under some mild analytical assumptions, regarding the cumulant function of the parametrisation.

2.1 Markov Random Fields

An intuitive and generic statistical framework to model causal relationship structures is given by probabilistic **graphical models**. Thereby the joint probability distribution is modelled by a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the vertices $v \in \mathcal{V}$ of the graph represent random variables X_v and the edges $(u, v) \in \mathcal{E}$ conditional probabilities $P(X_v \mid X_u)$ between the respective random variables. The strength of graphical modelling is revealed by the incorporation of latent variables, such that the modelled random vector $\boldsymbol{X}$ decomposes into observables $\boldsymbol{V}$ and latent random variables $\boldsymbol{H}$ by $\boldsymbol{X} = (\boldsymbol{V}, \boldsymbol{H})$. Then the joint probability distribution $P(\boldsymbol{V}, \boldsymbol{H})$ encodes structural uncertainties of indirect causalities, within the topology of the underlying graph $\mathcal{G}$, which in particular is highly desirable for observations in environments of strong structural uncertainty. Graphical modelling of principal manifolds requires that the errors of the variables are treated symmetrically and therefore, that the graph $\mathcal{G}$ is undirected. This property induces a topological structure to the random vector $\boldsymbol{X}$, which is substantiated by the terminology of **random fields**.

Definition (Random Field)**.** *Let (M, τ) be a topological space and $\boldsymbol{X}$ a random vector, such that M is an index set for the random variables in X. Then $\boldsymbol{X}$ is a random field over (M, τ).*

Notation. The terminology of random fields canonically applies to undirected graphs as follows: Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected graph. Then any edge $(v, u) \in \mathcal{E}$ can be identified with the unit interval in $\mathbb{R}$ and consequently any path $(v_i, u_i)_{1=1}^{n} \in \mathcal{E}^n$ with the length $n \in \mathbb{R}$. Let now for path connected vertices, the function $d\colon \mathcal{V} \times \mathcal{V} \to \mathbb{R}_0^+ \cup \{\infty\}$ denote the length of shortest paths between them and infinity for unconnected vertices. Then d is a distance measure over $\mathcal{V}$ and induces a canonical topology $\mathcal{T}_{\mathcal{G}}$ from $\mathbb{R}$ to $\mathcal{V}$, which makes $(\mathcal{V}, \mathcal{T}_{\mathcal{G}})$ a topological space. Therefore in the following a "*random field $\boldsymbol{X}$ over an undirected graph $\mathcal{G}$*" denotes a random field over the respective topological space $(\mathcal{V}, \mathcal{T}_{\mathcal{G}})$.

Besides the fundamental structural assumption **(S1)**, that the the observables $\boldsymbol{V}$ can be represented as the vertices of an undirected graph $\mathcal{G}$, the representation of the joint probability distribution $P(\boldsymbol{X})$ by an **undirected graphical model** also requires, that the conditional probabilities can be represented by the edges of $\mathcal{G}$. This

makes a further structural assumption **(S2)** necessary, which requires that $\boldsymbol{X}$ satisfies the local Markov property in $\mathcal{G}$. Then $\boldsymbol{X}$ is known as a **Markov random field**.

Definition (Markov Random Field)**.** *Let $\boldsymbol{X}$ be a random field over an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and P the joint probability distribution of $\boldsymbol{X}$. Then $\boldsymbol{X}$ is a Markov Random Field (MRF) over $\mathcal{G}$, iff $\forall v \in \mathcal{V}$, the random variable X_v conditionally only depends on random variables, indexed within the local neighbourhood $\mathcal{N}(v) \subseteq \mathcal{V}$, such that:*

$$P(X_v \mid \{X_i\}_{i \in V \setminus \{v\}}) = P(X_v \mid \{X_i\}_{i \in \mathcal{N}(v)}) \tag{2.1}$$

Although the structural assumptions **(S1)** and **(S2)** for $\boldsymbol{X}$, at first sight, seem very general and not restrictive, it turns out, that they have severe consequences for the joint probability distribution $P(\boldsymbol{X})$. As a preparatory step, it is necessary to reconstitute the notation cliques. Thereby for an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, a clique $c \subseteq \mathcal{V}$ in the graph denotes a completely connected subset of vertices $\mathcal{V}$, such that $\forall i, j \in c \Rightarrow (i, j) \in \mathcal{E}$ (see figure 2.1). This allows to define a **clique factorization** of P over $\mathcal{G}$.

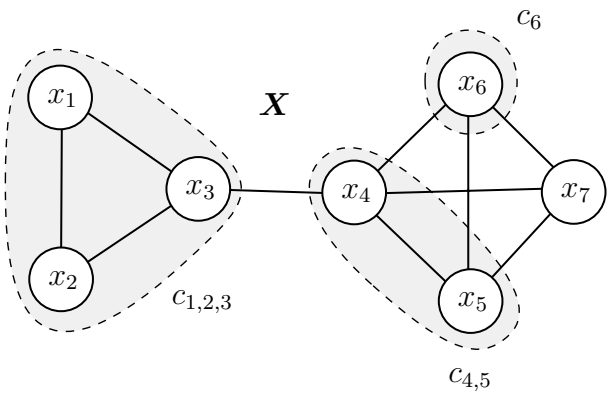

Figure 2.1: Examples for cliques in a Markov Random Field

Definition (Clique factorization)**.** *Let $\boldsymbol{X}$ be a random field over an undirected graph $\mathcal{G}$ with cliques $\mathcal{C}$ and P the probability distribution, which is induced by $\boldsymbol{X}$. Then P factorizes over $\mathcal{G}$, iff there exists a set of positive functions $\{\Phi_c\}_{c \in \mathcal{C}}$, termed **clique potentials**:*

$$\Phi_c \colon \prod_{i \in c} \operatorname{img}(X_i) \to \mathbb{R}_0^+, \ \forall c \in \mathcal{C} \tag{2.2}$$

and a normalization factor $Z \in \mathbb{R}_0^+$*, termed* ***partition function****, such that:*

$$P(\mathbf{X} = \boldsymbol{x}) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \Phi_c(\boldsymbol{x}_c), \text{ with } Z \coloneqq \int_X \prod_{c \in \mathcal{C}} \Phi_c(\boldsymbol{x}_c) \, \mathrm{d}\boldsymbol{x} \tag{2.3}$$

Lemma 2.1. *Let* $\boldsymbol{X}$ *be a random field over an undirected graph* $\mathcal{G}$ *and* P *the joint probability distribution of* $\boldsymbol{X}$*, such that* P *factorizes over* $\mathcal{G}$*. Then* $\boldsymbol{X}$ *is a Markov Random Field over* $\mathcal{G}$*.*

Proof of Lemma 2.1. Let $\mathcal{C}$ be the set of cliques in $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and for any vertex $i \in \mathcal{V}$ let $\mathcal{N}(i) \subset V$ denote the local neighbourhood of i in $\mathcal{G}$. Since P factorizes over $\mathcal{G}$, there exists a set of positive functions $\{\Phi_c\}_{c \in \mathcal{C}}$ with $\Phi_c \colon \prod_{i \in c} \operatorname{img}(X_i) \to \mathbb{R}_0^+$, $\forall c \in \mathcal{C}$, such that for $\mathcal{V}_i^- \coloneqq \mathcal{V} \backslash \mathcal{N}(i) \cup \{i\}$ it holds, that:

$$\begin{aligned} P(X_i \mid \{X_j\}_{j \in \mathcal{N}(i)}) &= \frac{P(X_i, \{X_j\}_{j \in \mathcal{N}(i)})}{P(\{X_j\}_{j \in \mathcal{N}(i)})} \\ &= \frac{\sum_{\mathcal{V}_i^-} \prod_{c \in \mathcal{C}} \Phi_c(\boldsymbol{x}_c)}{\int_{X_i} \sum_{\mathcal{V}_i^-} \prod_{C \in \mathcal{C}} \Phi_c(\boldsymbol{x}_c) \, \mathrm{d}\boldsymbol{x}} \end{aligned} \tag{2.4}$$

Thereupon for $i \in \mathcal{V}$ let $\mathcal{C}_i^+ \coloneqq \{c \in \mathcal{C} \mid i \in c\}$ denote the set of cliques in $\mathcal{G}$, that contain i and $\mathcal{C}_i^- \coloneqq \mathcal{C} \backslash \mathcal{C}_i$ the remaining set of cliques in $\mathcal{G}$, that do not contain i. Then for any $i \in \mathcal{V}$ it holds, that:

$$\prod_{c \in \mathcal{C}} \Phi_c(\boldsymbol{x}_c) = \prod_{c \in \mathcal{C}_i^+} \Phi_c(\boldsymbol{x}_c) \prod_{c \in \mathcal{C}_i^-} \Phi_c(\boldsymbol{x}_c) \tag{2.5}$$

Since by the definition of $\mathcal{V}_i^-$ and $\mathcal{C}_i$ for any clique $c \in \mathcal{C}_i$ it holds, that $c \subseteq \mathcal{V}_i$ and therefore:

$$\begin{aligned} \sum_{\mathcal{V}_i^-} \prod_{c \in \mathcal{C}} \Phi_c(\boldsymbol{x}_c) &\overset{2.5}{=} \sum_{\mathcal{V}_i^-} \prod_{c \in \mathcal{C}_i^+} \Phi_c(\boldsymbol{x}_c) \prod_{c \in \mathcal{C}_i^-} \Phi_c(\boldsymbol{x}_c) \\ &= \prod_{c \in \mathcal{C}_i^+} \Phi_c(\boldsymbol{x}_c) \sum_{\mathcal{V}_i^-} \prod_{c \in \mathcal{C}_i^-} \Phi_c(\boldsymbol{x}_c) \end{aligned} \tag{2.6}$$

Also the factor $\sum_{\mathcal{V}_i^-} \prod_{c \in \mathcal{C}_i^-} \Phi_c(\boldsymbol{x}_c)$ does not involve X_i and therefore can be factored

out, such that:

$$\sum_{\mathcal{V}_i^-} \prod_{c \in \mathcal{C}} \Phi_c(\boldsymbol{x}_c) \overset{2.6}{=} \left(\sum_{\mathcal{V} \setminus \mathcal{V}_i} \prod_{c \in \mathcal{C}_i^-} \Phi_c(\boldsymbol{x}_c) \right) \prod_{c \in \mathcal{C}_i^+} \Phi_c(\boldsymbol{x}_c) \tag{2.7}$$

This also applies to the denominator on the right side of equation ..., such that:

$$\int_{X_i} \sum_{\mathcal{V}_i^-} \prod_{c \in \mathcal{C}} \Phi_c(\boldsymbol{x}_c) \, \mathrm{d}\boldsymbol{x} \overset{2.7}{=} \left(\sum_{\mathcal{V}_i^-} \prod_{c \in \mathcal{C}_i^-} \Phi_c(\boldsymbol{x}_c) \right) \int_{X_i} \prod_{c \in \mathcal{C}_i} \Phi_c(\boldsymbol{x}_c) \, \mathrm{d}\boldsymbol{x} \tag{2.8}$$

From equations 2.4, 2.7 and 2.8 it then follows, that:

$$\begin{aligned} P(X_i \mid \{X_j\}_{j \in \mathcal{N}(i)}) &= \frac{\prod_{c \in \mathcal{C}_i^+} \Phi_c(\boldsymbol{x}_c)}{\int_{X_i} \prod_{c \in \mathcal{C}_i^+} \Phi_c(\boldsymbol{x}_c) \, \mathrm{d}\boldsymbol{x}} \\ &= \frac{\prod_{c \in \mathcal{C}_i^-} \Phi_c(\boldsymbol{x}_c)}{\prod_{c \in \mathcal{C}_i^-} \Phi_c(\boldsymbol{x}_c)} \frac{\prod_{c \in \mathcal{C}_i^+} \Phi_c(\boldsymbol{x}_c)}{\int_{X_i} \prod_{c \in \mathcal{C}_i^+} \Phi_c(\boldsymbol{x}_c) \, \mathrm{d}\boldsymbol{x}} \\ &= \frac{\prod_{c \in \mathcal{C}} \Phi_c(\boldsymbol{x}_c)}{\int_{X_i} \prod_{c \in \mathcal{C}} \Phi_c(\boldsymbol{x}_c) \, \mathrm{d}\boldsymbol{x}} \\ &= \frac{P(\boldsymbol{X})}{P(\{X_j\}_{j \in \mathcal{V}_i^-})} = P(X_i \mid \{X_j\}_{j \in \mathcal{V} \setminus \{i\}}) \end{aligned}$$

□

Theorem 2.1 (Hammersley-Clifford)**.** *Let $\boldsymbol{X}$ be a random field over an undirected graph $\mathcal{G}$ and P the probability distribution, which is induced by $\boldsymbol{X}$. Then the following statements are equivalent:*

(1) *$\boldsymbol{X}$ is a Markov Random Field over $\mathcal{G}$*

(2) *P factorizes over $\mathcal{G}$*

Proof of Theorem 2.1. Since the backward direction has already been proofed by lemma 2.1, it suffices to proof the forward direction. In the following for the undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ a set of functions $\{\Phi_s\}_{s \subseteq \mathcal{V}}$ is constructed, which is shown to satisfy the requirements for a clique factorization. Let P be a the joint probability,

which is induced by $\boldsymbol{X}$ and for any subset $z \subseteq \mathcal{V}$, let π_z be defined by:

$$\pi_z(\boldsymbol{x}_z) \coloneqq P(\{X_i = x_i\}_{i \in z} \mid \{X_i\}_{i \in \mathcal{V} \setminus z}) \mathrm{P}(\{X_i\}_{i \in \mathcal{V} \setminus z}) \tag{2.9}$$

Thereupon for any subset $s \subseteq \mathcal{V}$ let the function $\Phi_s \colon \prod_{i \in s} \mathrm{img}(X_i) \to \mathbb{R}$ be defined by:

$$\Phi_s(\boldsymbol{x}_s) \coloneqq \prod_{z \subseteq s} \pi_z(\boldsymbol{x}_z)^{\sigma(s,z)}, \text{ with } \sigma(s, z) \coloneqq (-1)^{|s|-|z|} \tag{2.10}$$

Then the set of functions $\{\Phi_s\}_{s \subseteq \mathcal{V}}$ satisfies the following requirements: (i) For any $s \subseteq \mathcal{V}$ the function Φ_s is positive, since any factor is positive (ii) The power $\sigma(s, z)$ is 1 if $|s| - |z|$ is even and -1 if $|s| - |z|$ is odd. Now let $z \subseteq \mathcal{V}$. Then π_z occurs as a factor within the function Φ_z, with the power $\sigma(z, z) = 1$. Furthermore π_z occurs as a factor in functions over any subsets of V that contain z and one additional element $a \in \mathcal{V}$ with the power $\sigma(z \cup \{a\}, z) = -1$. Continuing this process it can be seen by the binomial equation:

$$\sum_{k=0}^{n} \tbinom{n}{k} (-1)^k = (1-1)^n = 0, \ \forall 0 \leq k \leq n \tag{2.11}$$

that for any $z \subseteq \mathcal{V}$ the factors π_z are cancelled out within the product ϕ_V, except for the case $z = \mathcal{V}$. In this case $\pi_z = P(\boldsymbol{X})$, such that:

$$P(\boldsymbol{X} = \boldsymbol{x}) = \prod_{s \subseteq V} \Phi_s(\boldsymbol{x}_s) \tag{2.12}$$

Let now be $s \subseteq \mathcal{V}$, such that s is not a clique in $\mathcal{G}$. Then by definition there exist to vertices $i, j \in s$ within s, which are not connected, such that $(i, j) \notin \mathcal{E}$. Consequently the subsets $z \subseteq s$ of s can be distinguished by their inclusion of i and j, such that for any $w \subseteq s \setminus \{i, j\}$ there respectively exist the four subsets w, $w \cup \{i\}$, $w \cup \{j\}$ and $w \cup \{i, j\}$ within s. This allows to rewrite Φ_s as follows:

$$\Phi_s \overset{2.5}{=} \prod_{z \subseteq s} {\pi_z}^{\sigma(s,z)} = \prod_{w \subseteq s \setminus \{i,j\}} \left(\frac{\pi_w}{\pi_{w \cup \{i\}}} \frac{\pi_{w \cup \{i,j\}}}{\pi_{w \cup \{j\}}} \right)^{\sigma(s,z)} \tag{2.13}$$

In the following it will be shown, that $\Phi_s(\boldsymbol{x}_s) = 1$. In the purpose of a shorter representation of the proof, the notations of random variables and their conditional

probability are slightly abbreviated. First, for a subset of vertices $a \subseteq V$ let $\boldsymbol{X}_a \coloneqq \{X_i\}_{i \in a}$. Then the conditional probability is extended by an underline to mark fixed random variables, such that for any a, $b \subseteq V$ with $a \cap b = \emptyset$:

$$P(\boldsymbol{X}_a, \underline{\boldsymbol{X}}_b) \coloneqq P(\boldsymbol{X}_a \mid \boldsymbol{X}_b) P(\boldsymbol{X}_b) \tag{2.14}$$

Then by using Bayes´ Theorem, the factor π_w can be rewritten by the conditional dependency of X_i:

$$\begin{aligned} \pi_w &\overset{\text{def}}{=} P(\boldsymbol{X}_w, \underline{\boldsymbol{X}}_{V \setminus w}) = P(\boldsymbol{X}_w, \underline{\boldsymbol{X}}_{\{i\}}, \underline{\boldsymbol{X}}_{V \setminus (w \cup \{i\})}) \\ &\overset{\text{def}}{=} P(\boldsymbol{X}_w \mid \underline{\boldsymbol{X}}_{\{i\}}, \underline{\boldsymbol{X}}_{V \setminus (w \cup \{i\})}) P(\underline{\boldsymbol{X}}_{\{i\}}, \underline{\boldsymbol{X}}_{V \setminus (w \cup \{i\})}) \\ &\overset{\text{Bayes}}{=} P(\underline{\boldsymbol{X}}_{\{i\}} \mid \boldsymbol{X}_w, \underline{\boldsymbol{X}}_{V \setminus (w \cup \{i\})}) P(\boldsymbol{X}_w, \underline{\boldsymbol{X}}_{V \setminus (w \cup \{i\})}) \end{aligned} \tag{2.15}$$

This representation can also be applied to the factor $\pi_{w \cup \{i\}}$:

$$\begin{aligned} \pi_{w \cup \{i\}} &\overset{\text{def}}{=} P(\boldsymbol{X}_{w \cup \{i\}}, \underline{\boldsymbol{X}}_{V \setminus (w \cup \{i\})}) \\ &\overset{\text{def}}{=} P(\boldsymbol{X}_w, \boldsymbol{X}_{\{i\}} \mid \underline{\boldsymbol{X}}_{V \setminus (w \cup \{i\})}) P(\underline{\boldsymbol{X}}_{V \setminus (w \cup \{i\})}) \\ &\overset{\text{Bayes}}{=} P(\boldsymbol{X}_{\{i\}} \mid \boldsymbol{X}_w, \underline{\boldsymbol{X}}_{V \setminus (w \cup \{i\})}) P(\boldsymbol{X}_w, \underline{\boldsymbol{X}}_{V \setminus (w \cup \{i\})}) \end{aligned} \tag{2.16}$$

It appears, that the priors, given by the factors on the right sides of equations ... and ... are equal. Thereby the random variable X_j is fixed. Since i and j, however, are not connected, the Markov property, given by equation 2.1, requires that X_i and X_j are conditionally independent, if their neighbours are given. Consequently the random variables X_j in equations ... and ... can be chosen freely and it follows, that:

$$\begin{aligned} \frac{\pi_w}{\pi_{w \cup \{i\}}} &= \frac{P(\underline{\boldsymbol{X}}_{\{i\}} \mid \boldsymbol{X}_w, \underline{\boldsymbol{X}}_{V \setminus (w \cup \{i\})}) P(\boldsymbol{X}_w, \underline{\boldsymbol{X}}_{V \setminus (w \cup \{i\})})}{P(\boldsymbol{X}_{\{i\}} \mid \boldsymbol{X}_w, \underline{\boldsymbol{X}}_{V \setminus (w \cup \{i\})}) P(\boldsymbol{X}_w, \underline{\boldsymbol{X}}_{V \setminus (w \cup \{i\})})} \\ &= \frac{P(\underline{\boldsymbol{X}}_{\{i\}} \mid \boldsymbol{X}_{w \cup \{j\}}, \underline{\boldsymbol{X}}_{V \setminus (w \cup \{i,j\})}) P(\boldsymbol{X}_{w \cup \{j\}}, \underline{\boldsymbol{X}}_{V \setminus (w \cup \{i,j\})})}{P(\boldsymbol{X}_{\{i\}} \mid \boldsymbol{X}_{w \cup \{j\}}, \underline{\boldsymbol{X}}_{V \setminus (w \cup \{i,j\})}) P(\boldsymbol{X}_{w \cup \{j\}}, \underline{\boldsymbol{X}}_{V \setminus (w \cup \{i,j\})})} \\ &= \frac{P(\boldsymbol{X}_{w \cup \{j\}}, \underline{\boldsymbol{X}}_{V \setminus (w \cup \{j\})})}{P(\boldsymbol{X}_{w \cup \{j\}}, \underline{\boldsymbol{X}}_{V \setminus (w \cup \{i,j\})})} \overset{\text{def}}{=} \frac{\pi_{w \cup \{j\}}}{\pi_{w \cup \{i,j\}}} \end{aligned} \tag{2.17}$$

Thereupon it immediately follows that:

$$\frac{\pi_w}{\pi_{w\cup\{i\}}}\frac{\pi_{w\cup\{i,j\}}}{\pi_{w\cup\{j\}}} = 1 \tag{2.18}$$

And by equation 2.13 that $\Phi_s(\boldsymbol{x}_s) = 1$, if s is not a clique in $\mathcal{G}$. Let therefore $\mathcal{C}$ be the set of cliques in $\mathcal{G}$, then:

$$P(\boldsymbol{X} = \boldsymbol{x}) \overset{2.12}{=} \prod_{s\subseteq\mathcal{V}} \Phi_s(\boldsymbol{x}_s) = \prod_{c\in\mathcal{C}} \Phi_c(\boldsymbol{x}_c)$$

□

2.2 Energy Based Models

The distribution assumption of an $\mathcal{M}$-distribution, as defined in section 1.3, requires, that the probability distribution is strictly positive over $\mathcal{M}$. If furthermore it is assumed, that all observed random variables are subjected to a Gaussian random error with non-vanishing variances, then any outcome $\boldsymbol{x}$ in the sample space (S, Σ) of $\boldsymbol{X}$ has a strictly positive probability. This motivates a further structural assumption **(S3)**:

$$P(\boldsymbol{X} = \boldsymbol{x}) > 0, \forall \boldsymbol{x} \in S \tag{2.19}$$

By assuming **(S1)** and **(S2)**, and therefore that $\boldsymbol{X}$ is a Markov random field, the Hammersley-Clifford Theorem can be applied and it follows, that the probability distribution P, which is induced by $\boldsymbol{X}$ factorizes over a family of positive functions, given by the clique potentials $\{\Phi_c\}_{c\in\mathcal{C}}$. It can be concluded that, with respect to **(S1)** and **(S2)** the additional assumption **(S3)** is equivalent to requirement that all clique potentials are strictly positive. The strictly positivity of a clique potential, however, in turn is equivalent to the property, that it can be represented by an exponential function $\exp\colon \mathbb{R} \to (0, \infty)$ and therefore P by a **Boltzmann distribution**.

Definition (Boltzmann distribution)**.** *Let (S, Σ) be a measurable space and P a probability density over (S, Σ). Then P is a Boltzmann distribution, iff there exists a scalar function $E\colon S \to \mathbb{R}$, termed an* ***energy function****, and a normalization factor*

$Z \in \mathbb{R}^+$, *termed a* ***partition function****, such that:*

$$P(\boldsymbol{X} = \boldsymbol{x}) = \frac{1}{Z} e^{-E(\boldsymbol{x})}, \text{ with } Z \coloneqq \int_S e^{-E(\boldsymbol{x})} \, \mathrm{d}\boldsymbol{x} \tag{2.20}$$

Remark. The sign convention within the definition of Boltzmann distributions serves no functional purpose, except the compatibility with the terminology, known from statistical physics, which also is referenced by the "partition function" Z and the "energy function" E.

Corollary 2.1. *Let $\boldsymbol{X}$ be a random field over an undirected graph $\mathcal{G}$ and P the induced probability distribution. Then the following statements are equivalent:*

(1) *$\boldsymbol{X}$ is a Markov Random Field over $\mathcal{G}$, such that P is strictly positive*

(2) *P is a Boltzmann distribution, that factorizes over $\mathcal{G}$*

Proof of Corollary 2.1. The Corollary immediately follows as a consequence of the Hammersley-Clifford Theorem:

"$\Longrightarrow$" Let $\mathbf{X}$ be a Markov Random Field over $\mathcal{G}$ and $\mathcal{C}$ the set of cliques in $\mathcal{G}$. Then according to the Hammersley-Clifford Theorem it follows, that there exists a set of clique potentials $\{\Phi_c\}_{c \in \mathcal{C}}$ with $\Phi_c \colon \prod_{i \in c} \mathrm{img}(X_i) \to \mathbb{R}_0^+$, $\forall c \in \mathcal{C}$, such that P factorizes over $\mathcal{G}$:

$$P(\boldsymbol{X} = \boldsymbol{x}) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \Phi_c(\boldsymbol{x}_c) \tag{2.21}$$

If furthermore any outcome $\boldsymbol{x}$ has a positive probability $P(\boldsymbol{X} = \boldsymbol{x})$, then the clique potentials Φ_c have to be strictly positive. This allows to define an energy function as the sum of logarithmic clique potentials:

$$E(\boldsymbol{x}) \coloneqq -\sum_{c \in \mathcal{C}} \log \Phi_c(\boldsymbol{x}_c) \tag{2.22}$$

Since $P(\boldsymbol{X} = \boldsymbol{x}) = \exp(-E(\boldsymbol{x}))/Z$ it follows, that P is a Boltzmann distribution.

"$\Longleftarrow$" Let P factorize over $\mathcal{G}$, then according to the Hammersley-Clifford Theorem it follows, that $\boldsymbol{X}$ is a Markov Random Field over $\mathcal{G}$. If furthermore P is a Boltzmann distribution, then by definition P is strictly positive. □

By the use Corollary 2.1, the positivity assumption **(S3)** can immediately be incorporated within a statistical model: Let $\boldsymbol{X}$ be a MRF over an undirected graph $\mathcal{G}$, such that the induced probability P is strictly positive. Then according to corollary 2.1 P is given by a Boltzmann distribution, that factorizes over $\mathcal{G}$. Consequently the model space $\mathcal{M}$ of a statistical model for $\boldsymbol{X}$, can be described by a family of energy functions, which is therefore referred as an **Energy Based Model**.

Definition (Energy Based Model)**.** *Let $\boldsymbol{X}$ be an (S, Σ)-valued random field over a finite undirected graph $\mathcal{G}$ and $(S, \Sigma, \mathcal{P})$ a statistical model for $\boldsymbol{X}$. Then $(S, \Sigma, \mathcal{P})$ is an Energy Based Model (EBM) for $\boldsymbol{X}$ over $\mathcal{G}$, iff for all $P \in \mathcal{P}$ it holds, that (i) P factorizes over $\mathcal{G}$ and (ii) P is a Boltzmann distribution.*

Example 2.1 (Boltzmann Machine)**.** *In the contemporary literature about probabilistic graphical models the terminology of EBMs and Boltzmann Machines is frequently used equivalent to emphasize the underlying Boltzmann distribution. In the narrow sense of its original context, however, the **Boltzmann Machine** denotes a specific statistical model, that has been introduced to study the properties of stochastic Hopfield-Networks (Fahlman et al. 1983). In this context for a given undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with n vertices, any vertex $i \in \mathcal{V}$ represents a thresholded binary neuron with an activation threshold $b_i \in \mathbb{R}$ and any edge $(i, j) \in \mathcal{E}$ a synapse with a synaptic weight $w_{i,j} \in \mathbb{R}$. With regard to the parameters $W \in \mathbb{R}^{n \times n}$ and $\boldsymbol{b} \in \mathbb{R}^n$, the energy function is the given by :*

$$E_{W,\boldsymbol{b}}(\boldsymbol{x}) = -\left(\frac{1}{2}\boldsymbol{x}^T W \boldsymbol{x} + \boldsymbol{b}^T \boldsymbol{x}\right) \tag{2.23}$$

Thereupon the conditional probabilities can be derived from the energy function. Thereby the contribution of a single neuron $i \in \mathcal{V}$ to the total energy sums up to:

$$\Delta E_i = \frac{1}{2}\sum_{j=1} w_{i,j} x_j + b_i \tag{2.24}$$

Due to the underlying Boltzmann distribution it then follows, that:

$$P_{W,\boldsymbol{b}}(X_i = 1 \mid \{X_j = x_j\}_{j \neq i}) = \frac{1}{1 + \exp(\Delta E_i)} =: \operatorname{sigm}(\Delta E_i) \tag{2.25}$$

In comparison to more generic undirected graphical models, the structure of EBMs allows unconstrained statistical inference with the space of energy functions, which in difference to the clique potentials are not required to be positive. In order to take advantage of this circumstance, however, a parametrisation is required. This parametrisation is given by a **log-linear parametrisation**.

Definition (Log-linear parametrisation)**.** *Let* $(S, \Sigma, \mathcal{P})$ *be an EBM for* $\boldsymbol{X}$ *over* $\mathcal{G}$ *and* $P\colon \Theta \to \mathcal{P}$ *a parametrisation for* $\mathcal{P}$*. Then* P *is a* ***log-linear parametrisation*** *for* $\mathcal{P}$*, iff there exists a sufficient statistic*$\boldsymbol{\phi}$ *for* $\boldsymbol{X}$*, termed* ***feature function****, such that* $\forall \boldsymbol{\theta} \in \Theta$*:*

$$P_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x}) = \exp\left(\boldsymbol{\theta}^T \boldsymbol{\phi}(\boldsymbol{x}) - \Lambda(\boldsymbol{\theta})\right), \forall \boldsymbol{x} \in S \tag{2.26}$$

where the ***cumulant function*** $\Lambda\colon \Theta \to \mathbb{R}$ *is given by:*

$$\Lambda(\boldsymbol{\theta}) \coloneqq \log \int_S \exp\left(\boldsymbol{\theta}^T \boldsymbol{\phi}(\boldsymbol{x})\right) \mathrm{d}\boldsymbol{x} \tag{2.27}$$

The existence of the log-linear parametrization for EBMs over MRFs then immediately follows as a consequence of the Hamersley-Clifford theorem.

Lemma 2.2. *Let* $(S, \Sigma, \mathcal{P})$ *be an EBM for* $\boldsymbol{X}$ *over* $\mathcal{G}$ *and let* $\mathcal{P}$ *have a finite dimensional parametrisation* $P^\star\colon \Theta \to \mathcal{M}$*. Then there exists a log-linear parametrisation for* $\mathcal{P}$*.*

Proof of Lemma 2.2. The proof is constructive: Let $\mathcal{C}$ be the set of cliques in $\mathcal{G}$. According to equations 2.21 and 2.22 it follows, that any probability distribution $Q \in \mathcal{P}$ is identified by a unique energy function E_Q, which can be written as a sum of logarithmic clique potentials, such that:

$$E_Q(\boldsymbol{x}) = -\sum_{c \in \mathcal{C}} \log \Phi_{Q,c}(\boldsymbol{x}_c) \tag{2.28}$$

Consequently a parametrisation for $\mathcal{P}$ can be chosen, that decomposes in individual parametrisations for the cliques of the graph. Let now be $c \in \mathcal{C}$ a fixed clique and $n_{P^\star} \in \mathbb{N}$ the dimension of the finite dimension parametrisation $P^\star$. Then, since any $Q \in \mathcal{P}$ factorizes over $\mathcal{G}$, it can be concluded, that any family of clique potentials $\{\Phi_{Q,c}\}_{Q \in \mathcal{P}}$ has a minimal parametrisation of finite dimension $n_c < n_{P^\star}$. Further-

more since any potential function is strictly positive, there exists a positive function $\boldsymbol{\phi}_c \colon \prod_{i \in c} \operatorname{img}(X_i) \to \mathbb{R}^{n_c}$, and a parameter space $\Theta_c \subseteq \mathbb{R}^{n_c}$, such that for all $Q \in \mathcal{P}$, there exists a vector $\boldsymbol{\theta}_c \in \Theta_c$ with:

$$\Phi_{Q,c}(\boldsymbol{x}_c) = \exp(\boldsymbol{\theta}_c^{\mathrm{T}} \boldsymbol{\phi}_c(\boldsymbol{x}_c)) \tag{2.29}$$

Consequently there exist parameter spaces Θ_c and functions $\boldsymbol{\phi}_c$, such that for all $Q \in \mathcal{P}$ the corresponding energy function E_Q can be identified by a parameter $\boldsymbol{\theta} \in \prod_{c \in \mathcal{C}} \Theta_c$ with:

$$E_Q(\boldsymbol{x}) = -\sum_{c \in \mathcal{C}} \log \Phi_{Q,c}(\boldsymbol{x}_c) = -\sum_{c \in \mathcal{C}} \boldsymbol{\theta}_c^T \boldsymbol{\phi}_c(\boldsymbol{x}_c) \tag{2.30}$$

Since the energy functions E_Q in turn identify all $Q \in \mathcal{P}$, it follows, that for any $Q \in \mathcal{P}$, there exist parameters $\boldsymbol{\theta}_c \in \Theta_c$, termed **clique parameters**, such that $\forall \boldsymbol{x} \in S$:

$$Q(\boldsymbol{X} = \boldsymbol{x}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\left(\sum_{c \in \mathcal{C}} \boldsymbol{\theta}_c^T \boldsymbol{\phi}_c(\boldsymbol{x}_c) \right) \tag{2.31}$$

Let therefore $\Lambda(\boldsymbol{\theta}) \coloneqq \log Z(\boldsymbol{\theta})$, $\boldsymbol{\theta} \coloneqq \prod_{c \in C} \boldsymbol{\theta}_c$ and $\boldsymbol{\phi} \coloneqq \prod_{c \in C} \boldsymbol{\phi}_c$ and the parametrisation $P \colon \Theta \to \mathcal{P}$ be given by:

$$P_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x}) \coloneqq \exp\left(\boldsymbol{\theta}^T \boldsymbol{\phi}(\boldsymbol{x}) - \Lambda(\boldsymbol{\theta}) \right), \forall \boldsymbol{x} \in S$$

Then (i) P is a parametrisation for $\mathcal{M}$, since $\operatorname{img}(P) = \mathcal{P}$ and (ii) P is a log-linear parametrisation. □

In the following the log-linear parametrisation will be used for observation based statistical inference in EBMs. Let $\boldsymbol{X}$ be a MRF over a graph $\mathcal{G}$ with d vertices and let $\boldsymbol{x}$ be an i.i.d realization of $\boldsymbol{X}$ of length $n \in \mathbb{N}$, such that $\boldsymbol{x} \in \mathbb{R}^{n \times d}$. Furthermore let $(S, \Sigma, \mathcal{P})$ be an EBM for $\boldsymbol{X}$ with a log-linear parametrisation $P \colon \Theta \to \mathcal{P}$. Then the question arises, which parameters $\hat{\boldsymbol{\theta}} \in \Theta$ "best possibly" explain the observation of $\boldsymbol{x}$, which provides an estimation for the "true parameter". Thereby in the absence of further knowledge, given by a prior distribution, the estimation $\hat{\boldsymbol{\theta}}$ is required to

globally maximize the likelihood function:

$$L_{\boldsymbol{x}}(\boldsymbol{\theta}) \coloneqq \prod_{i=1}^{n} P_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x}^i)$$

Then the set of all maximum likelihood estimations constitutes a subset $\Theta_{\mathrm{ML}} \subseteq \Theta$. Nevertheless, such a global search can be quite challenging if $L_{\boldsymbol{x}}$ is non-concave. Then the determination of a global maximum would require to evaluate $L_{\boldsymbol{x}}$ for any single parameter in Θ. For the case, however, that $L_{\boldsymbol{x}} \in C^2(\Theta)$, then at least the local maximum likelihood estimations $\hat{\boldsymbol{\theta}} \in \Theta_{\max}$ are locally determined by the well-known criteria:

(ML1) $L_{\boldsymbol{x}}$ has a critical point at$\hat{\boldsymbol{\theta}}$, such that $\nabla L_{\boldsymbol{x}}(\hat{\boldsymbol{\theta}}) = 0$

(ML2) $L_{\boldsymbol{x}}$ has a negative curvature at $\hat{\boldsymbol{\theta}}$, such that $\nabla^2 L_{\boldsymbol{x}}(\hat{\boldsymbol{\theta}}) < 0$

These criteria, however, are more conveniently equated by the log-likelihood function $\log L_{\boldsymbol{x}}$, which has identical critical points and sign of curvature, but in difference to $L_{\boldsymbol{x}}$, decomposes into a sum of independent point realizations:

$$\begin{aligned} \log L_{\boldsymbol{x}}(\boldsymbol{\theta}) &\overset{\text{i.i.d.}}{=} \log \prod_{i=1}^{d} P_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x}^i) \\ &= \sum_{i=1}^{n} \log L_{\boldsymbol{x}^i}(\boldsymbol{\theta}),\ \forall \boldsymbol{\theta} \in \Theta \end{aligned} \tag{2.32}$$

Let now be $\mathcal{C}$ the set of cliques in $\mathcal{G}$ and the log-linear parametrisation P be given by clique parameters $\{\boldsymbol{\theta}_c\}_{c \in C}$, such that $\Theta = \prod_{c \in \mathcal{C}} \Theta_c$. Then the gradient $\nabla \log L_{\boldsymbol{x}}$ decomposes into a sum of partial derivatives $\partial_c \log L_{\boldsymbol{x}}$ with $\partial_c \coloneqq \partial / \partial \boldsymbol{\theta}_c$, such that:

$$\nabla \log L_{\boldsymbol{x}}(\boldsymbol{\theta}) = \sum_{c \in \mathcal{C}} \partial_c \log L_{\boldsymbol{x}}(\boldsymbol{\theta}) \tag{2.33}$$

The log-linear parametrisation can then be used to derive an energy representation of the partial derivatives $\partial_c \log L_{\boldsymbol{x}}$.

Proposition 2.1. *Let $(S, \Sigma, \mathcal{P})$ be an EBM for $\boldsymbol{X}$ over $\mathcal{G}$ with cliques $\mathcal{C}$ and $P\colon \Theta \to \mathcal{M}$ a differentiable log-linear parametrization with clique parameters $\{\boldsymbol{\theta}_c\}_{c \in \mathcal{C}}$ and energy functions $\{E_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta} \in \Theta}$. Then for a realization $\boldsymbol{x}$ of $\boldsymbol{X}$ of length $n \in \mathbb{N}$ and a clique*

*$c \in \mathcal{C}$ the partial derivative $\partial_c \log L_{\boldsymbol{x}}$ has an **energy representation**, which for $\boldsymbol{\theta} \in \Theta$ is given by:*

$$\partial_c \log L_{\boldsymbol{x}}(\boldsymbol{\theta}) = n \left\langle -\partial_c \log E_{\boldsymbol{\theta}} \right\rangle_{\mathrm{d}} - n \left\langle -\partial_c \log E_{\boldsymbol{\theta}} \right\rangle_{\mathrm{m}} \tag{2.34}$$

*where $\langle \cdot \rangle_{\mathrm{d}}$ denotes an expectation over S, with respect to the empirical **data distribution** $P_{\boldsymbol{x}}$, given by:*

$$P_{\boldsymbol{x}}(\boldsymbol{X}_c = \boldsymbol{y}_c) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(\boldsymbol{y}_c - \boldsymbol{x}_c^i), \text{ with } \mathbb{I}(\boldsymbol{x}) = \begin{cases} 1 & \boldsymbol{x} = \mathbf{0} \\ 0 & else \end{cases} \tag{2.35}$$

*and $\langle \cdot \rangle_{\mathrm{m}}$ an expectation over S, with respect to the **model distribution** $P_{\boldsymbol{\theta}}$*

Proof of Proposition 2.1. The Proposition is proofed directly by calculation.

Calculation of $\log L_{\boldsymbol{x}}$ For $\boldsymbol{\theta} \in \Theta$ the log likelihood of $\boldsymbol{\theta}$ with respect to $\boldsymbol{x}$ is calculated by:

$$\log L_{\boldsymbol{x}}(\boldsymbol{\theta}) \overset{2.32}{=} \sum_{i=1}^{n} \log L_{\boldsymbol{x}^i}(\boldsymbol{\theta}) \tag{2.36}$$

Thereupon for any single realization $\boldsymbol{x}^i$ the log-likelihood equates to:

$$\begin{aligned} \log L_{\boldsymbol{x}^i}(\boldsymbol{\theta}) &\overset{\text{def}}{=} \log P_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x}^i) && (2.37) \\ &\overset{\text{def}}{=} \log \exp \left(\sum_{c \in \mathcal{C}} \boldsymbol{\theta}_c^T \boldsymbol{\phi}_c(\boldsymbol{x}_c^i) - \Lambda(\boldsymbol{\theta}) \right) && (2.38) \\ &= \sum_{c \in \mathcal{C}} \boldsymbol{\theta}_c^T \boldsymbol{\phi}_c(\boldsymbol{x}_c^i) - \Lambda(\boldsymbol{\theta}) \end{aligned}$$

A substitution of equation 2.37 in 2.36 then gives:

$$\log L_{\boldsymbol{x}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{c \in \mathcal{C}} \boldsymbol{\theta}_c^T \boldsymbol{\phi}_c(\boldsymbol{x}_c^i) - \sum_{i=1}^{n} \Lambda(\boldsymbol{\theta}) \tag{2.39}$$

Calculation of $\partial_c \Lambda(\boldsymbol{\theta})$ Let now be $c \in \mathcal{C}$ a fixed clique in $\mathcal{G}$. Then $\partial_c \Lambda(\boldsymbol{\theta})$ is calculated by the chain rule:

$$\partial_c \Lambda(\boldsymbol{\theta}) = \partial_c \log Z(\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \partial_c Z(\boldsymbol{\theta}),\ \forall \boldsymbol{\theta} \in \Theta \tag{2.40}$$

With regard to the corresponding clique potential Φ_c, given by:

$$\Phi_c(\boldsymbol{x}_c,\, \boldsymbol{\theta}_c) \coloneqq \exp\left(\boldsymbol{\theta}_c^T \boldsymbol{\phi}_c(\boldsymbol{x}_c)\right),\ \forall \boldsymbol{x} \in S \tag{2.41}$$

it follows, that:

$$\begin{aligned} \prod_{z \in \mathcal{C} \setminus \{c\}} \Phi_z(\boldsymbol{x}_z,\, \boldsymbol{\theta}_z) &= \frac{Z(\boldsymbol{\theta})}{\phi_z(\boldsymbol{x}_z,\, \boldsymbol{\theta}_z)} \frac{1}{Z(\boldsymbol{\theta})} \prod_{z \in \mathcal{C}} \Phi_z(\boldsymbol{x}_z,\, \boldsymbol{\theta}_z) \\ &= \frac{Z(\boldsymbol{\theta})}{\Phi_z(\boldsymbol{x}_z,\, \boldsymbol{\theta}_z)} P_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x}_s) \end{aligned} \tag{2.42}$$

This allows to calculate $\partial_c Z(\boldsymbol{\theta})$ by:

$$\begin{aligned} \partial_c Z(\boldsymbol{\theta}) &= \int_S \partial_c \prod_{z \in \mathcal{C}} \Phi_z(\boldsymbol{x}_z,\, \boldsymbol{\theta}_z)\, \mathrm{d}\boldsymbol{x} \\ &= \int_S \left(\partial_c \Phi_c(\boldsymbol{x}_c,\, \boldsymbol{\theta}_c)\right) \prod_{z \in \mathcal{C} \setminus \{c\}} \Phi_z(\boldsymbol{x}_z,\, \boldsymbol{\theta}_z)\, \mathrm{d}\boldsymbol{x} \\ &= \int_S \left(\partial_c \Phi_c(\boldsymbol{x}_c,\, \boldsymbol{\theta}_c)\right) \frac{Z(\boldsymbol{\theta})}{\Phi_z(\boldsymbol{x}_z,\, \boldsymbol{\theta}_z)} P_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x}_s)\, \mathrm{d}\boldsymbol{x} \end{aligned} \tag{2.43}$$

This term can be simplified by a reverse application of the chain rule, such that:

$$\begin{aligned} \partial_c Z(\boldsymbol{\theta}) &= Z(\boldsymbol{\theta}) \int_S \partial_c \log \Phi_c(\boldsymbol{x}_c,\, \boldsymbol{\theta}_c) P_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x}_s)\, \mathrm{d}\boldsymbol{x} \\ &\overset{\text{def}}{=} Z(\boldsymbol{\theta}) \left\langle \partial_c \log \Phi_c(\boldsymbol{\theta}_c) \right\rangle_{\mathrm{m}} \\ &= Z(\boldsymbol{\theta}) \left\langle \boldsymbol{\phi}_c \right\rangle_{\mathrm{m}} \end{aligned} \tag{2.44}$$

A final substitution of equation 2.44 in equation 2.40 then gives:

$$\partial_c \Lambda(\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} Z(\boldsymbol{\theta}) \left\langle \boldsymbol{\phi}_c \right\rangle_{\mathrm{m}} = \left\langle \boldsymbol{\phi}_c \right\rangle_{\mathrm{m}} \tag{2.45}$$

Calculation of $\partial_c \log \mathrm{L}_{\boldsymbol{x}}$ From equation 2.39 it follows, that:

$$\begin{aligned} \partial_c \log L_{\boldsymbol{x}}(\boldsymbol{\theta}) &\overset{2.39}{=} \sum_{i=1}^{n} \boldsymbol{\phi}_c(\boldsymbol{x}_c^i) - \sum_{i=1}^{n} \partial_c \Lambda(\boldsymbol{\theta}) \\ &\overset{2.45}{=} \sum_{i=1}^{n} \boldsymbol{\phi}_c(\boldsymbol{x}_c^i) - n \left\langle \boldsymbol{\phi}_c \right\rangle_{\mathrm{m}} \end{aligned} \tag{2.46}$$

The sum on the right side of equation 2.46 can be rewritten by the empirical distribution:

$$\begin{aligned} \sum_{i=1}^{n} \boldsymbol{\phi}_c(\boldsymbol{x}_c^i) &= \int_S \boldsymbol{\phi}_c(\boldsymbol{x}_c) \mathbb{I}(\boldsymbol{x}_c - \boldsymbol{x}_c^i) \, \mathrm{d}\boldsymbol{x} \\ &\overset{\mathrm{def}}{=} \int_S \boldsymbol{\phi}_c(\boldsymbol{x}_c) n P_{\boldsymbol{x}}(\boldsymbol{X}_c = \boldsymbol{x}_c) \, \mathrm{d}\boldsymbol{x} \\ &\overset{\mathrm{def}}{=} n \left\langle \boldsymbol{\phi}_c \right\rangle_{\mathrm{d}} \end{aligned} \tag{2.47}$$

Then a substitution of equation 2.47 in equation 2.46 yields:

$$\partial_c \log \mathrm{L}_{\boldsymbol{x}}(\boldsymbol{\theta}) = n \left\langle \boldsymbol{\phi}_c \right\rangle_{\mathrm{d}} - n \left\langle \boldsymbol{\phi}_c \right\rangle_{\mathrm{m}} \tag{2.48}$$

Finally due to the identity:

$$\partial_c \log E_{\boldsymbol{\theta}}(\boldsymbol{x}) = -\boldsymbol{\phi}_c(\boldsymbol{x}_c)$$

it follows, that:

$$\partial_c \log L_{\boldsymbol{x}}(\boldsymbol{\theta}) = n \left\langle -\partial_c \log E_{\boldsymbol{\theta}} \right\rangle_{\mathrm{d}} - n \left\langle -\partial_c \log E_{\boldsymbol{\theta}} \right\rangle_{\mathrm{m}} \tag{2.49}$$

□

Example (*Supplement to Example 2.1*)**.** *With respect to its log-linear parametrisation it follows, that two kinds of cliques are to be distinguished in Boltzmann Machines: For any clique, that is given by a single vertex $c = \{i\} \subseteq \mathcal{V}$, the clique parameter is given by the activation threshold b_i and the feature function by $\phi_i(x_i) = x_i$, such that:*

$$-\frac{\partial}{\partial b_i} \log E_{W,\boldsymbol{b}}(\boldsymbol{x}) \overset{2.23}{=} x_i$$

Therefore according to Proposition 2.1 it follows, that:

$$\frac{\partial}{\partial b_i} \log L_{\boldsymbol{x}}(W, \boldsymbol{b}) = n \langle x_i \rangle_{\mathrm{d}} - n \langle x_i \rangle_{\mathrm{m}} \tag{2.50}$$

Furthermore for a clique, that is given by a pair $c = \{i, j\} \subseteq \mathcal{V}$, with $(i, j) \in \mathcal{E}$ the clique parameter is given by the synaptic weight w_{ij} and the feature function by $\phi_{ij}(x_i, x_j) = x_i x_j$, such that:

$$-\frac{\partial}{\partial w_{ij}} \log E_{W,\boldsymbol{b}}(\boldsymbol{x}) \overset{2.23}{=} x_i x_j$$

And therefore:

$$\frac{\partial}{\partial w_{ij}} \log L_{\boldsymbol{x}}(W, \boldsymbol{b}) = n \langle x_i x_j \rangle_{\mathrm{d}} - n \langle x_i x_j \rangle_{\mathrm{m}} \tag{2.51}$$

Due to the definition of the empirical data distribution $P_{\boldsymbol{x}}$ with respect to a realization $\boldsymbol{x}$ of $\boldsymbol{X}$, the calculation of $\partial_c \log L_{\boldsymbol{x}}$, by its energy representation of Proposition 2.1 requires, that all random variables, which are represented by the clique $c \in \mathcal{C}$ are realized. The mandatory next step therefore regards the generalization of this representation to latent random variables. In this purpose the underlying MRF $\boldsymbol{X}$ is decomposed into an observable component $\boldsymbol{V}$ and a latent component $\boldsymbol{H}$, such that $\boldsymbol{X} = (\boldsymbol{V}, \boldsymbol{H})$. Thereupon a straight forward approach to capture the latent component within the representation is to generalize the energy functions by **free energy functions**.

Corollary 2.2. *Let $(S, \Sigma, \mathcal{P})$ be an EBM for $\boldsymbol{X}$ over $\mathcal{G}$ with cliques $\mathcal{C}$ and a differentiable log-linear parametrization $P\colon \Theta \to \mathcal{P}$ with clique parameters $\{\boldsymbol{\theta}_c\}_{c \in \mathcal{C}}$ and energy functions $\{E_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta} \in \Theta}$. Furthermore let the $\boldsymbol{X} = (\boldsymbol{V}, \boldsymbol{H})$ with $\mathrm{img}\boldsymbol{V} = (S_V, \Sigma_V)$ and $\mathrm{img}\boldsymbol{H} = (S_H, \Sigma_H)$. Then for any $\boldsymbol{\theta} \in \Theta$ let the **free energy function** $F_{\boldsymbol{\theta}}\colon S_V \to \mathbb{R}$ be defined by:*

$$F_{\boldsymbol{\theta}}(\boldsymbol{v}) \coloneqq -\log \int_{S_H} e^{-E_{\boldsymbol{\theta}}(\boldsymbol{v}, \boldsymbol{h})} \, \mathrm{d}\boldsymbol{h}, \ \forall \boldsymbol{v} \in S_V \tag{2.52}$$

*Then for an i.i.d realization $\boldsymbol{v}$ of $\boldsymbol{V}$ of length n and for a clique $c \in \mathcal{C}$ the partial derivative $\partial_c \log L_{\boldsymbol{v}}$ has a **free energy representation**, which at $\boldsymbol{\theta} \in \Theta$ is given by:*

$$\partial_c \log L_{\boldsymbol{v}}(\boldsymbol{\theta}) = n \langle -\partial_c \log F_{\boldsymbol{\theta}} \rangle_{\mathrm{d}} - n \langle -\partial_c \log F_{\boldsymbol{\theta}} \rangle_{\mathrm{m}} \tag{2.53}$$

Proof of Corollary 2.2. The likelihood $L_{\boldsymbol{v}}(\boldsymbol{\theta})$ is identified by the marginal probability $P_{\boldsymbol{\theta}}(\boldsymbol{V} = \boldsymbol{v})$, which by the law of total probability is given by:

$$\begin{aligned} P_{\boldsymbol{\theta}}(\boldsymbol{V} = \boldsymbol{v}) &= \int_{S_H} P_{\boldsymbol{\theta}}(\boldsymbol{V} = \boldsymbol{v},\, \boldsymbol{H} = \boldsymbol{h})\, \mathrm{d}\boldsymbol{h} \\ &\overset{\text{def}}{=} \frac{1}{Z(\boldsymbol{\theta})} \int_{S_H} e^{-E_{\boldsymbol{\theta}}(\boldsymbol{v},\boldsymbol{h})}\, \mathrm{d}\boldsymbol{h}, \text{ with } Z(\boldsymbol{\theta}) \coloneqq \int_S e^{-E_{\boldsymbol{\theta}}(\boldsymbol{x})}\, \mathrm{d}\boldsymbol{x} \\ &\overset{2.52}{=} \frac{1}{Z'(\boldsymbol{\theta})} e^{-F_{\boldsymbol{\theta}}(\boldsymbol{v})}, \text{ with } Z'(\boldsymbol{\theta}) \coloneqq \int_V e^{-F_{\boldsymbol{\theta}}(\boldsymbol{v})}\, \mathrm{d}\boldsymbol{v} \end{aligned} \tag{2.54}$$

This shows, that the marginal distributions are Boltzmann distributions, where the energy functions are given by the free energy of the joint probability distribution $P_{\boldsymbol{\theta}}(\boldsymbol{V},\, \boldsymbol{H})$. Since furthermore the parametrisation $P_{\boldsymbol{\theta}}(\boldsymbol{V},\, \boldsymbol{H})$ is assumed to be differentiable over Θ it follows, that also $P_{\boldsymbol{\theta}}(\boldsymbol{V})$ is differentiable over Θ and Proposition 2.1 can be applied, such that:

$$\partial_c \log L_{\boldsymbol{v}}(\boldsymbol{\theta}) = n \left\langle -\partial_c \log F_{\boldsymbol{\theta}} \right\rangle_{\mathrm{d}} - n \left\langle -\partial_c \log F_{\boldsymbol{\theta}} \right\rangle_{\mathrm{m}} \tag{2.55}$$

□

The free energy representation for EBMs allows to derive the partial derivatives $\partial_c \log L_{\boldsymbol{v}}$ with respect to all clique parameters $\boldsymbol{\theta}_c$. Let therefore P be a differentiable log-linear parametrisation with $\Theta = \prod_{c \in \mathcal{C}} \Theta_c$. Then according to equations 2.33 and 2.53 the gradient $\nabla \log L_{\boldsymbol{v}}$ has a free energy representation, which is given by:

$$\nabla \log L_{\boldsymbol{v}}(\boldsymbol{\theta}) = n \sum_{c \in \mathcal{C}} \left(\left\langle -\partial_c \log F_{\boldsymbol{\theta}} \right\rangle_{\mathrm{d}} - \left\langle -\partial_c \log F_{\boldsymbol{\theta}} \right\rangle_{\mathrm{m}} \right) \tag{2.56}$$

This representation can be used to evaluate the criteria **(ML1)** and **(ML2)** for local maximum likelihood. Furthermore, however, the gradient $\nabla \log L_{\boldsymbol{v}}$ identifies the direction that maximizes the increase of $\log L_{\boldsymbol{v}}$, which motivates a traversal of the model space $\mathcal{P}$ along the gradient. For a discretization of the curve in a finite number of steps, which are proportional to the norm of the gradient, this strategy provides the **Steepest Gradient Descent** (see algorithm 2.1).

Theorem 2.2 (Convergence of `GD` in EBMs)**.** *Let* $(S, \Sigma, \mathcal{P})$ *be an EBM for* $\boldsymbol{X}$ *over*

Algorithm 2.1 Steepest Gradient Descent in EBMs

parameters
prior parameter $\boldsymbol{\theta} \in \Theta$
step size $\eta > 0$
number of steps $N \in \mathbb{N}$

procedure $\textsc{gd}(\boldsymbol{\theta}, \eta, N)$
  $\boldsymbol{\theta}^{(1)} \leftarrow \boldsymbol{\theta}$
  for $i \in \{1, \dots, N-1\}$ **do**
    for $c \in \mathcal{C}$ **do**

$$\boldsymbol{\theta}_c^{(i+1)} \leftarrow \boldsymbol{\theta}_c^{(i)} - \eta d \left(\langle -\partial_c \log F_{\boldsymbol{\theta}} \rangle_{\mathrm{d}} - \langle -\partial_c \log F_{\boldsymbol{\theta}} \rangle_{\mathrm{m}} \right)$$

    end for
  end for
  return $\boldsymbol{\theta}^{(N)}$
end procedure

$\mathcal{G}$ with a log-linear parametrization $P\colon \Theta \to \mathcal{M}$ with a cumulant function Λ, that satisfies:

(A1) *$\Lambda$is two times continuously differentiable, i.e. $\Lambda \in C^2(\Theta)$*

(A2) *Λ is pseudo concave*

(A3) *$\nabla\Lambda$ is Lipschitz continuous*

Furthermore for $\boldsymbol{X} = (\boldsymbol{V}, \boldsymbol{H})$ let $\boldsymbol{v}$ be an i.i.d. realization of $\boldsymbol{V}$ and let $\Theta_{\min}$ and $\Theta_{\max}$ respectively denote the critical points of $L_{\boldsymbol{v}}(\boldsymbol{\theta}) \coloneqq P_{\boldsymbol{\theta}}(\boldsymbol{v})$. Then for all prior parameters $\boldsymbol{\theta} \in \Theta \backslash \Theta_{\min}$, there exists a step size $\eta > 0$ and a local ML estimation $\hat{\boldsymbol{\theta}} \in \Theta_{\max}$, such that:

$$\textsc{gd}(\boldsymbol{\theta}, \eta, N) \to \hat{\boldsymbol{\theta}}, \text{ for } N \to \infty$$

Proof of Theorem 2.2. According to **(A1)** the gradient $\nabla\Lambda$ and the Hessian $\nabla^2\Lambda$ are well defined over Θ, and therefore also $\nabla \log L_{\boldsymbol{v}}$ and $\nabla^2 \log L_{\boldsymbol{v}}$. Let $\boldsymbol{\theta} \in \Theta$ such that $\boldsymbol{\theta}$ does not locally minimize $L_{\boldsymbol{v}}$. Furthermore for the set of cliques $\mathcal{C}$ let the clique parameters be given by $\{\boldsymbol{\theta}_c\}_{c \in \mathcal{C}}$. Then for the case, that $\boldsymbol{\theta}$ locally maximizes $L_{\boldsymbol{v}}$, it holds that $\partial_c \log L_{\boldsymbol{v}}(\boldsymbol{\theta}) = 0$, $\forall c \in \mathcal{C}$. Thereupon the Corollary is trivially satisfied,

since $\forall i \in \mathbb{N}$:

$$\begin{aligned}\boldsymbol{\theta}_c^{(i+1)} &\overset{\text{def}}{=} \boldsymbol{\theta}_c^{(i)} - \eta n \left(\langle -\partial_c \log F_{\boldsymbol{\theta}} \rangle_{\text{d}} - \langle -\partial_c \log F_{\boldsymbol{\theta}} \rangle_{\text{m}} \right) \\ &\overset{2.55}{=} \boldsymbol{\theta}_c^{(i)} + \eta \partial_c \log L_{\boldsymbol{v}} = \boldsymbol{\theta}_c^{(i)}\end{aligned}$$

Such that $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}$, $\forall i \in \mathbb{N}$ and thus converges in $\boldsymbol{\theta}$, which locally maximizes $L_{\boldsymbol{v}}$. Let therefore $\boldsymbol{\theta}$ be assumed, not to locally minimize or maximize the likelihood, such that $\nabla \log L_{\boldsymbol{v}} \neq 0$. Then according to **(A2)** for $f \coloneqq -\log L_{\boldsymbol{v}}$ it follows, that f is pseudo convex, such that a maximal $U \subseteq \Theta$ with $\boldsymbol{\theta} \in U$ can be chosen with $f \mid_U$ is convex. Let $\hat{\boldsymbol{\theta}} \in U$ minimize f in U, then $\hat{\boldsymbol{\theta}}$ maximizes $L_{\boldsymbol{v}}$ in U. In order to proof the Corollary, it therefore suffices to show that $\lim_{i \to \infty} f\left(\boldsymbol{\theta}^{(i)}\right) = f(\hat{\boldsymbol{\theta}})$, or equivalently:

$$\forall \varepsilon < 0 \, \exists \eta > 0, \, N \in \mathbb{N}: \, \left\| f\left(\boldsymbol{\theta}^{(N)}\right) - f(\hat{\boldsymbol{\theta}}) \right\| < \varepsilon$$

Local Inequalities The properties f inherits from Λ can be used to formulate inequalities in U. Since $f \mid_U$ is convex, it follows that:

$$f(\boldsymbol{\theta}) \geq f(\boldsymbol{\vartheta}) + \nabla f(\boldsymbol{\vartheta})^T (\boldsymbol{\theta} - \boldsymbol{\vartheta}), \, \forall \boldsymbol{\theta}, \, \boldsymbol{\vartheta} \in U \tag{2.57}$$

Furthermore since $L_{\boldsymbol{d}} \in C^2(\Theta)$ it follows, that also $f \in C^2(\Theta)$ and according to **(A3)** since $\nabla \Lambda$ is Lipschitz continuous also ∇f is Lipschitz continuous, such that the largest Eigenvalue of the Hessian $\nabla^2 f$ is upper bounded by a Lipschitz constant $\delta > 0$. Consequently:

$$f(\boldsymbol{\theta}) \leq f(\boldsymbol{\vartheta}) + \nabla f(\boldsymbol{\vartheta})^T (\boldsymbol{\theta} - \boldsymbol{\vartheta}) + \frac{\delta}{2} \left\| \boldsymbol{\theta} - \boldsymbol{\vartheta} \right\|^2, \, \forall \boldsymbol{\theta}, \, \boldsymbol{\vartheta} \in U \tag{2.58}$$

Upper bound for $f(\boldsymbol{\theta}^{(i+1)}) - f(\hat{\boldsymbol{\theta}})$ The local inequalities can be applied the compare $f(\boldsymbol{\theta}^{(i)})$ with $f(\boldsymbol{\theta}^{(i+1)})$. It follows, that:

$$\begin{aligned} f(\boldsymbol{\theta}^{(i+1)}) &\overset{2.58}{\leq} f(\boldsymbol{\theta}^{(i)}) + \nabla f(\boldsymbol{\theta}^{(i)})^T(\boldsymbol{\theta}^{(i+1)} - \boldsymbol{\theta}^{(i)}) + \frac{\delta}{2}\left\|\boldsymbol{\theta}^{(i+1)} - \boldsymbol{\theta}^{(i)}\right\|^2 \\ &\overset{\text{def}}{=} f(\boldsymbol{\theta}^{(i)}) + \nabla f(\boldsymbol{\theta}^{(i)})^T(-\eta\nabla f(\boldsymbol{\theta}^{(i)})) + \frac{\delta}{2}\left\|-\eta\nabla f(\boldsymbol{\theta}^{(i)})\right\|^2 \\ &= f(\boldsymbol{\theta}^{(i)}) - \eta\left\|\nabla f(\boldsymbol{\theta}^{(i)})\right\|^2 + \frac{\eta^2\delta}{2}\left\|\nabla f(\boldsymbol{\theta}^{(i)})\right\|^2 \\ &= f(\boldsymbol{\theta}^{(i)}) - \eta\left(1 - \frac{\eta\delta}{2}\right)\left\|\nabla f(\boldsymbol{\theta}^{(i)})\right\|^2 \end{aligned} \tag{2.59}$$

Let now the step size η satisfy the inequality$\eta \leq \frac{1}{\delta}$, then according to equation 2.59 it can be concluded, that:

$$\begin{aligned} f(\boldsymbol{\theta}^{(i+1)}) &\leq f(\boldsymbol{\theta}^{(i)}) - \frac{\eta}{2}\left\|\nabla f(\boldsymbol{\theta}^{(i)})\right\|^2 \\ &\overset{2.57}{\leq} f(\hat{\boldsymbol{\theta}}) + \nabla f(\boldsymbol{\theta}^{(i)})^T(\boldsymbol{\theta}^{(i)} - \hat{\boldsymbol{\theta}}) - \frac{\eta}{2}\left\|\nabla f(\boldsymbol{\theta}^{(i)})\right\|^2 \\ &= f(\hat{\boldsymbol{\theta}}) + \frac{1}{2\eta}\left(\left\|\boldsymbol{\theta}^{(i)} - \hat{\boldsymbol{\theta}}\right\|^2 - \left\|(\boldsymbol{\theta}^{(i)} - \eta\nabla f(\boldsymbol{\theta}^{(i)})) - \hat{\boldsymbol{\theta}}\right\|^2\right) \\ &= f(\hat{\boldsymbol{\theta}}) + \frac{1}{2\eta}\left(\left\|\boldsymbol{\theta}^{(i)} - \hat{\boldsymbol{\theta}}\right\|^2 - \left\|\boldsymbol{\theta}^{(i+1)} - \hat{\boldsymbol{\theta}}\right\|^2\right) \end{aligned} \tag{2.60}$$

And therefore:

$$f(\boldsymbol{\theta}^{(i+1)}) - f(\hat{\boldsymbol{\theta}}) \overset{2.60}{\leq} \frac{1}{2\eta}\left(\left\|\boldsymbol{\theta}^{(i)} - \hat{\boldsymbol{\theta}}\right\|^2 - \left\|\boldsymbol{\theta}^{(i+1)} - \hat{\boldsymbol{\theta}}\right\|^2\right) \tag{2.61}$$

Upper bound for $f(\boldsymbol{\theta}^{(N)}) - f(\hat{\boldsymbol{\theta}})$ The upper bound for $f(\boldsymbol{\theta}^{(i+1)}) - f(\hat{\boldsymbol{\theta}})$ can be used to determine an upper bound for the N-th step error $f(\boldsymbol{\theta}^{(N)}) - f(\hat{\boldsymbol{\theta}})$. For $N \geq 2$

it follows, that:

$$
\begin{aligned}
f(\boldsymbol{\theta}^{(N)}) - f(\hat{\boldsymbol{\theta}}) &\leq \frac{1}{N-1} \sum_{i=1}^{N-1} \left(f(\boldsymbol{\theta}^{(i)}) - f(\hat{\boldsymbol{\theta}}) \right) \\
&\leq \frac{1}{N-1} \sum_{i=1}^{N-1} \left(f(\boldsymbol{\theta}^{(i+1)}) - f(\hat{\boldsymbol{\theta}}) \right) \\
&\overset{??}{\leq} \frac{1}{2\eta(N-1)} \sum_{i=1}^{N-1} \left(\left\| \boldsymbol{\theta}^{(i)} - \hat{\boldsymbol{\theta}} \right\|^2 - \left\| \boldsymbol{\theta}^{(i+1)} - \hat{\boldsymbol{\theta}} \right\|^2 \right) \\
&= \frac{1}{2\eta(N-1)} \left(\left\| \boldsymbol{\theta} - \hat{\boldsymbol{\theta}} \right\|^2 - \left\| \boldsymbol{\theta}^{(N)} - \hat{\boldsymbol{\theta}} \right\|^2 \right) \\
&\leq \frac{1}{2\eta(N-1)} \left\| \boldsymbol{\theta} - \hat{\boldsymbol{\theta}} \right\|^2
\end{aligned}
\tag{2.62}
$$

Let now be $\varepsilon > 0$. Then for $0 < \eta \leq \frac{1}{\delta}$ and

$$
N > \frac{1}{2\eta\varepsilon} \left\| \boldsymbol{\theta} - \hat{\boldsymbol{\theta}} \right\|^2 + 1 \tag{2.63}
$$

it follows, that:

$$
\begin{aligned}
\left\| f(\boldsymbol{\theta}^{(N)}) - f(\hat{\boldsymbol{\theta}}) \right\| &= f(\boldsymbol{\theta}^{(N)}) - f(\hat{\boldsymbol{\theta}}) \\
&\overset{2.62}{\leq} \frac{1}{2\eta(N-1)} \left\| \boldsymbol{\theta} - \hat{\boldsymbol{\theta}} \right\|^2 \overset{2.63}{<} \varepsilon
\end{aligned}
$$

□

Chapter 3

Approximative tractable inference in Energy Based Models

Overview

As the previous chapter closes by showing, that Energy based models allow a local maximum likelihood estimation of principal manifolds, the current chapter is concerned with a tractable approximation of this estimation. The first section starts by showing, that the log-likelihood gradient, as given by the free energy representation, for non-trivial cases, can not be solved analytically due to the underlying Boltzmann distribution. To address this issue, in a first step, a Monte Carlo (MC) approximation is applied to define a stochastic gradient descent, known as Mini-Batch Gradient Descent. Thereupon it is shown, that this algorithm under some additional statistical assumptions, almost surely converges in a local ML estimation. As the MC approximation, however, requires, that exact samples have to be drawn from the model it can not directly be applied to EBMs. Therefore, in a second step, the MC approximation by itself is approximated by a Markov Chain Monte Carlo (MCMC) approximation of the log-likelihood gradient which provides a corresponding stochastic gradient descent, known as Contrastive Divergence (CD) learning. The first section closes by showing, that CD learning under some supplementary analytical assumptions, regarding the function space, almost surely converge "close" to a local ML estimation, where the bias is quantified by an upper bound given by a respective KL -Divergence. Since the

theoretical convergence, however, it not sufficient for tractable inference, the second section relates the tractability of CD learning to the mixing time of the Markov chain of the MCMC approximation and provides theoretical bounds. Finally with Restricted Boltzmann Machines (RBM) and Deep Boltzmann Machines (DBM) specifically structured EBMs are introduced that for an additional sparsity assumption assure the rapid mixing property and therefore tractable inference by CD learning.

3.1 Computability and Sampling in EBMs

Due to the structure of EBMs, the number of potential states increases exponentially with the number of vertices in the underlying graph. This property has significant consequences for the computability of the Steepest Gradient Descent in EBMs: By reconstituting its definition in Corollary 2.2, it appears, that for any clique $c \in \mathcal{C}$ the update for the clique parameter $\boldsymbol{\theta}_c$ is proportional to the partial derivative $\partial_c \log L_{\boldsymbol{v}}$, which according to equation 2.53 decomposes in two parts. These are respectively referred to as the **positive phase** and the **negative phase**:

$$\partial_c \log L_{\boldsymbol{v}}(\boldsymbol{\theta}) = n \underbrace{\langle -\partial_c \log F_{\boldsymbol{\theta}} \rangle_{\mathrm{d}}}_{=:\text{positive phase}} - n \underbrace{\langle -\partial_c \log F_{\boldsymbol{\theta}} \rangle_{\mathrm{m}}}_{=:\text{negative phase}} \tag{3.1}$$

Thereby the naming reflects the respective effects for statistical inference: Whereas the positive phase increases the probability for a realization $\boldsymbol{v}$ of $\boldsymbol{V}$, the negative phase decreases the probability of predictions, generated by the model distribution. With respect to the empirical data distribution $P_{\boldsymbol{v}}$ the positive phase thereby equates to:

$$\begin{aligned} \langle -\partial_c \log F_{\boldsymbol{\theta}} \rangle_{\mathrm{d}} &\overset{\mathrm{def}}{=} -\int_S P_{\boldsymbol{v}}(\boldsymbol{V} = \boldsymbol{v}) \partial_c \log F_{\boldsymbol{\theta}}(\boldsymbol{v}) \, \mathrm{d}\boldsymbol{v} \\ &= -\frac{1}{n} \sum_{i=1}^{n} \partial_c \log F_{\boldsymbol{\theta}}(\boldsymbol{v}^i) \end{aligned} \tag{3.2}$$

With regard to a log-linear parametrisation, the partial derivative of the free energy, then can be rewritten as an expectation of the feature $\boldsymbol{\phi}_c$ with respect to the condi-

tional model distribution $P_{\boldsymbol{\theta}}(\boldsymbol{H} \mid \boldsymbol{V})$. In order to derive a complete representation of the positive phase as an integral over $\boldsymbol{\phi}_c$, the empirical distribution $P_{\boldsymbol{v}}(\boldsymbol{V})$ can be used to define the **extended data distribution**:

$$P_{\boldsymbol{v},\boldsymbol{\theta}}(\boldsymbol{V}, \boldsymbol{H}) \coloneqq P_{\boldsymbol{v}}(\boldsymbol{V}) P_{\boldsymbol{\theta}}(\boldsymbol{H} \mid \boldsymbol{V}) \tag{3.3}$$

Then the log-likelihood gradient has a feature representation, which is given by:

$$\partial_c \log L_{\boldsymbol{v}}(\boldsymbol{\theta}) = n \underbrace{\mathbb{E}_{\boldsymbol{X} \sim P_{\boldsymbol{v},\boldsymbol{\theta}}}(\boldsymbol{\phi}_c(\boldsymbol{X}_c))}_{\text{positive phase}} - n \underbrace{\mathbb{E}_{\boldsymbol{X} \sim P_{\boldsymbol{\theta}}}(\boldsymbol{\phi}_c(\boldsymbol{X}_c))}_{\text{negative phase}} \tag{3.4}$$

In order to calculate the positive phase, it has to be distinguished, if the clique c represents only observables, or also latent variables. In the first case the feature $\boldsymbol{\phi}_c$, by definition, does not depend on any latent random variables, such that:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{X} \sim P_{\boldsymbol{v},\boldsymbol{\theta}}}(\boldsymbol{\phi}_c(\boldsymbol{X}_c)) &= -\int_S \boldsymbol{\phi}_c(\boldsymbol{x}_c) P_{\boldsymbol{v},\boldsymbol{\theta}}(\boldsymbol{X}_c = \boldsymbol{x}_c) \, \mathrm{d}\boldsymbol{x} \\
&= -\frac{1}{n} \sum_{i=1}^{n} \int_{S_H} \boldsymbol{\phi}_c(\boldsymbol{v}_c^i, \boldsymbol{h}_c) P_{\boldsymbol{\theta}}(\boldsymbol{H} = \boldsymbol{h} \mid \boldsymbol{V} = \boldsymbol{v}^i) \, \mathrm{d}\boldsymbol{h} \\
&= -\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\phi}_c(\boldsymbol{v}_c^i) \int_{S_H} P_{\boldsymbol{\theta}}(\boldsymbol{H} = \boldsymbol{h} \mid \boldsymbol{V} = \boldsymbol{v}^i) \, \mathrm{d}\boldsymbol{h} \\
&= -\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\phi}_c(\boldsymbol{v}_c^i)
\end{aligned} \tag{3.5}$$

If, however, the clique c also comprises latent random variables, then the corresponding feature $\boldsymbol{\phi}_c$ also depends on latent random variables and therefore can not be pulled out of the integral over S_H. Then due to the underlying Boltzmann distribution, this integral, even for quite simple feature functions - like quadratic terms - is not analytically solvable. This motivates an approximation by **Monte Carlo** (MC) integration: Let $\boldsymbol{v}$ by an i.i.d. realization of $\boldsymbol{V}$ and $\boldsymbol{X}_{\mathrm{d}} \sim P_{\boldsymbol{v},\boldsymbol{\theta}}$ an S-valued i.i.d random sample of length m. Then for any $c \in \mathcal{C}$ also $\boldsymbol{\phi}_c(\boldsymbol{X}_{\mathrm{d},c})$ is a random sample of length m and describes data distributed realizations of the feature functions $\boldsymbol{\phi}_c$. Furthermore let:

$$\hat{\mathbf{MC}}_{\mathrm{d},c}^{(m)} \coloneqq \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{\phi}_c(\boldsymbol{X}_{\mathrm{d},c})^i, \; \forall m \in \mathbb{N} \tag{3.6}$$

Since the expectation of the random sample $\boldsymbol{\mu}_c \coloneqq \mathbb{E}(\boldsymbol{\phi}_c(\boldsymbol{X}_{\mathrm{d},c}))$ by its definition equals the positive phase it can be concluded from the law of large numbers, that:

$$\hat{\mathbf{MC}}_{\mathrm{d},c}^{(m)} \overset{\text{a.s.}}{\to} \mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{v},\boldsymbol{\theta}}}(\boldsymbol{\phi}_c(\boldsymbol{X})), \text{ for } m \to \infty \tag{3.7}$$

Consequently $\hat{\mathbf{MC}}_{\mathrm{d},c}^{(m)}$ is an unbiased estimator for $\mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{v},\boldsymbol{\theta}}}(\boldsymbol{\phi}_c)$. Let now be $\boldsymbol{X}_{\mathrm{m}}$ an S-valued i.i.d random sample with $\boldsymbol{X}_{\mathrm{m}} \sim \mathrm{P}_{\boldsymbol{\theta}}$ and let:

$$\hat{\mathbf{MC}}_{\mathrm{m},c}^{(m)} \coloneqq \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\phi}_c(\boldsymbol{X}_{\mathrm{m},c})^i, \; \forall m \in \mathbb{N} \tag{3.8}$$

Then analogues to the positive phase, the sequence $(\mathbf{MC}_{\mathrm{m},c}^{(m)})_{m\in\mathbb{N}}$ almost surely converges towards the negative phase:

$$\hat{\mathbf{MC}}_{\mathrm{m},c}^{(m)} \overset{\text{a.s.}}{\to} \mathbb{E}_{\boldsymbol{X}\sim \mathrm{P}_{\boldsymbol{\theta}}}(\boldsymbol{\phi}_c(\boldsymbol{X})), \text{ for } m \to \infty \tag{3.9}$$

This motivates the definition of an MC approximation for $\partial_c \log L_{\boldsymbol{v}}(\boldsymbol{\theta})$ by:

$$\hat{\mathbf{MC}}_{c}^{(m)} \coloneqq \underbrace{n\,\frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\phi}_c(\boldsymbol{X}_{\mathrm{d},c})^i}_{\hat{\mathbf{MC}}_{\mathrm{d},c}^{(m)}} - \underbrace{n\,\frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\phi}_c(\boldsymbol{X}_{\mathrm{m},c})^i}_{\hat{\mathbf{MC}}_{\mathrm{m},c}^{(m)}}, \; \forall m \in \mathbb{N} \tag{3.10}$$

Then according to equations 3.1, 3.7 and 3.9 it follows, that for all $c \in \mathcal{C}$:

$$\hat{\mathbf{MC}}_{c}^{(m)} \overset{\text{a.s.}}{\to} \partial_c \log L_{\boldsymbol{v}}(\boldsymbol{\theta}), \text{ for } m \to \infty \tag{3.11}$$

such that $\hat{\mathbf{MC}}_{c}^{(m)}$ yields an unbiased estimator for $\partial_c \log L_{\boldsymbol{v}}(\boldsymbol{\theta})$. Let now $m \in \mathbb{N}$ be fixed. Then the convergence motivates to use the MC approximation:

$$\partial_c \log L_{\boldsymbol{v}}(\boldsymbol{\theta}) \approx \hat{\mathbf{MC}}_{c}^{(m)}$$

to define an update rule for a stochastic gradient descent. This is known as a **Mini-Batch Gradient Descent** (see Algorithm 3.1).

The practical usability of the MC approximation, however, depends on its efficiency in terms of its convergence rate. For the positive phase let $\boldsymbol{\phi}_c(\boldsymbol{X}_{\mathrm{d},c})$ have a

Algorithm 3.1 Mini-Batch Gradient Descent in EBMs

parameters
prior parameter $\boldsymbol{\theta} \in \Theta$
step size $\eta > 0$
mini-batch size $m \in \mathbb{N}$
number of steps $N \in \mathbb{N}$

procedure MBGD($\boldsymbol{\theta}$, η, m, N)
  $\boldsymbol{\theta}^{(1)} \leftarrow \boldsymbol{\theta}$
  for $i \in \{1, \dots, N-1\}$ **do**
    Sample $\boldsymbol{x}_{\mathrm{d}}$ from $P_{\boldsymbol{v},\boldsymbol{\theta}}$ of length m
    Sample $\boldsymbol{x}_{\mathrm{m}}$ from $P_{\boldsymbol{\theta}}$ of length m
    for $c \in \mathcal{C}$ **do**

$$\boldsymbol{\theta}_c^{(i+1)} \leftarrow \boldsymbol{\theta}_c^{(i)} - \eta \frac{N_d}{m} \sum_{j=1}^{m} \left(\boldsymbol{\phi}_c(\boldsymbol{x}_{\mathrm{d},c}^j) - \boldsymbol{\phi}_c(\boldsymbol{x}_{\mathrm{m},c}^j) \right)$$

    end for
  end for
  return $\boldsymbol{\theta}^{(N)}$
end procedure

finite variance $\boldsymbol{\sigma}_c^{\mathbf{2}}$. Then the Central Limit Theorem postulates, that the sequence

$$(\sqrt{m}(\hat{\mathbf{MC}}_{\mathrm{d},c}^{(m)} - \boldsymbol{\mu}_c))_{m \in \mathbb{N}}$$

has a stationary distribution, which is given by a normal distribution $\mathcal{N}(0, \boldsymbol{\sigma}_c^{\mathbf{2}})$, such that:

$$m\mathrm{Var}(\hat{\mathbf{MC}}_{\mathrm{d},c}^{(m)}) \overset{\text{a.s.}}{\to} \boldsymbol{\sigma}_c^{\mathbf{2}}, \text{ for } m \to \infty \tag{3.12}$$

It therefore can be concluded, that for $m \to \infty$ the variance decreases by the factor $1/m$ and the convergence rate is given by $1/\sqrt{m}$. For the case, that $\boldsymbol{\phi}_c(\boldsymbol{X}_{\mathrm{d},c})$ also has a finite third absolute moment $\boldsymbol{\rho}_c$, the Berry-Esseen Theorem allows to extend this result to the individual distributions, that approximate $\mathcal{N}(0, \boldsymbol{\sigma}_c^{\mathbf{2}})$. In this case the Monte Carlo integration $\hat{\mathbf{MC}}_{\mathrm{d},c}^{(m)}$ efficiently approximates the positive phase with a convergence rate $1/\sqrt{m}$. And if furthermore $\boldsymbol{\phi}_c(\boldsymbol{X}_{\mathrm{m},c})$ has a finite second and a finite third absolute moment, then also $\hat{\mathbf{MC}}_{\mathrm{d},c}^{(m)}$ and therefore $\hat{\mathbf{MC}}_c^{(m)}$ has a convergence rate $1/\sqrt{m}$. Since the standard error of this approximation is furthermore proportional to $\|\partial_c \log L_{\boldsymbol{v}}(\boldsymbol{\theta})\|$, the intuition suggests, that in "sufficiently harmless" likelihood

landscapes for some $m \in \mathbb{N}$ the Mini-Batch Gradient Descent should converge towards a parameter, that locally maximizes $L_{\boldsymbol{v}}$. Thereby, however, it has to be considered, that the approximation can mislead by a consistent realization of very improbable examples, such that - in difference to the Steepest Gradient Descent `GD` - the best possible result, that can be achieved by `MBGD` is almost sure convergence.

Theorem 3.1 (Convergence of `MBGD` in EBMs)**.** *Let* $(S, \Sigma, \mathcal{M})$ *be an EBM for* $\boldsymbol{X} = (\boldsymbol{V}, \boldsymbol{H})$ *over* $\mathcal{G}$ *with cliques* $\mathcal{C}$ *and a log-linear parametrization* $P\colon \Theta \to \mathcal{M}$*, that additionally to* ***(A1)****,* ***(A2)*** *and* ***(A3)*** *satisfies:*

(A4) *for* $k \in \{2, 3, 4\}$ *there exist* $a_k, b_k \geq 0$*, such that:*

$$\int_S \|\nabla P_{\boldsymbol{\theta}}(\boldsymbol{x})\|^k \, \mathrm{d}\boldsymbol{x} < a_k + b_k \|\boldsymbol{\theta}\|^k, \, \forall \boldsymbol{\theta} \in \Theta$$

Furthermore for a realization $\boldsymbol{v}$ *of* $\boldsymbol{V}$ *let the critical points of* $L_{\boldsymbol{v}}\colon \boldsymbol{\theta} \mapsto P_{\boldsymbol{\theta}}(\boldsymbol{v})$ *respectively be given by* $\Theta_{\min}$ *and* $\Theta_{\max}$*. Then for all mini-batch sizes* $m \in \mathbb{N}$ *and all prior parameters* $\boldsymbol{\theta} \in \Theta \backslash \Theta_{\min}$*, there exists a step size* $\eta > 0$ *and a local maximum likelihood estimation* $\hat{\boldsymbol{\theta}}_{\mathrm{ML}} \in \Theta_{\max}$*, such that:*

$$\texttt{MBGD}(\boldsymbol{\theta}, \eta, m, N) \xrightarrow{\text{a.s.}} \hat{\boldsymbol{\theta}}_{\mathrm{ML}}, \text{ for } N \to \infty$$

Notes on the proof of Theorem 3.1. The Theorem has been proven by (Bottou 1998, p26 ff.) and uses the framework of stochastic approximation theory as introduced by (Kushner et al. 1978) and (Ljung et al. 1983). Thereby the proof essentially utilizes the assumption **(A4)** to provide an upper bound for the standard error of the approximation $\hat{\mathbf{MC}}_c^{(m)}$ for the case $m = 1$. Since the standard error, however, is proportional to $1/\sqrt{m}$, the case $m = 1$ already describes the "worst case". Consequently the a.s. convergence is in particular assured for an $m \in \mathbb{N}$ with $m > 1$. □

The Mini-Batch Gradient Descent efficiently solves the computability issues of the Steepest Gradient Descent. Nevertheless, it can not directly be applied to EBMs, due to the intractable task to sample $\boldsymbol{X}_{\mathrm{d}}$ and $\boldsymbol{X}_{\mathrm{m}}$ from the extended data distribution and the model distribution: This can be understood as follows: Usually one would try to start with a set of random variables, that respectively are statistically independent from all other variables. For an underlying acyclic directed graph, this would

be given by those random variables that are represented by the roots of the graph. Then in a series of subsequent steps all further random variables could successively be sampled from previous realizations, until all random variables are realized. For EBMs, however, the situation is much more complicated, since due to the undirected edges no random variable can be chosen, that are independent from the remaining, such that exact inference is generally intractable. This motivates approximative inference in EBMs, where the two main classes of algorithms can be distinguished (Wainwright et al. 2007): The first class comprises variational methods, such as mean-field approximations. Thereby the strategy of the methods is, to replace the model space by a "tractable" family of probability distributions, like fully-factorized distributions, which typically allows a fast and accurate approximation. Nevertheless this approximation is restricted to the "tractable" family, such that with respect to the original model space the estimates can be correspondingly worse. A very different strategy is applied by **Markov chain Monte Carlo** (MCMC) based algorithms, which introduce a **Markov chain** within the sampling procedure, which stationary distribution equals the desired target distribution. Thereby the idea is to define an initial S-valued random vector $\boldsymbol{X}^{(0)}$ with a known probability distribution $P^{(0)}$, e.g. a uniform distribution, and thereupon to repeatedly apply a **transition function** $T\colon (S, \Sigma) \to (S, \Sigma)$, which generates a sequence of random vectors $(\boldsymbol{X}^{(t)})_{t\in\mathbb{N}}$ by:

$$\boldsymbol{X}^{(0)} \xrightarrow{T} \boldsymbol{X}^{(1)} \xrightarrow{T} \boldsymbol{X}^{(2)} \xrightarrow{T} \dots \tag{3.13}$$

Then the sequence, by its definition is a Markov chain, since every random vector $\boldsymbol{X}^{(t+1)}$ conditionally only depends on its predecessor $\boldsymbol{X}^{(t)}$. Such Markov chains have an intuitive representation within the function space $\mathcal{P} \subseteq L^1(S, \Sigma)$ of probability distributions over (S, Σ). By defining the **transition operator** $\tau\colon \mathcal{P} \to \mathcal{P}$ as the push forward operator $\tau(P) \coloneqq T_*P$, $\forall P \in \mathcal{P}$, the Markov chain 3.13 is accompanied by a sequence of probability distributions $(P^{(t)})_{t\in\mathbb{N}}$ in $\mathcal{P}$:

$$P^{(0)} \xrightarrow{\tau} P^{(1)} \xrightarrow{\tau} P^{(2)} \xrightarrow{\tau} \dots$$

Due to the L_1-norm of the embedding space $L^1(S, \Sigma)$, it follows, that $\mathcal{P}$ is a normed space with respect to the induced norm, and therefore in particular a topological

space. Furthermore, with respect to the equivalence of measures in L^1, $\mathcal{P}$ is also a vector space and therefore a topological vector space. These properties can be used to show, that the transition operator is a continuous linear operator on $\mathcal{P}$ (Zucca 2002), which motivates to regard $(P^{(t)})_{t\in\mathbb{N}}$ as a trajectory within a dynamical system, where the **evolution function** $\tau\colon \mathbb{N}\times\mathcal{P}\to\mathcal{P}$ is given by::

$$\tau^t(P) \coloneqq (\underbrace{\tau\circ\cdots\circ\tau}_{t})(P),\ \forall P\in\mathcal{P} \tag{3.14}$$

The challenge for sampling in EBMs can therefore be described by the task to determine an evolution function τ and an initial point $P^{(0)}$, such that:

(MC1) The Markov chain $(\boldsymbol{X}^{(t)})_{t\in\mathbb{N}}$ has a stationary distribution π

(MC2) The stationary distribution π equals the desired target distribution

In the first step, it has is shown, that the Markov chain has a stationary distribution.

Proposition 3.1. *Let (S, Σ) be a measurable space and $\boldsymbol{X}$ an S-valued MRF over an undirected connected graph $\mathcal{G}$. Then for all measurable functions $T\colon (S,\Sigma)\to(S,\Sigma)$ the transition operator τ which is induced by T has a unique fixed point.*

Proof of Proposition 3.1. For the discrete case, that considers a finite state space, the proof is given by (Gelfand et al. 1990) and comprises two steps: In the first step the existence of a fixed point for τ is shown as a consequence of a fixed point theorem. In the second step the property of the underlying graph to be connected is used to show its uniqueness. Thereupon according to (Casella et al. 1992, p170ff) the continuous case follows by a generalization of the respective discrete concepts by their continuous pendants.

Existence Let P be a probability distribution over a finite measurable space (S, Σ) with $S = \{1, \ldots, , n\}$ and $\Sigma = \mathscr{P}(S)$. Then any state $i \in S$ has a probability:

$$p_i \coloneqq P(\boldsymbol{X} = i) \in [0, 1] \tag{3.15}$$

such that P can be represented by a vector $\boldsymbol{p} \in [0, 1]^n$. Let now be $\mathcal{P} \subseteq L^1(S, \Sigma)$ the set of all probability distributions over (S, Σ). Then by the normalization condition

$$\sum_{i=1}^{n} p_i = 1 \tag{3.16}$$

the set $\mathcal{P}$ forms a convex and compact subset of $\mathbb{R}^n$, which is termed a **probability simplex**. Since furthermore the transition operator τ, that is induced by the transition function T is a continuous linear operator on $\mathcal{P}$, the Brouwer Fixed Point Theorem can be applied and it follows that τ at least has one fixed point in $\mathcal{P}$.

Uniqueness For any states $i, j \in S$, the transition from state i to j has a unique transition probability:

$$p_{i,j} \coloneqq P(T \circ \boldsymbol{X} = j \mid \boldsymbol{X} = i) \in [0, 1] \tag{3.17}$$

such that, the transition operator τ can be represented by a transition matrix $A \coloneqq (p_{i,j})_{i,j \in S}$. Then the trajectory of an initial distribution $P^{(0)} \in \mathcal{P}$ with a representation $\boldsymbol{p}^{(0)} \in \mathbb{R}^n$, has a representation, given by:

$$\boldsymbol{p}^{(t)} = A^t \boldsymbol{p}^{(0)}, \, \forall t \in \mathbb{N} \tag{3.18}$$

Therefore by assuming that $\boldsymbol{p}$ represents a fixed point of τ, then $\boldsymbol{p}$ has to satisfy the relationship $\boldsymbol{p} = A\boldsymbol{p}$, such that $\boldsymbol{p}$ is an Eigenvector of A. Thereby the matrix A by definition is non-negative. Thereupon, however, by the requirement, that the underlying undirected graph $\mathcal{G}$ is complete, it follows that A is indeed strictly positive, since all vertices within $\mathcal{G}$ are path connected such that $p_{i,j} > 0$ for all states $i, j \in S$. This allows an application of the Perron-Frobenius theorem, which states, that in this case the largest Eigenvalue λ_1 has an algebraic multiplicity of 1 and equals the value 1. Consequently there exists a unique Eigenvector $\boldsymbol{v}_1$ to the Eigenvalue 1 and for all other Eigenvectors $\boldsymbol{v}_2, \ldots, \boldsymbol{v}_n$ it holds, that their respective eigenvalues $\lambda_2 \ldots, \lambda_n$ are strictly smaller then 1, such that their respective Eigenspaces decay to 0 for $t \to \infty$. This shows that τ has a unique fixed point. □

According to Proposition 3.1 for an underlying connected graph any Markov chain $(\boldsymbol{X}^{(t)})_{t\in\mathbb{N}}$, which is induced by a measurable function T, has a unique stationary distribution π. Furthermore, however, it shows that the stationary distribution does not depend on the the initial distribution $P^{(0)}$, such that requirement **(MC1)** is always satisfied. This allows e.g. to draw the initial random sample from a uniform distributions by:

$$X^{(0)} \sim P^{(0)} \coloneqq \mathcal{U}_V \otimes \mathcal{U}_H$$

Thereupon in order to also satisfy requirement **(MC2)** in the next step a transition function T has to be determined, such that the stationary distribution π equals the desired target distribution. For a probability distribution, which is completely known by its conditional probabilities, such a transition function is given by the **Gibbs Sampler** (Geman et al. 1984).

Definition (Gibbs Sampler)**.** *Let (S, Σ) be a measurable space with $S \subseteq \mathbb{R}^d$ and $\boldsymbol{X}$ an S-valued random vector. Then a measurable function $T\colon (S, \Sigma) \to (S, \Sigma)$ is a systematic sweep Gibbs sampler for a probability distribution P over (S, Σ), iff for $\boldsymbol{X}' \coloneqq T \circ \boldsymbol{X}$ it holds, that:*

$$X_i' \sim P(X_i \mid \{X_j'\}_{1\leq j<i}, \{X_j\}_{i<j\leq d}), \forall i \in \{1, \dots, d\} \tag{3.19}$$

The recursive definition of the systematic sweep Gibbs sampler can easily be entangled by iterating over the indices: For a given S-valued random vector $\boldsymbol{X}$ the application of the measurable function T provides a further S-valued random vector $\boldsymbol{X}'$. Thereby according to its definition, the random variable X_1', conditionally only depends on the random variables $\{X_j\}_{1<j\leq d}$ and therefore can be sampled at first. Since any subsequent steps, do not depend on X_1, but only on X_1', the random variable X_1' can be thought as an update of X_1. Thereupon iteratively, the respective i-th random variable X_i' is sampled from the previously updated random variables $\{X_j'\}_{1\leq j<i}$, and the remaining initial random variables $\{X_j\}_{i<j\leq d}$ until the last random variable X_d' is sampled from the previously updated random variables $\{X_j'\}_{1\leq j<d}$. Let now be τ the transition operator, which is induced by a Gibbs sampler T for a probability distribution P. Then the above iteration can be used to show that P is a fixed point forτ.

Corollary 3.1. *Let $(S, \Sigma, \mathcal{P})$ be an EBM for $\boldsymbol{X}$ over an undirected connected graph $\mathcal{G}$ and $P\colon \Theta \to \mathcal{P}$ a parametrization of $\mathcal{P}$. Then for all $\boldsymbol{\theta} \in \Theta$, the Gibbs sampler $T_{\boldsymbol{\theta}}\colon (S, \Sigma) \to (S, \Sigma)$ for $P_{\boldsymbol{\theta}}$ induces a Markov chain that converges in probability towards $P_{\boldsymbol{\theta}}$.*

Proof of Corollary 3.1. Let $\tau = T_{\boldsymbol{\theta}*}$ be the transition operator, induced by $T_{\boldsymbol{\theta}}$. Then for a discrete sample space according to Proposition 3.1 τ has a unique fixed point. It therefore suffices to show, that $P_{\boldsymbol{\theta}}$ is a fixed point of τ. This can be proofed by complete induction. Let $\boldsymbol{X}' \coloneqq T_{\boldsymbol{\theta}} \circ \boldsymbol{X}$. Then:

Induction Start For $i = 1$, it holds that:

$$\begin{aligned}
\tau(P_{\boldsymbol{\theta}})(\boldsymbol{X})_1 &\overset{\text{def}}{=} (T_{\boldsymbol{\theta}*}P_{\boldsymbol{\theta}})(\boldsymbol{X})_1 \\
&\overset{\text{def}}{=} \int_{\boldsymbol{x}\in S} P_{\boldsymbol{\theta}}(X_1 \mid \{X_j\}_{1<j\leq d}) P_{\boldsymbol{\theta}}(\{X_j = x_j\}_{1<j\leq d})\,\mathrm{d}\boldsymbol{x} \\
&\overset{\text{ltp}}{=} P_{\boldsymbol{\theta}}(X_1) \\
&\Rightarrow X_1' \sim X_1
\end{aligned} \tag{3.20}$$

Induction Step Let $X_j' \sim X_j$, $\forall j \leq i$. Then for $i \mapsto i+1$, it holds, that:

$$\begin{aligned}
\tau(P_{\boldsymbol{\theta}})(\boldsymbol{X})_{i+1} &\overset{\text{def}}{=} \int_S P_{\boldsymbol{\theta}}(X_{i+1} \mid \{Y_j\}_{1\leq j\leq i},\, \{X_j\}_{i+1<j\leq d}) \\
&\qquad P_{\boldsymbol{\theta}}(\{Y_j = x_j\}_{1\leq j\leq i},\, \{X_j = x_j\}_{i+1<j\leq d})\,\mathrm{d}\boldsymbol{x} \\
&= \int_S P_{\boldsymbol{\theta}}(X_{i+1} \mid \{X_j = x_j\}_{j\neq i+1}) \\
&\qquad P_{\boldsymbol{\theta}}(\{X_j = x_j\}_{j\neq i+1})\,\mathrm{d}\boldsymbol{x} \\
&\overset{\text{ltp}}{=} P_{\boldsymbol{\theta}}(X_{i+1}) \\
&\Rightarrow X_{i+1}' \sim X_{i+1}
\end{aligned} \tag{3.21}$$

Equations 3.20 and 3.21 show, that $\tau(P_{\boldsymbol{\theta}})(\boldsymbol{X})_i = P_{\boldsymbol{\theta}}(X_i)$ for all $i \in \{1, \ldots, d\}$, such that $\tau(P_{\boldsymbol{\theta}}) = P_{\boldsymbol{\theta}}$. □

In the following Corollary 3.1 is to be applied to the desired target distributions, which are given by the extended data distribution $P_{\boldsymbol{v},\boldsymbol{\theta}}$ an the model distribution $P_{\boldsymbol{\theta}}$. In the case of the model distribution for an initial distribution $P_{\boldsymbol{\theta}}^{(0)}$, which is chosen

by a uniform distribution $\mathcal{U}_V \otimes \mathcal{U}_H$, it follows that the Markov Chain $(\boldsymbol{X}_{\mathrm{m}}^{(k)})_{k\in\mathbb{N}}$ with:

$$\boldsymbol{X}_{\mathrm{m}}^{(k)} \sim T_{\boldsymbol{\theta}*}^{k}\left(\mathcal{U}_V \otimes \mathcal{U}_H\right), \forall i \in \mathbb{N} \tag{3.22}$$

has a stationary distribution $P_{\boldsymbol{\theta}}$. For the extended data distribution $P_{\boldsymbol{v},\boldsymbol{\theta}}$, however, some further considerations have to be taken into account. Due to its definition by:

$$P_{\boldsymbol{v},\boldsymbol{\theta}}(\boldsymbol{V},\, \boldsymbol{H}) \coloneqq P_{\boldsymbol{v}}(\boldsymbol{V})\mathrm{P}_{\boldsymbol{\theta}}(\boldsymbol{H} \mid \boldsymbol{V})$$

it follows, that the observables are determined by the empirical data distribution $P_{\boldsymbol{v}}(\boldsymbol{V})$ and therefore independent from the latent random variables, such that the conditional probabilities $P_{\boldsymbol{v},\boldsymbol{\theta}}(\boldsymbol{V} \mid \boldsymbol{H})$ are given by uniform distributions. If therefore the initial distribution $P_{\boldsymbol{v},\boldsymbol{\theta}}^{(0)}$ of the Markov chain is chosen according to:

$$P_{\boldsymbol{v},\boldsymbol{\theta}}^{(0)}(\boldsymbol{V},\, \boldsymbol{H}) = P_{\boldsymbol{v}}(\boldsymbol{V}) \otimes \mathcal{U}_H(\boldsymbol{H})$$

then the Gibbs sampler for $P_{\boldsymbol{v},\boldsymbol{\theta}}$ does not alter the marginal distribution of the observables, given by $P_{\boldsymbol{v}}$. It therefore suffices to only update the realizations of the latent random variables. Since the conditional probabilities $P_{\boldsymbol{v},\boldsymbol{\theta}}(\boldsymbol{H} \mid \boldsymbol{V})$ in turn are given by $P_{\boldsymbol{\theta}}(\boldsymbol{H} \mid \boldsymbol{V})$ these updates can be applied by the Gibbs sampler $T_{\boldsymbol{\theta}}$ for the model distribution $P_{\boldsymbol{\theta}}$, where the observables are **clamped** to their initial values, which is denoted by $T_{\boldsymbol{\theta},H}$. Thereupon for $i \in \mathbb{N}$ let $T_{\boldsymbol{\theta},H*}^{k}$ denote the pushforward measure, which is induced by k applications of $T_{\boldsymbol{\theta},H}$. Then the Markov chain $(\boldsymbol{X}_{\mathrm{d}}^{(k)})_{k\in\mathbb{N}}$ with:

$$\boldsymbol{X}_{\mathrm{d}}^{(k)} \sim T_{\boldsymbol{\theta},H*}^{k}\left(P_{\boldsymbol{v}} \otimes \mathcal{U}_H\right), \forall k \in N \tag{3.23}$$

converges in probability towards $P_{\boldsymbol{v},\boldsymbol{\theta}}$ for $k \to \infty$. It therefore can be concluded that samples $\boldsymbol{X}_{\mathrm{d}} \sim P_{\boldsymbol{v},\boldsymbol{\theta}}$ and $\boldsymbol{X}_{\mathrm{m}} \sim P_{\boldsymbol{\theta}}$ can be approximated by the respective Markov Chains $(\boldsymbol{X}_{\mathrm{d}}^{(k)})_{k\in\mathbb{N}}$ and $(\boldsymbol{X}_{\mathrm{m}}^{(k)})_{k\in\mathbb{N}}$. The next step therefore is to apply this approximation to the Monte Carlo approximations of the positive and the negative phase. In this purpose let the MCMC approximation for the positive phase be given by:

$$\widehat{\mathbf{MCMC}}_{\mathrm{d},c}^{(m,k)} \coloneqq \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\phi}_c(\boldsymbol{X}_{\mathrm{d},c}^{(k)})^i, \forall m,\, k \in \mathbb{N} \tag{3.24}$$

Then according to Corollary 3.1 and equation 3.7 it follows, that:

$$\hat{\mathbf{MCMC}}_{\mathrm{d},c}^{(m,k)} \overset{\text{a.s.}}{\to} \mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{v},\boldsymbol{\theta}}}(\boldsymbol{\phi}_c), \text{ for } m,\, k \to \infty \tag{3.25}$$

Furthermore let:

$$\hat{\mathbf{MCMC}}_{\mathrm{m},c}^{(m,k)} \coloneqq \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\phi}_c(\boldsymbol{X}_{\mathrm{m},c}^{(k)})^i,\ \forall m,\, k \in \mathbb{N} \tag{3.26}$$

Then according to Corollary 3.1 and equation 3.9 it follows, that:

$$\hat{\mathbf{MCMC}}_{\mathrm{m},c}^{(m,k)} \overset{\text{a.s.}}{\to} \mathbb{E}_{\boldsymbol{X}\sim P_{\boldsymbol{\theta}}}(\boldsymbol{\phi}_c), \text{ for } m,\, k \to \infty \tag{3.27}$$

Summarized by equation 3.4 it then follows, that for the definition:

$$\hat{\mathbf{MCMC}}_{c}^{(m,k_\mathrm{d},k_\mathrm{m})} \coloneqq n \underbrace{\frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\phi}_c(\boldsymbol{X}_{\mathrm{d},c}^{(k_\mathrm{d})})^i}_{\hat{\mathbf{MCMC}}_{\mathrm{d},c}^{(m,k_\mathrm{d})}} - n \underbrace{\frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\phi}_c(\boldsymbol{X}_{\mathrm{m},c}^{(k_\mathrm{m})})^i}_{\hat{\mathbf{MCMC}}_{\mathrm{m},c}^{(m,k_\mathrm{m})}},\ \forall m,\, k_\mathrm{d},\, k_\mathrm{m} \in \mathbb{N} \tag{3.28}$$

According to equations 3.25 and 3.27 it holds, that:

$$\hat{\mathbf{MCMC}}_{c}^{(m,k_\mathrm{d},k_\mathrm{m})} \overset{\text{a.s.}}{\to} \partial_c \log L_{\boldsymbol{v}}(\boldsymbol{\theta}), \text{ for } m,\, k_\mathrm{d},\, k_\mathrm{m} \to \infty \tag{3.29}$$

The almost sure convergence towards $\partial_c \log L_{\boldsymbol{v}}(\boldsymbol{\theta})$ motivates an approximation:

$$\partial_c \log L_{\boldsymbol{v}}(\boldsymbol{\theta}) \approx \hat{\mathbf{MCMC}}_{c}^{(m,k_\mathrm{d},k_\mathrm{m})}$$

within the update rule of a stochastic gradient descent. Thereby, in difference to the Monte Carlo approximation $\hat{\mathbf{MC}}_c^{(m)}$, it has to be noted, that for any fixed $k_\mathrm{d},\, k_\mathrm{m} \in \mathbb{N}$ the probability distributions $P_{\boldsymbol{v},\boldsymbol{\theta}}^{(k_\mathrm{d})}$ and $P_{\boldsymbol{\theta}}^{(k_\mathrm{m})}$, which are generated by the Markov Chains, are generally different from their respective equilibrium distributions $P_{\boldsymbol{v},\boldsymbol{\theta}}$ and $P_{\boldsymbol{\theta}}$, such that $\hat{\mathbf{MCMC}}_{c}^{(m,k_\mathrm{d},k_\mathrm{m})}$ is a biased estimator for $\partial_c \log L_{\boldsymbol{v}}(\boldsymbol{\theta})$. One approach to make this bias tangible is to identify the gradient by the Kullback-Leibler divergence.

Lemma 3.1. *For any i.i.d realization $\boldsymbol{v}$ of $\boldsymbol{V}$ and any parameter $\boldsymbol{\theta} \in \Theta$ the following statements are equivalent:*

(1) *The likelihood $L_{\boldsymbol{v}}(\boldsymbol{\theta})$ is locally maximized by $\boldsymbol{\theta}$*

(2) *The KL-Divergence $D_{\mathrm{KL}}(P_{\boldsymbol{v},\boldsymbol{\theta}} \| P_{\boldsymbol{\theta}})$ is locally minimized by $\boldsymbol{\theta}$*

Proof of Lemma 3.1. The KL-divergence is defined by:

$$D_{\mathrm{KL}}(P_{\boldsymbol{v},\boldsymbol{\theta}} \| P_{\boldsymbol{\theta}}) \coloneqq \int_S P_{\boldsymbol{v},\boldsymbol{\theta}}(\boldsymbol{x}) \log \frac{P_{\boldsymbol{v},\boldsymbol{\theta}}(\boldsymbol{x})}{P_{\boldsymbol{\theta}}(\boldsymbol{x})} \,\mathrm{d}\boldsymbol{x} \tag{3.30}$$

Such that:

$$\begin{aligned} D_{\mathrm{KL}}(P_{\boldsymbol{v},\boldsymbol{\theta}} \| P_{\boldsymbol{\theta}}) &\overset{3.30}{=} \int_S P_{\boldsymbol{v},\boldsymbol{\theta}}(\boldsymbol{v},\, \boldsymbol{h}) \log \frac{P_{\boldsymbol{v}}(\boldsymbol{v}) P_{\boldsymbol{\theta}}(\boldsymbol{h} \mid \boldsymbol{v})}{P_{\boldsymbol{\theta}}(\boldsymbol{v}) P_{\boldsymbol{\theta}}(\boldsymbol{h} \mid \boldsymbol{v})} \,\mathrm{d}(\boldsymbol{v},\, \boldsymbol{h}) \\ &= \int_{S_V} \left(\int_{S_H} P_{\boldsymbol{v},\boldsymbol{\theta}}(\boldsymbol{v},\, \boldsymbol{h}) \,\mathrm{d}\boldsymbol{h} \right) \log \frac{P_{\boldsymbol{v}}(\boldsymbol{v})}{P_{\boldsymbol{\theta}}(\boldsymbol{v})} \,\mathrm{d}\boldsymbol{v} \\ &\overset{\mathrm{ltp}}{=} \int_{S_V} P_{\boldsymbol{v}}(\boldsymbol{v}) \log P_{\boldsymbol{v}}(\boldsymbol{v}) \,\mathrm{d}\boldsymbol{v} - \int_{S_V} P_{\boldsymbol{v}}(\boldsymbol{v}) \log P_{\boldsymbol{\theta}}(\boldsymbol{v}) \,\mathrm{d}\boldsymbol{v} \end{aligned} \tag{3.31}$$

Since the first term on the right side does not depend on $\boldsymbol{\theta}_c$, it follows, that:

$$\begin{aligned} \partial_c D_{\mathrm{KL}}(P_{\boldsymbol{v},\boldsymbol{\theta}} \| P_{\boldsymbol{\theta}}) &\overset{3.31}{=} -\partial_c \int_{S_V} P_{\boldsymbol{v}}(\boldsymbol{v}) \log P_{\boldsymbol{\theta}}(\boldsymbol{v}) \,\mathrm{d}\boldsymbol{v} \\ &= -\partial_c \sum_{i=1}^{n} \log P_{\boldsymbol{\theta}}(\boldsymbol{v}^i) \\ &\overset{\mathrm{def}}{=} -\partial_c \log L_{\boldsymbol{v}}(\boldsymbol{\theta}) \end{aligned} \tag{3.32}$$

It follows, that **(ML1)** has an equivalent formulation by:

$$\nabla L_{\boldsymbol{v}}(\boldsymbol{\theta}) = 0 \overset{3.32}{\Leftrightarrow} \nabla D_{\mathrm{KL}}(P_{\boldsymbol{v},\boldsymbol{\theta}} \| P_{\boldsymbol{\theta}}) = 0$$

and furthermore **(ML2)** by:

$$\nabla^2 L_{\boldsymbol{v}}(\boldsymbol{\theta}) < 0 \overset{3.32}{\Leftrightarrow} \nabla^2 D_{\mathrm{KL}}(P_{\boldsymbol{v},\boldsymbol{\theta}} \| P_{\boldsymbol{\theta}}) > 0$$

□

The KL-divergence $D_{\mathrm{KL}}(P \| Q)$ measures the discriminative information of P with respect to Q and may intuitively be thought as the distance of Q from P from the perspective of P. This allows an intuitive interpretation of Lemma 3.1: The likelihood

function $L_{\boldsymbol{v}}$ is locally maximized, if and only if the distance between the observed data $P_{\boldsymbol{v},\boldsymbol{\theta}}$ and the model prediction $P_{\boldsymbol{\theta}}$ is locally minimized. By an MCMC approximation, for given k_{d}, $k_{\mathrm{m}} \in \mathbb{N}$, the probability distributions $P_{\boldsymbol{v},\boldsymbol{\theta}}$ and $P_{\boldsymbol{\theta}}$ are then respectively substituted by their approximations $P_{\boldsymbol{v},\boldsymbol{\theta}}^{(k_{\mathrm{d}})}$ and $P_{\boldsymbol{\theta}}^{(k_{\mathrm{m}})}$. It follows, that:

$$\lim_{m\to\infty} \mathbf{MC\hat{M}C}_c^{(m,k_{\mathrm{d}},k_{\mathrm{m}})} \overset{\text{a.s.}}{=} -\partial_c D_{\mathrm{KL}}(P_{\boldsymbol{v},\boldsymbol{\theta}}^{(k_{\mathrm{d}})} \| P_{\boldsymbol{\theta}}^{(k_{\mathrm{m}})}),\ \forall c \in \mathcal{C} \tag{3.33}$$

Lemma 3.2. *The bias of a stochastic gradient descent, that uses the MCMC approximation for $m \to \infty$ is upper bounded by:*

$$\mathrm{bias}_{\mathrm{MCMC}} \leq 2D_{\mathrm{KL}}(P_{\boldsymbol{\theta}}^{(k_{\mathrm{m}})} \| P_{\boldsymbol{\theta}})$$

Proof of Lemma 3.2. By Lemma 3.1 and equation 3.33 it can be concluded, that in a stochastic gradient descent, that uses the MCMC approximation, for $m \to \infty$ the fixed points of the gradient descent are given by parameters $\boldsymbol{\theta}_{\mathrm{CD}} \in \Theta$, that satisfy the requirement:

$$\partial_c D_{\mathrm{KL}}(P_{\boldsymbol{v},\boldsymbol{\theta}}^{(k_{\mathrm{d}})} \| P_{\boldsymbol{\theta}}^{(k_{\mathrm{m}})}) = 0,\ \forall c \in \mathcal{C} \tag{3.34}$$

Thereby the bias of an estimation $\boldsymbol{\theta}_{\mathrm{MCMC}}$ in the function space is upper bounded with respect to the corresponding local ML estimation $\boldsymbol{\theta}_{\mathrm{ML}}$. This follows as a consequence of the Pythagorean theorem for dually flat manifolds (Amari 2016). It then follows, that:

$$\begin{aligned} \mathrm{bias}_{\mathrm{MCMC}} &= D_{\mathrm{KL}}(P_{\boldsymbol{v},\boldsymbol{\theta}}^{(k_{\mathrm{d}})} \| P_{\boldsymbol{\theta}}^{(k_{\mathrm{m}})}) - D_{\mathrm{KL}}(P_{\boldsymbol{v},\boldsymbol{\theta}} \| P_{\boldsymbol{\theta}}) \\ &\leq \underbrace{D_{\mathrm{KL}}(P_{\boldsymbol{v},\boldsymbol{\theta}} \| P_{\boldsymbol{v},\boldsymbol{\theta}}^{(k_{\mathrm{d}})})}_{=:\mathrm{bias}_{\mathrm{d}}} + \underbrace{D_{\mathrm{KL}}(P_{\boldsymbol{\theta}}^{(k_{\mathrm{m}})} \| P_{\boldsymbol{\theta}})}_{=:\mathrm{bias}_{\mathrm{m}}} \end{aligned} \tag{3.35}$$

By the definition of the extended data distribution, however, it follows that sufficiently close to the ML estimate $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ it holds, that $\mathrm{bias}_{\mathrm{d}} \leq \mathrm{bias}_{\mathrm{m}}$, such that:

$$\mathrm{bias}_{\mathrm{MCMC}} \leq 2D_{\mathrm{KL}}(P_{\boldsymbol{\theta}}^{(k_{\mathrm{m}})} \| P_{\boldsymbol{\theta}}) \tag{3.36}$$

□

Consequently in order to minimize the bias of the MCMC based stochastic gradi-

ent descent, the discriminative information between $P_{\boldsymbol{\theta}}^{(k_\mathrm{m})}$ and $P_{\boldsymbol{\theta}}$ is to be minimized. A very efficient approach is, to initialize the Markov Chain by a prior $P_{\boldsymbol{\theta}}^{(0)}$, that is already "close" to its equilibrium distribution $P_{\boldsymbol{\theta}}$. Then there is justified hope, that even a small number of sampling steps suffices to generate a posterior $P_{\boldsymbol{\theta}}^{(k_\mathrm{m})}$ that sufficiently approximates $P_{\boldsymbol{\theta}}$. A key idea in this direction was, that near the fixed points of a gradient descent, the model distribution $P_{\boldsymbol{\theta}}$ will be found "close" to the extended data distribution $P_{\boldsymbol{v},\boldsymbol{\theta}}$. Consequently in this case the posterior data distribution $P_{\boldsymbol{v},\boldsymbol{\theta}}^{(k_\mathrm{d})}$, which approximates $P_{\boldsymbol{v},\boldsymbol{\theta}}$ will presumably provide a "good prior" for the model distribution. This consideration motivated an MCMC based stochastic gradient descent, which is referred as **contrastive divergence** learning (see algorithm 3.2, **(author?)** 2002).

Since its first publication (Hinton 2002) CD learning gained a lot of attention due to its empirically success. Thereby in many domains of application the `CD` algorithm appeared to generate sequences $(\boldsymbol{\theta}^{(t)})_{t\in\mathbb{N}}$ that converge "close" to local ML estimates $\hat{\boldsymbol{\theta}}_\mathrm{ML}$. Due to its high applicability the pragmatic consensus prevailed, to start the density estimation by CD learning with very short Markov chains and, in the case of convergence, to increase the chain length to get closer to a local ML estimate. This pragmatic approach was accompanied by a great enthusiasm in the machine learning community to further investigations on "good priors" which provided many variants of CD learning, including *Persistent CD* (Tieleman 2008), *Fast Persistent CD* (Tieleman et al. 2009) and *Tempered MCMC* (Desjardins et al. 2010). The fundamental question, if the algorithm converges, however, has only been answered recently and still requires assumptions, that are difficult to verify. Thereby analogous to the Mini-batch Gradient Descent, the convergence can be traced back to the standard error of the approximation of the gradient.

Theorem 3.2 (Convergence of `CD` in EBMs)**.** *Let* $(S, \Sigma, \mathcal{P})$ *be an EBM for* $\boldsymbol{X} = (\boldsymbol{V}, \boldsymbol{H})$ *over* $\mathcal{G}$ *with cliques* $\mathcal{C}$ *and a log-linear parametrization* $P\colon \Theta \to \mathcal{M}$*, that satisfies* ***(A1)***, ***(A2)***, ***(A3)***, ***(A4)*** *and:*

(A5) *The graph* $\mathcal{G}$ *is connected*

(A6) *For any Gibbs sampler* $T_{\boldsymbol{\theta}}$*, which is defined by* $P_{\boldsymbol{\theta}}$*, let the induced transition operator* τ *have an* $\mathcal{L}_2$*-spectral gap* $1 - \lambda_2 > 0$

Algorithm 3.2 Contrastive Divergence learning in EBMs

parameters

prior distribution parameter $\boldsymbol{\theta} \in \Theta$
step size $\eta > 0$
mini-batch size $m \in \mathbb{N}$
positive phase chain length k_{d}
negative phase chain length k_{m}
number of steps $N \in \mathbb{N}$

procedure $\mathtt{CD}(\boldsymbol{\theta}, \eta, m\, k_{\mathrm{d}}\, k_{\mathrm{m}}\, N)$
  $\boldsymbol{\theta}^{(1)} \leftarrow \boldsymbol{\theta}$
  for $i \in \{1, \ldots, N-1\}$ **do**
    Sample $\boldsymbol{x}_{\mathrm{d}}$ from $P_{\boldsymbol{v}}(\boldsymbol{V}) \otimes \mathcal{U}(\mathrm{img}\boldsymbol{H})$ of length m
    for $j \in \{1, \ldots, k_{\mathrm{d}}\}$ **do**
      Sample $\boldsymbol{x}_{\mathrm{d}}$ from $T_{\boldsymbol{\theta},\boldsymbol{H}}(\boldsymbol{x}_{\mathrm{d}})$
    end for
    Sample $\boldsymbol{x}_{\mathrm{m}} \sim \boldsymbol{x}_{\mathrm{d}}$ of length m
    for $j \in \{1, \ldots, k_{\mathrm{m}}\}$ **do**
      Sample $\boldsymbol{x}_{\mathrm{m}} \sim T_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathrm{m}})$
    end for
    for $c \in \mathcal{C}$ **do**

$$\boldsymbol{\theta}_c^{(i+1)} \leftarrow \boldsymbol{\theta}_c^{(i)} - \eta \frac{d}{m} \sum_{j=1}^{d} \left(\boldsymbol{\phi}_c(\boldsymbol{x}_{\mathrm{d},c}^j) - \boldsymbol{\phi}_c(\boldsymbol{x}_{\mathrm{m},c}^j) \right)$$

    end for
  end for
  return $\boldsymbol{\theta}^{(N)}$
end procedure

(A7) *Define a metric on the set of transition operators by:*

$$\rho(\tau_{\boldsymbol{\theta}},\, \tau_{\boldsymbol{\vartheta}}) \coloneqq \sup_{\pi\colon\, |\pi|\leq 1} \sup_{\boldsymbol{x}\in S} |\tau_{\boldsymbol{\theta}}(\pi)(\boldsymbol{x}) - \tau_{\boldsymbol{\vartheta}}(\pi)(\boldsymbol{x})|$$

and assume, that $\exists \xi > 0$*, such that:*

$$\rho(\tau_{\boldsymbol{\theta}},\, \tau_{\boldsymbol{\vartheta}}) \leq \xi \left\|\boldsymbol{\theta} - \boldsymbol{\vartheta}\right\|, \forall \boldsymbol{\theta},\, \boldsymbol{\vartheta} \in \Theta$$

Then for all realizations $\boldsymbol{v}$ *of* $\boldsymbol{V}$*, prior parameters* $\boldsymbol{\theta} \in \Theta$ *and mini-batch sizes* $m \in \mathbb{N}$*, there exists a step size* $\eta > 0$*, Markov Chain lengths* k_{d}*,* $k_{\mathrm{m}} \in \mathbb{N}$ *and a parameter* $\hat{\boldsymbol{\theta}}_{\mathrm{CD}} \in \Theta$*, such that the sequence* $(\boldsymbol{\theta}^{(t)})_{t\in\mathbb{N}}$ *which is generated by* `CD` *satisfies that:*

$$\limsup_{t\to\infty} \left\| \frac{1}{t} \sum_{s=1}^{t} \boldsymbol{\theta}^{(t)} - \hat{\boldsymbol{\theta}}_{\mathrm{CD}} \right\| \xrightarrow{P} 0$$

Notes on the proof of Theorem 3.2. The Theorem has been proven by (Wu et al. 2016) and uses the terminology of stochastic approximation theory. Thereby the enumeration of the assumptions as given in (Wu et al. 2016) differs from the above. This is because, in the above case, the assumptions are gradually expanded from the algorithms `GD` to `MBGD` and from `MBGD` to `CD`. Also the proof in (Wu et al. 2016) is given for generic exponential families, where in particular the above assumption **(A5)** is formulated as a positivity constraint of the transition density function. However, by reconstituting Proposition 3.1, it can be concluded that the given assumption **(A5)** of a connected graph $\mathcal{G}$, is sufficient in EBMs for the transition operator $\tau_{\boldsymbol{\theta}}$ of a Gibbs Sampler $T_{\boldsymbol{\theta}}$, to have a unique fixed point. In this case the Markov Chain converges as the components of all eigenvalues of $\tau_{\boldsymbol{\theta}}$ decay to zero, except for the largest Eigenvalue, which according to the Perron-Frobenius Theorem is identical to 1. Consequently the convergence rate is then determined by the **spectral gap** $1-\lambda_2$, which denotes the difference between the largest Eigenvalue 1 and the second largest Eigenvalue λ_2. Therefore in order to enforce a strictly positive convergence rate, it has to be assumed, that **(A6)** $1 - \lambda_2(\boldsymbol{\theta}) > 0$, $\forall \boldsymbol{\theta} \in \Theta$. Finally it has to be avoided, that the sequence $(\boldsymbol{\theta}_t)_{t\in\mathbb{N}}$, which is generated by the gradient descent, causes arbitrary large jumps in the function space, which in **(A7)** is ensured by a corresponding Lipschitz continuity with respect to the function space. The proof given by (Wu et

al. 2016) treats the case of an infinite mini-batch size $m \to \infty$ and an exact extended data distribution, such that $P^{(k_\mathrm{d})}_{\boldsymbol{v},\boldsymbol{\theta}} = P_{\boldsymbol{v},\boldsymbol{\theta}}$. This result, however, immediately is generalized, by an incorporation of assumption **(A4)** and the corresponding results from Theorem 3.1 as well as an additional Markov chain for the positive phase, defined by equation 3.23. □

3.2 Structured EBMs for Efficient Sampling

In the previous section, it has been shown, that with CD learning a stochastic gradient descent in EBMs, which uses an MCMC approximation of the gradient, under some generic assumptions is assured to converge. Nevertheless the theoretical existence of such a limit does not implicate, that its approximation is practically tractable. Thereby the decisive factor, which determines this behaviour, is the number of iterations which is required for the Markov chain to get "close" to its stationary distribution. This number has a vivid meaning in the respective L^1-space: Since the stationary Markov chain may be regarded as a dynamical system with a unique fixed point, this quantity equals the minimum number of steps, which is required for an arbitrary probability distribution to enter an ε-ball about the stationary distribution. This provides the notation of the **mixing time** of a Markov chain.

Definition (Mixing Time)**.** *Let $(\boldsymbol{X}^{(t)})_{t\in\mathbb{N}}$ be an irreducible aperiodic Markov chain with a state space (S, Σ), a stationary distribution π and a transition operator τ. Furthermore let $\mathcal{P} \subset L^1(S, \Sigma)$ be the set of all probability distributions over (S, Σ). Then for all $t \in \mathbb{N}$ the maximal variation distance to the stationary distribution is defined by:*

$$\delta_{\max}(t) \coloneqq \max_{\mu\in\mathcal{P}} \frac{1}{2} \left\| \tau^t(\mu) - \pi \right\|_1 \tag{3.37}$$

Thereupon the for an $\varepsilon > 0$ the ε-mixing time of $(\boldsymbol{X}^{(t)})_{t\in\mathbb{N}}$ is defined by the smallest number of steps t, that is required for $\delta_{\max}(t) \leq \varepsilon$:

$$t_{\mathrm{mix}}(\varepsilon) \coloneqq \arg\min_{t\in\mathbb{N}} \left(\delta_{\max}(t) \leq \varepsilon\right) \tag{3.38}$$

A Markov chain is then referred to be **torpid mixing**, if its mixing time increases exponential with the state space and **rapid mixing** if the mixing time maximally

has a polynomial growth. Since by an application to arbitrary EBMs the mixing time apparently can vary between these cases (Liu et al. 2014) it got all the more important to study this behaviour for specific classes of EBMs, or the other way around: To find EBMs where the Gibbs sampler is rapid mixing. It is therefore important to provide bounds for the mixing time, that depend on the underlying structure of the respective EBM. An important assumption, that has been incorporated within the previous section was **(A5)**, which requires, that the underlying graph is connected. This assumption assures, that the Markov chain has a unique stationary distribution, since for an EBM over a complete graph any state transition has a strictly positive probability. Then according to the Perron-Frobenius Theorem the largest eigenvalue λ_1 of the transition operator τ equals 1 and has an algebraic multiplicity of 1, such that its corresponding eigenvector comprises the only component, that resists a repeated application of τ, whereas all others sooner or later decay to zero. At this point it is getting apparent, that the second (absolutely) largest eigenvalue λ_2 and therefore the spectral gap $1 - \lambda_2$ has some considerable influence on the mixing time. This property has been exploited in Theorem 3.2 by assumption **(A6)**, which requires that $1 - \lambda_2 > 0$, to ensure a positive convergence rate and therefore a finite mixing time. Moreover, however, the spectral gap can also be used to quantitatively bound the mixing time.

Proposition 3.2. *Let $(\boldsymbol{X}^{(t)})_{t\in\mathbb{N}}$ be an irreducible aperiodic Markov chain with a state space (S, Σ), an absolutely positive stationary distribution π and a transition operator τ. Furthermore let τ have a spectral gap $1-\lambda_2$. Then for all $0 < \varepsilon < 1$ it holds, that:*

$$(t_{\text{rel}} - 1) \ln \frac{1}{2\varepsilon} \leq t_{\text{mix}}(\varepsilon) \leq \left\lceil t_{\text{rel}} \left(\frac{1}{2} \ln \frac{1}{\pi_{\min}} + \ln \frac{1}{2\varepsilon} \right) \right\rceil \tag{3.39}$$

where:

$$t_{\text{rel}} \coloneqq \frac{1}{1-\lambda_2}, \text{ and } \pi_{\min} \coloneqq \min_{\boldsymbol{x}\in S} \pi(\boldsymbol{X} = \boldsymbol{x})$$

Notes on the proof of Proposition 3.2. The Proposition has originally been proven for continuous-time reversible Markov processes by (Aldous 1982). On this basis the upper bound for time-discrete Markov chains has been proven by (Diaconis et al. 1991) and the lower bound by (Levon et al. 2009, Theorem 12.4). □

A fundamental disadvantage of the spectral bounds, as given by Proposition 3.2

is, that the spectrum of the transition operator is usually not tractable. Thereupon, however, the spectral gap also provides no intuitive interpretation with respect to the underlying structure. The next step is therefore to find an invariant quantity of the underlying Markov chain with an intuitive meaning for its structure. For this purpose, it is closer to the intuition, to consider the Markov chain as a random walk within a graph over the state space. Thereby the states can be regarded as the vertices of a graph and the transitions between the states as the edges. The mixing time may then identifies the minimum number of steps, which are required, until the probability of any position is sufficiently close to its stationary probability. It is clear, that this property requires, that the random walk is able to traverse the whole state space, such that its representation by a graph requires that the graph is connected. Thereupon, however, it also follows intuitively, that a "bottleneck" in the graph locally impedes a traversal and therefore delays the mixture. In order to transfer this idea to the whole state space, a measure is required, that quantifies its "global connectivity". This quantity is given by the **conductance** (King 2003, p1).

Definition (Conductance)**.** *Let* $(\boldsymbol{X}^{(t)})_{t\in\mathbb{N}}$ *be an irreducible aperiodic Markov chain with a state space* (S, Σ)*, a strictly positive stationary distribution* π *and a transition operator* τ *with a density* $\mathrm{d}\tau$*. Then for any* $s \subset S$ *and* $\overline{s} \coloneqq S \backslash s$ *the conductance of the cut* $(s, \overline{s})$ *is defined by:*

$$\Phi_s \coloneqq \frac{\mathrm{vol}(\partial s)}{\min(\mathrm{vol}(s),\, \mathrm{vol}(\overline{s}))} \tag{3.40}$$

where:

$$\mathrm{vol}(s) \coloneqq \sum_{(i,j)\in s\times\mathcal{V}} \mathrm{d}\tau(i,\, j), \text{ and } \mathrm{vol}(\partial s) \coloneqq \sum_{(i,j)\in s\times\overline{s}} \mathrm{d}\tau(i,\, j)$$

Thereupon let $\mathcal{S}^\star$*be the set of all* $s \subset S$ *with* $\pi(s) \leq \frac{1}{2}$*, then the conductance* Φ *of* τ *is defined by:*

$$\Phi \coloneqq \min_{s\in\mathcal{S}^\star} \Phi_s \tag{3.41}$$

For a Markov chain, the conductance Φ_s of a cut $(s, \overline{s})$ can be thought as the conditional probability, that in its stationary distribution the Markov chain will leave the region s within a single application of the transition operator. Consequently, if for some region $s \subset S$ the conductance Φ_s of the cut $(s, \overline{s})$ is very small, then

there necessarily exists an eigenvalue λ_2 of the transition operator τ, such that the components in the corresponding eigenspace only very slow decay to zero. This, however, is only the case, if λ_2 is close to 1 such that the transition operator τ has a small spectral gap $1 - \lambda_2$. This consideration is formalized by Cheeger's Inequality for Markov chains.

Theorem 3.3 (Cheeger's Inequality for Markov chains)**.** *Let $(\boldsymbol{X}^{(t)})_{t\in\mathbb{N}}$ be an irreducible aperiodic and reversible Markov chain with a state space (S, Σ), a strictly positive stationary distribution π and a transition operator τ. Furthermore let Φ be the conductance and λ_2 the second largest eigenvalue of τ. Then the spectral gap $1-\lambda_2$ is bounded by the conductance:*

$$\frac{1}{2}\Phi^2 \leq 1 - \lambda_2 \leq 2\Phi \tag{3.42}$$

Notes on the proof of Theorem 3.3. The Theorem has independently been proven by (Lawler et al. 1988) and (Jerrum et al. 1989). In both cases the proof essentially generalizes Cheeger's inequality, known from differential geometry. □

Since the goal of these considerations is, to define an upper bound for the mixing time, which is associated with an intuitive meaning, the previous results can be combined as follows:

Corollary 3.2. *Let $(\boldsymbol{X}^{(t)})_{t\in\mathbb{N}}$ be an irreducible aperiodic Markov chain with a state space (S, Σ), a stationary distribution π and a transition operator τ. Furthermore let Φ be the conductance of τ. Then for all $\varepsilon > 0$ it holds, that:*

$$t_{\text{mix}}(\varepsilon) \leq \left\lceil \frac{2}{\Phi^2} \ln \frac{1}{\varepsilon \pi_{\min}} \right\rceil \tag{3.43}$$

Notes on the proof of Corollary 3.2. The Corollary has been proven by (Jerison 2013, p6). The proof essentially combines the upper bound of Theorem 3.3 with the lower bound of Proposition 3.2. □

The application of the upper bound for the mixing time as given by Corollary 3.2, however, for each class of underlying EBM requires a separate and usually sophisticated mathematical analysis, that relates the conductance with the respective

parametrisation (Koller et al. 2009, p520). Consequently this effort has only been carried out for specific classes of EBMs which prove themselves in practice. One of the first classes of that kind, which attracted a broad attention, were **Restricted Boltzmann Machines** (**RBM**).

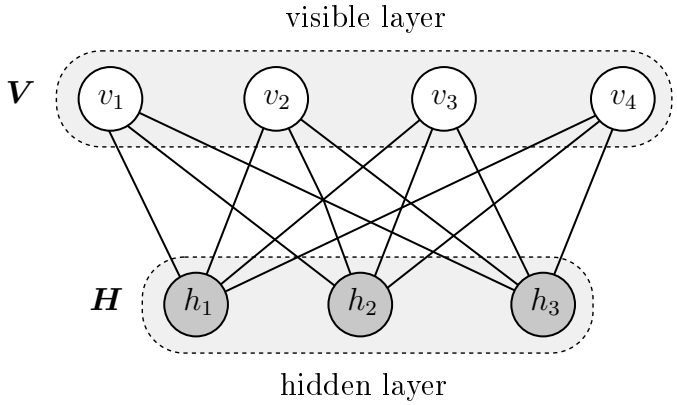

Figure 3.1: Graphical Model of a Restricted Boltzmann Machine

Example 3.1 (Restricted Boltzmann Machine)**.** *Soon after Boltzmann Machines have been proposed, with the "Harmonium" a more restrictive version was introduced as a design study for an artificial cognitive system (Smolensky 1986). The fundamental idea was, to distinguish a layer of observables from a layer of representational features, and to restrict the graph to edges between those two layers (see figure 3.1). Thereby for a* ***visible layer*** $\boldsymbol{V}$*, consisting of* d *observables and a* ***hidden layer*** $\boldsymbol{H}$*, consisting of* m *latent random variables the parametrisation comprises activation thresholds* $\boldsymbol{a} \in \mathbb{R}^d$ *and* $\boldsymbol{b} \in \mathbb{R}^l$ *and synaptic weights* $W \in \mathbb{R}^{d \times l}$ *between the layers, such that:*

$$E_{W,\boldsymbol{a},\boldsymbol{b}}(\boldsymbol{v},\, \boldsymbol{h}) = -(\boldsymbol{v}^{\mathrm{T}} W \boldsymbol{h} + \boldsymbol{a}^{\mathrm{T}} \boldsymbol{v} + \boldsymbol{b}^{\mathrm{T}} \boldsymbol{h}) \tag{3.44}$$

Then a Gibbs Sampler for the model distribution $P_{W,\boldsymbol{a},\boldsymbol{b}}$ *is determined by its conditional probabilities, which due to the bipartite graph structure for an observable* V_i *is given by:*

$$P_{W,\boldsymbol{a},\boldsymbol{b}}(V_i = 1 \mid \boldsymbol{H} = \boldsymbol{h}) \overset{2.25}{=} \operatorname{sigm} \sum_{j=1}^{l} (w_{i,j} h_j + b_j) \tag{3.45}$$

and for a latent random variables H_j by:

$$P_{W,\boldsymbol{a},\boldsymbol{b}}(H_j = 1 \mid \boldsymbol{V} = \boldsymbol{v}) \overset{2.25}{=} \operatorname{sigm} \sum_{i=1}^{d} (w_{i,j} v_i + a_i) \tag{3.46}$$

Furthermore all observables V_i, V_j with $i \neq j$ are pairwise conditionally independent, such that:

$$P_{W,\boldsymbol{a},\boldsymbol{b}}(\boldsymbol{V} \mid \boldsymbol{H}) \overset{3.45}{=} \prod_{i=1}^{d} P_{W,\boldsymbol{a},\boldsymbol{b}}(V_i \mid \boldsymbol{H}) \tag{3.47}$$

The same conclusion can be drawn for the latent random variables:

$$P_{W,\boldsymbol{a},\boldsymbol{b}}(\boldsymbol{H} \mid \boldsymbol{V}) \overset{3.46}{=} \prod_{i=1}^{l} P_{W,\boldsymbol{a},\boldsymbol{b}}(H_i \mid \boldsymbol{H}) \tag{3.48}$$

CD learning mainly owes its great popularity to its empirically successful, when applied to RBMs. The advantages compared to general BMs were particularly evident with respect to its mixing behaviour. For this purpose, however, it is necessary to reconstitute the sampling scheme in RBMs. For a partition of sets, which respectively comprise conditionally independent random variables, a Gibbs sampler can be formulated to parallelly update all random variables within a set "at once". This provides a **Block Gibbs Sampler**.

Definition (Block Gibbs Sampler)**.** *Let $\boldsymbol{X}$ be an (S, Σ)-valued MRF over an undirected graph $G = (\mathcal{V}, \mathcal{E})$ and $(\mathcal{V}_i)_{i \in I}$ a partition of $\mathcal{V}$, such that for any $i \in I$ the random variables in $\mathcal{V}_i$ are pairwise conditionally independent. Then a measurable function $T \colon (S, \Sigma) \to (S, \Sigma)$ is a block Gibbs sampler for a probability distribution P over (S, Σ), iff for $\boldsymbol{X}' \coloneqq T \circ \boldsymbol{X}$ it holds, that:*

$$\boldsymbol{X}'_{\mathcal{V}_i} \sim P(\boldsymbol{X}_{\mathcal{V}_i} \mid \{\boldsymbol{X}_{\mathcal{V}_j}\}_{1 \leq j < i}, \{\boldsymbol{X}_{\mathcal{V}_j}\}_{i < j \leq |I|}), \forall i \in I \tag{3.49}$$

By applying the block Gibbs Sampler to RBMs the sampling scheme as used in CD learning then obtains a remarkably simple structure: The sampling starts by a Markov chain for the extended data distribution $P_{\boldsymbol{v},\boldsymbol{\theta}}(\boldsymbol{V}, \boldsymbol{H})$. Thereby the observables are initially drawn from the marginal distribution $P_{\boldsymbol{v},\boldsymbol{\theta}}(\boldsymbol{V})$, which by definition equals

the data distribution $P_{\boldsymbol{v}}(\boldsymbol{V})$, such that:

$$\boldsymbol{v}_{\mathrm{d}}^{(0)} \sim P_{\boldsymbol{v}}$$

Thereupon, the latent random variables are not required to be drawn from a uniform distribution, since the conditional probability $P_{\boldsymbol{v},\boldsymbol{\theta}}(\boldsymbol{H} \mid \boldsymbol{v}_{\mathrm{d}}^{(0)})$, which equals $P_{\boldsymbol{\theta}}(\boldsymbol{H} \mid \boldsymbol{v}_{\mathrm{d}}^{(0)})$, is already completely known by equation 3.48, such that:

$$\boldsymbol{h}_{\mathrm{d}}^{(0)} \sim P_{\boldsymbol{\theta}}(\boldsymbol{H} \mid \boldsymbol{v}_{\mathrm{d}}^{(0)})$$

At this point, it is important to notice, that no further iterations in the Markov chain for $P_{\boldsymbol{v},\boldsymbol{\theta}}$ are required, since the drawn sample $(\boldsymbol{v}_{\mathrm{d}}^{(0)}, \boldsymbol{h}_{\mathrm{d}}^{(0)})$ already is an exact sample of $P_{\boldsymbol{v},\boldsymbol{\theta}}$, such that $P_{\boldsymbol{v},\boldsymbol{\theta}} = P_{\boldsymbol{v},\boldsymbol{\theta}}^{(0)}$. Thereupon the Markov chain for the model distribution $P_{\boldsymbol{\theta}}$ is initialised by the last drawn sample for the data distribution, such that $\boldsymbol{v}_{\mathrm{m}}^{(0)} \coloneqq \boldsymbol{v}_{\mathrm{d}}^{(0)}$ and $\boldsymbol{h}_{\mathrm{m}}^{(0)} \coloneqq \boldsymbol{h}_{\mathrm{d}}^{(0)}$. In any following iteration step the $(i+1)^{\mathrm{th}}$ observables are updated from the i^{th} latent random variables:

$$\boldsymbol{v}_{\mathrm{m}}^{(i+1)} \sim P_{\boldsymbol{\theta}}(\boldsymbol{V} \mid \boldsymbol{h}_{\mathrm{m}}^{(i)}),\ \forall i \in \mathbb{N}$$

and then the $(i+1)^{\mathrm{th}}$ latent random variables from the $(i+1)^{\mathrm{th}}$ observables:

$$\boldsymbol{h}_{\mathrm{m}}^{(i+1)} \sim P_{\boldsymbol{\theta}}(\boldsymbol{H} \mid \boldsymbol{v}_{\mathrm{m}}^{(i+1)}),\ \forall i \in \mathbb{N}$$

Consequently when applied to RBMs, the sampling scheme of CD learning reduces to a single Markov chain, which alternates between the observables and latent random variables and therefore is referred as **alternating Gibbs sampling**. Thereby for any alternation, the respective random variables can be computed in parallel, which in particular for large models allows considerable time savings by the use of parallel computing. Nevertheless, there is a further and maybe even more important aspect, that justifies the success of CD learning, when applied to RBMs: In the first decade of the century, CD learning on RBMs has extensively been tested in numerous domains, where in many cases even for a single step approximation, the algorithm succeeded to converge in appropriate time. This empirical success of CD learning on RBMs received a theoretical explanation by evaluating the conductance for RBMs. Then

the weight matrix determines a tangible upper bound for the mixing time.

Theorem 3.4 (Mixing time in RBMs)**.** *Let* $(S, \Sigma, \mathcal{P})$ *be an RBM with a visible layer* $\boldsymbol{V}$*, that comprises* d *observables and a hidden layer* $\boldsymbol{H}$*, consisting of* l *latent random variables, such that the model distribution* $P_{W,\boldsymbol{a},\boldsymbol{b}}$ *is given by parameters* $\boldsymbol{a} \in \mathbb{R}^d$*,* $\boldsymbol{b} \in \mathbb{R}^l$ *and* $W \in \mathbb{R}^{d \times l}$*. Furthermore let:*

$$\|W\|_1 \|W^{\mathrm{T}}\|_1 < 4 \tag{3.50}$$

Then the mixing time t_{mix} *of the alternating Gibbs Sampler for* $P_{W,\boldsymbol{a},\boldsymbol{b}}$ *satisfies, that:*

$$t_{\mathrm{mix}}(\varepsilon) \leq \frac{1}{\ln(4) - \ln\left(\|W\|_1 \|W^{\mathrm{T}}\|_1\right)} \ln\left(\frac{\min(d,\, l)}{\varepsilon}\right) \tag{3.51}$$

where the vectorial induced ℓ_1*-norm of the matrix is given by:*

$$\|W\|_1 \coloneqq \max_{1 \leq j \leq l} \sum_{i=1}^{d} |W_{i,j}| \tag{3.52}$$

Notes on the proof of Theorem 3.4. The Theorem has been proven by (Tosh 2016, p4, Lemma 4). Thereby the proof utilizes Markovian couplings to provide an upper bound for the total variation distance and therefore to the mixing time (Tosh 2016, p3, Lemma 2). A subsequent application of the bounds to the individual Markov chains of (i) the observables and (ii) the latent random variables of an alternating Gibbs sampler then provides a generic upper bound for the mixing time (Tosh 2016, p4, Theorem 3). This upper bound, however, requires a restriction (Tosh 2016, p4) of the respective Markov chains which in the case of an RBMs is shown to be satisfied by the assumption of equation 3.50 (Tosh 2016, p4, Lemma 4) and (Tosh 2016, p18, Theorem 17) with reference to (Jerrum et al. 1993). This allows a derivation of the upper bound as given by equation 3.51 (Tosh 2016, p4, Corollary 5). □

Consequently for the case, that a the weight matrix of a given RBM satisfies the condition, given by equation 3.50, then the Markov chain of the alternative Gibbs sampling is polynomially bounded and therefore rapid mixing. In order to assure, that CD learning profits from rapid mixing, however, it has to be taken into account, that during each step of CD learning the weight matrix is updated. A possible approach

to enforce the rapid mixing property is therefore to regularize the ℓ_1-norm within the update rule, such that the weights are bounded by the above criterion. At this point, however, it is important to notice, that an application of the ℓ_1-regularization is only justified for the structural assumption of a respective sparsity, which of course is not satisfied for arbitrary observed data. Nevertheless, it can be concluded that RBMs in many cases allow tractable inference. Much before these theoretical results, however, the empirical success of RBMs already motivated their use as building blocks for more complicated "stacked" graphical models (Hinton et al. 2006). Thereby in particular the **Deep Boltzmann Machines** (**DBM**) (Salakhutdinov et al. 2009) may be regarded as a generalization of RBMs.

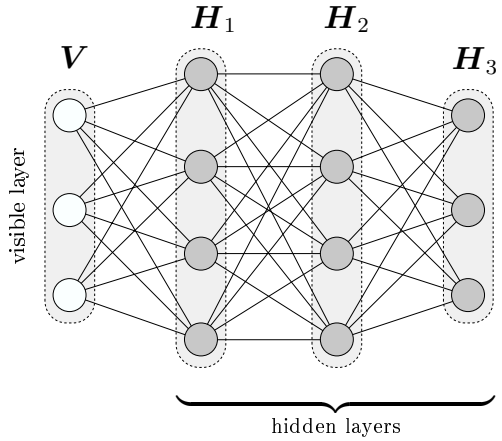

Figure 3.2: Graphical Model of a Deep Boltzmann Machine

Example 3.2 (Deep Boltzmann Machine)**.** *Deep Boltzmann Machines (DBM) generalize RBMs by decomposing the hidden component* $\boldsymbol{H}$ *in* $L \in \mathbb{N}$ *hidden layers* $\boldsymbol{H}_1, \boldsymbol{H}_2, \boldsymbol{H}_3 \ldots \boldsymbol{H}_L$*, that form a hierarchical deep structure (see figure 3.2). Thereby the visible component* $\boldsymbol{V}$ *and the first hidden layer* $\boldsymbol{H}_1$ *are connected by weights* $W_1 \in \mathbb{R}^{d \times l_1}$ *and any hidden layer* $\boldsymbol{H}_i$ *with the subsequent layer* $\boldsymbol{H}_{i+1}$ *by corresponding weights* $W_i \in \mathbb{R}^{l_i \times l_{i+1}}$*. With respective thresholds* $\boldsymbol{a} \in \mathbb{R}^d$ *and* $\boldsymbol{b}_i \in \mathbb{R}^{l_i}$ *the energy function of the DBM is therefore:*

$$E(\boldsymbol{v}, \boldsymbol{h}) = \underbrace{E(\boldsymbol{v}, \boldsymbol{h}_1)}_{RBM} + \sum_{i=1}^{L-1} E(\boldsymbol{h}_i, \boldsymbol{h}_{1+1}) \tag{3.53}$$

where the energy term $E(\boldsymbol{v}, \boldsymbol{h}_1)$ *equals that of an RBM:*

$$E(\boldsymbol{v}, \boldsymbol{h}_1) \coloneqq -(\boldsymbol{v}^{\mathrm{T}} W_1 \boldsymbol{h}_1 + \boldsymbol{a}^{\mathrm{T}} \boldsymbol{v} + \boldsymbol{b}_1^{\mathrm{T}} \boldsymbol{h}_1) \tag{3.54}$$

and the following terms respectively add contributions from additional connections and vertices of the $i+1$*th hidden layer:*

$$E(\boldsymbol{h}_i, \boldsymbol{h}_{1+1}) \coloneqq -(\boldsymbol{h}_i^{\mathrm{T}} W_{i+1} \boldsymbol{h}_{i+1} + \boldsymbol{b}_{i+1}^{\mathrm{T}} \boldsymbol{h}_{i+1}),\ \forall i \in \{1, \dots, L-1\} \tag{3.55}$$

The structure of DBMs then forms a bipartite graph (see figure 3.3) which allows an application of an alternating Gibbs sampler.

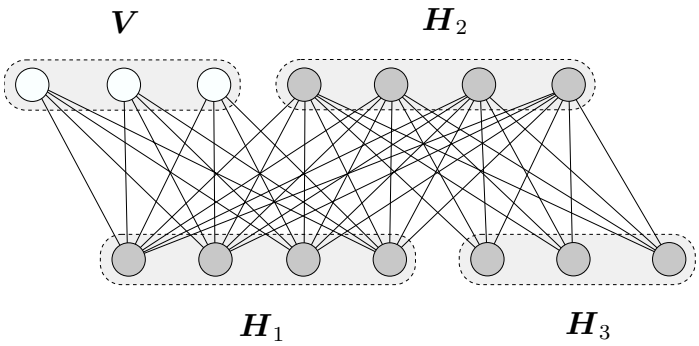

Figure 3.3: Bipartite graph structure of a Deep Boltzmann Machine

Corollary 3.3 (Mixing time in DBMs)**.** *Let* $(S, \Sigma, \mathcal{P})$ *be an DBM with a visible layer* $\boldsymbol{V}$*, that comprises* d *observables, and* l *hidden layers* $\boldsymbol{H}_i$*,* $i \in \{2, \dots, L+1\}$*, that respectively comprise* l_i *latent random variables. Then the model distribution* $P_{\boldsymbol{\theta}}$ *is parametrised by (i) activation thresholds* $\boldsymbol{a} \in \mathbb{R}^d$ *and* $\boldsymbol{b}_i \in \mathbb{R}^{l_i}$ *and (ii) weight matrices* $W_1 \in \mathbb{R}^{d \times l_1}$ *and* $W_i \in \mathbb{R}^{l_i \times l_{i+1}}$*. Let* W *be given by:*

$$W \coloneqq \begin{pmatrix} W_1 & 0 & 0 & 0 \\ W_2^{\mathrm{T}} & W_3 & 0 & 0 \\ 0 & W_4^{\mathrm{T}} & \ddots & 0 \\ 0 & 0 & \ddots & \ddots \end{pmatrix}, \text{ with } \|W\|_1 \|W^{\mathrm{T}}\|_1 < 4$$

Then the mixing time t_{mix} *of the alternating Gibbs Sampler for* $P_{\boldsymbol{\theta}}$ *satisfies, that:*

$$t_{\text{mix}}(\varepsilon) \leq \frac{1}{\ln(4) - \ln\left(\|W\|_1 \|W^{\mathrm{T}}\|_1\right)} \ln\left(\frac{\min(m_{\text{even}},\, m_{\text{odd}})}{\varepsilon}\right) \tag{3.56}$$

where $m_{\text{even}} \coloneqq d + l_2 + l_4 + \ldots$ *is the total number of vertices in the even layers and and* $m_{\text{odd}} \coloneqq l_1 + l_3 + \ldots$ *the total number of vertices in the odd layers.*

Notes to the proof of Corollary 3.3. The Corollary has been proven by (Tosh 2016, p5, Corollary 6). Thereby due to the bipartite graph structure (see figure 3.3), the Corollary follows as an immediate consequence of Theorem 3.4. □

Chapter 4

Application to Gene Regulation Analysis

Overview

Energy based models, as introduced in chapters 2 and 3, prove themselves to provide a suitable class of statistical models for the approximation of principal manifolds. Thereby, however, various structural assumptions had to be be incorporated, which in the first section are evaluated in the domain of gene expression data. Afterwards biological structural knowledge is incorporated to define an deep structured EBM for the approximation of principal manifolds in gene expression data. This model is used in the second section to model a certain gene regulatory network upon gene expression profiles taken from Glioblastoma Multiforme (GBM) cells as well as of normal cells of identical type. Thereby the modelling comprises three steps: In the first step the assumed gene regulatory network is used to estimate parameters for the visible layer, the first hidden layer and their connecting weights. The estimated parameters are used as a prior to approximate a principal manifold, that by stratification equally considers GBM and normal cells. This joint model is then used as a prior to "finetune" the approximations in one case for GBM cells and in the other case for normal cells. Thereby the prior joint model is important to assure comparable results with respect to local ML estimates. Finally the differences in gene regulation are quantified by differences within the respective $\mathcal{M}$-Correlations. These results are compared to other

approaches.

4.1 Statistical Modelling of Gene Regulatory Networks by EBMs

With respect to an application to gene regulation, maybe the foremost question that arises, is if an undirected probabilistic graphical model, that treats errors in variables equally is suitable. This question in particular appears by considering, that biochemical reactions as the building blocks of gene regulation, indeed have some considerate direction, which at certain conditions is given by the sign of the reaction rate. Nevertheless on the scale of gene regulation, the causality usually acts in both directions, to allow cells to preserve homoeostasis: Gene regulatory networks usually stabilize and adapt the production of proteins as an intracellular resource, or as a cellular response to changing environmental conditions, like the available nutrition. A well studied example for such an adaptive behaviour is the Lac operon model.

Example 4.1 (Lac operon)**.** *Glucose is an important energy source for many bacteria, like E. coli. If, however, glucose is not available within the cellular environment, then those bacteria also import the more complex lactose molecules. An increasing intracellular concentration of lactose in turn increases the probability for lactose repressor genes to be transcribed from the bacterial DNA, which, ceteris paribus, causes an increase in the concentration of lactose repressors. Thereupon an increasing concentration of lactose repressors, increases the probability for the degradation of lactose to glucose, such that, ceteris paribus, the concentration of lactose decreases and the concentration of glucose increases. Both mechanisms taken together, can be summarizes by an overall effect, in which intracellular lactose induces its own degradation to glucose.*

The Lac operon gives an example for a causal relationship between concentration levels, i.e. lactose and lactose repressors, which has no designated direction. Although this example may not be regarded as universal, it at least motivates to represent the concentration levels of individual genes by an undirected graph, which is the structural assumption **(S1)**. Let therefore $\boldsymbol{V}$ denote observable concentration levels of genes in

a gene regulatory network. Thereupon the regulatory network may also involve interactions by further unobserved genes, which concentration levels are comprised by a latent random vector $\boldsymbol{H}$. Consequently the assumption that $\boldsymbol{X} = (\boldsymbol{V}, \boldsymbol{H})$ covers all gene regulatory interactions of a given regulatory network, equals the structural assumption **(S2)**, that any X_i satisfies the local Markov property in an undirected graph $\mathcal{G}$. Let now $\boldsymbol{V}$ be realized by a finite number of observations. Then for any outcome $\boldsymbol{v}$, which is NOT realised, it can not be distinguished, if the marginal probability $P(\boldsymbol{V} = \boldsymbol{v})$ is vanishing or not. As this consideration virtually extends to the latent random vector, without loss of generality, it can be assumed, that $P(\boldsymbol{X} = \boldsymbol{x}) > 0$ for all outcomes $\boldsymbol{x}$, which is the structural assumption **(S3)**. This shows, that the choice of EBMs for the statistical modelling of gene expression is considerate arguable. Thereupon the next step is to implement such a statistical model.

For this purpose a good starting point is to substantiate the random variables X_i. Gene expression data appears in multifarious variants due to their underlying screening technologies. Thereby in particular high-throughput technologies like DNA Microarray or Next Generation Sequencing allow a statistical treatment by quantitatively comparable variances. In the following the focus is given to gene expression profiles, given by cDNA microarrays. These microarrays consist of thousands of DNA sequences, referred as probes, that are printed in a high density array on a glass slide. Thereupon the sample, that is to be quantified, as well as a standardized reference sample, respectively are reverse-transcribed into cDNA and labelled with red and green fluorescent dyes. These target samples are jointly hybridized with the arrayed probes on the glass slide, which afterwards is discharged from the unbound residues. Finally a laser beam of defined wavelength is used to locally determine the intensities $I_R(i)$ and $I_G(i)$ of each spot i in the array. This allows to use the reference intensity $I_G(i)$ to quantify the relative abundance of the respective DNA sequence in the sample by the ratio:

$$R_i \coloneqq \frac{I_R(i)}{I_G(i)} \tag{4.1}$$

Thereupon a normal distributed additive Gaussian error is obtained by applying a logarithm to R_i. Thereby the most common used logarithm is given by the basis 2.

The cDNA **log-ratios** are then given by:

$$X_i \coloneqq \log_2 R_i = \log_2 I_R(i) - \log_2 I_G(i) \tag{4.2}$$

By its definition it immediately follows, that the standardized reference sample adds an offset to the log-ratio X_i. Thereby it is common to choose this offset, with respect to the resolution of experiment, such that negative log-ratios may be regarded as measurement errors (see figure 4.4). The logarithm transformation allows to decompose any log-ratio X_i into a systematic component $X_i^\star$ which is determined by interactions and normal distributed random component, such that:

$$X_i = X_i^\star + \varepsilon_i, \text{ with } \varepsilon_i \sim \mathcal{N}(0,\, \sigma_i^2) \tag{4.3}$$

The next step is therefore to describe the systematic component $X_i^\star$ by **direct causal interactions**. By equation ... and ... an error free ratio can be defined by $R_i^\star \coloneqq R_i \exp(-\varepsilon_i)$, such that:

$$X_i^\star = \log_2(R_i^\star) \tag{4.4}$$

Let now $R_i^\star$ represent the relative abundance of a gene G_i, then $R_i^\star$ by definition is proportional to the concentration level $[\mathrm{G}_i]$. It therefore can be concluded that $\exists \lambda_i > 0$ with $R_i^\star = \lambda_i[\mathrm{G}_i]$ and by $c_i \coloneqq \log_2(\lambda_i)$:

$$X_i^\star \overset{4.4}{=} \log_2(\lambda_i[\mathrm{G}_i]) = \log_2([\mathrm{G}_i]) + c_i \tag{4.5}$$

Gene regulatory interactions occur in multifarious manifestations, which in their entirety can only hardly be considered in a tractable family of functions. It is therefore required to incorporate simplifying assumptions. A popular approach to provide a commensurate simplification is to describe direct causal interactions by **Hill-type interactions**. Thereby the concentration levels of a ligand L and an enzymes E under the under the assumption of cooperative binding is parametrised by a microscopic *dissociation constant* $K_A \in \mathbb{R}_0^+$ and a *Hill coefficient* $n \in \mathbb{N}$, such that:

$$\ln([\mathrm{E}]) = \frac{1}{1 + \exp(nK_A - n\ln([\mathrm{L}]))} \tag{4.6}$$

If it is therefore assumed, that all direct gene interactions may be described by Hill-type interactions, it can be concluded, that for interacting genes G_i and G_j with respective log-ratios $X_i^\star$ and $X_j^\star$there exist $\alpha, \beta \in \mathbb{R}$, such that:

$$X_i^\star = \frac{1}{1 + \exp(\alpha + \beta X_j^\star)} = \mathrm{sigm}(\alpha + \beta X_j^\star) \tag{4.7}$$

The relationship given by equation 4.7 shall now be used to describe conditional distributions in the EBM. For observables V_i, V_j this can be obtained by **conditioning** over an additional latent random variables H:

Definition (Conditioning)**.** *Let $\boldsymbol{V}$ be an MRF over an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Then for adjacent vertices $i, j \in \mathcal{V}$ the corresponding random variables V_i, V_j are said to be conditioned on a random variable H, iff their joint probability P satisfies:*

$$P(V_i, V_j \mid H) = P(V_i \mid H)\mathrm{P}(V_j \mid H) \tag{4.8}$$

Since conversely the marginal probability of the observables $P(\boldsymbol{V})$ is given by a marginalisation of P over $\boldsymbol{H}$, conditioning may be regarded as an inverse operation to marginalisation. In the following, conditioning is used to locally derive a representation for Hill-type interactions by a n EBM with appropriate feature functions.

Lemma 4.1. *Let G_i and G_j denote interacting genes and V_i and V_j their respective log-ratios. Then V_i and V_j can be conditioned on a latent random variable H, such that $P(V_i, \mid V_j)$ and $P(V_j, \mid V_i)$ represent Hill-type interactions.*

Proof of Lemma 4.1. The proof is constructive: Let H be a binary random variable, such that the joint probability $P(V_i, V_j, H)$ satisfies: (i) V_i is Gaussian distributed, such that for some $w_i, a_i \in \mathbb{R}$:

$$V_i \sim \mathcal{N}\left(V_i^\star, \sigma_i^2\right), \text{ with } V_i^\star = w_i H + a_i$$

(ii) V_j is Gaussian distributed, such that for some $w_j, a_j \in \mathbb{R}$:

$$V_j \sim \mathcal{N}\left(V_j^\star, \sigma_j^2\right), \text{ with } V_j^\star = w_j H + a_j$$

(iii) H is Bernoulli distributed, such that for some $b \in \mathbb{R}$:

$$H \sim \mathcal{B}\left(\text{sigm}\left(\frac{w_i}{\sigma_i}V_i + \frac{w_j}{\sigma_j}V_j + b\right)\right)$$

Then $P(V_i, \mid V_j)$ and $P(V_j, \mid V_i)$ represent Hill-type interactions. Furthermore it holds, that $P(V_i \mid H, V_j) = P(V_i \mid H)$ and $P(V_j \mid H, V_i) = P(V_j \mid H)$ and therefore $P(V_i, V_j \mid H) = P(V_i \mid H)\mathrm{P}(V_j \mid H)$. □

The modelling of individual Hill-type interactions, by conditioning on latent random variables, can immediately be extended to gene regulatory networks (GRN). Thereby the model, could follow the design principle to respectively assign every edge of the GRN with a single latent variable, that represents the respective type of interaction by an appropriate feature function. In this way, the believed dependency structures can be applied to the observation, by a respective Maximum likelihood estimation.

Corollary 4.1. *Let the random vector $\boldsymbol{V}$ denote observed log-ratios of a gene regulatory network, where all interactions are given by Hill-type interactions. Then there exists a canonical EBM, that models $\boldsymbol{V}$ by conditioning (see figure 4.1).*

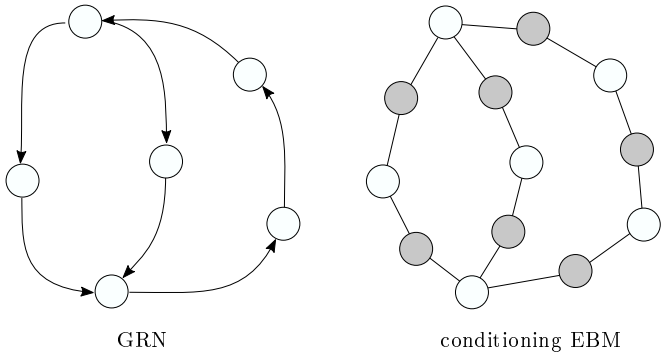

Figure 4.1: Modelling Gene Regulatory Networks (GRN) by conditioning EBMs

Proof of Corollary 4.1. For any edge $(i, j) \in \mathcal{E}$ of the GRN let V_i and V_j be conditioned on a random variable H_{ij} as introduced in lemma 4.1. Then the joint probability can be represented by a Boltzmann distribution, where the energy function for

any outcome $(\boldsymbol{v}, \boldsymbol{h})$ is given by:

$$E(\boldsymbol{v}, \boldsymbol{h}) = -\left(\sum_{i=1}^{d} \frac{(v_i - a_i)^2}{2\sigma_i^2} + \sum_{i=1}^{d} \sum_{(i,j)\in\mathcal{E}} w_{ij} \frac{v_i}{\sigma_i} h_{ij} + \sum_{(i,j)\in\mathcal{E}} h_{ij} b_{ij}\right) \tag{4.9}$$

□

It is important to notice, that the graph layout of an EBM that models random vector $\boldsymbol{V}$ is not an intrinsic property of the observed random vector, but a matter of choice with respect to additional latent random variables, that are introduced within the model, to address structural uncertainties. In the context of traditional graphical models latent variables are typically used with some intended semantic, that models the dependency structure between observable variables (see figure 4.1). Nevertheless, this approach only accounts for uncertainties within believed dependencies, but not within the connection structure itself. A possible approach to overcome this restriction is to relax the connection structure of a conditioning EBM by a deep structured EBM. In this sense, the local ML estimate, which is obtained from the conditioning EBM is used as a prior for the deep structured EBM. This approach, however, requires a natural embedding in the parameter space. With respect to the conditioning of Hill-type interactions, the observables are represented by Gaussian distributions and the latent random variables by Bernoulli distributions. The parametrisation then can naturally be embedded in a Gauss-Bernoulli DBM (see figure 4.2).

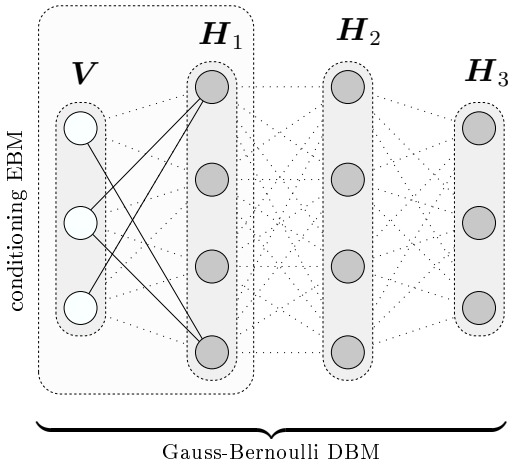

Figure 4.2: Initialise Gauss-Bernoulli DBMs by conditioning EBMs

Example 4.2 (Gauss-Bernoulli Deep Boltzmann Machine)**.** *By identifying the observables of a DBM with Gaussian distributions and the latent random variables with Bernoulli distribution the resulting deep structured EBM is referred as a Gauss-Bernoulli DBM (GBDBM). Thereby the observables* $\boldsymbol{V} = (V_1, V_2, \dots, V_l)$ *are assumed to contain Gauss distributed errors* $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_l)$ *with:*

$$\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$$

Thereupon $\boldsymbol{V}$ *and the first hidden layer* $\boldsymbol{H}_1$ *are connected by weights* $W_1 \in \mathbb{R}^{d \times l_1}$ *and any hidden layer* $\boldsymbol{H}_i$ *with the subsequent layer* $\boldsymbol{H}_{i+1}$ *by corresponding weights* $W_i \in \mathbb{R}^{l_i \times l_{i+1}}$*. With respective thresholds* $\boldsymbol{a} \in \mathbb{R}^d$ *and* $\boldsymbol{b}_i \in \mathbb{R}^{l_i}$ *the energy function of the GBDBM is then given by:*

$$E(\boldsymbol{v}, \boldsymbol{h}) = E(\boldsymbol{v}, \boldsymbol{h}_1) + \sum_{i=1}^{L-1} E(\boldsymbol{h}_i, \boldsymbol{h}_{1+1}) \tag{4.10}$$

where the energy term $E(\boldsymbol{v}, \boldsymbol{h}_1)$ *equals:*

$$E(\boldsymbol{v}, \boldsymbol{h}_1) \coloneqq -\left(\sum_{i=1}^{d} \frac{(v_i - a_i)^2}{2\sigma_i^2} + \sum_{i=1}^{d} \sum_{j=1}^{l_1} W_{1ij} \frac{v_i}{\sigma_i^2} h_{1j} + \sum_{j=1}^{l_1} b_{1j} h_{1j} \right) \tag{4.11}$$

and the following terms respectively add contributions from additional connections and vertices of the $i+1$*th hidden layer:*

$$E(\boldsymbol{h}_i, \boldsymbol{h}_{1+1}) \coloneqq -(\boldsymbol{h}_i^{\mathrm{T}} W_{i+1} \boldsymbol{h}_{i+1} + \boldsymbol{b}_{i+1}^{\mathrm{T}} \boldsymbol{h}_{i+1}), \forall i \in \{1, \dots, L-1\} \tag{4.12}$$

Consequently the conditional distribution of V_i *is given by:*

$$P(V_i \mid \boldsymbol{H}_1) = \mathcal{N}\left(\sum_{j=1}^{l_1} w_{ij} H_{1j} + a_i, \sigma_i^2 \right)$$

The latent random variables $\boldsymbol{H} = (\boldsymbol{H}_1, \boldsymbol{H}_2, \dots, \boldsymbol{H}_L)$ *with* $\boldsymbol{H}_k = (H_{k1}, H_{k2}, \dots, H_{kl_k})$ *are assumed to be Bernoulli distributed over a sigmoid activation function. Then the conditional distribution of* H_{1j} *is given by:*

$$P(H_{1j} \mid \boldsymbol{V},\, \boldsymbol{H}_2) = \mathcal{B}\left(\operatorname{sigm}\left(\sum_{i=1}^{d} W_{1ij}\frac{V_i}{\sigma_i^2} + \sum_{k=1}^{l_2} W_{2jk}H_k^{(2)} + b_j^{(1)}\right)\right)$$

$$P(H_{kj} \mid \boldsymbol{H}_{k-1},\, \boldsymbol{H}_{k+1}) = \mathcal{B}\left(\operatorname{sigm}\left(\sum_{i=1}^{l_{k-1}} W_{kij}H_{(k-1)i} + \sum_{i=1}^{l_{l+1}} W_{(k+1)ji}H_{(k+1)i}\right)\right)$$

$$for\ k \in \{2,\, \ldots\, ,\, L-1\}$$

$$P(H_{Lj} \mid \boldsymbol{H}_{L-1}) = \mathcal{B}\left(\operatorname{sigm}\left(\sum_{i=1}^{l_{L-1}} W_{(L-1)ij}H_{(L-1)i}\right)\right)$$

In contrast to traditional graphical models, deep learning approaches like GBDBMs usually do not prescribe this connection structure, but allow the model to adapt itself to the underlying observations. Thereby complicated non-linear dependencies between the observables are accomplished by indirect connections, that flow through multiply stacked latent variables. Accordingly it is all the more important to incorporate network characteristics, that restrict the model space to reasonable structured models. A particular characteristic of gene regulatory networks is a very small connectivity between 1.5 and 2.5 (Leclerc 2008, p1). This structural knowledge can be incorporated by enforcing this connectivity by an ℓ_1-regularization in the update rule of the CD learning algorithm. This approach, also has positive effects on the tractability in of the local ML estimates: The weight matrix of a GBDBM is comprehensively given by:

$$W := \begin{pmatrix} W_1 & 0 & 0 & 0 \\ W_2^{\mathrm{T}} & W_3 & 0 & 0 \\ 0 & W_4^{\mathrm{T}} & \ddots & 0 \\ 0 & 0 & \ddots & \ddots \end{pmatrix}$$

Thereupon a respective ℓ_1-regularization assures that $\|W\|_1\|W^{\mathrm{T}}\|_1 < 4$. By Corollary 3.3 it then can be concluded, that the Markov Chains in the MCMC approximation is rapid mixing.

4.2 Gene Regulation Analysis of Glioblastoma Multiforme

Glioblastoma multiforme is the most common malignant brain tumor in adults. At the scale of gene regulation Glioblastoma is charcterised by gene deletions which in most cases affect the tumor suppressor gene TP53, the retinoblastoma suppressor gene RB-1 and further deletions in chromosome 22 as well as chromosome 10. Thereby the genetic damages often occur in combination. In newly developed primary glioblastomas more frequent losses of the PTEN gene occur by amplification of the EGFR gene (Ohgaki et al. 2004). In the secondary glioblastomas mutations of the TP53 gene often occur. In addition, point mutations in IDH1 and IDH2 genes coding for an isocitrate dehydrogenase are more common in this group, particularly the R132H mutation in the IDH1 gene (Watanabe et al. 2009). A further potential key player in the is given by Glycosaminoglycans (GAGs) (Afratis et al. 2012). Thereby GAGs play an intricate role in the extracellular matrix by specific interactions with growth factors and other transient components. The accumulated evidence regarding an altered structure of GAG in Glioblastoma indicates their importance disease progression. It may therefore be assumed, that the gene regulation of the GAG degradation pathway is altered by Glioblastoma. To investigate this question from the KEGG database a gene regulatory network has been derived (see figure 4.3).

Thereby the GAG gene regulary network (see figure 4.3) comprises a large list of further candiates, that are assumed to indirectly interact with the GAG degradation pathway. (For convenience these candidates are represented by dots in the figure). Thereupon the challenge is to analyse the gene regulation with respect to cDNA log-ratios of a GBM dataset (see figure 4.4). This sample decomposes into three classes (i) “normal”, comprising a small subgroup of 10 patients, which are diagnosed with no GBM, (ii) “wildtype”, with a subgroup of 487 patients, that are diagnosed with GBM and finally (iii) “mutated” with a subgroup of 19 patients, that are diagnosed with GBM, where the TP53 gene occurs to be mutated.

In a first step the set of candidates has been reduced to genes with high correlation to the GAG degradation pathway. These have been identified by strong low dimensional L-correlations (see figure 4.5).

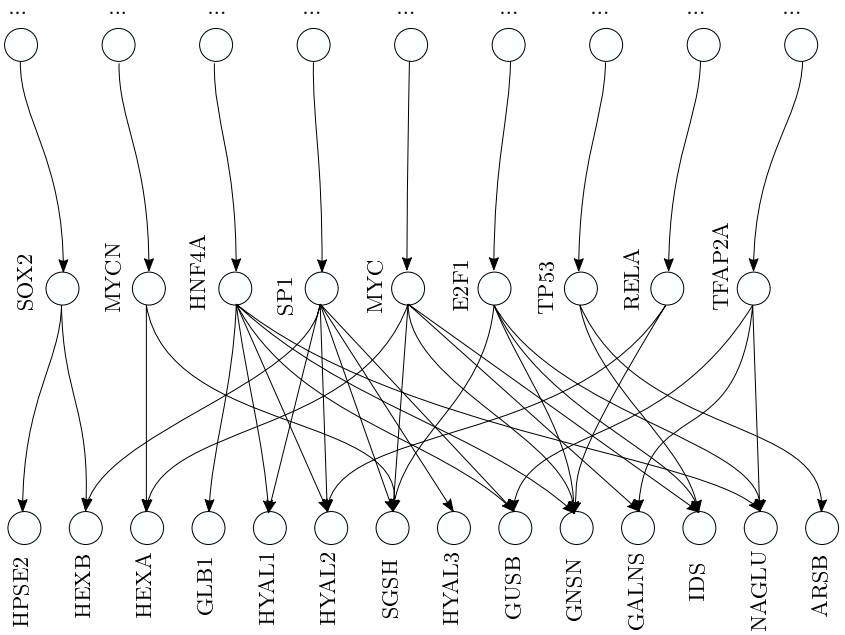

Figure 4.3: Gene Regulatory Network of GAG degradation pathway

Afterwards the GRN was reduced to these respective high correlating genes and embedded within a conditioning EBM (see figure 4.1) and a local ML estimate was derived for all samples by CD learning. The respective posterior in turn was used to initialise a GBDBM (see figure 4.2) which for reasons of tractability has been defined with only three hidden layers, where any subsequent hidden layer only comprised the half number of vertices than its predecessor. Thereupon a local ML estimate for this GBDBM was derived for all samples by CD learning. In order to calculate the $\mathcal{M}$-correlation upon the approximated local ML estimate, the reliabilities R_i have been estimated by the estimated variations $\hat{\sigma}_i^2$, such that for the finite relalization $\boldsymbol{v}$ of $\boldsymbol{V}$, the estimated reliability of X_i is given by:

$$\hat{R}_i = 1 - \frac{\hat{\sigma}_i^2}{\mathrm{Var}_i(\boldsymbol{v})}$$

In order to approximate the integral of $\mathcal{M}$, a MCMC approximation has been used. Thereby a multiplicity of Markov Chains have been initialized by the empirical distribution and repeatedly pushed forward by an alternating Gibbs sampler. The gen-

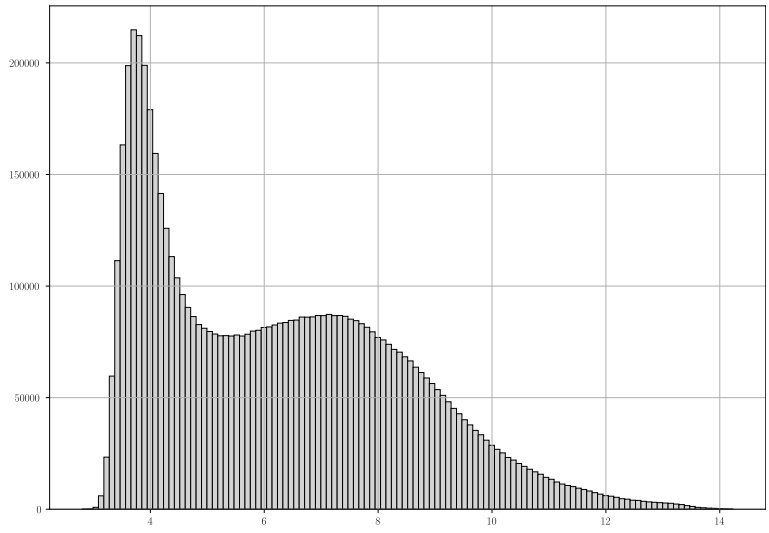

Figure 4.4: Histogram of cDNA log-ratios of GBM dataset

erated samples, that approximate the model distribution afterwards have been used to sample $\boldsymbol{v}_{\mathcal{M}}$ from $\mathcal{M}$ by deterministic sampling of the observables. Then by the law of large numbers it follows, that:

$$\frac{1}{m}\sum_{i=1} S_{i,j}(\boldsymbol{v}^i_{\mathcal{M}})S_{j,i}(\boldsymbol{v}^i_{\mathcal{M}}) \overset{a.s}{\to} \int_{\mathcal{M}} S_{i,j}(\boldsymbol{v})S_{j,i}(\boldsymbol{v})\,\mathrm{d}P_{\mathcal{M}}, \text{ for } m \to \infty$$

Furthermore the local sensitivities $S_{i,j}$ have been approximated by the "bruteforce" approximation:

$$\hat{S}_{i,j} = \partial_j \mathbb{E}(V_i \mid \mathbb{E}(\boldsymbol{H} \mid V_j))$$

This allowed an approximation of the $\mathcal{M}$-Correlation $\rho^2_{V_i,V_j|\mathcal{M}}$. The result has been used to define a graph structure, between the genes where strong $\mathcal{M}$-Correlations are represented by short distances of the respective vertices (see figure 4.6).

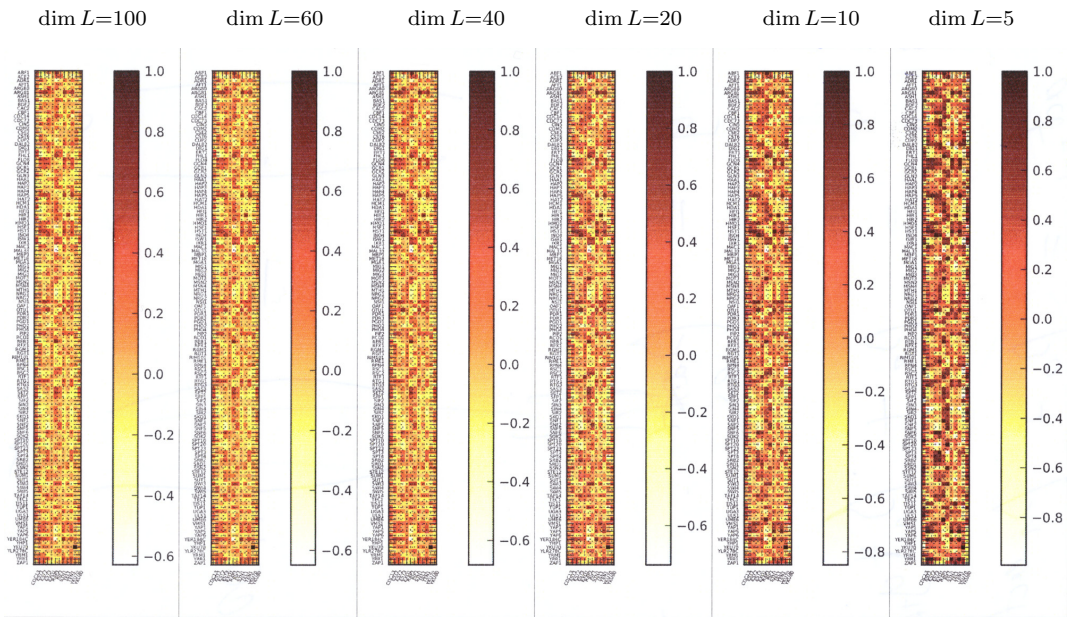

Figure 4.5: L-Correlation for selected genes in GBM dataset

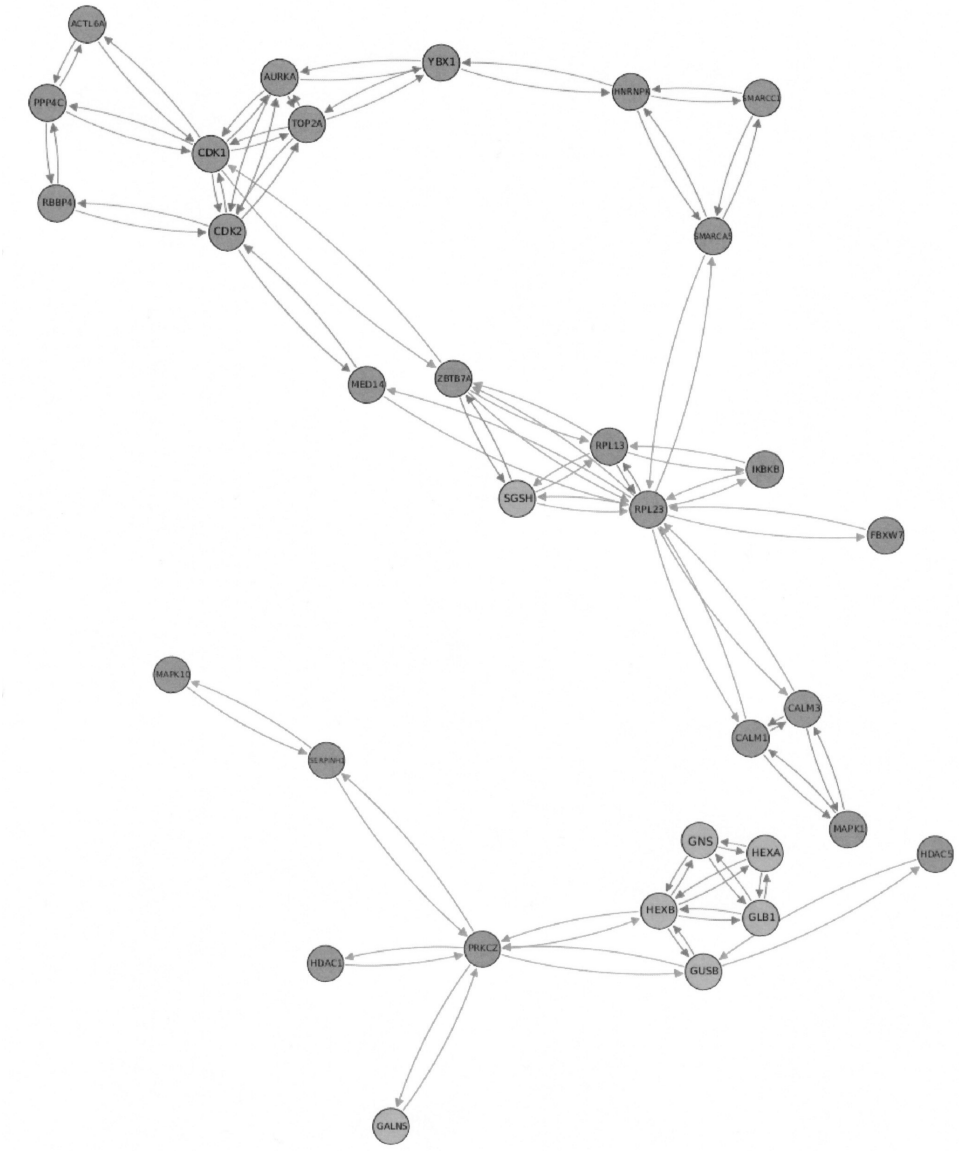

Figure 4.6: Predicted Gene Regulation of GAG degradation pathway in GBM by $\mathcal{M}$-Correlations

Discussion

During the last decade deep structured graphical models motivated many different applications, ranging from classification, regression and dimensionality reduction tasks. The $\mathcal{M}$-Correlation shows, that the scope of their applicability also comprises the quantification of association structures. A great obstacle within the thesis, however, was that an application of the $\mathcal{M}$-Correlation to gene expression data can only hardly be evaluated with respect to its performance. Accordingly a commensurate quantification and comparison to other methods like mutual information based estimates is still pending. Since the opportunities in this direction are multifarious, further research in this area which considers differential and topological proximity structures is required.

References

[1] Nikos Afratis, Chrisostomi Gialeli, Dragana Nikitovic, Theodore Tsegenidis, Evgenia Karousou, Achilleas D. Theocharis, Mauro S. Pavão, George N. Tzanakakis, and Nikos K. Karamanos. Glycosaminoglycans: Key players in cancer cell biology and treatment. *FEBS Journal*, 279(7):1177–1197, 2012.

[2] D J Aldous. Some inqualities for reversible Markov chains. *Journal of the London Mathematical Society*, 25(2):564–576, 1982.

[3] Shun-Ichi Amari. *Information Geometry and Its Applications*. Springer, Tokyo, 1 edition, 2016.

[4] Christof Angermueller, Tanel Pärnamaa, Leopold Parts, and Oliver Stegle. Deep learning for computational biology. *Molecular Systems Biology*, 12(7):878, 2016.

[5] Leon Bottou. Online Learning and Stochastic Approximations. pages 1–34, 1998.

[6] George Casella and Edward I. George. Explaining the Gibbs Sampler. *The American Statistician*, 46(3):167–174, 1992.

[7] Yifei Chen, Yi Li, Rajiv Narayan, Aravind Subramanian, and Xiaohui Xie. Gene expression inference with deep learning. *Bioinformatics*, 32(12):1832–1839, 2016.

[8] Guillaume Desjardins, Aaron Courville, Yoshua Bengio, Pascal Vincent, and Olivier Delalleau. Parallel Tempering for Training of Restricted Boltzmann Machines. *International Conference on Artificial Intelligence and Statistics*, 9:145–152, 2010.

[9] Persi Diaconis and Daniel Stroock. Geometric Bounds for Eigenvalues of Markov Chains. *The Annals of Applied Probabilty*, 1(1):36, 1991.

[10] Scott E Fahlman, Scott E Fahlman, Geoffrey E Hinton, Geoffrey E Hinton, Terrence J Sejnowski, and Terrence J Sejnowski. Massively Parallel Architectures for AI: NETL, THISTLE and Boltzmann Machines. *Proceedings of the 8th International Joint Conference on Artificial Intelligence*, pages 109–113, 1983.

[11] Alan E. Gelfand and Adrian F M Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409, 1990.

[12] Stuart Geman and Donald Geman. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741, 1984.

[13] a N Gorban and a Y Zinovyev. Principal Graphs and Manifolds. *Arxiv preprint arXiv08090490*, page 36, 2008.

[14] Alexander N Gorban and Andrei Y Zinovyev. Elastic Maps and Nets and Their Application to Microarray Data Visualization. pages 1–35.

[15] Trevor Hastie and Werner Stuetzle. Principal Curves. *Journal of the American Statistical Association*, 84(406):502–516, 1989.

[16] G. Hinton and R. Salakhutdinov. Reducing the Dimensionality of Data with Neural Networks. *Science*, 313(July):504–507, 2006.

[17] Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

[18] Daniel Jerison. General mixing time bounds for finite Markov chains via the absolute spectral gap. page 20, 2013.

[19] Marc Jerrum and Alistair Sinclair. Approximating the Permanent. *SIAM Journal on Computing*, 18(6):1149–1178, 1989.

[20] Mark Jerrum and Alistair Sinclair. Polynomial-time approximation algorithms for the ising model. *SIAM Journal on Computing*, 22(5):1087–1116, 1993.

[21] JF Kenney and ES Keeping. Linear regression and correlation. *Mathematics of statistics*, 1:252–285, 1962.

[22] Jamie King. Conductance and Rapidly Mixing Markov Chains. 2003.

[23] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models*. The MIT Press, 2009.

[24] H. J. Kushner and D. S. Clark. Stochastic Approximation for Constrained and Unconstrained Systems. *Applied Math. Sci.*, 26, 1978.

[25] Gregory F. Lawler and Alan D. Sokal. Bounds on the L2 Spectrum for Markov Chains and Markov Processes : A Generalization of Cheeger ' s Inequality. *Transactions of the American Mathematical Society*, 309(2):557–580, 1988.

[26] Robert D Leclerc. Survival of the sparsest: robust gene networks are parsimonious. *Molecular Systems Biology*, 4(213):2004–2009, 2008.

[27] David a. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov Chains and Mixing Times*. 2009.

[28] Xianghang Liu and Justin Domke. Projecting Markov Random Field Parameters for Fast Mixing. *Neural Information Processing Systems*, pages 1–9, 2014.

[29] L. Ljung and T. Söderström. *Theory and Practice of recursive identification*. MIT Press, Cambridge, 1983.

[30] H. Ohgaki and P. Dessen. Genetic pathways to glioblastoma: a population-based study. *Cancer Research*, 64(19):6892–6899, 2004.

[31] Karl Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(1):559–572, 1901.

[32] Ruslan Salakhutdinov and Geoffrey Hinton. Deep Boltzmann Machines. *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, 1(3):448–455, 2009.

[33] Paul Smolensky. Information processing in dynamical systems: Foundations of harmony theory. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations*, pages 194–281. MIT Press, 1986.

[34] G. W. Snedecor and W. G. Cochran. *Statistical Methods*. Iowa State University Press, Ames, 6 edition, 1967.

[35] Arida Ferti Syafiandini, Ito Wasito, Setiadi Yazid, Aries Fitriawan, and Mukhlis Amien. Multimodal Deep Boltzmann Machines for Feature Selection on Gene Expression Data.

[36] Tijmen Tieleman. Training Restricted Boltzmann Machines using Approximations to the Likelihood Gradient. *Proceedings of the 25th International Conference on Machine Learning*, 307:7, 2008.

[37] Tijmen Tieleman and Geoffrey Hinton. Using Fast Weights to Improve Persistent Contrastive Divergence. *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, (8):1033–1040, 2009.

[38] Christopher Tosh. Mixing Rates for the Alternating Gibbs Sampler over Restricted Boltzmann Machines and Friends. *International Conference on Machine Learning*, pages 1–7, 2016.

[39] Martin J Wainwright and Michael I Jordan. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2007.

[40] T. Watanabe and S. Nobusawa. IDH1 mutations are early events in the development of astrocytomas and oligodendrogliomas. *The American Journal of Pathology*, 174(4):1149–1153, 2009.

[41] Tung-Yu Wu, Bai Jiang, Yifan Jin, and Wing H. Wong. Convergence of Contrastive Divergence Algorithm in Exponential Family. pages 1–26, 2016.

[42] Shurong Zheng, Ning-zhong Shi, and Zhengjun Zhang. Generalized Measures of Correlation. *Manuscript*, pages 1–45, 2010.

[43] Fabio Zucca. On Some Properties of Transition Operators. *Extracta Mathematicae*, 17(2):201–209, 2002.

Erklärung

Hiermit erkläre ich, dass ich meine Arbeit selbstständig unter Anleitung verfasst habe, dass ich keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe, und dass ich alle Stellen, die dem Wortlaut oder dem Sinne nach anderen Werken entlehnt sind, durch die Angabe der Quellen als Entlehnungen kenntlich gemacht habe.

Heidelberg, den 25. September 2017 Patrick Michl