DISSERTATION

submitted to the

Combined Faculty for the Natural Sciences and Mathematics

of

Heidelberg University, Germany

for the degree of
Doctor of Natural Sciences

put forward by
M.Sc. Fabian Rathke
born in Berlin

Date of oral examination: .....

Probabilistic Graphical Models for Medical Image Segmentation

Advisor: Prof. Dr. Christoph Schnörr

Zusammenfassung

Die Segmentierung von Bildern ist eine der grundlegenden Aufgaben der Bildverarbeitung. Es existieren viele Abwandlungen, wovon eine die Segmentierung von Schichten mit einer natürlich vorgegebenen *Reihenfolge* ist. Die Zellschichten in der menschlichen Retina stellen eine Instanz dieses Problems dar. Die vorliegende Doktorarbeit untersucht einen Segmentierungsansatz für diese Problemklasse, welcher auf *probabilistischen grafischen Modellen* basiert. Diese Modelle beinhalten das Problem der Infererenz: Wie kann man - gegeben einen Scan der Retina - eine einzelne Vorhersage oder, falls möglich, eine *Verteilung* über Segmentierungen dieses Scans erhalten. Exakte Inferenz ist im Allgmeinen nicht praktikabel, weswegen wir einen approximativen Ansatz untersuchen, der auf *variationeller Inferenz* basiert. Dieser erlaubt die effiziente Approximierung der *vollen* A-posteriori-Wahrscheinlichkeit. Eine charakteristische Eigenschaft unseres Ansatzes ist die Integration einer A-priori-Verteilung über Retinakonturen, welche *nicht* auf lokale Information beschränkt ist. Wir evaluieren unseren Ansatz anhand verschiedener unter anderem auch pathologischer Datensätze. Dabei können wir zeigen, dass globale Konturinformation Segmentierungsergebnisse nach dem Stand der Technik liefert. Da wir die volle A-posteriori-Verteilung inferrieren, ist es uns weiterhin möglich, sowohl die Qualität unserer Vorhersage als auch den Grad der Anomalie des vorliegenden Scans zu bewerten. Motiviert durch unsere Problemstellung haben wir außerdem die nicht-parametrische Dichteschätzung unter der Nebenbedingung der Log-Konkavität untersucht. Diese Klasse von Dichtefunktionen ist auf die konvexe Hülle der empirischen Daten beschränkt. Dies liefert automatisch Konturverteilungen, die die Reihenfolge der Retinaschichten beachten, indem sie ungültigen Konturkonfigurationen keine Wahrscheinlichkeitsmasse zuweisen. Wir untersuchen einen bekannten Ansatz aus der Literatur, zeigen die Erweiterung von 2-D auf N-D und wenden ihn auf Daten der Retina an.

Abstract

Image segmentation constitutes one of the elementary tasks in computer vision. Various variations exists, one of them being the segmentation of layers that entail a natural *ordering* constraint. One instance of that problem class are the cell layers in the human retina. In this thesis we study a segmentation approach for this problem class, that applies the machinery of *probabilistic graphical models*. Linked to probabilistic graphical models is the task of inference, that is, given an input scan of the retina, how to obtain an individual prediction or, if possible, a *distribution* over potential segmentations of that scan. In general, exact inference is unfeasible which is why we study an approximative approach based on *variational inference*, that allows to efficiently approximate the *full* posterior distribution. A distinguishing feature of our approach is the incorporation of a prior shape model, which is *not* restricted to local information. We evaluate our approach for different data sets, including pathological scans, and demonstrate how global shape information yields state-of-the-art segmentation results. Moreover, since we approximatively infer the full posterior distribution, we are able to assess the quality of our prediction as well as rate the scan in terms of its abnormality. Motivated by our problem we also investigate non-parametric density estimation with a log-concavity constraint. This class of density functions is restricted to the convex hull of the empirical data, which naturally leads to shape distributions that comply with the ordering constraint of retina layers, by not assigning any probability mass to invalid shape configurations. We investigate a prominent approach from the literature, show its extensions from 2-D to N-D and apply it to retina boundary data.

Acknowledgments

It was very enjoyable at times and frustrating at others, but retrospectively this PhD was a fantastic experience. Of course this thesis would not have been possible without the support of many people that I worked with and met during my PhD time.

First of all I want to express my gratitude towards my supervisor Christoph Schnörr, whose calm, helpful and inspiring way of providing assistance and guidance was of great help. His kindness, fairness and his absolute commitment to his students have deeply impressed me. And though it may have took some time, I learned a lot.

Next come my colleagues, who made my time in Heidelberg so much more enjoyable. I first of all thank Fabian for sharing the office with me and many pleasant hours laughing, playing Kicker or going to our favorite pub. Next come Berni, Markus and Kira, for having a lot of unforgettable gaming nights with "Frauenbier", good music and Ligretto. The rather absent illumination may have ruined my eyes, but it was worth it. Special thanks also go to Eva for helping me out finishing this thesis by sharing my fate and making spending the weekends at the library much more enjoyable. I also thank Barbara, for many inspiring discussions about life in general and our favorite books/music/movies in particular. The running/swimming/biking/bouldering group, among them Andreea and Florian, for keeping me fit. Andi for providing shelter and "Dosenbier" and playing the drums during our exhaustive band hero nights. Boris and Dajana for at least sometimes going out with us and also for having me for dinner at their place. Karsten for constantly providing good mood. Robert for becoming so fast so good in playing Kicker. "The Dueck" and Dominic from the other side of the Neckar for our absolutely fantastic RTG trips (with one more to come hopefully). Jörg for his entertaining Kicker rules. Evelyn for her warmhearted administrative support. Dorothea for joining our fabulous bike trip across the Alps. Frank for never letting the coffee stream dry up. Paul for his enlightenments about the Protestant church in the 19th century and his very enjoyable humor. Stefania for organizing and providing the main ingredient of our "Wurstparties". Bogdan for making funny sounds while playing Kicker. Henrik for letting me drink his whiskey. And Vera for laughing out loud. My final thanks go to Gabriel, Agnes, Annette, Bernard X., Johannes, Tobias, Tabea, Ecaterina, Mattia and whoever else I may have forgotten during writing that text for being such great colleagues.

Furthermore, there are my friends whom I left behind in Berlin but who kept the contact alive over all those years, welcomed me every time I came to Berlin and visited me in Heidelberg: Annemi, Milan, Kasi, Fiete, Nico, Sarah and many more.

Finally I thank my family (Mutti, Vati and Pia), for preparing delicious meals (Danke Mutti) and having a good time whenever I came for a visit, sending cookies for

Christmas (Danke Pia and Mutti), optimizing my diet (Danke Vati) and supporting my in every possible way during and especially before my PhD. I would not be where I am without you.

Funding by the Deutsche Forschungsgemeinschaft via the research training group 1653 Spatio / Temporal Graphical Models and Applications in Image Analysis is also gratefully acknowledged.

List of Figures

List of Publications

- Fabian Rathke, Stefan Schmidt, and Christoph Schnörr. Order preserving and shape prior constrained intra-retinal layer segmentation in optical coherence tomography. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2011)*, volume 6893, pages 370–377. Springer, 2011.
- Fabian Rathke, Stefan Schmidt, and Christoph Schnörr. Probabilistic intra-retinal layer segmentation in 3-D OCT images using global shape regularization. *Med. Image Anal.*, 18(5):781–794, 2014.

Contents

1 Introduction

1.1 Motivation

Segmentation tasks arise in many areas of computer vision and methodically come in many different flavors. In video surveillance systems segmentation techniques are utilized to detect objects of interest, for example pedestrians and other objects in a street scene [AO11]. Another wide field for the application of segmentation approaches is the analysis of satellite images, with one example being the tracking of sand storms in desert areas [BGF13]. Finally, the umbrella term *medical imaging* entails plenty of different imaging modalities and corresponding tools for evaluation and interpretation. They seek to reveal the internal structures of the human body hidden by skin and bones and ultimately assist in the detection and treatment of diseases.

Optical coherence tomography (OCT) is an imaging modality that measures the delay and magnitude of backscattered light. It is able to generate cross-sectional or three-dimensional images of optical scattering media such as biological tissue. OCT is especially well suited for ophthalmic imaging since naturally, the ocular media allows light to travel with almost no interference, thus enabling micrometer resolution and millimeter penetration depth into the retinal tissue itself [DF08]. Since no other method can perform noninvasive imaging with such a resolution, OCT has become a standard tool in clinical ophthalmology [SPF04]. The recent introduction [dBCP+03, WLK+02] of spectral-domain OCT dramatically increased the resolution as well as the imaging speed and enabled the acquisition of 3-D volumes composed of hundreds of 2-D scans.

Since the manual segmentation of retina scans is tedious and time-consuming, automated segmentation methods becomes evermore important given the growing amount of gathered data. Several studies have shown, that accurate segmentations can facilitate the detection of many common diseases such as glaucoma or age-related macular degeneration [BZB+01, ZNO+07, TLL+08]. Ideally, this is carried out independently of any user interaction, in order to enable the automatic screening of large databases. A probabilistic representation of inferred segmentations is desirable to facilitate subsequent assessments by health professionals. Ultimately, an integrated warning system for the detection of pathologies would be particular valuable for the everyday usage in a clinical environment.

Automatic segmentation approaches face several challenges: The scan quality can be impaired for several reasons. For example blood vessels, located in the outermost cell layer, can cause a shadowing of subsequent layers and thereby a blurred appearance of layer boundaries. Other scan artifacts can arise by the scan process itself. Thus, a segmentation approach that relies solely on appearance

information would in many cases yield unsatisfying results. Furthermore, since texture only constitutes a local feature, texture-only based approaches would lack awareness about the global configuration of the retina.

In general these issues are tackled by adding prior shape knowledge to the segmentation process, that helps gluing together local features in a meaningful way and provides shape-driven hypotheses in regions with poor texture quality. In addition, a shape prior with *global* information allows to draw conclusions about the shape configuration.

1.2 Related Work

We begin with a formal definition of the task at hand:

Definition 1.1 *(Image labeling problem).* Let $\Omega \subset \mathbb{R}^d$ be the image domain ($d = 2, 3$) and $I(x)$ be an image defined on that domain. The image labeling problem consists in finding a partition $\mathcal{P}_K(\Omega)$ into K mutually disjoint subregions Ω_k, that is

$$\mathcal{P}_K(\Omega) \in \left\{ \Omega_k \,\middle|\, \Omega = \bigcup_{k=1}^{K} \Omega_k,\ \Omega_j \bigcap \Omega_i = \emptyset,\ \forall i \neq j \right\} \tag{1.2.1}$$

Any $\mathcal{P}_K(\Omega)$ is called a *segmentation* of I.

The retina segmentation task entails an additional constraint on the *ordering* of partitions Ω_k, corresponding to the given ordering of retina layers. Let x be composed of x_{pos}, the position on the retina surface and x_{depth}, its depth (the direction of incoming light). For any two points $x^j \in \Omega_j$ and $x^k \in \Omega_k$ with $j < k$,

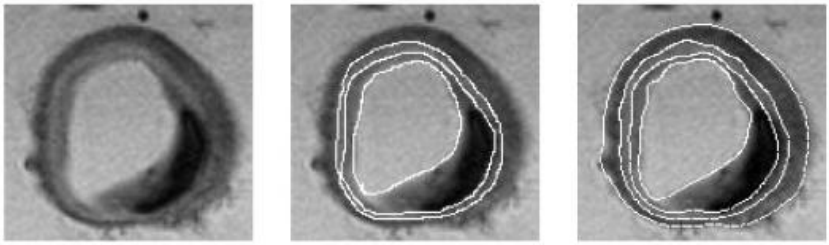
$$x^j_{\text{pos}} = x^k_{\text{pos}} \quad \Longrightarrow \quad x^j_{\text{depth}} < x^k_{\text{depth}} \tag{1.2.2}$$

has to hold. Note that there exist many other instances of that problem class, for example the multi-surface segmentation of arterial walls in vascular MR images[1] [YHSB+03] (upper panel Figure 1.1) or the segmentation of tree rings [CHKM07] (lower panel Figure 1.1).

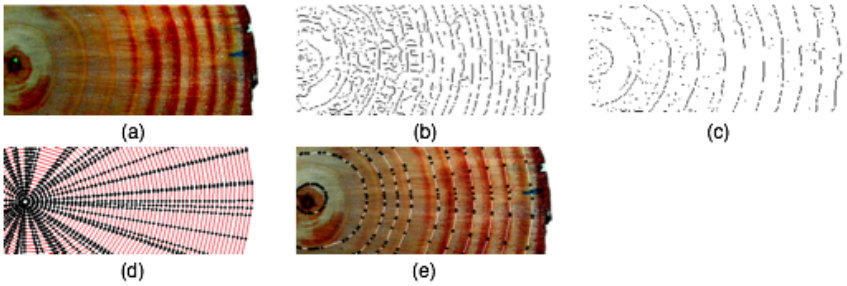
In general, one differentiates between binary segmentation task ($K = 2$) used to differentiate foreground from background in object detection scenarios and those where $K > 2$. For the application considered in this thesis, the resolution of current scanning devices easily allows the meaningful differentiation of $K = 10$ or more distinct cell layers (c.f. Figure 1.2).

Segmentation approaches rely on shape information to deal with missing low-level information, as pointed out above. Although for some industrial applications it may be sufficient to include just one single template shape, in general one needs to include information about shape variations too. A common approach is to gather shape characteristics from an annotated set of training samples and to build a *statistical shape model*.

[1] In terms of polar coordinates (r, φ).



(a) Segmentation of arterial walls



(b) Segmentation of tree rings

Figure 1.1 - Two instances of the segmentation task with *ordering constraint*: (a) The segmentation of arterial walls in vascular MR images (taken from [SHB14]) and (b) the segmentation of tree rings (taken from [CHKM07]). For the former example one has to switch to the polar coordinate system.

The field of different approaches developed in the last decades is vast. Our selection is guided by the set of approaches successfully applied to the retina segmentation problem. In general we can differentiate between *spatially continuous* and *spatially discrete* graph-based approaches. We will first review representatives of the former type.

1.2.1 Spatially Continuous Segmentation Approaches

Active Contour Models. Methods of this group feature an explicit parametric representation of the segmentation curve, which they evolve towards image gradients using partial differential equations. Development started with the seminal paper of [KWT88]. Here a parametric curve $C : [0, 1] \to \Omega$ is driven by a force that pulls the curve towards gradient-based image features while another one controls the length and rigidity of the curve.

An important extension by a (sparse) global shape prior, called the *active shape model*, was proposed by [CTCG95]. Their statistical shape model is governed by a set of control points with associated lower-dimensional latent space found by PCA. During the energy minimization process the movement of the curve is restricted by limiting the variance of the control points projected into the latent space.

An extension of active shape models called *active appearance models* (AAMs) [CET98] evaluates texture information from inside the segmented region. As set of texture control points is defined inside a mean shape and each training image is

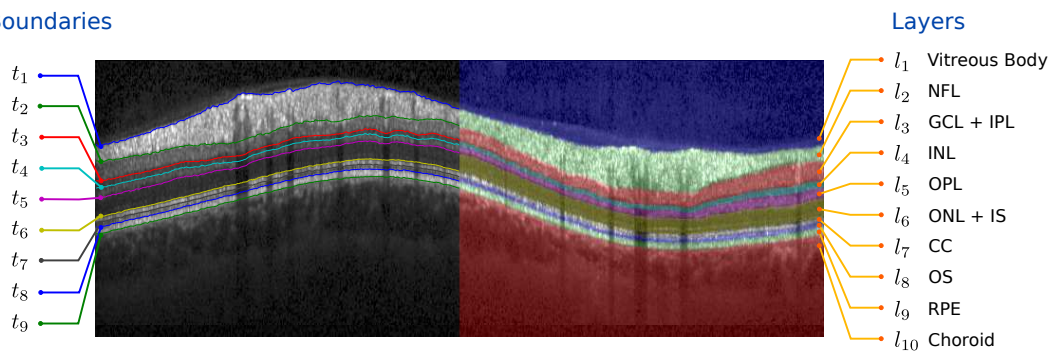


Figure 1.2 - The retinal layers segmented by our approach and their corresponding anatomical names: Nerve fibre layer (NFL), ganglion cell layer and inner plexiform layer (GCL + IPL), inner nuclear layer (INL), outer plexiform layer (OPL), outer nuclear layer and inner segment (ONL + IS), connecting cilia (CC), outer segment (OS), retinal pigment epithelium (RPE). See Section 2.5.2 for more information on some of these layers and the cells they are composed of.

warped to that mean shape to sample texture features and build a latent appearance model, again using PCA. During optimization, the model compares texture features as estimated by the current fit to those found in the image, and proposes new positions for the shape control points.

Although active contour models are very popular for segmentation, they lack a meaningful probabilistic interpretation. Moreover the explicit representation does not easily carry over to $K > 2$ partitions and the optimization of the gradient-based approaches (thus excluding AAM) is plagued by many local minima.

[KPH+10] adapted active appearance models to the retina segmentation task. They use only sparsely sampled landmark points for their statistical shape model, potentially losing information along the way. Furthermore, caused by the inherent properties of the AMM, only a point estimate is inferred.

Level-Set Methods. Level-set methods [OS88], another family of spatially continuous approaches, represent the segmentation curve C implicitly as zero level-lines $C = \{x \in \Omega \,|\, \phi(x) = 0\}$ of some time-evolving embedding function $\phi(x, t)$. The evolution of ϕ is given by the partial differential equation $\frac{\partial \phi}{\partial t} = -|\nabla \phi| F$ where F is the speed function. Alternatively there exist variational formulations based on an energy $E(\phi)$ with corresponding Euler-Langrange equation $\frac{\partial \phi}{\partial t} = -\frac{\partial E(\phi)}{\partial \phi}$ [CRD07].

Like in the case of AAMs, level-set formulations can be extended to include region-based terms, that are governed by intensity, texture or color of objects and background [CRD07]. [LGF00] where the first to integrate shape priors into the level-set framework, driving the embedding function towards a PCA-based representation of sample shapes[2]. Many other shape representations where proposed, see the review of [CRD07] for more details.

Applications of level-set methods to the retina segmentation task [MWBC09, YHSS11, NVT+13] utilize a multiphase formulation with one ϕ_i for each partition

[2] The number of required eigenmodes may be higher than for explicit contours, since PCA captures the variance of the contour C only indirectly via the variance of the embedding function ϕ [CRD07].

boundary. No shape prior [MWBC09], simple distances between layers [NVT+13] or a shape prior enforcing circular-shaped contours [YHSS11] are applied. While the assumption of simple circular contours was justified for the data used in [YHSS11], in general much more complicated shapes are observed.

1.2.2 Spatially Discrete Segmentation Approaches

A very big class of spatially discrete segmentation approaches is based on Markov Random Fields (see Section 2.3.4). Let $\mathcal{G} = (V, E)$ be an undirected graph composed of nodes $i \in V$ corresponding to pixels in I and edges $(i, j) \in E \subset V \times V$ determining the neighborhood of pixels in I, and let $f = \{f_i \in \{1, \dots K\} \,|\, i \in V\}$ denote the labeling corresponding to $\mathcal{P}_K(\Omega)$. The problem of finding a partition $\mathcal{P}_K(\Omega)$ can then be formulated as minimization of the energy function

$$E(f) = \sum_{i \in V} D_i(f_i) + \sum_{(i,j) \in E} W_{ij}(f_i, f_j), \tag{1.2.3}$$

where $D_i(f_i)$ is the cost of assigning label f_i to pixel i and accordingly for $W_{ij}(f_i, f_j)$. Complexity of the problem is determined, among other things, by the type of pairwise cost functions W_{ij} and the complexity of the neighborhood structure denoted by E.

There exist many different optimization techniques to solve (1.2.3), among them message-passing approaches that are covered in Section 2.4, linear-relaxation techniques, combinatorial methods or graph-cuts. Depending on the model structure, any of these techniques may perform best [KAH+13]. The latter technique was previously applied to the retina segmentation task.

Graph-cut methods. Segmentation of graph-based energy minimization schemes based on minimum graph cuts became popular after the seminal work of [BVZ01]. A *cut* $C = (S, T)$ is a partition of the nodeset V into two disjoint subsets by removing all edges that connect these two partitions called the *cut-set*. Each edge has an associated weight, and the cost of the cut is defined as the sum over all edges (that is their respective weights) in the cut-set.

A series of papers [GAW+09, SBG+13, DCA+13] use graph-cut based approaches. They take into account the interaction of neighboring boundaries to mutually restrict their relative positions. This shape prior information is encoded into the graph as hard constraints [GAW+09] or, as recently introduced by [SBG+13] and subsequently extended by [DCA+13], as probabilistic soft constraints. However, due to limitations on complexity, only *local* shape information is included and boundaries are segmented sequentially.

More retina segmentation approaches. Other approaches can not be clearly assigned to any of the segmentation approaches presented above. Several employ rule-based heuristic techniques [ASG+08, FSP05, ISW+05, MHMT10], which for example use outlier detection along with linear interpolation to account for erroneous segmentations. Others [BFT07, YRW+10] use dynamic programming for single Markov chains per boundary and constrain the maximal vertical distance between neighboring boundary positions. [VvdSLdB11] classify pixels using support

vector machines and regularize the output using level-set techniques. None of these approaches incorporates shape prior information.

1.3 Contribution

As pointed out in the previous section, there exists no retina segmentation approach that utilizes a *full global* shape prior. To our knowledge [KPH+10] is the only work that utilizes global shape features, but their active appearance model uses landmark points that represent only 5% of all boundary positions, thereby possibly neglecting crucial information. All other approaches apply at best local marginal shape distributions, that do not take into account long-range interactions.

In this thesis we demonstrate the various benefits of using global shape information, which become apparent by

a) a **state-of-the-art segmentation performance**, outperforming approaches that only utilize local or no shape information,

b) the ability to **assess the quality of the segmentation** and

c) the ability to judge the degree of abnormality of the global retina configuration, that is the **detection of pathologies**. Here, too, we can demonstrate superior performance over state-of-the-art approaches, that only rely on local shape information.

We tackle the retina segmentation problem by combining probabilistic appearance models with a shape prior distribution that takes into account *all* boundary positions and their possible interactions. Both model parts, that are themselves probabilistic graphical models, are merged in a hierarchical probabilistic graphical model. Difficulties arise about how to utilize the global shape information while keeping the computation of posterior distributions over all possible partitions $\mathcal{P}_K(\Omega)$ or modes thereof *tractable*.

Two different schemes were proposed and published separately: [RSS11] and [RSS14]. The former conference paper [RSS11] can be seen as a predecessor of the latter journal paper [RSS14]. While in [RSS11] the observed segmentation performance was satisfactory, the model had several minor and major shortcomings that we addressed in [RSS14]. For example, during inference the former model took only *modes* of the shape prior into account, thereby not utilizing all available information. Also, the probabilistic interpretability of that model was not clear, due to a somewhat unorthodox inference process. This hampered the evaluation of the posterior distribution.

In [RSS14] we completely remodeled the interplay between the shape and appearance components and adopted a more sophisticated inference framework based on variational inference. This enabled us to incorporate full conditional distributions of the shape prior and resulted in a sound probabilistic framework as well. We also extended the approach to 3-D. As a result of that recomposition, we observed a more robust performance for scans of low quality. Moreover, the new inference part enabled

us to infer an approximation to the *full* posterior distribution over segmentations. This in turn made it possible to implement features b) and c) mentioned above. Last but not least, by exploiting the inherent sparsity of the model as well as implementing the crucial part of the model in C, we were able to constrain the time requirements, such that the approach became applicable in a clinical environment.

One problem still remained: The support of our shape prior distribution is *not* constrained to partitions that are valid in a biological sense. In the models above we dealt with that issue in a sub-optimal way by simply ignoring probability mass assigned to shape configurations violating the natural order of layers. As a preparatory step to rectify this, we investigated the class of non-parametric log-concave density estimators. As one of their crucial characteristics, their support is constrained to the sample data set. And since samples represent only valid configurations, we obtain a density that obeys the ordering constraint.

Investigating the approach of [KM10], we proposed an alternative optimization in terms of the primal which yields similar performance as the Mosek-based implementation of Koenker et al. We furthermore elaborated the extension from 2-D to N-D mentioned in [KM10] and present results for the 3-D case. The integration into our segmentation model remains as future work though.

In order to ensure a clear and well-structured presentation, we decided to only present the journal paper [RSS14] in this thesis, while we refer to [RSS11] for details of our first approach. Moreover we compactly confined the segment about log-concave density estimation to the second last chapter.

1.4 Thesis Outline

Preliminaries are covered in **Chapter 2**. We first address the basics of convex analysis, then present probability theory from a measure-theoretic point of view. After we examined some necessary concepts of graph theory, we are in a position to introduce probabilistic graphical models, providing the theory underlying our retina segmentation model. The next section sheds light on inference methods in graphical models with the focus on the methodology of variational inference. Finally, the last section will introduce physiological background information about the retina segmentation task.

Chapter 3 introduces our probabilistic graphical model for retina segmentation and presents each component separately. We then point out the intractability of directly inferring probability distributions and outline how variational inference can be applied to obtain approximative posterior distributions in a deterministic fashion.

Chapter 4 presents the three different data sets of retina scans used in this work. Afterwards we evaluate segmentation performance and demonstrate how the inferred approximative posterior distributions can be utilized to recognize pathological scans and also determine the quality of the segmentations.

In **Chapter 5** we address the problem of the Gaussian shape prior not being faithful to the ordering of the retinal layer boundaries. We introduce, extend and implement an existing approach to estimate log-concave densities on the support

of the sample set, thereby automatically taking into account the natural ordering constraints for retina boundaries.

Finally, we draw a conclusion in chapter **Chapter 6** and discuss possible directions of future work.

An **Index**, located before the bibliography, collects all keywords that are marked bold throughout this thesis for reference.

1.5 Notation

The following table gives an overview of relevant notation. This serves as a reference, while each section introduces the notation it requires separately.

Table 1.1 - Relevant notation used throughout this thesis.

	Convex Analysis (Section 2.1)
$\overline{\mathbb{R}}$	Extended reals: $\mathbb{R} \cup \{\infty, \infty\}$
$\operatorname{int} S$	Interior of the set S
$\bar{S}$	Closure of the set S
$\operatorname{bd} S$	Boundary of the set S, $\operatorname{bd} S = \bar{S} \setminus \operatorname{int} S$
$\operatorname{conv} S$	Convex hull of the set S
$\mathcal{S}^n$	Space of symmetric $n \times n$ matrices
$\mathcal{S}^n_+$ $[\mathcal{S}^n_{++}]$	Cone of symmetric, positive semidefinite [definite] $n \times n$ matrices
$\preceq_K$ $[\prec_K]$	Generalized inequality: $x \preceq_K y \iff y - x \in K$ $[y - x \in \operatorname{int} K]$
$\succeq$ $[\succ]$	$X \succeq 0 \iff X \in \mathcal{S}^n_+$ $[X \succ 0 \iff X \in \mathcal{S}^n_{++}]$
$\delta_C(x)$	Indicator function of the convex set C
$\operatorname{epi} f$	Epigraph of the function f
	Measure and probability theory (Section 2.2)
Ω	Sample space
ω	Outcome of a random experiment, $\omega \in \Omega$
A	Event, $A \subseteq \Omega$
$\mathcal{F}$	σ-algebra; collection of subsets of Ω
P	Measure on the measurable space $(\Omega, \mathcal{F})$
X, Y	Random variables, $X : \Omega \to \mathbb{R}$ (or $X : \Omega \to \mathbb{R}^n$)
x, y	Realizations of X, Y (elements of the codomain)
$\mathcal{B}$	Borel algebra on $\mathbb{R}$ (or $\mathbb{R}^n$)
P_X	Measure on the measurable space $(\mathbb{R}, \mathcal{B})$ induces by X
$p(x)$	Density function corresponding to P_X
$H[p]$	Shannon entropy of $p(x)$
$d\nu(x)$	Base measure (counting measure or Lebesgue measure)
	Probabilistic graphical models (Section 2.3)
$\mathcal{G} = (V, E)$	Graph composed of a set of nodes V and edges $E \subseteq V \times V$
$(i \to j)$	Directed edge from node i to node j
(i, j)	Undirected edge between nodes i and j

$\mathrm{pa}(i)$	Parents of node i: $j \in \mathrm{pa}(i) \Leftrightarrow (j \to i) \in E$
$\mathrm{ch}(i)$	Children of node i: $j \in \mathrm{ch}(i) \Leftrightarrow (i \to j) \in E$
$\mathrm{ne}(i)$	Neighbors of node i: $j \in \mathrm{ne}(i) \Leftrightarrow (i, j) \in E$
$X \perp\!\!\!\perp Y \mid Z$	Conditional independence of X and Y given Z
	Variational inference (Section 2.4.2)
$\phi(x)$	Sufficient statistics
θ	Canonical parameters
$A(\theta)$	Log partition function
p_θ	Density from the exponential family: $p_\theta(x) = \exp\{\langle\theta, \phi(x)\rangle - A(\theta)\}$
μ	Mean parameters (dual to θ)
$A^*(\mu)$	Conjugate of $A(\theta)$
	Retina segmentation model (Chapter 3)
N, M	OCT scan dimensions (rows, columns)
N_b	Number of segmented boundaries; $N_b = 9$ in this paper
i, j, k	Indices of N, M and N_b: $i = 1, \ldots, N,\ j = 1, \ldots, M,\ k = 1, \ldots, N_b$
$b_{k,j} \in \mathbb{R}$	Real-valued location of boundary k in column j (relative depth)
$c_{k,j} \in \{1, \ldots, N\}$	Integer-valued boundary variables analogous to b, but specifying row-positions on the pixel grid
$x_{i,j} \in \mathcal{X}$	Class variables indicating membership to layer or transition classes
$y_{i,j}$	Observed data; here patches around pixel (i, j)
q_c, q_b	Approximating densities of the posterior: $p(b, c\|y) \approx q_c(c)q_b(b)$
	Log-concave density estimation (Chapter 5)
$f(x)$	Log-concave density function
$g(x)$	Convex function, such that $f(x) = -\log g(x)$
X	Set of sample points $x^i \in \mathbb{R}^d$
$\mathcal{G}(X)$	Cone of polyhedral convex functions on $\mathrm{conv}\, X$
$C(X)$	Space of continuous functions on $\mathrm{conv}\, X$
$C^*(X)$	Space of signed, finite, regular Borel measures, dual to $C(X)$
$\mathcal{K}(X)$	Cone of closed (lower semicontinuous) convex functions on $\mathrm{conv}\, X$
$\mathcal{K}^*(X)$	Polar cone of $\mathcal{K}(X)$
P_n	Empirical measure corresponding to X
$\mathcal{L}^n$	Lorentz cone: $\mathcal{L}^n = \{(x, z) \in \mathbb{R}^{n+1} \mid z \geq \|x\|_2\}$

2 Preliminaries

This chapter will introduce all the necessary tools which are required in the remainder of this thesis. Convex analysis lies at the heart of many mathematical areas and we will give a short introduction in Section 2.1. Section 2.2 follows with a treatise of probability theory from a measure-theoretic point of view. A significant part of our graphical model relies on the multivariate normal distributions, which we introduce at the end of that section. There, it is a pivotal question how to obtain regularized estimates for the covariance matrices, and we will address this issue by presenting two approaches from the literature. Section 2.3 and Section 2.4 present the main theory behind probabilistic graphical models and how inference is performed with the focus on variational inference. Finally, Section 2.5 comprises some background material regarding the image data used in this thesis, human retina scans acquired using Optical Coherence Tomography. We also give a short survey about retinal anatomy and how one of its most numerous diseases, glaucoma, affects its structure.

2.1 Convex Analysis

In this section the necessary terminology of convex analysis will be presented. For a much more thorough treatment we refer to the books [BV04] and [Roc70]. To ease presentation, all definitions will be given in terms of the Euclidean space $\mathbb{R}^n$, but can readily be rewritten in terms of a generic vector space X.

We denote the extended real numbers $\mathbb{R} \cup \{-\infty, +\infty\}$ by $\overline{\mathbb{R}}$ and the space of symmetric $n \times n$ matrices by $\mathcal{S}^n$.

Definition 2.1 *(Convex set).* A set $C \subseteq \mathbb{R}^n$ is convex, if for any $x, y \in C$

$$\lambda x + (1 - \lambda)y \in C, \qquad \forall \lambda \in [0, 1]. \tag{2.1.1}$$

Definition 2.2 *(Indicator function of a convex set).* We the define the convex **indicator function**[1] of the convex set C as

$$\delta_C(x) = \begin{cases} 0 & x \in C \\ +\infty & x \notin C. \end{cases} \tag{2.1.2}$$

Definition 2.3 *(Cones).* A set $K \subseteq \mathbb{R}^n$ is called a cone, if $\lambda x \in K$ for all $x \in K$ and $\lambda \geq 0$. If K is also convex it is called a **convex cone**.

We call K a **proper cone**, if it is closed, convex, has nonempty interior and is pointed, that is $x \in K \Rightarrow -x \notin K$, except if $x = 0$. Two important proper cones are

[1] The concept of a convex function will be introduced in Definition 2.7.

$\mathcal{S}^n_+$ and $\mathcal{S}^n_{++}$, the cones of symmetric, positive semidefinite and definite matrices.

Definition 2.4 *(Polyhedron and polytope).* We call the intersection of a finite number of half-spaces

$$C = \{x \,|\, x^T b_i \leq \xi_i,\ i = 1, \ldots, n\}, \tag{2.1.3}$$

a **polyhedron**. If C is bounded we call it a **polytope**.

Definition 2.5 *(Convex Hull).* Given a set $C \subset \mathbb{R}^n$, we call the set of all convex combinations of points in C,

$$\operatorname{conv} C = \left\{ \lambda_1 x_1 + \ldots + \lambda_k x_k \,|\, x_i \in C,\ \lambda_i \geq 0,\ i = 1, \ldots, k,\ \sum_i \lambda_i = 1 \right\}, \tag{2.1.4}$$

the **convex hull** of C.

A convex hull generated by a *finite* set of points is a polyhedron. Both representations, by a set of points and by a set of half-spaces, are *dual* to each other and serve as an example for conjugate duality, described below[2].

Definition 2.6 *(Generalized inequality).* We define a **generalized inequality** as the partial ordering on $\mathbb{R}^n$ associated with the proper cone K

$$x \preceq_K y \iff y - x \in K, \tag{2.1.5}$$

and the strict partial ordering

$$x \prec_K y \iff y - x \in \operatorname{int} K. \tag{2.1.6}$$

We will use $X \succeq 0$ as a short form for $-X \preceq_{\mathcal{S}^n_+} 0$ and $X \succ 0$ for $-X \prec_{\mathcal{S}^n_+} 0$.

Definition 2.7 *(Convex function).* A function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is called convex, if it satisfies Jensen's inequality for any $x, y \in \mathbb{R}^n$, that is

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y), \qquad \forall \lambda \in [0,1]. \tag{2.1.7}$$

A dual characterization of a convex function is in terms of its **epigraph**:

$$\operatorname{epi} f := \{(x, \mu) \in \mathbb{R}^n \times \mathbb{R} \,:\, \mu \geq f(x)\}. \tag{2.1.8}$$

Convexity of the set $\operatorname{epi} f$ is equivalent to convexity of f.

An elementary property of convex sets is the fact that they possess a dual representation in terms of the intersection of all half-spaces that contain them. Let f be a convex function. For $\operatorname{epi} f$ to be contained in the half-space corresponding to the hyperplane $h(x) = x^T \xi - \mu$, we require that

$$\sup_{x \in \operatorname{dom} f} \{h(x) - f(x)\} \leq 0 \quad \iff \quad \sup_{x \in \operatorname{dom} f} \{x^T \xi - f(x)\} \leq \mu$$

This determines a function whose graph describes the set of all half-spaces tangent

[2] The indicator function δ_C and the support function δ^*_C of a set C are conjugate to each other.

to the graph of f:

Definition 2.8 *(Conjugate duality).* Let $f : \mathbb{R}^n \to \mathbb{R}$. The function $f^* : \mathbb{R}^n \to \mathbb{R}$ defined as

$$f^*(\xi) = \sup_{x \in \operatorname{dom} f} \{x^T \xi - f(x)\} \tag{2.1.9}$$

is called the **conjugate function**.

It immediately follows from the definition that $x^T \xi \leq f(x) + f^*(\xi)$, known as *Fenchel's inequality*. In this inequality the role of both functions seems interchangeable. This is true iff f is a convex function in which case we can equally express f in terms of f^*:

$$f(x) = f^{**}(x) = \sup_{\xi \in \operatorname{dom} f^*} \{x^T \xi - f^*(\xi)\}. \tag{2.1.10}$$

Thus we gained an alternative representation of f in terms of an optimization formulation. This is also called the *variational* representation of f and will become important again when we deal with variational inference in Section 2.4.2.

Definition 2.9 *(Convex minimization problem).* The convex minimization problem subject to a set of generalized inequality constraints is given by

$$\text{minimize}_{x \in \mathcal{D}} \; f_0(x) \quad : \quad f_i(x) \preceq_{K_i} 0, \quad \forall i = 1, \ldots, n \tag{2.1.11}$$

with convex domain $\mathcal{D} = \cap_{i=0}^n \operatorname{dom} f_i$, convex function $f_0 : \mathbb{R}^n \to \mathbb{R}$, K_i-convex functions[3] $f_i : \mathbb{R}^n \to \mathbb{R}^{k_i}$ and proper cones $K_i \subseteq \mathbb{R}^{k_i}$.

Using **logarithmic barrier functions**, problem (2.1.11) can be readily transformed into an unconstrained optimization problem. Barrier functions $-\psi(-f_i(x))$ can be understood as differentiable approximations to indicator functions

$$I_{K_i}(f_i(x)) = \begin{cases} 0 & f_i(x) \preceq_{K_i} 0, \\ \infty & \text{else.} \end{cases}$$

We require that $-\psi$ is a convex, closed and continuously differentiable function, $\operatorname{dom} \psi(-f_i(x)) = \operatorname{int} K_i$ and $\nabla^2 \psi(y) \prec 0$ for any $y \in \operatorname{int} K_i$ [BV04].

Example 2.10 *(Barrier function for the positive semidefinite cone).* Let $A \in \mathcal{S}^n$, then $\psi(A) = \log\det A$ is the logarithmic barrier function for the generalized inequality $-A \preceq_{\mathcal{S}^n_+} 0$, since

$$\log\det A = \log \prod_i \lambda_i = \sum_i \log \lambda_i. \tag{2.1.12}$$

Using barrier functions, we can reformulate (2.1.11) into the unconstrained convex problem

$$\text{minimize}_{x \in \mathcal{D}} \; t f_0(x) - \sum_{i=1}^n \psi(-f_i(x)) := \pi(x). \tag{2.1.13}$$

Note that we introduced the factor t which governs the relative weight between the

[3] K_i-convexity amounts to replacing $\leq$ with $\preceq_{K_i}$ in Equation (2.1.7).

objective term f_0 and the log-barrier terms. A valid strategy to optimize $\pi(x)$ is to initialize x with a feasible x_0 and set $t = 1$, and then to solve a series of unconstrained optimization problems while resetting $t := \mu t$ after each step, where $\mu > 1$. This approach is called the **barrier method**, which is guaranteed to converge towards the global optimum of problem (2.1.11) as $t \to \infty$ [BV04].

For the inner optimization steps one usually uses **Newton's method**, which is a step-wise descent method

$$x^{k+1} := x^k + \lambda \Delta x^k, \tag{2.1.14}$$

with step size λ and Newton step Δx^k, defined as

$$\Delta x^k := -[\nabla^2 \pi(x^k)]^{-1} \nabla \pi(x^k), \tag{2.1.15}$$

consisting of the inverse Hessian and gradient of $\pi(x)$. We will make use of logarithmic barrier functions and the barrier method in Section 5.3.3.

2.2 Probability Theory

Length, area, volume are all different instances of the concept of *measure*. Also the theory of probability is build upon measure theory and this section will develop the necessary concepts. It orients itself towards the book [ADD00], to which we refer for a more detailed treatment of the topic.

2.2.1 Probability Space

We denote by Ω the **sample space**, that is the set of all possible outcomes ω of a random experiment. The set Ω can be finite, e.g. $\Omega = \{\omega_1, \ldots, \omega_n\}$, countably infinite, e.g. $\Omega = \mathbb{N}$ or uncountably infinite, e.g. $\Omega = \mathbb{R}$. We call any subset of Ω an **event** and denote it by A. Finally, a **measure** P is a set function that assigns a number $P(A)$ to each set A.

For a measure to be well-defined, we need to impose certain constraints on the collection of subsets of Ω:

Definition 2.11 *(σ-algebra).* Let $\mathcal{F}$ be a collection of subsets of Ω. $\mathcal{F}$ is called a **σ-algebra** or σ-field iff $\Omega \in \mathcal{F}$ and $\mathcal{F}$ is closed under complementation and countable union, that is:

a) $\Omega \in \mathcal{F}$.

b) If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$ (where A^c is the complement of A that is $A^c := \Omega \setminus A$).

c) If $A_1, A_2, \ldots, A_n \in \mathcal{F}$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{F}$.

It follows by the DeMorgan laws, that $\mathcal{F}$ is closed under countable intersection, since $\cap_{i=1}^{\infty} A_i = (\cup_{i=1}^{\infty} A_i^c)^c \in \mathcal{F}$.

We call the pair $(\Omega, \mathcal{F})$ a **measurable space** and any $A \in \mathcal{F}$ is called **$\mathcal{F}$-measurable**.

The smallest σ-algebra consists of the two sets Ω and $\emptyset$. Contrary, we denote by 2^{Ω} the **power set** of ω, that is the set of all subsets of Ω. Given a collection $\mathcal{C}$ of

subsets of Ω, there exists a *unique smallest* σ-algebra containing $\mathcal{C}$, that is said to be generated by $\mathcal{C}$. It is defined as the intersection of all σ-algebras containing $\mathcal{C}$.

A very important σ-algebra, generated by all open intervals (a, b), is the **Borel algebra** on $\mathbb{R}$ denoted by $\mathcal{B}$. Elements of $\mathcal{B}$ are called **Borel sets** and one can show that $\mathcal{B}$ contains all types of intervals such as $[a, b]$ or $[a, b)$ and consequently (by means of countable union and intersection) all open and all closed sets.

Given a measurable space $(\Omega, \mathcal{F})$, we now want to assign a probability to each event $A \in \mathcal{F}$.

Definition 2.12 *(Probability measure).* Given a non-empty set Ω and a σ-algebra $\mathcal{F}$. We call the mapping $P : \mathcal{F} \to [0, 1]$ a **probability measure** iff the three Kolmogorov axioms are satisfied:

1. $P(A) \in \mathbb{R}$, $P(A) \geq 0 \quad \forall A \in \mathcal{F}$.

2. $P(\Omega) = 1$.

3. (Countable additivity) For a sequence of disjoint sets $\{A_i\}$ we require $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

We call the triple $(\Omega, \mathcal{F}, P)$ a **probability space** and the number $P(A)$ the probability of A. If only the first and third condition holds, we call P a **measure**.

2.2.2 Random Variables

We continue with one of the central concept of probability theory, that of a **random variable** X. From a measure-theoretic point of view, X is a function that maps each outcome $\omega \in \Omega$ to the reals or extended reals. Intuitively it is a quantity that is measured in connection with a random experiment, for example the height of a person who is randomly drawn from the sample space Ω, containing all inhabitants of this planet.

Suppose we are interested in probabilities of the form $a \leq X(\omega) \leq b$ for all $a, b \in \mathbb{R}$, that is we want to compute $P(\{\omega \,|\, X(\omega) \in B\})$ for events of the form $B = [a, b]$. For this to be possible, sets $X^{-1}(B)$ must be $\mathcal{F}$-measurable for each interval B.

Definition 2.13 *(Random variable).* A function $X : \Omega \to \mathbb{R}$ is a random variable if the set $\{\omega \,|\, a \leq X(\omega) \leq b\}$ is $\mathcal{F}$-measurable for all $a, b \in \mathbb{R}$.

Since the set of intervals $\{[a, b] \,|\, a, b \in \mathbb{R}\}$ generates the Borel algebra $\mathcal{B}$, it can be shown that $X^{-1}(B)$ is $\mathcal{F}$-measurable for every Borel set. This induces that the probability $P(\{\omega \,|\, X(\omega) \in B\})$ is well-defined for all $B \in \mathcal{B}$.

Definition 2.14 *(The probability law of a random variable).* Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $X : \Omega \to \mathbb{R}$ be a random variable. For every Borel set B we define

$$P_X(B) = P(\{\omega \,|\, X(\omega) \in B\}), \qquad B \in \mathcal{B}. \tag{2.2.1}$$

The resulting function $P_X : \mathcal{B} \to [0, 1]$ is called the **probability law** of X.

The measurable map X induces a *push-forward* operation that takes the measure P on $(\Omega, \mathcal{F})$ to the measure P_X on $(\mathbb{R}, \mathcal{B})$. The measure P_X may be characterized by a single function from $\mathbb{R}$ to $\mathbb{R}$.

Definition 2.15 *(Cumulative distribution function).* The **cumulative distribution function** (cdf) of a random variable X is the function $F : \mathbb{R} \to [0, 1]$, given by

$$F(x) = P(\{\omega \,|\, X(\omega) \leq x\}), \qquad x \in \mathbb{R}. \tag{2.2.2}$$

It can be shown that $F(x) \to 0$ as $x \to -\infty$ and $F(x) \to 1$ as $x \to \infty$.

In practice the original probability space $(\Omega, \mathcal{F}, P)$ generally remains in the background, while one works with the more accessible probability space $(\mathbb{R}, \mathcal{B}, P_X)$[4].

2.2.3 Probability Density Functions

We say that a random variable X is **discrete**, if its range $X(\Omega)$ is countable.

Definition 2.16 *(Probability mass function).* Given a discrete random variable X, we can define its **probability mass function** (pmf) $p_X : \mathbb{R} \to [0, 1]$ as

$$p_X(x) = P(\{\omega \,|\, X(\omega) = x\}), \qquad x \in \mathbb{R}.$$

The measure corresponding to X is the countable sum

$$P_X(B) = \sum_{x \in B} p_X(x), \tag{2.2.3}$$

as p_X is 0 except at a countable set $\{x_n, n = 1, 2, \ldots\}$.

The definition of a continuous random variable is more subtle:

Definition 2.17 *(Probability density function).* A random variable is called **(absolutely) continuous**, iff there exists a nonnegative Borel measurable function $f : \mathbb{R} \to [0, \infty)$ such that

$$F(x) = \int_{-\infty}^{x} f(t)dt, \qquad x \in \mathbb{R}, \tag{2.2.4}$$

with $F(x)' = f(x)$ almost everywhere. We call f the **(probability) density function** (pdf) of X. It follows that

$$P_X(B) = \int_B f(x)dx \quad \text{for each} \quad \mathrm{B} \in \mathcal{B}. \tag{2.2.5}$$

[4] Most of the time it is the other way round: Given a distribution function F, the measure P_X it induces and no reference to the underlying probability space, one can construct a canonical probability space by taking $\Omega = \mathbb{R}, \mathcal{F} = \mathcal{B}$ and define X as the identity map $X(\omega) = \omega$, which induces the measure $P(B) = P_X(B)$.

2.2.4 Random Vectors

If one associates more than one random variable with the same experiment, we speak of a n-dimensional **random vector** $X : \Omega \to \mathbb{R}^n$, which is a n-tuple $(X_1, \ldots, X_n)$ of random variables such that each X_i is Borel measurable. The probability measure induced by X is $P_X(B) = P(\{\omega \,|\, X(\omega) \in B\})$ for all $B \in \mathcal{B}$, where $\mathcal{B}$ now denotes the σ-algebra generated by all open sets in $\mathbb{R}^n$.

Definition 2.18 *(Joint distribution function).* We call $F(x_1, \ldots, x_n)$ the **joint distribution function** defined by

$$F(x) = P_X((-\infty, x]) = P(\{\omega \,|\, X_i(\omega) \leq x_i,\ i = 1, \ldots, n\}). \tag{2.2.6}$$

Much of the previous development carries over, with analogous definitions of the joint pmf and pdf.

Before we continue, *some remarks about terminology* are in order: By virtue of the Radon-Nikodym theorem[5], probability mass functions can be considered density functions, by replacing the Lebesgue measure with the counting measure. This motivates a unified treatment of pdfs and pmfs and we denote both of them by $p(x)$ for the remainder of this thesis. We will also sometimes use the term **probability distribution** or just distribution to mutually refer to both, pdfs and pmfs. The base measure, referring to the Lebesgue measure or the counting measure, will be denoted by $d\nu(x)$.

Given $X : \Omega \to \mathbb{R}^n$, we call $p(x)$ the **joint distribution** of X and the distribution $p(x_A)$ of any subset X_A, $A \subset \{1, \ldots, n\}$ the **marginal distribution** of X_A.

Definition 2.19 *(Independent random variables).* Random variables $X_1, \ldots, X_n$ are called **independent**, iff their joint distribution $p(x_1, \ldots, x_n)$ fully factorizes:

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i) \tag{2.2.7}$$

If independence between two random variables X, Y does not hold, then the knowledge about the outcome of Y *does* change the probabilities related to the possible outcomes of X. Thus we need to replace the marginal distribution $p(x)$ by a revised distribution.

Definition 2.20 *(Conditional probability).* Let (X, Y) be a random vector of arbitrary dimension. We denote by $p(x|y)$ the **conditional distribution** of X given that we know the realization of Y. It holds that

$$p(x|y) = \frac{p(x, y)}{p(y)}, \tag{2.2.8}$$

provided $p(y) > 0$.

Given that definition, one of the most important tools in statistical inference arises: **Bayes theorem** or Bayes rule.

[5] A fundamental theorem of measure theory, see Chapter 2 of [ADD00].

Theorem 2.21 *(Bayes theorem).* Given two random variables X, Y, we can express the conditional distribution $p(x|y)$ in terms of the conditional distribution $p(y|x)$ as follows:

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} = \frac{p(y|x)p(x)}{\int p(y|x)p(x)d\nu(x)}. \tag{2.2.9}$$

The second equality applies the **law of total probability**, a useful tool to calculate marginal distributions. In statistical inference, the terms in Bayes theorem are often denoted by

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}}, \tag{2.2.10}$$

and we will from now on frequently make use of them.

The final definitions will be useful to obtain compact characterizations of probability distributions.

Definition 2.22 *(Expectation).* Let $X = (X_1, \ldots, X_n)$ be a random vector and let g be a Borel measurable function from $\mathbb{R}^n$ to $\mathbb{R}$. We define the **expectation** of g in terms of the density function $p(x)$ as

$$E[g(X)] = \int g(x)p(x)d\nu(x), \tag{2.2.11}$$

provided the integral exists.

Definition 2.23 *(Shannon entropy).* Let X be a random variable and $p(x)$ its associated density function. Setting $g(x) = -\log p(x)$ yields the **Shannon entropy** or simply the **entropy** of X, defined as

$$H[p] := E[-\log p(X)] = -\int p(x) \log p(x)d\nu(x). \tag{2.2.12}$$

Definition 2.24 *(Moments of a random variable).* Let $X : \Omega \to \mathbb{R}$ be a random variable, then $E[X^k]$ is called the kth **moment** of X and $E[(X - E[X])^k]$ is called the kth **central moment** of X.

The first moment ($k = 1$) is often called the **mean** of X, whereas the second central moment is called the **variance** of X, sometimes written as Var(X) and abbreviated by σ^2. Furthermore, the square root of the variance, denoted by σ, is called the **standard deviation** of X and the inverse variance, denoted by τ, the **precision** of X. Finally the **covariance** of random variables X and Y is defined by

$$\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]. \tag{2.2.13}$$

2.2.5 Multivariate Normal Distribution

The **multivariate normal distribution**, also called multivariate Gaussian distribution, is one of the workhorses of statistical inference and will be used extensively throughout this thesis.

Definition 2.25 *(Multivariate normal distribution).* A continuous random vector

$X = (X_1, \dots, X_n)$ is said to be normally distributed, denoted by $X \sim \mathcal{N}(\mu, \Sigma)$, if its joint density function is

$$p(x) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\Big(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\Big) \tag{2.2.14}$$

with mean vector $\mu \in \mathbb{R}^n$ and (symmetric, positive definite) **covariance matrix** $\Sigma \in \mathbb{R}^{n\times n}$, $\Sigma_{ij} = \mathrm{Cov}(X_i, X_j)$. The inverse $K := \Sigma^{-1}$ is called the **precision matrix**.

The marginal and conditional distribution of a normal distribution are again normal. Let us define the following partitions

$$X = \begin{pmatrix} X_A \\ X_B \end{pmatrix}, \qquad \mu = \begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix},$$

and accordingly

$$\Sigma = \begin{pmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{pmatrix}, \qquad K = \begin{pmatrix} K_{AA} & K_{AB} \\ K_{BA} & K_{BB} \end{pmatrix}.$$

Proposition 2.26 *(Marginal and conditional normal distribution [RW06]).* The marginal distribution of X_A is given by

$$X_A \sim \mathcal{N}(x_A; \mu_A, \Sigma_{AA}). \tag{2.2.15}$$

The distribution of X_A conditioned on X_B is

$$\begin{aligned} X_A|X_B &\sim \mathcal{N}(x_A; \mu_{A|B}, \Sigma_{A|B}), \\ \mu_{A|B} = \mu_A - K_{AA}^{-1}K_{AB}(x_B - \mu_B), &\quad \Sigma_{A|B} = K_{AA}^{-1}. \end{aligned} \tag{2.2.16}$$

Alternatively, the covariance matrix of the conditional distribution can be calculated via $\Sigma_{A|B} = \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}$, due to the Schur complement.

Another useful property is that products of Gaussian distributions are Gaussian again, see for example the appendix of [RW06].

2.2.5.1 Unregularized Covariance Estimation

Assume we are given N realizations $\mathcal{D} = \{x^i\}_{i=1}^N \in \mathbb{R}^n$ of a random vector $(X_1, \dots, X_n)$, that we believe to be normally distributed, i.e. $X \sim \mathcal{N}(\mu, \Sigma)$, and we want to obtain estimates of μ and Σ.

Definition 2.27 *(Likelihood function and maximum likelihood).* The **likelihood function** $L(\theta; \mathcal{D})$ of the parameter vector θ given data $\mathcal{D}$ under the assumption of *independent and identically distributed* (i.i.d.) samples x^i is

$$L(\theta; \mathcal{D}) = p(\mathcal{D}|\theta) = \prod_{i=1}^{N} p_\theta(x^i). \tag{2.2.17}$$

Taking the logarithm of (2.2.17) yields the **log-likelihood function** $\ell(\theta; \mathcal{D})$. Finally,

$$\hat{\theta} = \arg\max_{\theta} \ell(\theta; \mathcal{D}) \tag{2.2.18}$$

is called the **maximum likelihood** (ML) estimate of θ.

Maximizing $\ell(K, \mu; \mathcal{D})$ with respect to μ yields the sample mean $\bar{x} = 1/N \sum_{i=1}^{N} x^i$ as an estimate. To obtain $\hat{\Sigma}$, we consider the problem

$$\arg\max_{K} \log p(\mathcal{D}|K, \bar{x}) = \arg\max_{K} \log\det K - \operatorname{tr}(KS), \tag{2.2.19}$$

where S is the sample covariance matrix, with $S_{ij} = 1/N \sum_{k=1}^{N} (x_i^k - \bar{x}_i)(x_j^k - \bar{x}_j)$. Taking the derivative of both terms with respect to K yields [PP12]

$$\frac{\partial \log\det K}{\partial K} = \Sigma, \qquad \frac{\partial \operatorname{tr}[KS]}{\partial K} = S.$$

Thus setting $\partial \ell(K; \mathcal{D})/\partial K \stackrel{!}{=} 0$ we obtain $\hat{\Sigma} = S$.

Estimating Σ involves the estimation of $n(n-1)/2$ parameters. If $N \ll n^2$, the maximum likelihood estimate is ill-conditioned and may perform poorly. Here *regularization techniques* come into play, which introduce additional information into the problem. They may come in different flavors: As application of the principle of Occam's razor[6] or from a Bayesian perspective by adding a prior distribution. Below we present two regularization techniques that we will utilize for our segmentation model, and each can be seen as a representative of one of these types of regularization.

2.2.5.2 ℓ^1-Regularized Covariance Estimation

One popular regularization approach enforces sparsity on the precision matrix K [MB06, BEGd08, FHT08], thereby reducing the amount of estimated parameters. This is achieved by augmenting the ML problem (2.2.19) with ℓ^1-regularization acting on K. From a Bayesian point of view, this can motivated by adding a prior distribution $p(K)$:

Definition 2.28 *(Maximum a posteriori).* Given a data set $\mathcal{D}$ and a parameter vector θ that we want to estimate. The mode of the posterior distribution, that is

$$\hat{\theta} = \arg\max_{\theta} p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta), \tag{2.2.20}$$

is known as **maximum a posteriori** (MAP) estimate of θ.

An alternative motivation arises from the viewpoint of graphical models, since entries with $K_{ij} = 0$ imply absent edges between nodes i and j in the graph associated with X, see Example 2.44. As we see in the next section, this corresponds to additional conditional independence assumptions and thereby a less complex distribution $p(x)$.

[6] Occam's razor states that one should always choose, from a set of hypotheses, the one with the fewest assumptions.

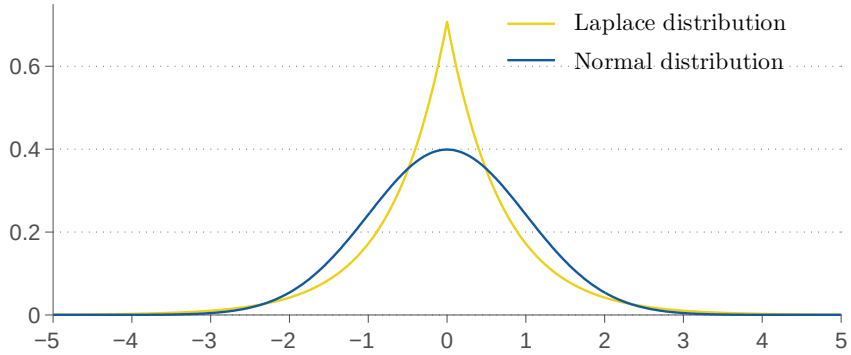


Figure 2.1 - Comparison of Laplace distribution (2.2.21) and univariate normal distribution (2.2.14) both with mean zero and variance one ($\lambda = \sqrt{2}$ for Laplace), that illustrates the difference in the allocation of probability mass especially around the mean.

We now treat each K_{ij} as a random variable and assume pairwise independence, that is $p(K) = \prod_{ij} P(K_{ij})$ (Definition 2.19). The prior distribution that gives rise to the ℓ^1-norm is the Laplace distribution with density function

$$p(K_{ij}) = \frac{\lambda}{2} \exp\left(-\lambda |K_{ij} - \mu|\right), \tag{2.2.21}$$

with mean $\mu = 0$ and variance $\sigma^2 = 2/\lambda^2$. Figure 2.1 illustrates the difference between a Laplace distribution and normal distribution both with zero mean and variance one, and how the Laplace distribution much more emphasizes values around its mean. Taking the logarithm of $p(K)$ gives

$$\log p(K) = \sum_{i,j}^{n} \log p(K_{ij}) = n^2(\log \lambda - \log 2) - \lambda \|K\|_1, \tag{2.2.22}$$

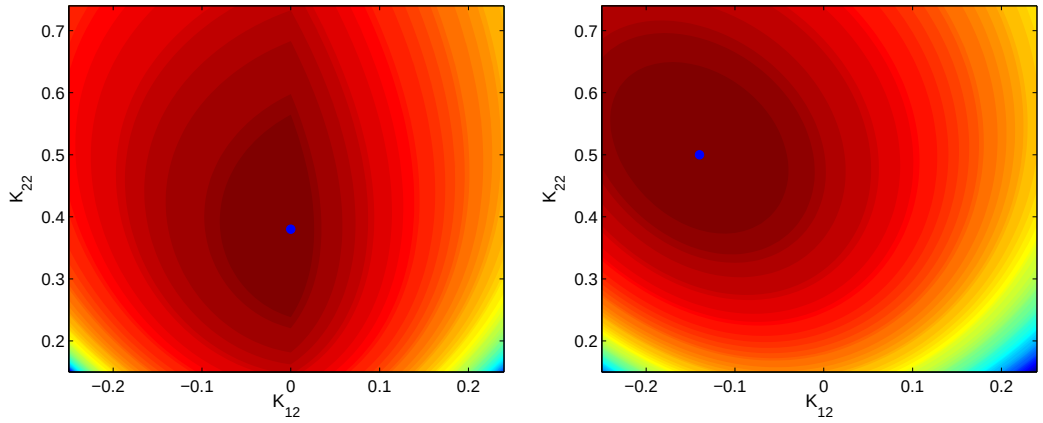
where $\|K\|_1 = \sum_{ij} |K_{ij}|$ is the ℓ^1-norm of K. Combining the prior (2.2.22) with the likelihood function (2.2.19) yields the concave optimization problem

$$\underset{K \succ 0}{\arg\max} \;\; \log\det K - \operatorname{tr}(KS) - \lambda \|K\|_1, \tag{2.2.23}$$

which yields the MAP estimate of K as discussed in Definition 2.28. Note that positive semidefiniteness of K is automatically enforced by the term $\log\det K$, which is the logarithmic barrier function for the positive semidefinite cone (see Example 2.1).

A regularization using ℓ^1-norm is often called **lasso regularization** [Tib96]. Assuming normally distributed K_{ij} would result in the regularization term $-\frac{1}{2}\tau\|K\|_2^2$, also known as **Tikhonov regularization** in the machine learning literature, with precision parameter τ and $\|K\|_2^2 = \sum_{ij} K_{ij}^2$. Both types of regularization are compared in Figure 2.2.

To derive the dual problem of (2.2.23), one replaces $\|K\|_1$ by $\max_{\|U\|_\infty \leq 1} \operatorname{tr}(KU)$, where $\|\cdot\|_\infty$ is the maximum norm $\max_{ij}(|U_{ij}|)$. Exchanging max and min, solving the inner optimization problem with $K = (S+U)^{-1}$ and setting $K^{-1} = W = S + U$



(a) $-\lambda\|K\|_1,\ \lambda = \sqrt{2}$ (b) $-\frac{1}{2}\tau\|K\|_2^2,\ \tau = 1$

Figure 2.2 - Comparison of ℓ^1 and ℓ^2 regularization. Contours for optimal K_{11} are shown. Parameters λ and τ are chosen as in Figure 2.1. ℓ^1-regularization enforces $K_{12} = 0$. i.e. sparsity on the precision matrix K.

yields the dual

$$\hat{\Sigma} := \arg\max_W \Big\{ \log\det W : \|W - S\|_\infty \leq \lambda \Big\}. \tag{2.2.24}$$

[FHT08] efficiently solve (2.2.24) by applying a block-coordinate descent approach. We implemented their approach, called the **graphical lasso**, in C and use it to regularize the covariance matrices of our texture models (c.f. Section 3.1.1).

2.2.5.3 Probabilistic Principle Component Analysis

An alternative approach for regularized covariance estimation is based upon Principle Component Analysis (PCA). PCA is a well-known technique for dimensionality reduction that, given a data set $\mathcal{D} = \{x^i\}_{i=1}^N \in \mathbb{R}^n$, consists of a linear projection into a lower-dimensional subspace of dimension $q \ll n$. PCA finds q *principal axes* that are orthonormal and retain maximum variance of the projected data [Hot33]. It can be easily shown, that these axes correspond to the q dominant eigenvectors of the sample corvariance matrix S (defined in (2.2.19)). The variance of the projected data is given by the sum of the corresponding eigenvalues.

[TB99] embedded PCA into a density framework, which they called **Probabilistic Principle Component Analysis** (PPCA). PPCA assumes that the random vector X is generated from a latent random vector Z via

$$X = WZ + \mu + \epsilon, \tag{2.2.25}$$

where $W \in \mathbb{R}^{n\times q}$, $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ is isotropic Gaussian noise and $Z \sim \mathcal{N}(0, I)$ is assumed to have standard normal distribution[7]. Thus, $X|Z \sim \mathcal{N}(WZ + \mu, \sigma^2 I)$. The

[7] PPCA can be considered as a generalisation of PCA, which assumes the deterministic relation $X = WZ + \mu$.

marginal distribution of X is a product of two Gaussians and therefore Gaussian itself [RW06]. Calculating the moments of X (see Definition 2.24) gives

$$\begin{aligned} E[X] &= WE[Z] + \mu + E[\epsilon] = \mu, \\ E[XX^T] - E[X]E[X]^T &= WE[ZZ^T]W^T + E[\epsilon\epsilon^T] = WW^T + \sigma^2 I, \end{aligned} \tag{2.2.26}$$

thus $X \sim \mathcal{N}(\mu, \sigma^2 I + WW^T)$. Using the Woodbury identity [PP12, p. 18], the precision matrix is given by

$$K = \sigma^{-2} I - \sigma^{-2} W M W^T, \qquad M = (\sigma^2 I + W^T W)^{-1}. \tag{2.2.27}$$

To estimate the parameters μ, σ^2 and W, one can use the ML approach mentioned earlier. [TB99] go on to show, that $\hat{\mu}$ is the sample mean $\bar{x}$ and

$$\hat{W} = U_q(\Lambda_q - \sigma^2 I)^{1/2}, \qquad \hat{\sigma}^2 = \frac{1}{n-q} \sum_{i=q+1}^{n} \lambda_i,$$

where U_q contains the q principle eigenvectors of S, and Λ_q is a diagonal matrix with eigenvalues $\lambda_1 \dots \lambda_q$ on its diagonal. Note, that although the maximum likelihood framework is used, we obtain a regularized estimate of Σ, since the regularization is implicitly given by the low-rank decomposition $\Sigma = WW^T + \sigma^2 I$. This reduces the numbers of parameters to be estimated from $n(n-1)/2$ to $nq + 1$, which is linear in the dimension of the data.

The ability to decompose both Σ and K into W and $\sigma^2 I$ is especially useful in case of high dimensional data, since one only has to store W and σ^2. The full covariance matrix for the shape prior in our 3-D data set would be of size 45 GB, whereas W with $q = 25$ only requires 15 MB of memory. Furthermore, operations related to Σ and K can be rewritten in terms of W and σ, thereby significantly reducing their complexity.

2.3 Graphical Models

This section presents probabilistic graphical models, which combine the concepts of graph theory and probability theory. They provide an intuitive visual representation of probability distributions, that allow insights into their structure. They also facilitate a unified treatment of various existing approaches (hidden Markov models, Kalman Filter, etc.) in terms of inference.

We start with some terminology related to graph theory in 2.3.1, followed in 2.3.2 by a general definition of probabilistic graphical models, combining the notions of a graph and that of a probability distribution. In 2.3.3 and 2.3.4 we present two different types of graphical models which arise depending in the type of edges used in the graph. To unify the treatment of inference in Section 2.4, we show in 2.3.4.1 how to convert a directed graphical model into an undirected one.

References for this chapter are the excellent overview about graphical models in [Bis06] as well as the more thorough treatments of [Lau96] and [KF09].

2.3.1 Graph Theory

This section gives a short overview of some basic notions of graph theory, that will be used in the subsequent sections.

Definition 2.29 *(Graph).* A **graph** is an ordered pair $\mathcal{G} = (V, E)$ comprising a set of **nodes** (or vertices) $V = \{1, \dots, n\}$ and a set of **edges** $E \subset V \times V$.

Edges may be **undirected**, in which case no distinction is made between edges (s, t) and (t, s) and s is called the **neighbor** of t and vice versa. We denote the set of all neighbors of a node i by $\mathrm{ne}(i)$. Alternatively, edges may be **directed**, with the direction of the edge indicated by $(s \to t)$ and s being the **parent** of the **child** t. For a node i, $\mathrm{pa}(i)$ denotes its set of parents and $\mathrm{ch}(i)$ its set of children. We call a graph $\mathcal{G}$ composed of the former type of edges an **undirected graph**, otherwise a **directed graph**.

Definition 2.30 *(Path).* A **path** from node i_1 to node i_m is a sequence $\{i_1, i_2, \dots, i_m\}$ of distinct nodes in V, such that either $(i_j, i_{j+1}) \in E$ or $(i_j \to i_{j+1}) \in E$ for $j = 1, \dots, m-1$. If all edges are directed we call it a **directed path**.

We call a node j an **descendant** of node i, if there is a directed path from i to j.

Definition 2.31 *(Trail).* A **trail** is a path such that the directionality of arrows is ignored.

Definition 2.32 *(Cycle).* A **cycle** is a directed path with $i_1 = i_m$.

Definition 2.33 *(Loop).* A **loop** is a trail with $i_1 = i_m$.

A directed graph without cycles is called a **directed acyclic graph** (DAG), a notion that will become important when we introduce directed graphical models. If we add the constraint, that there can be only one trail between every pair of nodes, we obtain a **polytree**. Finally, if every node has exactly one parent (except the *source* node which has no parent), we call $\mathcal{G}$ a **tree**. It follows that

$$\text{directed graph} \supset \text{directed acyclic graph} \supset \text{polytree} \supset \text{tree}.$$

All four types of directed graphs are illustrated in Figure 2.3. Similarly, we call an *undirected* graph without loops a tree.

Definition 2.34 *(Subgraph).* The **subgraph** induced by a set of vertices $V' \subset V$ is given by $\mathcal{G}' = (V', E')$, $E' = \{(i, j) : i \in V', j \in V'\}$.

We call a graph $\mathcal{G}$ **complete**, if $E = V \times V$. A complete subgraph is called a **clique**. A **maximal clique** is a complete subgraph $\mathcal{G}'$, such that by adding any additional vertex from $V \setminus V'$ it ceases to be complete.

2.3.2 Probabilistic Graphical Models

Previously we introduced the notion of a probability measure (or probability distribution) P and that of a graph $\mathcal{G}$. The following definition combines these two:

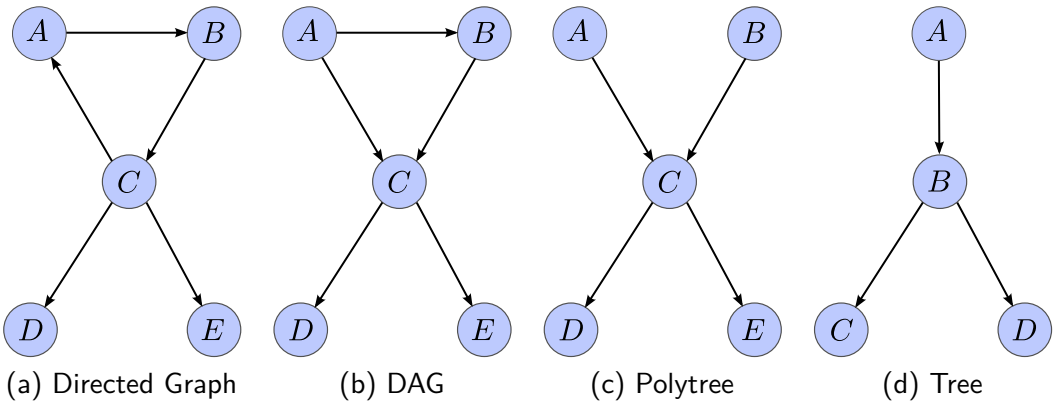


Figure 2.3 - Visualization of (a) a general directed graph containing the cycle $\{A, B, C\}$, (b) a directed acyclic graph where the cycle is removed, but there are still two trails between several pairs of nodes, (c) a polytree where now only one trail exists between every pair of nodes but C has still two parents and (d) a tree, where each node has only one parent except the source node A.

Definition 2.35 *(Probabilistic graphical model).* A **probabilistic graphical model** corresponds to a tuple $(\mathcal{G}, P)$ of a graph $\mathcal{G} = (E, V)$ and a probability distribution P. Each node $i \in V$ is associated with a random variable X_i, taking values in some space $\mathcal{X}_i$, which may be either continuous (e.g. $\mathcal{X}_i = \mathbb{R}$) or discrete (e.g. $\mathcal{X}_i = \{0, 1, \ldots, r-1\}$). The random vector X is distributed according to the distribution P.

We will use lower-case letters (e.g. $x_i \in \mathcal{X}_i$) to denote realizations of the random variable X_i. Subsets of X will be denoted by X_A or correspondingly by x_A with $A \subset V$. Realizations x of the random vector X take values in the **Cartesian product space** $\mathcal{X}^n = \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \mathcal{X}_n$.

Missing edges and the configuration of directed edges encode **conditional independences** (CI) between components of the random vector X associated with $\mathcal{G}$, thereby determining the factorization of P:

Definition 2.36 *(Conditional independence).* Two random variables X_i and X_j are said to be conditionally independent given the random vector X_A, denoted by

$$X_i \perp\!\!\!\perp X_j \mid X_A,$$

iff their joint conditional distribution factorizes according to

$$p(x_i, x_j | x_A) = p(x_i | x_A) p(x_j | x_A).$$

Using the following factorization criterion, it is easy to determine conditional independence:

$$X_i \perp\!\!\!\perp X_j \mid X_A \quad \Longleftrightarrow \quad p(x_i, x_j, x_A) = f(x_i, x_A)\, g(x_j, x_A), \tag{2.3.1}$$

for some functions f and g.

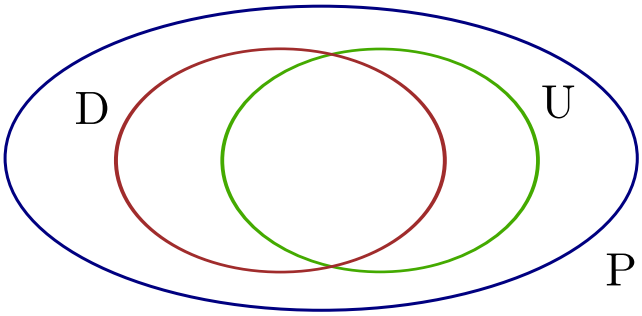


Figure 2.4 - Venn diagram that illustrates the relationship between the set of all distributions P and those which can be expressed as perfect maps by undirected (U) and directed (D) graphical models [Bis06].

In terms of the graph $\mathcal{G}$, conditional independence $X_i \perp\!\!\!\perp X_j \mid X_A$ implies that all trails from node i to node j are **blocked** by the nodes in A. Conditions when a trail is blocked differ between directed and undirected graphical models and will be introduced below.

Equipped with the concept of conditional independence, we now can specify more precisely the connection between $\mathcal{G}$ and its associated distribution P (complementing the definition of a probabilistic graphical model given above): $\mathcal{G}$ is an I-map of P.

Definition 2.37 *(Independency map).* A graph $\mathcal{G}$ is said to be an independency map or short **I-map** of a distribution P, if all CI assumptions reflected in the structure of $\mathcal{G}$ also hold in P.

This means that P factorizes according to $\mathcal{G}$, but may have additional independencies not reflected by $\mathcal{G}$. The complete graph $\mathcal{G}$ encoding no CI assumption is a trivial I-map for any distribution P. A stricter connection between $\mathcal{G}$ and P is the following:

Definition 2.38 *(Perfect map).* We say that a graph $\mathcal{G}$ is a **perfect map** for a distribution P, if all CI assumptions encoded in $\mathcal{G}$ are also reflected in P and vice versa.

We will see, that undirected and directed graphical models have different sets of distributions, for which they are perfect maps. Furthermore there exist distributions, for which there exists no perfect map in both types of graphical models. The Venn diagram in Figure 2.4 illustrates these facts.

Definition 2.39 *(Markov blanket).* We define the **Markov blanket** [Pea88] of a node i as the set of nodes A, such that

$$p(x_i|x_A, x_{V\setminus\{A,i\}}) = p(x_i|x_A).$$

Thus learning the state of a node not in the set A, will tell us nothing new about X_i.

2.3.3 Directed Graphical Models

The notion of a blocked trail and therefore conditional independence is quite subtle for directed graphical models. There exist three different possibilities (see also Figure

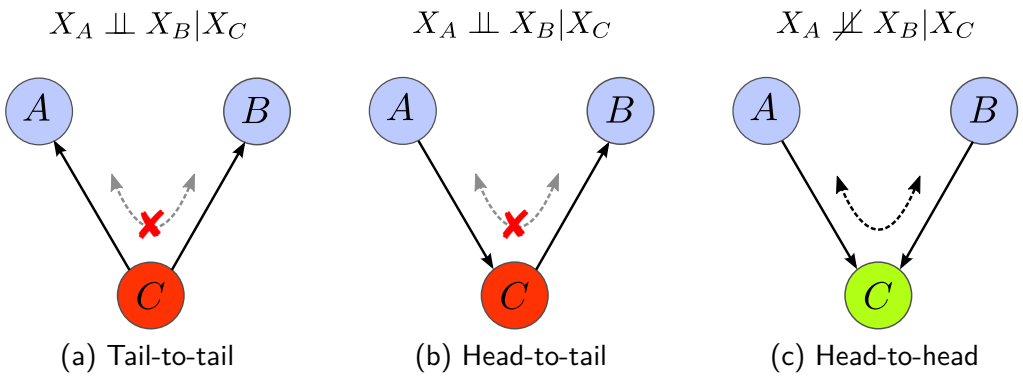


Figure 2.5 - The effects of observing X_C on the three possible trails from A to C to B. Trails that run tail-to-tail and head-to-tail over node C become blocked, therefore the observation of X_C renders X_A and X_B independent. Contrary, head-to-head connections become unblocked and thus X_A and X_B become dependent, denoted by $X_i \not\perp\!\!\!\perp X_j \mid X_A$.

2.5) for the connectivity of node i within a trail:

(i) tail-to-tail ($\leftarrow i \rightarrow$),

(ii) head-to-tail ($\rightarrow i \rightarrow$) or

(iii) head-to-head ($\rightarrow i \leftarrow$).

If the random variable X_i is observed, all trails including node i are blocked if the connectivity at i is either of type (i) or (ii). Head-to-head connections behave contrary, here the trail becomes *unblocked* if X_i (or any of its descendants) is observed. That renders the random variables X_{pa_i} dependent, and the observation of any if its members will decrease the probability of the remaining ones, thereby **explaining away** these other possible causes of X_i.

More general, given non-intersecting sets of nodes A, B, C, sets A and B are conditional independent with respect to C, if all trails from A to B are blocked. Whether a certain trail is blocked, can be deduced by investigating all nodes along that trail in terms of the rules introduced above. This technique is called **d-seperation** [Pea88].

Definition 2.40 *(I-equivalence).* Two graphs $\mathcal{G}, \mathcal{G}'$ that encode the same set of CI assumptions are said to be **I-equivalent**.

For example the graphs (a) and (b) in Figure 2.5 are I-equivalent.

As pointed out in the definition of probabilistic graphical models, the structure of $\mathcal{G}$ encodes a set of CI assumptions that are also expressed by the corresponding distribution P. This is reflected in the following definition:

Definition 2.41 *(Directed Graphical Model).* A directed graphical model or **Bayesian network** (BN) is a tuple $(\mathcal{G}, P)$, where $\mathcal{G} = (E, V)$ is a directed acyclic graph and P is a distribution whose factorization reflects the structure of $\mathcal{G}$ and is given by

$$p(x_1, x_2, \ldots, x_n) = \prod_{i \in V} p(x_i | x_{\mathrm{pa}(i)}). \tag{2.3.2}$$

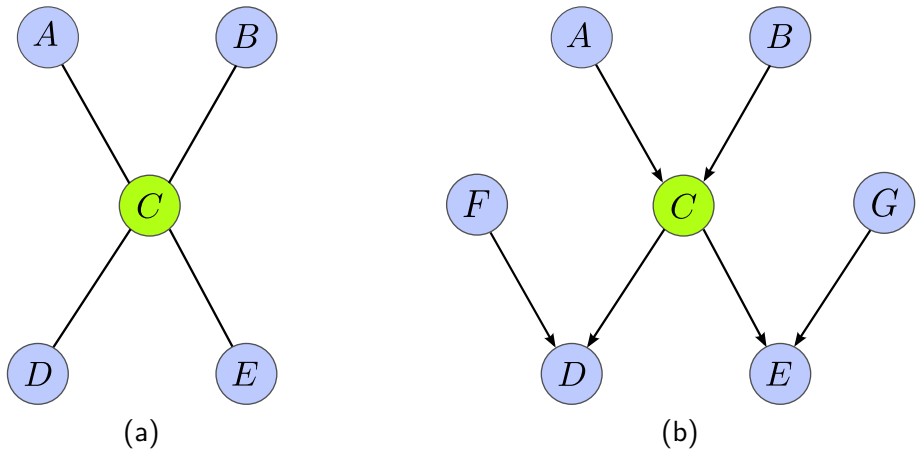


Figure 2.6 - The Markov blankets of the node C for undirected (a) and directed graphical models (b). For undirected graphical models only the direct neighbors of C are in the blanket. But since for directed graphical models, conditioning on a child nodes D and E renders node C dependent on their parents F and G (head-to-head connection, see Figure 2.5 (c)), one must include them.

For directed graphical models, the Markov blanket of node i consists of the sets $\mathrm{pa}(i)$ and $\mathrm{ch}(i)$, as well as the set of **co-parents**, the latter being all parents of $\mathrm{ch}(i)$ except i itself. They have to be included, since the observation of a node $j \in \mathrm{ch}(i)$ unblocks the head-to-head connections to the parents of j and thereby renders them dependent given i. See Figure 2.6 (a) for an illustration, where F and G are the co-parents of C.

Example 2.42 *(Linear-Gaussian models).* Linear-Gaussian models are an important class of continuous directed graphical models. Here each node represents a random variable X_i that has a normal distribution with mean μ_i, which is a weighted linear combination of its parents states and variance σ_i^2:

$$p(x_i|\mathrm{pa}_i) = \mathcal{N}\Big(x_i; \sum_{j\in\mathrm{pa}_i} W_{ij}x_j + b_i, \sigma_i^2\Big).$$

Since $p(x)$ is the product of several normal distributions, it is normally distributed too. Comparing the formula above with (2.2.25), shows that PPCA is a linear-Gaussian model, where the node i of each X_i is the child of all nodes j belonging to latent variables Z_j.

2.3.4 Undirected Graphical Models

An undirected graphical model is composed of edges (s, t) with no directional information. This significantly simplifies the detection of conditional independence: A node whose random variable is observed, blocks every trail that includes this node. Two nodes i and j are said to be conditionally independent given a set A of nodes, if all trails that connect these two nodes contain at least one observed node $a \in A$.

More general, if the set C of nodes was observed, and all trails from nodes in a set A to nodes in a set B include nodes of C, then $A \perp\!\!\!\perp B \mid C$. The Markov blanket of node i consists of the set $\mathrm{ne}(i)$, see Figure 2.6 (b).

The cliques of $\mathcal{G}$ constitute subgraphs where no CI assumptions are expressed. We therefore use them as building blocks of the factorization of $p(x)$:

Definition 2.43 *(Undirected graphical models).* An undirected graphical model or **Markov Random Field** (MRF) is a tuple $(\mathcal{G}, P)$ of an undirected graph $\mathcal{G} = (E, V)$ and a distribution P. Given the set $\mathcal{C}$ of cliques of $\mathcal{G}$, the distribution P factorizes according to

$$p(x_1, x_2, \ldots, x_n) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(x_C), \qquad Z = \sum_x \prod_{C \in \mathcal{C}} \psi_C(x_C), \tag{2.3.3}$$

where $x_C \in V$ denotes the subset of nodes belonging to the clique C. Z is the normalization constant or **partition function**[8], ensuring that $p(x)$ is a valid probability distribution. The ψ_C are called **potential functions**, and are assumed to be non-negative.

One can without loss of generality restrict the set C to all maximal cliques [WJ08]. However, inference algorithms, such as those presented in Section 2.4, may be able to exploit the non-maximal representation, which is why we stick with the definition given above.

Opposed to directed graphical models, potential function have no interpretation in terms of marginal distributions. Therefore, the factorization (2.3.2) can be viewed as a special case of (2.3.3). Again we can state equivalence between the set of all CI assumptions expressed by the factorization (2.3.3) and those encoded in the corresponding graph $\mathcal{G}$, which in the case of undirected graphical models is called the **Hammersley-Clifford theorem** [Cli90][9].

As indicated by the Venn diagramm in Figure 2.4, there are cases where the CI assumptions expressed by an undirected model cannot be encoded by a directed graphical model and vice versa. While for BNs there exists the concept of head-to-head connections, that renders nodes *dependent* when observing other nodes, for MRFs observing a node may only result in additional independencies between other nodes. On the other hand, an undirected graph $\mathcal{G}$ over nodes $\{a, b, c, d\}$ with edge set $E = \{(a, b), (b, c), (c, d), (d, a)\}$ expresses the CI assumptions $X_a \perp\!\!\!\perp X_c \mid X_b, X_d$ and $X_b \perp\!\!\!\perp X_d \mid X_a, X_c$, which can not expressed simultaneously using Bayesian networks.

Example 2.44 *(Gaussian Markov Random Fields).* A **Gaussian Markov Random Field** (GMRF) [RH05] is the tuple $(\mathcal{G}, P)$ of an undirected graph $\mathcal{G}$ together with a normal distribution P, i.e. $X \sim \mathcal{N}(\mu, \Sigma)$. Its connectivity, and thereby the conditional independence structure, can be read from the precision matrix K. It holds that

$$X_i \perp\!\!\!\perp X_j \mid X \setminus \{X_i, X_j\} \quad \Longleftrightarrow \quad K_{ij} = 0 \quad \Longleftrightarrow \quad (i, j) \wedge (j, i) \notin E, \tag{2.3.4}$$

[8] In case of continuous random variables, summation is replaced by integration.
[9] For this theorem to hold, all potential functions have to be strictly positive.

and a dense matrix K implies a fully connected graph $\mathcal{G}$. While off-diagonal zero entries of Σ indicate marginal independence $X_i \perp\!\!\!\perp X_j \mid \emptyset$, those of K denote conditional independence $X_i \perp\!\!\!\perp X_j \mid X \setminus \{X_i, X_j\}$.

2.3.4.1 Directed Graphical Models → Undirected Graphical Models

It will turn out convenient to restrict the presentation of inference algorithms in the next section to one class of graphical models. As discussed above, the factorization (2.3.3) containts (2.3.2) as a special case, thus is much more expressive, which is why we will discuss inference from the perspective of undirected graphical models. Here we describe how to transform directed to undirected graphical models.

We want to associate a potential function ψ_C with every factor $p(x_i|x_{\mathrm{pa}(i)})$, in order to directly convert the factorization (2.3.2), i.e.

$$\psi_C(x_i, x_{\mathrm{pa}(i)}) := p(x_i|x_{\mathrm{pa}(i)}).$$

In order to do that, the node set $\{i \cup \mathrm{pa}(i)\}$ must form a clique. For tail-to-tail and head-to-tail connections this is naturally the case, since nodes that are connected like this only have one parent and C consists of just two nodes.

But replacing edges in head-to-head connections by undirected edges yields no clique. Instead additional nodes have to be added between all parents. This process is called **moralization**, and the resulting graph is called the **moral graph**. This hides the CI assumptions expressed by head-to-head connections in the fully connected clique. Certainly the potentials themselves still express these CI assumptions. It holds by construction that $Z = 1$.

Inspecting Figure 2.3 reveals that only directed trees can be transformed such that no edges have to be added and the transformed graph remains a tree, whereas for polytrees loops emerge, such that the resulting undirected graph is no tree.

2.4 Inference on Graphical Models

Given a graphical model $(\mathcal{G}, P)$ with joint distributions function $p(x)$, **probabilistic inference** comprises the task of computing *conditional distributions* $p(x_A|x_B)$ over a set $A \subset V$ of random variables, given another disjoint set $B \subset V$ of observed random variables. Another task is the determination of the *marginal distribution* $p(x_A)$. Yet another important inference problem is the calculation of the *mode* of $p(x_A)$, that is the element $\hat{x}_A = \arg\max_{x_A} p(x_A)$. The inferred random variables can be either unobserved latent variables and/or model parameters that are treated as random variables.

The inherent challenges of inference problems become apparent when we consider the discrete random vector $X : \Omega \to \mathcal{X}^n$ with state space $\mathcal{X}_i = \{0, \dots, r-1\}$ for all i. Naively calculating the marginal distribution $p(x_i)$ requires summing $p(x)$ over the product set of configurations $\mathcal{S} = \{x' \in \mathcal{X}^n | x'_i = x_i\}$:

$$p(x_i) = \sum_{x' \in \mathcal{S}} p(x_1, \dots, x_n).$$

This approach requires the summation over r^{n-1} configurations x', which becomes quickly intractable with increasing n and r. The situation for continuous random variables most often is even harder as they require the computation of integrals. A notable exception are GMRFs respectively normal distributions, that provide analytical solutions for the tasks mentioned above (see Section 2.2.5).

Practical inference techniques are either of *exact* or (in most cases) of *approximative* nature. All have in common that they rely upon the factorization of $p(x)$ expressed by $\mathcal{G}$ (c.f. Equations (2.3.2) and (2.3.3)), to reduce the complexity over the naive approach. Various approximate inference approaches exist, among them *graph cuts* [BVZ01] that correspond to max-flow algorithms, *Markov Chain Monto Carlo* methods [Nea93] that rely on stochastic sampling or *linear programming relaxation techniques* [DFJ54] such as dual decomposition.

Below we will present two important families of inference techniques that are relevant for this thesis: *message-passing* approaches (which encompass exact inference techniques for trees) and *variational inference*, a class of inference techniques that yield deterministic approximative solutions.

2.4.1 Message-Passing Approaches

Message-passing algorithms follow the paradigm of dynamic programming: the summation or integration over a set A of random variables is broken down into several subproblems over smaller sets $A_i \subset A$. In case of overlapping subproblems this reduces the amount of operations performed. The most important representative, introduced in the next section, is the **sum-product algorithm** or **belief propagation**, that yields exact marginals when applied to trees[10]. Closely related is the max-product algorithm, which infers the mode of the distribution with summation replaced by maximization.

When applied to general graphs with cycles, sum-product message passing is known as **loopy belief propagation**. Although there are no guarantees for convergence, loopy belief propagation has performed reasonably well for many problems, see for example [MWJ99]. Several variations of the sum-product message passing scheme were proposed, in order to yield better approximations and/or better convergence properties. Among them are generalized belief propagation [YFW+01], tree-reweighted belief propagation [WJW03] or the popular sequential tree-reweighted belief propagation [Kol06].

An exact inference algorithm for general graphs is the **junction tree algorithm** [LS88], which applies the sum-product method to a modified version of the input graph, known as the junction tree. But since complexity increases exponentially with growing **treewidth** of the modified graph, a measure of the size of the largest clique, the approach quickly becomes intractable.

Numerous toolboxes for inference on general graphical models exist. For example, the OpenGM library [ABK12, KAH+13] offers plenty of different inference techniques as well as wrappers for several programming languages.

[10] Undirected and directed trees, including polytrees, if represented as directed or factor graphs.

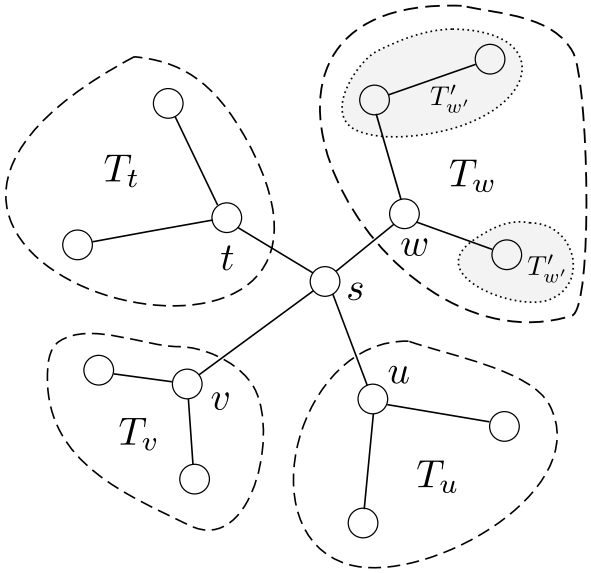


Figure 2.7 - Decomposition of a tree into subtrees, rooted at node s. Each subtree is rooted at its labeled node, which are all neighbors of s. The illustration is adapted from [WJ08].

2.4.1.1 Message-Passing on Trees

We will now give a description of the sum-product algorithm when applied to undirected trees[11]. An alternative interpretation from the viewpoint of variational inference is given in the next section.

We first observe, that for a tree-structured graph $T = (V, E)$ the cliques are given by the individual nodes and edges. Thus, following (2.3.3), any undirected graphical model without cycles has the following factorization:

$$p(x_1, x_2, \ldots, x_n) = \frac{1}{Z} \prod_{s \in V} \psi_s(x_s) \prod_{(s,t) \in E} \psi_{st}(x_s, x_t). \tag{2.4.1}$$

We are interested in computing marginal distributions $p(x_s)$ for all $s \in V$ and $p(x_s, x_t)$ for all $(s, t) \in E$. For future reference, we will denote them by μ_s and $\mu_{s,t}$.

If we consider an arbitrary node $s \in V$, we can define a subgraph $T_u = (V_u, E_u)$ for every node $u \in \mathrm{ne}(s)$. Each subgraph T_u is composed of nodes and edges that can be reached from u without traveling over s. Since T is a tree, every subgraph T_u is a tree too. Thus each node $u \in \mathrm{ne}(s)$ can be viewed as the root of its own subgraph, as illustrated in Figure 2.7.

All subgraphs are mutually disjoint and together comprise all nodes except the node s and all edges except those that connect s with its neighbors. We collect all terms from (2.4.1) connected to edges and nodes in T_u into the following product

$$p(x_{V_u}; T_u) \propto \prod_{t \in V_u} \psi_t(x_t) \prod_{(t,v) \in E_u} \psi_{tv}(x_t, x_v). \tag{2.4.2}$$

[11] This excludes polytrees, since their conversion introduces loops as discussed in Section 2.3.4.1. This constitutes no limitation as they are not used in this work.

The calculation of the marginal μ_s can then be expressed in terms of these products:

$$\begin{aligned}\mu_s(x_s) &= \sum_{x_t : t \neq s} p(x_1, \ldots, x_n) \\ &= \sum_{x_t : t \neq s} \kappa \cdot \psi_s(x_s) \prod_{u \in \mathrm{ne}(s)} \psi_{su}(x_s, x_u)\, p(x_{V_u}; T_u) \\ &= \kappa \psi_s(x_s) \prod_{u \in \mathrm{ne}(s)} \underbrace{\sum_{x_u \in \mathcal{X}_{V_u}} \psi_{su}(x_s, x_u)\, p(x_{V_u}; T_u)}_{M^*_{us}(x_s)} .\end{aligned} \tag{2.4.3}$$

The positive scalar κ ensures normalization of μ_s. Each subgraph T_u in turn can be split into smaller subgraphs $T'_{u'}$, as illustrated in Figure 2.7 for the subgraph T_w. Therefore the subproblems of computing $M^*_{us}(x_s)$ can be broken down recursively until only elementary subgraphs consisting of single nodes remain.

Usually one is interested in the complete set of marginals μ_s for all $s \in V$. The sum-product algorithm computes the marginals for all nodes simultaneously. At each iteration, every node t passes a **message** to all its neighbors $s \in \mathrm{ne}(t)$, denoted by $M_{ts}(x_s)$, in total $2|E|$ messages. Message updates are calculated according to the following recursion

$$M_{ts}(x_s) \leftarrow \kappa \sum_{x_t} \left\{ \psi_{st}(x_s, x_t) \psi_t(x_t) \prod_{u \in \mathrm{ne}(t) \setminus s} M_{ut}(x_t) \right\}. \tag{2.4.4}$$

Again κ denotes a normalization constant.

For tree-structured graphs, the approach converges after two iterations of message-passing using the following schedule: In the first iteration messages are passed from leaf nodes, nodes with only one neighbor, inwards to some arbitrarily chosen root node s. In the second iteration, messages are passed back to the leaves. This yields a fixed point $M^* = \{M^*_{st}, M^*_{ts}, (s,t) \in E\}$, with components M^*_{st} equal to those defined in (2.4.3) up to a constant. Then marginals μ_s are given by Formula (2.4.3), while pairwise marginals μ_{st} can be calculated via

$$\mu_{st} = \kappa \psi_{st}(x_s, x_t) \psi_t(x_t) \psi_s(x_s) \prod_{u \in \mathrm{ne}(s) \setminus t} M^*_{us}(x_s) \prod_{v \in \mathrm{ne}(t) \setminus s} M^*_{vt}(x_t). \tag{2.4.5}$$

Replacing the summation in (2.4.4) with maximization, yields the max-product algorithm. In case of a general graph with loops, messages are passed along the graph until convergence (which is not guaranteed), resulting in the loopy belief propagation algorithm mentioned earlier. The resulting fixed points may only represent approximations of the true marginals [Bis06].

The next section will give an interpretation of message-passing approaches from the perspective of variational inference and energy minimization.

2.4.2 Variational Inference

The phrase "variational" is an umbrella term referring to the reformulation of an inference task into an optimization problem. This new problem formulation then is relaxed until the optimization can be performed efficiently. These relaxiations amount to a *deterministic* approximation of the original problem as opposed to, for example, Monte Carlo based stochastic methods.

The following exposure is inspired by the excellent survey about variational inference by Wainwright and Jordan [WJ08]. We will introduce exponential families, the parameterized family of probability distributions underlying variational inference. Many well known probability distributions are members of this family, and we will give examples for the normal distribution and Markov Random Fields with discrete random variables. We then demonstrate how the variational methodology can be used to derive tractable approximations for probabilistic inference.

2.4.2.1 Exponential Family

Definition 2.45 *(Exponential family).* Let $X = (X_1, \ldots, X_n)$ be a random vector taking values in some space $\mathcal{X}^n = \otimes_{s=1}^n \mathcal{X}_s$. Furthermore, given a vector of **sufficient statistics** $\phi = (\phi_\alpha : \mathcal{X}^n \to \mathbb{R}, \alpha \in \mathcal{I})$ and the associated vector of **canonical** or **exponential** parameters $\theta = (\theta_\alpha \in \mathbb{R}, \alpha \in \mathcal{I})$. Here $\mathcal{I}$ is an index set with $d = |\mathcal{I}|$ elements, such that $\phi : \mathcal{X}^n \to \mathbb{R}^d$ and $\theta \in \mathbb{R}^d$.

The **exponential family** associated with $\phi(x)$ is given by

$$p_\theta(x_1, \ldots, x_n) = \exp\{\langle \theta, \phi(x) \rangle - A(\theta)\}, \tag{2.4.6}$$

with convex **log partition function** (or **cumulant function**)

$$A(\theta) = \log \int_{\mathcal{X}^n} \exp\langle \theta, \phi(x) \rangle d\nu(x). \tag{2.4.7}$$

The integral is taken with respect to some base measure $d\nu(x)$[12]. The corresponding measure is defined by $dP(x) = p_\theta(x) d\nu(x)$.

The cumulant function $A(\theta)$ in general cannot be expressed in closed form, and computing the integral turns out to be intractable for most graphical models. We will see during the course of this treatise, that we can derive a variational representation of $A(\theta)$. By itself this will lead to no simplifications, but will enable us to derive tractable relaxations.

Definition 2.46 *(Minimal and overcomplete representations).* We speak of a **minimal representation**, if there are no linear dependencies between the sufficient statistics $\phi_\alpha(x)$. Otherwise, we call it an **overcomplete representation**.

With ϕ fixed, we define the set of valid canonical parameters θ as

$$\Omega := \{\theta \in \mathbb{R}^d \,|\, A(\theta) < +\infty\}. \tag{2.4.8}$$

[12] Counting measure or Lebesgue measure depending on $\mathcal{X}$.

An alternative parametrization, *dual* to the canonical parameters is in terms of a vector of **mean parameters** μ:

Definition 2.47 *(Mean parameters).* We can associate a mean parameter μ_α with every sufficient statistic ϕ_α via

$$\mu_\alpha = E[\phi_\alpha(X)] = \int \phi_\alpha(x) p(x) d\nu(x). \tag{2.4.9}$$

We define the set of realizable mean parameters

$$\mathcal{M} := \{\mu \in \mathbb{R}^d \,|\, \exists\, p \text{ s.t. } E[\phi_\alpha(X)] = \mu_\alpha,\ \forall\, \alpha \in \mathcal{I}\}. \tag{2.4.10}$$

By definition, elements of $\mathcal{M}$ are convex combinations of sufficient statistics, therefore the set $\mathcal{M}$ is convex.

For discrete random variables with finite state space $\mathcal{X}^n$, Definition (2.4.10) describes a *finitely generated* bounded convex set called polytope (see Definition 2.4). Two alternative representations exists: The first one is by a set of extreme points (zero-dimensional faces of the set [Roc70]), which are the sufficient statistics $\phi(x)$:

$$\mathcal{M} = \text{conv}\{\phi(x), x \in \mathcal{X}^n\}. \tag{2.4.11a}$$

The dual representation is in terms of a finite collection of linear inequality constraints

$$\mathcal{M} = \{\mu \in \mathbb{R}^d \,|\, \langle a_j, \mu\rangle \geq b_j\ \forall j \in \mathcal{J}, |\mathcal{J}| \text{ finite}\}. \tag{2.4.11b}$$

for suitable parameters a_j, b_j from a constraint set $\mathcal{J}$.

Definition 2.48 *(Marginal polytope).* In the context of discrete pairwise MRFs, the set $\mathcal{M}$ is called the **marginal polytope** $\mathbb{M}(\mathcal{G})$, with respect to the graph $\mathcal{G}$.

For the graphical models relevant in this thesis, discrete Markov Random Fields and Gaussian Markov Random Fields, we give their presentation in terms of the exponential family.

Example 2.49 *(Discrete Markov Random Fields).* Consider an MRF where each random variable X_s takes values in the discrete label space $\mathcal{X} := \{0, 1, \ldots, r-1\}$ for some integer $r \geq 2$. As sufficient statistics $\phi(x)$, we define indicator functions for nodes $s \in V$ and states $j \in \mathcal{X}$ as

$$\mathbb{I}_{s;j}(x_s) = \begin{cases} 1 & \text{if } x_s = j, \\ 0 & \text{otherwise,} \end{cases}$$

with associated parameter vector $\theta_s = \{\theta_{s;j}\} \in \mathbb{R}^r$, and for edges $(s,t) \in E$ and pair of states $(j,k) \in \mathcal{X} \times \mathcal{X}$ as

$$\mathbb{I}_{st;jk}(x_s, x_t) = \begin{cases} 1 & \text{if } x_s = j \wedge x_t = k, \\ 0 & \text{otherwise,} \end{cases}$$

with associated parameter vector $\theta_{st} = \{\theta_{st;jk}\} \in \mathbb{R}^{r \times r}$. One can easily show, that this representation is overcomplete [WJ08]. Let us introduce the shorthand notation

$$\theta_s(x_s) := \sum_j \theta_{s;j} \mathbb{I}_{s;j}(x_s), \qquad \theta_{st}(x_s, x_t) := \sum_{j,k} \theta_{st;jk} \mathbb{I}_{st;jk}(x_s, x_t).$$

We can now write $p(x)$ as

$$p_\theta(x) = \exp\Big\{ \sum_{s \in V} \theta_s(x_s) + \sum_{(s,t) \in E} \theta_{st}(x_s, x_t) - A(\theta) \Big\}, \tag{2.4.12}$$

with log partition function

$$A(\theta) = \log \sum_{x \in \mathcal{X}^n} \exp\Big\{ \sum_{s \in V} \theta_s(x_s) + \sum_{(s,t) \in E} \theta_{st}(x_s, x_t) \Big\}. \tag{2.4.13}$$

The mean parameters of a discrete MRF correspond to its singleton and pairwise marginal distributions

$$\mu_{s;j} = E[\mathbb{I}_{s;j}(X_s)] = \mathbb{P}[X_s = j], \quad \forall j \in \mathcal{X}_s, \tag{2.4.14a}$$

$$\mu_{st;jk} = E[\mathbb{I}_{st;jk}(X_s, X_t)] = \mathbb{P}[X_s = j, X_t = k], \quad \forall (j,k) \in \mathcal{X}_s \times \mathcal{X}_t. \tag{2.4.14b}$$

For later reference, we define $\mu_s(x_s)$ and $\mu_{st}(x_s, x_t)$ as shorthand notations in the same manner as above.

Example 2.50 *(Gaussian Markov Random Fields).* Let $(X_1, \ldots, X_n)$ be a normally distributed random vector. The vector of sufficient statistics is given by

$$\phi(x) = \{x_s, x_s^2, s \in V;\ x_s x_t, (s,t) \in E\}. \tag{2.4.15}$$

We associate the vector $\theta \in \mathbb{R}^n$ with $x = (x_1, \ldots, x_n)$ and the symmetric matrix $\Theta \in \mathbb{R}^{n \times n}$ with xx^T. The density function is given by

$$p_\theta(x) = \exp\Big\{ \langle \theta, x \rangle - \frac{1}{2} \langle \Theta, xx^T \rangle - A(\theta, \Theta) \Big\}, \tag{2.4.16}$$

where $\langle \Theta, xx^T \rangle$ denotes the inner product over the vector space of matrices $\mathbb{R}^{n \times n}$. This form is also called the *canonical representation* of normal distributions [RH05, Def. 2.2]. The relation to the standard representation (2.2.14) is $\theta = K\mu$ and $\Theta = K$. Since $A(\theta, \Theta)$ is finite only for $\Theta \succ 0$, the valid parameter space Ω is given by

$$\Omega = \{(\theta, \Theta) \in \mathbb{R}^n \times \mathcal{S}_{++}^n\}. \tag{2.4.17}$$

The corresponding parametrization in terms of the mean parameters $\mu = E[X]$ and $\Gamma := E[XX^T]$ is

$$\mathcal{M} = \{(\mu, \Gamma) \in \mathbb{R}^n \times \mathcal{S}_+^n \,|\, \Gamma - \mu\mu^T \succeq 0\}. \tag{2.4.18}$$

Here $\Gamma - \mu\mu^T$ corresponds to the covariance matrix Σ as defined in (2.2.14).

2.4.2.2 Forward Mapping: $\theta \mapsto \mu$

Proposition 2.51 *([WJ08]).* The cumulant function $A(\theta)$ (2.4.7) is a convex function on its domain Ω (strictly in case of a minimal representation). Its gradient and Hessian matrix are the cumulants of the random vector $\phi(X)$:

$$\nabla A(\theta) = E[\phi(x)] := \int \phi_\alpha(x) p(x) \nu(dx), \tag{2.4.19a}$$

$$\nabla^2 A(\theta) = E[\phi(x)\phi(x)^T] - E[\phi(x)]E[\phi(x)]^T. \tag{2.4.19b}$$

We see from Proposition 2.51, that ∇A maps from the set of canonical parameters Ω to the set of realizable mean parameters $\mathcal{M}$. This mapping is one-to-one in case of a minimal representation or one-to-one from an affine subspace in case of a overcomplete representation. Furthermore, one can show that for minimal representations, $\nabla A(\theta)$ covers the whole interior of $\mathcal{M}$, denoted by $\operatorname{int} \mathcal{M}$ [WJ08].

2.4.2.3 Variational Representation of $A(\theta)$

Conjugate duality (see Definition 2.8) is the central technique for formulating variational representations of complex functions. For the log-partition function $A(\theta)$ (2.4.7), we define the conjugate dual function

$$A^*(\mu) := \sup_{\theta \in \Omega} \{\langle \mu, \theta \rangle - A(\theta)\}. \tag{2.4.20}$$

The choice of μ for the dual variable is deliberate, as μ and θ are coupled by conjugate duality. Any maximizer θ^* of (2.4.20) is uniquely determined by the correspondence

$$\mu = \nabla A(\theta^*) = E_\theta[\phi(X)]. \tag{2.4.21}$$

Since $A(\theta)$ is a convex function, it holds that $A(\theta) = A^{**}(\theta) = (A^*(\mu))^*$ [BV04]. Therefore the conjugate of A^* yields a variational representation of A:

Theorem 2.52 *(Variational representation of A).* The variational representation of the log-partition function $A(\theta)$ in terms of the dual A^* is given by

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\}. \tag{2.4.22}$$

This representation of A is one of the cornerstones of variational inference. The following theorem links the dual function A^* to the entropy of $p(x)$:

Theorem 2.53 *(Shannon entropy and A^*).* For any $\mu \in \mathcal{M}$ and the unique $\theta(\mu)$ satisfying (2.4.21), the dual function A^* corresponds to the negative entropy (c.f. Definition 2.23):

$$A^*(\mu) = \begin{cases} -H[p_{\theta(\mu)}], & \mu \in \operatorname{int} \mathcal{M}, \\ +\infty & \mu \notin \overline{\mathcal{M}}. \end{cases} \tag{2.4.23}$$

For any boundary point $\mu \in \operatorname{bd} \mathcal{M}$ we have $A^*(\mu) = \lim_{n \to +\infty} A^*(\mu^n)$, taken over any sequence $\mu^n \in \operatorname{int} \mathcal{M}$ converging to μ.

This is another important result, since it establishes two facts: First of all any optimization problem including A^* can be confined to the convex set $\mathcal{M}$. Furthermore, we may replace the indirect variational formulation of A^* (2.4.20) by the explicit formulation of the negative entropy. The downside of this is that the entropy is defined in terms of θ. Furthermore since $H[p]$ involves summing or integrating over the entire state space $\mathcal{X}^n$, the same issues over intractability may emerge as for the cumulant function $A(\theta)$.

Thus, although from a superficial point of view solving the convex objective function (2.4.22) over the convex set $\mathcal{M}$ seams tractable, there exist some serious obstacles:

- The set $\mathcal{M}$ may be difficult to express explicitly. For the marginal polytope $\mathbb{M}(\mathcal{G})$, the number of facets can grow exponentially, depending on the structure of $\mathcal{G}$.
- The indirect representation of $A^*(\mu)$ in terms of θ, thus its evaluation for even a single μ is computationally very expansive.

Variational methods can be classified according to the type of approximations to $\mathcal{M}$ and A^*. Below we will introduce two instances, message-passing (this time from a variational perspective) and mean field methods.

2.4.2.4 Message Passing Revisited

We will now revisit the message passing approach that we presented in Section 2.4.1.1 from a variational perspective. It relies on the **Bethe approximation**, an outer polyhedral bound to the marginal polytope $\mathbb{M}(\mathcal{G})$ and an approximation to the dual function $A^*(\mu)$, described here for the case of *pairwise* MRFs, such that every clique contains at most two variables[13].

Definition 2.54 *(Local polytope).* We call a polyhedral outer bound to $\mathbb{M}(\mathcal{G})$ the **local polytope** $\mathbb{L}(\mathcal{G})$, if it consists of locally consistent non-negative functions $\tau_s(x_s)$ and $\tau_{st}(x_s, x_t)$, that satisfy normalization constraints

$$\sum_{x_s} \tau_x(x_s) = 1, \ \forall x_s \in \mathcal{X}_s, \tag{2.4.24}$$

and marginalization constraints

$$\sum_{x_t} \tau_{st}(x_s, x_t) = \tau_s(x_s), \ \forall x_s \in \mathcal{X}_s, \quad \sum_{x_s} \tau_{st}(x_s, x_t) = \tau_t(x_t), \ \forall x_t \in \mathcal{X}_t. \tag{2.4.25}$$

Since any realization of mean parameters μ has to fulfill these constraints, the following proposition holds:

Proposition 2.55 *([WJ08]).* The inclusion $\mathbb{M}(\mathcal{G}) \subseteq \mathbb{L}(\mathcal{G})$ holds for any graph. For any tree-structured graph T both sets are equal, while for general graphs with cycles $\mathbb{L}(\mathcal{G})$ is a strict outer bound of $\mathbb{M}(\mathcal{G})$.

[13] This constitutes a rather general definition, since every undirected graphical model can be transformed into a pairwise MRF [WJ08]

As the second ingredient of the Bethe approximation, we approximate the dual function or negative entropy $A^*(\mu)$. While the entropy in general lacks a closed-form expression for MRFs with cycles, for a tree-structured MRF we can state an explicit expression in terms of the mean parameters μ that is exact:

$$H[p_\mu] = \sum_{s \in V} \underbrace{- \sum_{x_s \in \mathcal{X}_s} \mu_s(x_s) \ln \mu_s(x_s)}_{:= H_s(\mu_s)} - \sum_{(s,t) \in E} \underbrace{\sum_{x_s, x_t} \mu_{st}(x_s, x_t) \ln \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s)\mu_t(x_t)}}_{:= I_{st}(\mu_{st})} . \tag{2.4.26}$$

It decomposes into two terms, which are the **singleton entropy** $H_s(\mu_s)$ and **mutual information** $I_{st}(\mu_{st})$. For an MRF with loops (2.4.26) constitutes an approximation.

Combining both ingredients yields the **Bethe variational problem**

$$\max_{\tau \in \mathbb{L}(G)} \left\{ \langle \theta, \tau \rangle + \sum_{s \in V} H_s(\tau_s) - \sum_{(s,t) \in E} I_{st}(\tau_{st}) \right\}, \tag{2.4.27}$$

which gives exact results for tree-structured graphs. Setting the derivatives of the corresponding Lagrangian to zero, one can recover the sum-product updates (2.4.4), where the Lagrange multipliers λ_{st} associated with the marginalization constraints (2.4.25) assume the role of messages $M_{st} := \exp(\lambda_{st}(x_s))$ [WJ08].

In Section 2.4.1 we mentioned several message-passing approaches for dealing with general graphs, such as generalized belief propagation or tree-reweighted belief propagation. When viewed from a variational perspective, it turns out that they correspond to tighter approximations of the marginal polytope $\mathbb{M}(\mathcal{G})$ and the entropy function $H(p_{\theta(\mu)})$.

2.4.2.5 Mean Field Methods

An essential tool for the motivation of mean field methods is:

Definition 2.56 *(Kullback-Leibler divergence).* The **Kullback-Leibler (KL) divergence** or relative entropy between two probability mass functions $p(x)$ and $q(x)$ is defined as

$$\begin{aligned} \mathrm{KL}(q||p) &= - \sum_{x \in \mathcal{X}} q(x) \ln \frac{p(x)}{q(x)} \\ &= -E_q \left[\ln p(x) \right] - H[q]. \end{aligned} \tag{2.4.28}$$

The KL divergence is a measure of a distance between p and q. One can use Jensen's inequality to show that $\mathrm{KL}(q||p) \geq 0$ for all probability mass functions $q(x)$ and $p(x)$ with equality if and only if $p(x) = q(x)$ for all x [CT06]. For continues random variables X we simply replace summation by integration.

Mean field methods can be motivated from two perspectives: The first one is as a lower bound to $A(\theta)$. For any mean parameter $\mu \in \operatorname{int} \mathcal{M}$, the following inequality holds

$$A(\theta) \geq \langle \theta, \mu \rangle - A^*(\mu), \tag{2.4.29}$$

which is a direct consequence of (2.4.22). Mean field methods restrict the set $\mathcal{M}$ by only considering *tractable* subgraphs F of $\mathcal{G}$, such that it is feasible to perform exact calculations. As for the Bethe approximation, this includes approximations to both A^* and $\mathcal{M}$.

Associated with the cliques of F is a subset of sufficient statistics ϕ and parameters θ, indexed by $\mathcal{I}(F)$. The set of distributions that are Markov with respect to F is parametrized by

$$\Omega(F) := \{\theta \in \Omega \,|\, \theta_\alpha = 0, \;\; \forall \alpha \in \mathcal{I} \setminus \mathcal{I}(F)\}. \tag{2.4.30}$$

Naturally, this restricts the set of realizable mean parameters $\mathcal{M}$ to a proper subset denoted by $\mathcal{M}_F(\mathcal{G})$, which constitutes an *inner approximation* to $\mathcal{M}(\mathcal{G})$. We denote the dual function A^* restricted to $\mu \in \mathcal{M}_F(\mathcal{G})$ by $A^*_F(\mu)$. Mean field methods then find the best approximation to $A(\theta)$ by maximizing the lower bound

$$\max_{\mu \in \mathcal{M}_F(\mathcal{G})} \{\langle \theta, \mu \rangle - A^*_F(\mu)\}. \tag{2.4.31}$$

Tractability of mean field methods comes at a price though, as one can show that the set $\mathcal{M}_F(\mathcal{G})$ as well as the objective function (2.4.31) may be non-convex [WJ08].

The alternative perspective is given as minimization of the KL divergence between the approximating distribution $q \in \mathcal{Q}$ and the target distribution p_θ. Here $\mathcal{Q}$ denotes a tractable family of distributions, exhibiting additional CI assumptions compared to p_θ. There is a direct correspondence between $\mathcal{Q}$ and $\mathcal{M}_F(\mathcal{G})$.

Denote by X the set of observed random variables and by Z the set of hidden latent random variables. We wish to find the distribution $q(Z)$ closest to $p(Z|X)$ in terms of the KL divergence (2.4.28), that is

$$\min_{q \in \mathcal{Q}} \mathrm{KL}(q||p) = \min_{q \in \mathcal{Q}} \; -\sum_Z q(Z) \ln \frac{p_\theta(Z|X)}{q(Z)}. \tag{2.4.32}$$

We can absorb the effect of observing X by updating the parameters θ, such that $p_\theta(Z|X) = p_{\tilde{\theta}}(Z)$. Using the reparameterized form, we continue from (2.4.32):

$$\begin{aligned}
&\min_{q \in \mathcal{Q}} \; -\sum_Z q(Z) \ln p_{\tilde{\theta}}(Z) + \sum_Z q(Z) \ln q(Z) \\
=&\min_{q \in \mathcal{Q}} \; -E_q[\langle \tilde{\theta}, \phi(X) \rangle - A(\tilde{\theta})] - H[q] \\
=&\min_{\mu \in \mathcal{M}_F(\mathcal{G})} -\langle \tilde{\theta}, \mu \rangle + A(\tilde{\theta}) + A^*_F(\mu),
\end{aligned}$$

where we used the fact that $-H[q] = A^*_F(\mu)$. Since $A(\theta)$ is constant with respect to μ, we end up with the problem (2.4.31).

2.5 Retina Imaging

2.5.1 Optical Coherence Tomography

Optical Coherence Tomography (OCT) is a comparatively new technique for acquiring cross-sectional images of internal structures in biological tissues, first demonstrated by [HSL+91]. Soon the first tests with human retina followed [FHD+93, SIH+93]. Since then OCT has found an ever-growing application in ophthalmic diagnosis. It enabled ophthalmologists to obtain high-resolution images of retinal layers, not possible with techniques previously used.

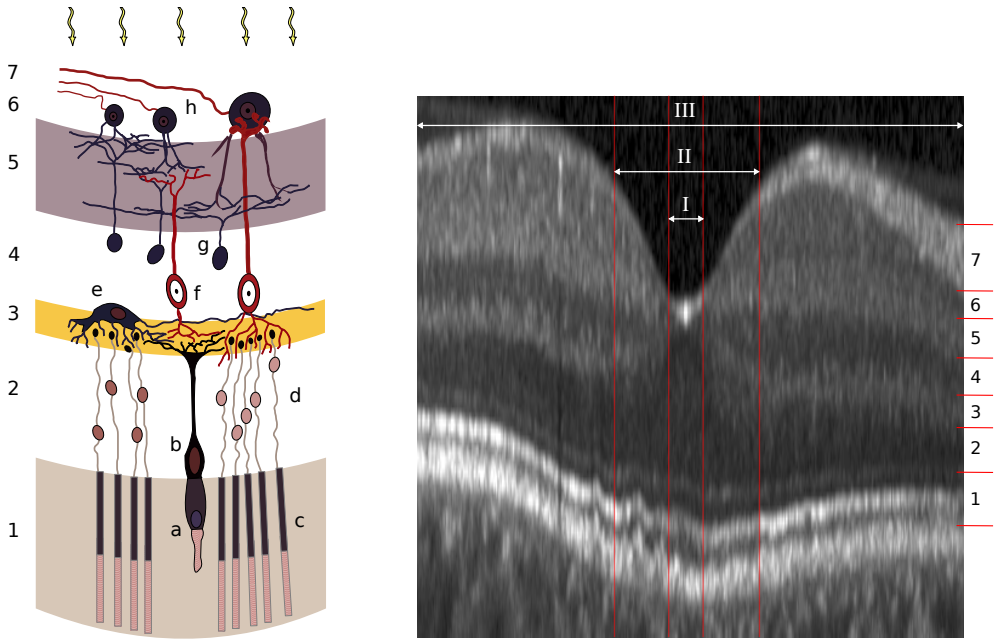
OCT is a low coherence interferometry technique, using a Michelson interferometer. Here, the light source is split into two beams by the use of a partial reflective mirror. One beam is reflected by the so called reference mirror. The other beam is focused on the sample tissue, using a mirror and an objective lens. While parts of the beam are absorbed or scattered, some amount is reflected back. Both beams are then recombined to produce an interference pattern, visible to the observer. Comparing the echo time delay and intensity of the reflected light with that from the reference arm, a so-called **A-scan** is obtained, an axial gray-scale plot of the interferometric signal strength. By translating the optical beam laterally, 2-D scans (**B-scan**) or 3-D scans (n B-scans) of the sample are obtained.

There exist two main types of optical coherence tomography: **Time-domain OCT** and **Fourier-domain OCT**. In time-domain OCT the reference reflector position is translated to obtain information from various depths in the sample. Contrary, in Fourier-domain OCT the reference arm is held fixed, but broadband light source and a spectrometer are used. Through Fourier transformation, the spectrum of the backscattered light is transformed into the sample reflectance as a function of depth. Avoiding the limitations of the mechanical component, Fourier-domain OCT features a much higher acquisition speed and a lower signal-to-noise ratio [CSYI03, LHF+03].

2.5.2 Retinal Anatomy

The vertebrate retina, anatomically a part of the central nervous system, translates incoming light into electrical signals, pre-processes them and relays them via the optic nerve to the visual cortex for visual perception. It is composed of several layers that contain different types of neurons and the synapses that interconnect them, c.f. Figure 2.8. Of these neurons only **photoreceptor cells**, located at the back of the retina, are directly sensitive to light. There exist two types of photoreceptor cells, rods and cones. Cones are responsible for color vision and work best in bright light, while rods are very sensitive to light and therefore are saturated in bright light, but provide vision in conditions of dim light.

Next come layers containing three different types of neuronal cells: **bipolar cells**, **horizontal cells** and **amacrine cells**. These cells act as pre-processing units, that relay the input they receive from the photoreceptor cells to the **ganglion cells**, which are situated in the outermost layer. Horizontal and amacrine cells provide lateral connectivity, the former from receptors to bipolar cells and the latter from bipolar cells to ganglion cells. Each ganglion cell has a *receptive field*, i.e. is influenced



(a) (b)

Figure 2.8 - Both figures illustrate the axial organization of cells in the retina. In Figure (a) [RyC11] letters label different cells types: **a,b**: cones and cone nuclei, **c,d**: rods and rod nuclei, **e**: horizontal cells, **f**: bipolar cells, **g**: amacrine cells, **h**: ganglion cells. Numbers correspond to different layers: **1**: rod and cone layer, **2**: outer nuclear layer (ONL) **3**: outer plexiform layer (OPL), **4**: inner nuclear layer (INL), **5**: inner plexiform layer (IPL), **6**: ganglion cell layer (GCL), **7**: nerve fiber layer (NLF). (Source: Wikipedia) (b) OCT-scan from the central slice of a macula-centered 3-D volume, showing the locations of layers **1**-**7**. Regions labeled by roman numbers denote foveola (**I**), fovea (**II**) and macula (**III**) and correspond to those depicted in the fundus image in Figure 2.9 (a).

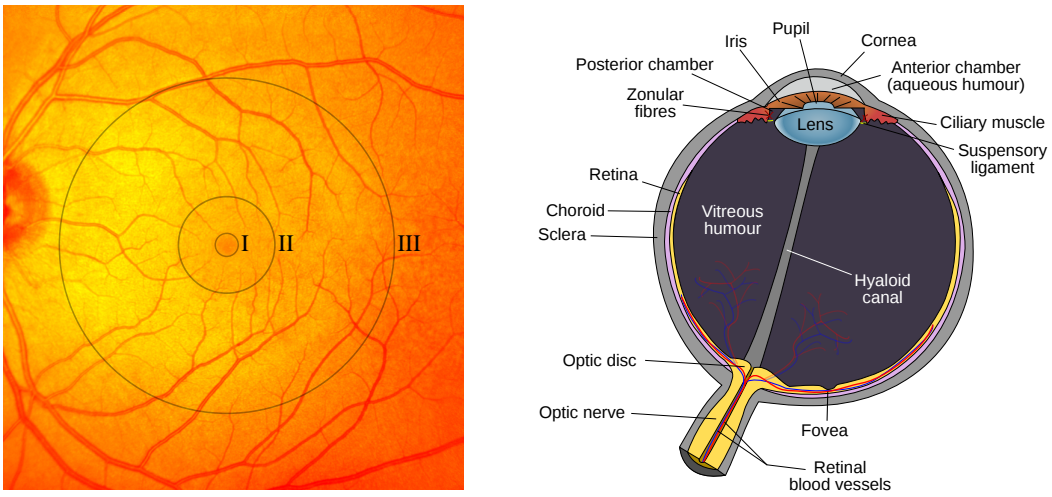


Figure 2.9 - (a) Fundus image that shows the foveola (**I**), fovea (**II**) and macula (**III**). Visible to the left is the optic disc (or optic nerve head). See also Figure 2.8 (b) for an example of an OCT scan of that region. (b) Schematic diagram of the human eye, illustrates the way the light travels after entering through the lens, through the vitreous humour towards retinal layers located around the fovea. (Source: Wikipedia)

by a certain compact region of photoreceptor cells. There are about 125 million photoreceptor cells but only 1.2 to 1.5 million ganglion cells, so each ganglion cell receives on average input from 100 photoreceptor cells [Hec87]. The axons of the ganglion cells form the **optic nerve**, which leaves the retina and transmits the visual information to the next processing instance inside the brain.

The **macula** (III in Figure 2.9 (a)) is the region of the retina in responsible for visual acuity. It measures about $5.5\,\mu m$ in diameter and is defined anatomically as having two or more layers of ganglion cells [Sch99]. Located at its center is the **fovea** (II), a $1.5\,\mu m$ wide region which in turn includes the **foveola** (I), a $0.35\,\mu m$ wide area where the upper neuronal layers disappear, such that light can directly hit the photoreceptor cells [Alf06]. The foveola also features a much lower ratio of photoreceptor cells per ganglion cell, allowing a much higher resolution [Hub95]. Figure 2.9 (b) illustrates the eye ball and the position of the foveal pit therein.

2.5.3 Glaucoma

The term *glaucoma* describes a group of ocular disorders, that are the second leading cause of blindness in the United States [KFKA09]. Around 66 million individuals are estimated to be affected by glaucoma world wide [WK04]. The most common type is **open-angle glaucoma**, accounting for roughly 90 % of all glaucoma cases. Glaucoma is characterized by the loss of ganglion cells and their axons, as well as tissue remodelling involving the optic nerve head and the retina [MGF08]. This is believed to be caused mainly by a diminished aqueous outflow of the eye, often accompanied by a slow build-up of intraocular pressure (IOP). Nevertheless, since

not all patients with glaucoma exhibit an increased IOP, also other factors contribute to the progression of the disease [WK04].

Symptoms are the loss of peripheral vision and, if left untreated, the irreversible loss of vision. The progress of vision loss can be determined by a visual field test. Nevertheless, these symptoms occur comparatively late in the course of the disease, such that at the point of detection as many as 50 % of all ganglion cells may have been lost [Qui99]. This emphasizes the necessity to use other diagnostic methods, which are able to detect glaucoma in a much earlier stage. Measurement of IOP, although simple and fast, has only limited clinical benefit, since it is hampered by a high false positive rate [KHH+02].

As pointed out earlier, glaucoma affects ganglion cells as well as the optic nerve head also known as optic disc. Regarding the latter, one can for example measure the cup-to-disc-ratio, which compares the size of the white cup, an area within the optic disc having no nerve fibers, with the size of the optic disc itself. Glaucoma is correlated with an increase of this ratio and several studies underlined the value of this measurement [QKD+92, WGHH+98, KVGH+99]. *Measuring the thickness of the nerve fiber layer (NFL)* constitutes another structural indicator for glaucoma. Early studies used fundus photographs as the one shown in Figure 2.9 (a) to localize NFL defects [HFN73, SMP+77, AN85]. But fundus photographs only provide a top view of the NFL, therefore indication has to rely on the subjective assessment of typical texture variations. It was shown by [QA+82], that up to 50 % of the thickness of the NFL may be lost until the defect is visible in the fundus image.

The recent advent of high-resolution OCT enabled the accurate measurement of NFL thickness. Several studies demonstrated the applicability of NFL thickness evaluation for the detection of glaucoma [BZB+01, LCC+05, CKFB09, LRZ+11]. Additionally, recent research suggests, that glaucoma not only manifests itself in transformations of the NFL, but additionally also influences the ganglion cell layer (GCL), inner plexiform layer (IPL) and to a lesser extent the inner nucleus layer (INL) (numbers 4-6 in Figure 2.8) [TLL+08, TCL+09, KHH+11]. This seems reasonable, since glaucoma affects ganglion cells, whose axons are located in the NFL, but their nuclei reside inside the GCL and their dendrites, connecting them to bipolar cells and amacrine cells, lie inside the IPL.

Many other diseases of the retina exists, which all manifest themselves in different cell layers. This stresses the need for a segmentation approach which yields accurate delineations of as many cell layer boundaries as possible. The approach presented in this thesis segments eight different inner cell layers, and can easily be extended to segment more, if training data in form of labeled OCT scans becomes available.

3 A Probabilistic Graphical Model for Retina Segmentation

This chapter will present our retina segmentation model. In Section 3.1 we outline the parts that constitute our graphical model: the appearance models $p(y|c)$ modeling texture of partitions and their boundaries, the shape prior $p(b)$ and the prior for discrete boundary assignments $p(c|b)$. In Section 3.2 we present the approximative probabilistic inference framework, based on variational inference. We derive explicit update formulas for the sufficient statistics of the approximating distributions $q_c(c)$ and $q_b(b)$ in Section 3.3.

3.1 Graphical Model

This section presents our probabilistic graphical model, statistically modeling an OCT scan y and its segmentations b and c respectively. We introduce c, the discretized version of the continuous boundary vector b, to make mathematically explicit the connection between the discrete pixel domain of y and the continuous boundary domain of b. Our ansatz is given by

$$p(y, c, b) = p(y|c)p(c|b)p(b), \tag{3.1.1}$$

where the factors are

$p(y\|c)$	appearance, data likelihood term,
$p(c\|b)$	Markov Random Field regularizer, determined by the shape prior and
$p(b)$	global shape prior.

Moreover, we introduce the vector x that holds class labels for all pixel, indicating their affiliation to either one of the cell layers or their corresponding boundaries. Note that x is directly determined by c. We will sometimes make this connection explicit by writing $p(y|x(c))$.

Notation. Figure 3.1 displays the notation for components of vectors x, b and c: We will use the subindex $k \in \{1, \dots, N_b\}$ to differentiate between boundaries and the subindices $j = \{1, \dots, M\}$ and $i = \{1, \dots, N\}$ to differentiate between image columns and rows. The symbol $\bullet$ will denote the set of all elements of the respective index, for example $b_{k,\bullet} \in \mathbb{R}^M$ is the vector holding real-valued positions of boundary k in all image columns, and $c_{k,\bullet} \in \{1, \dots, N\}^M$ is its discretized counterpart. By $b_{\setminus j}$ we denote the subset $\{b \setminus b_{\bullet,j}\}$ and will use an similar notation for μ and Σ. We will sometimes drop the bullet-symbol to improve readability, thus $b_{\bullet,j}$ will become b_j. Moreover, we will not emphasize the difference between random variables and their

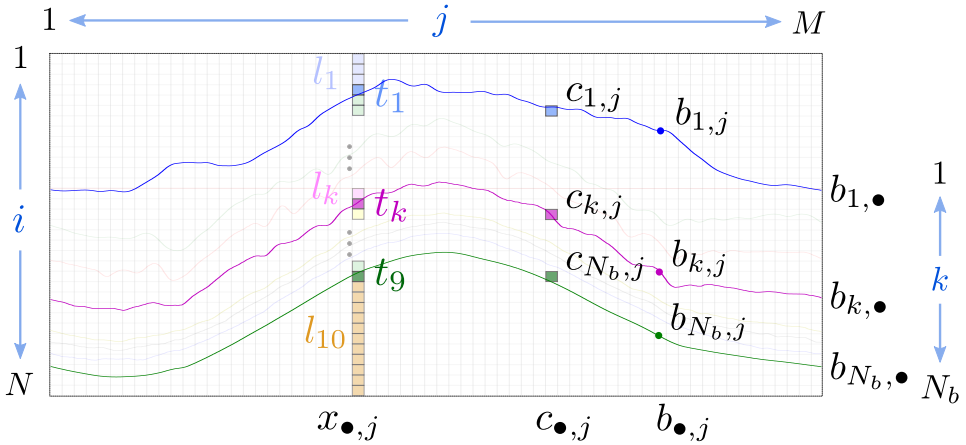


Figure 3.1 - Important variables used throughout this section. Note the difference between real valued boundary position $b_{k,j}$ and its discretized counterpart $c_{k,j}$.

realizations, but use lower case letter throughout our development. Section 1.5 recapitulates the notation in compact form.

In what follows we will detail each component, thereby completing the definition of our graphical model. The illustration of the graph in terms the individual layers in Figure 3.2 accompanies this presentation.

3.1.1 Appearance Models

We utilize Gaussian distributions to model the appearance of retinal layers as well as their boundaries. Given a segmentation hypothesis c, we can assign class labels $x_{i,j} \in \mathcal{X}$ to each pixel, their range being given by

$$\mathcal{X} = \{\mathcal{X}_l, \mathcal{X}_t\}, \qquad \mathcal{X}_l = \{l_1, \ldots, l_{10}\}, \quad \mathcal{X}_t = \{t_1, \ldots, t_9\}.$$

Thus, labels denote membership of observed pixels to either tissue layers l_k or transitions (boundaries) t_k that separate them. To obtain a valid mapping $c \mapsto x$, we require c to satisfy the ordering constraint

$$1 \leq c_{1,j} < c_{2,j} < \cdots < c_{N_b,j} \leq N, \quad \forall j = 1, \ldots, M, \tag{3.1.2}$$

and point out that the real-valued counterpart b may violate this constraint.

Since OCT scans display a large inter-scan as well as intra-scan variability in terms of their brightness and contrast, each patch $y_{i,j}$ is first normalized by subtracting its mean. We then project each patch $y_{i,j}$ onto a low-dimensional manifold, applying the technique of PCA [Hot33]. To this end, we randomly draw patches from the training set independently of their class affiliation, and estimate their empirical covariance matrix and calculate its eigenvalues and eigenvectors. The projection can then be carried out using the first q_{pca} eigenvectors sorted by their eigenvalues. See Section 4.1.3 for a discussion on how we set that parameter during evaluation.

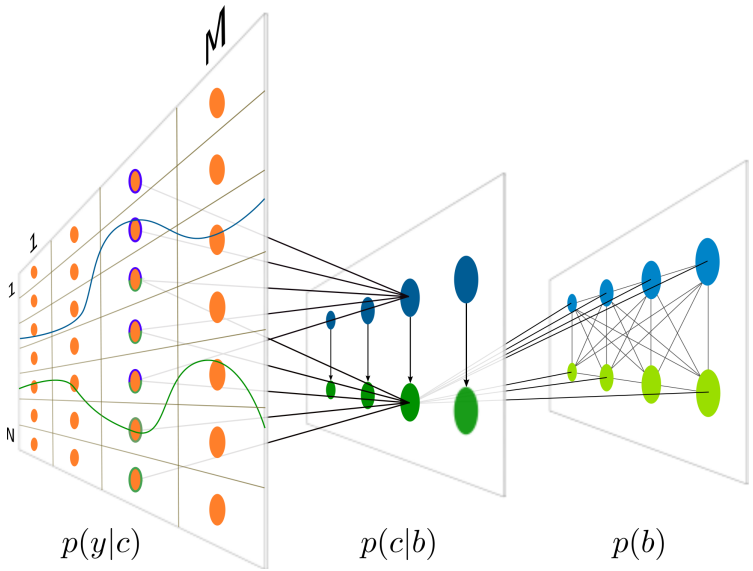


Figure 3.2 - Illustration of our probabilistic graphical model for $M = 4$, $N = 7$ and $N_b = 2$. The connectivity from b to c is only displayed for node $c_{2,3}$. Similarly, connectivity for c to y via x is only displayed for nodes in the third image column and additionally illustrated by the edge color of the y-nodes.

We define the probability of the projected patch $y_{i,j}$[1] at pixel (i, j) belonging to the class $x_{i,j}$ as

$$p(y_{i,j}|x_{i,j}(c)) = \mathcal{N}(y_{i,j}; \mu_{x_{i,j}}, \Sigma_{x_{i,j}}) \,. \tag{3.1.3}$$

The class-specific moments μ_x, Σ_x, $\forall x \in \mathcal{X}$ are learned offline using patches from the respective class. Regularized estimates for Σ_x are obtained by utilizing the graphical lasso approach [FHT08], see Section 2.2.5.2. This leads to sparse estimates for K, where the degree of sparsity is governed by the parameter α_{glasso}. Again confer Section 4.1.3 for details on how we set that parameter.

We define patches $y_{i,j}$ to be conditionally independent given a segmentation c, that is

$$p(y|c) = \prod_{j=1}^{M} \prod_{i=1}^{N} p(y_{i,j}|x_{i,j}(c)) \,. \tag{3.1.4}$$

Finally, we introduce switches $\beta^t \in \{0, 1\}$ and $\beta^l \in \{0, 1\}$, that turn on and off all terms belonging to the corresponding transition class t_k or layer class l_k, which yields the appearance model

$$p(y|c) = \prod_{j=1}^{M} \prod_{i:x_{i,j} \in \mathcal{X}_l} p(y_{i,j}|x_{i,j}(c))^{\beta^l} \prod_{i:x_{i,j} \in \mathcal{X}_t} p(y_{i,j}|x_{i,j}(c))^{\beta^t} . \tag{3.1.5}$$

As we point out in the next section, our model can handle discriminative terms as well. We can convert generative terms (3.1.3) into discriminative ones by renormaliz-

[1] For ease of notation, we will make no difference between a patch and its low-dimensional projection and denote both by $y_{i,j}$.

ing:

$$p(x_{i,j}(c)|y_{i,j}) = \frac{p(y_{i,j}|x_{i,j}(c))p(x_{i,j}(c))}{\sum_{x_{i,j}\in\mathcal{X}} p(y_{i,j}|x_{i,j}(c))p(x_{i,j}(c))}, \tag{3.1.6}$$

where we use a uniform prior $p(x_{i,j}(c))$. Also $p(c|y)$ factorizes as a product distribution.

3.1.2 Shape Prior

As a model of the typical shape variation of layers due to both biological variability as well as to the image formation process, we adopt a joint Gaussian distribution. For 2-D circular scans (see for example Figure 1.2), a wave-like distortion pattern is observed due to the conic scanning geometry and the spherical shape of the retina, which we capture statistically rather than modeling it explicitly.

We denote the continuous height values of all boundaries k over all image columns j by the N_bM-dimensional vector $b = (b_{k,j})_{k=1,\dots,N_b;\, j=1,\dots,M}$. Hence,

$$p(b) = \mathcal{N}(b; \mu, \Sigma), \tag{3.1.7}$$

where parameters μ and Σ are learned offline from labeled training data. We regularize the estimation of Σ by Probabilistic Principal Component Analysis (PPCA), presented in Section 2.2.5.3. PPCA assumes that the high-dimensional observation b was generated from a low-dimensional latent source $s \in \mathbb{R}^q$ via

$$b = Ws + \mu + \epsilon,$$

where $s \sim \mathcal{N}(0, I)$ and $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ is isotropic Gaussian noise. The moments of $p(b)$ are given by $E[b] = \mu$ and $E[bb^T] = WW^T + \sigma^2 I = \Sigma$. The precision matrix $K = \Sigma^{-1}$ can be decomposed into W and $\sigma^2 I$ as well (c.f. (2.2.27)), thereby reducing complexity as well as memory requirements of most operations related to Σ and K.

Figure 3.3 shows samples drawn from $p(b)$, modeling fovea-centered 3-D volumes (left panel, with the fovea clearly visible) and circular scans (right panel).

3.1.3 Shape-Induced Regularizers

The third component of our model is a prior for discrete boundary assignments c, that regularizes the data likelihood term $p(y|c)$. We define $p(c|b)$ as a collection of column-wise acyclic graphs

$$p(c|b) = \prod_{j=1}^{M} p(c_{\bullet,j}|b), \qquad p(c_{\bullet,j}|b) = p(c_{1,j}|b) \prod_{k=2}^{N_b} p(c_{k,j}|c_{k-1,j}, b). \tag{3.1.8}$$

That means the communication *between* image columns is governed by the shape prior $p(b)$.

In order to define the conditional distributions in (3.1.8), we need a couple of notational prerequisites. Recall that $b_{\setminus j}$ denotes the sub-vector of b after removing $b_{\bullet,j}$, that is all boundary positions in image column j. The conditional distributions are

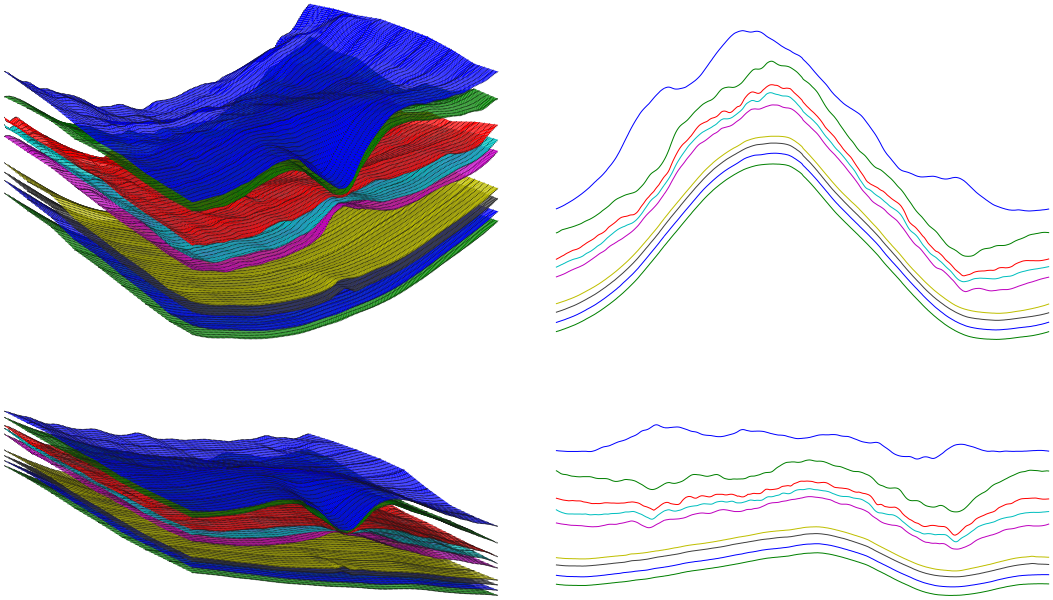


Figure 3.3 - Samples drawn from the the shape prior distribution $p(b)$ trained on volumes (left) and circular scans (right). Only one half of the volume is shown. In the volume samples the fovea, that is the region of the retina where the upper layers disappear, is clearly visible.

specified in terms of b:

$$\begin{aligned} p(c_{1,j}=n|b) &= \Pr\left(n-\frac{1}{2} \leq b_{1,j} \leq n+\frac{1}{2}\right), \\ p(c_{k,j}=n|c_{k-1,j}=m,b) &= \\ &\Pr\left(n-\frac{1}{2} \leq b_{k,j} \leq n+\frac{1}{2} \,\middle|\, m-\frac{1}{2} \leq b_{k-1,j} \leq m+\frac{1}{2}\right), \end{aligned} \tag{3.1.9}$$

where the probabilities on the right-hand side are computed using the conditional distributions $p(b_{1,j}|b_{\setminus j})$ and $p(b_{k,j}|b_{\setminus j})p(b_{k,j}|b_{k-1,j})$ respectively, for all configurations of c conforming to (3.1.2). Since $p(b)$ is a normal distribution, these computations are straightforward, see Section 2.2.5. From a modeling-perspective, the conditional distribution $p(b_{k,j}|b_{\setminus j})$ provides a way to introduce global shape knowledge into the column-wise Markov random fields $p(c_{\bullet,j}|b)$.

3.1.4 2-D vs. 3-D

Our description so far considered OCT scans of dimension two. Nevertheless, our approach is equally applicable to 3-D volumes. We can use the very same notation, since adding additional B-Scans will only increase the number of image columns M. Similarly, the connectivity of the graphical model $p(y,c,b)$ can be transferred one-to-one.

The shape prior $p(b)$ which is fully connected since K is dense, can be extended to an arbitrary dimension. We exploit the fact that both, Σ and K, have an explicit low-rank decomposition (as discussed in Section 3.1.2), such that memory consumption is

not an issue and complexity of operations is reduced as well. For the regularization term $p(c|b)$, each node $c_{k,j}$ is connected to nodes $b_{\setminus j}$ of all columns except the current one, which now additionally includes columns of all other B-scans. Finally, the data likelihood $p(y|c)$ continues to fully factorize over pixels (i, j). Each pixel (i, j) remains connected to at most two nodes $c_{k,j}$ from the same column j, determining it's label $x_{i,j}$. Finally, we use separate sets of appearance models for each B-scan in the volume to capture variations across different regions of the retina.

3.2 Variational Inference

Based on the model presented in the last section and given observed data y, we wish to infer the posterior distribution

$$p(b, c|y) = \frac{p(y|c)p(c|b)p(b)}{p(y)}. \tag{3.2.1}$$

Here, one major obstacle is the calculation of the marginal likelihood $p(y)$, which requires the integration respective summation of $p(y, c, b)$ over b and c. But since we lack a closed form solution and the problem at hand is high-dimensional, this is intractable.

We cope with this problem by applying an established variational method: approximating the posterior by a tractable distribution $q(b, c)$ by minimizing the Kullback-Leibler (KL) distance $\mathrm{KL}(q\|p)$ with respect to q. This type of inference approximations is called the mean-field approach and was discussed in Section 2.4.2.5. We point out that unlike in related work (e.g. [MTRP09]) where the subproblem of inferring the discrete decision variables has to be approximated as well, our model has been designed such that by choosing q properly, all subproblems are tractable and can be solved efficiently.

We choose the factorized approximating distribution

$$q(b, c) = q_b(b)q_c(c). \tag{3.2.2}$$

This merely decouples the continuous shape prior and the discrete order-preserving segmentation component of the overall model, but otherwise will represent both components exactly. The Kullback-Leibler distance between q and p is given by

$$\begin{aligned}\mathrm{KL}\left(q(b, c)\|p(b, c|y)\right) &= \int_b \sum_c q(b, c) \log \frac{q(b, c)}{p(b, c|y)} db \\ &= -\int_b \sum_c q(b, c)\Big(\log\left(p(y|c)p(c|b)p(b)\right) - \log p(y) - \log q(b, c)\Big) db\,. \end{aligned} \tag{3.2.3}$$

Dropping the constant term $\log p(y)$, we may obtain our objective function.

Optionally, we can use the marginal likelihood $\log p(y)$ to introduce *discriminative*

appearance terms into the model, using

$$\log \frac{p(y|c)}{p(y)} = \log \frac{p(y|c)p(c)}{p(y)} - \log p(c) = \log p(c|y) - \log p(c).$$

Since $p(b)$ already contains prior knowledge about the shape of boundary positions, we assume an uninformative prior for c. Hence dropping $p(c)$ and taking into account the factorization of q, we obtain the objective function

$$J(q_b, q_c) = -\int_b \sum_c q_b(b) q_c(c) \log \Big(p(c|y) p(c|b) p(b) \Big) db - H[q_b] - H[q_c], \tag{3.2.4}$$

where $H[q_c]$ and $H[q_c]$ denotes the entropies of the approximating distributions, c.f. Definition 2.23. It turned out that discriminative appearance terms yielded much better performance, and we discuss that issue in Section 4.1.2. We will therefore focus on the discriminative case in the subsequent derivation.

In what follows, we make the expectations with respect to q_c and q_b explicit. This will provide us below with a closed-form expression of the objective function $J(q_b, q_c)$. We begin by defining q_c and q_b.

3.2.1 Definitions of q_c and q_b

For $q_c(c)$ we adopt the same factorization as for $p(c|b)$, that is, written in a slightly different but equivalent form

$$q_c(c) = \prod_{j=1}^{M} q_{c;1,j}(c_{1,j}) \prod_{k=2}^{N_b} \frac{q_{c;k\wedge k-1,j}(c_{k,j}, c_{k-1,j})}{q_{c;k-1,j}(c_{k-1,j})}, \tag{3.2.5}$$

where $q_{c;k,j}$ are discrete probability distributions, such that the normalization constraints (2.4.24) are satisfied. Similarly, by $q_{c;k\wedge k-1,j}$ we denote discrete probability distributions over pairs of variables $c_{k-1,j}, c_{k,j}$. To enhance readability, we will subsequently omit indices k, j of q_c, if they are determined by their input variables.

For $q_c(c)$ to be a valid distribution, additional marginalization constraints have to be satisfied for all $q_{c;k\wedge k-1,j}$, c.f. (2.4.25). Note that we ignore here the set of valid configurations (3.1.2), because this has already been taken into account when defining $p(c|b)$. As for q_c and $p(c|b)$, we let q_b adopt the same factorization as $p(b)$, thus

$$q_b(b) = \mathcal{N}(b; \bar{\mu}, \bar{\Sigma}), \tag{3.2.6}$$

where the bar-notation helps to distinguish the sufficient statistics of q_b from those of $p(b)$.

3.2.2 First Summand $\log p(c|y)$ of $J(q_b, q_c)$

The term $p(c|y)$ does not depend on b, so q_b integrates out. Moreover, q_c factorize over image columns j and $p(x(c)|y)$ is a factor distribution. Hence we can rewrite

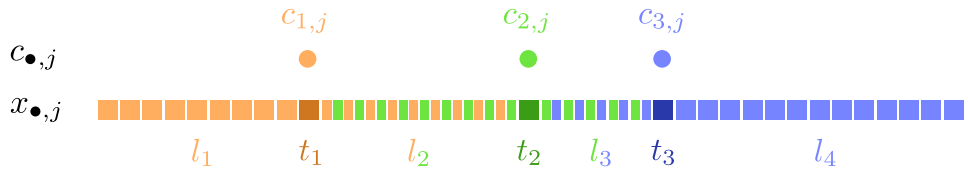


Figure 3.4 - Label vector $x_{\bullet,j}$ and boundary position vector $c_{\bullet,j}$ for $N_b = 3$ and $N = 40$. The coloring of each label $x_{i,j}$ denotes dependency on $c_{\bullet,j}$. Each label depends at most on two elements of $c_{\bullet,j}$, and we can use that fact when calculating expectations with respect to q_c.

the first summand of (3.2.4) as

$$-\int_b \sum_c q_b(b) q_c(c) \log p(c|y) = -\sum_{j=1}^{M} \sum_{c_{\bullet,j}} q_c(c_{\bullet,j}) \sum_{i=1}^{N} \log p(x_{i,j}(c_{\bullet,j})|y_{i,j}), \tag{3.2.7}$$

where the second sum ranges over all combinations of boundary assignments for $c_{\bullet,j}$. We can further simplify this equation by noting that each label $x_{i,j}$ depends at most on two $c_{k,j}$, as illustrated in Figure 3.4.

This enables us to split the inner sum over rows $i = 1, \dots, N$ into $k+1$ sums. One for each pair of labels (l_k, t_k) that are dependent only on $c_{k-1,j}$ and $c_{k,j}$ and one for l_{N_b+1} respectively. For each of theses sums we can therefore sum out all $q_c(c_{\bullet,j})$ that are independent of the respective labels. Continuing from (3.2.7) we obtain

$$\begin{aligned} = -\sum_{j=1}^{M} \Big(& \sum_{c_{1,j}} q_c(c_{1,j}) \sum_{i=1}^{c_{1,j}} \log p(x_{i,j}(c_{1,j})|y_{i,j}) \\ & + \sum_{c_{1,j}} \sum_{c_{2,j}} q_c(c_{1,j}, c_{2,j}) \sum_{i=c_{1,j}+1}^{c_{2,j}} \log p(x_{i,j}(c_{1,j}, c_{2,j})|y_{i,j}) + \dots \Big). \end{aligned} \tag{3.2.8}$$

For each pair (l_k, t_k) of labels we define matrices $\Psi_{k,j}$, whose entries equal the sum over pixel $y_{i,j}$ with $x_{i,j} \in \{t_k, l_k\}$:

$$(\Psi_{k,j})_{m,n} = \left(\sum_{i=m+1}^{n-1} \beta^l \log p(x_{i,j} = l_k|y_{i,j}) \right) + \beta^t \log p(x_{n,j} = t_k|y_{n,j}),$$

for $k = 2, \dots, N_b$, $j = 1, \dots, M$ and $1 \le m \le n \le N$. Entries for $n \le m$ are not defined and set to negative infinity. Note that throwing away probability mass is a rather rude approach. We will discuss a more elegant approach in Section 5, by estimating a prior density that is directly restricted to the support of the training data and thereby to the cone of correctly ordered boundaries.

We now can write for each of the summands in (3.2.8) (except for those only related to $c_{1,j}$ and $c_{N_b,j}$ respectively):

$$\sum_{c_{k-1,j}} \sum_{c_{k,j}} q_c(c_{k,j}, c_{k-1,j}) (\Psi_{k,j})_{c_{k-1,j}, c_{k,j}} = \langle q_{c;k \wedge k-1,j}, \Psi_{k,j} \rangle,$$

where $\langle A, B \rangle$ denotes the inner product of matrices A and B. This also determines

the form of $q_{c;k\wedge k-1,j}$. Accordingly, we introduce vectors $(\psi_{1,j})_n$ and $(\psi_{N_b,j})_n$, representing sums over pixels with labels l_1, t_1 and l_{N_b+1} depending on $c_{1,j}$ and $c_{N_b,j}$ respectively.

We now can state the final form for the first term in $J(q_b, q_c)$:

$$-\sum_{j=1}^{M}\Big((q_{c;1,j})^T\psi_{1,j}+\sum_{k=2}^{N_b}\langle q_{c;k\wedge k-1,j},\Psi_{k,j}\rangle+(q_{c;N_b,j})^T\psi_{N_b,j}\Big)\,. \tag{3.2.9}$$

3.2.3 Second Summand $\log p(c|b)$ of $J(q_b, q_c)$

The second term in $J(q_b, q_c)$ is

$$-\int_b\sum_c q(b,c)\log p(c|b) = -E_{q_c}\Big[E_{q_b}[\log p(c|b)]\Big]. \tag{3.2.10}$$

We will first calculate the expectation with respect to q_b.

Expectation with respect to q_b. In (3.1.9) we defined the terms of $p(c|b)$ as conditional distributions of $p(b)$. By the standard rule for conditional normal distributions (see Section 2.2.5) we obtain for $p(b_j|b_{\setminus j})$, the distribution of boundary positions in column j conditional on boundary positions in all other image columns:

$$\begin{gathered}p(b_j|b_{\setminus j}) = \mathcal{N}(b_j;\mu_{j|\setminus j},\Sigma_{j|\setminus j}),\\ \mu_{j|\setminus j} = \mu_j - \Sigma_{j|\setminus j}K_{j,\setminus j}(b_{\setminus j}-\mu_{\setminus j}),\qquad \Sigma_{j|\setminus j} = (K_{jj})^{-1}\,,\end{gathered} \tag{3.2.11}$$

The univariate density $p(b_{k,j}|b_{\setminus j})$ is obtained by marginalizing over (3.2.11), and we denote its mean by $(\mu_{j|\setminus j})_k$ and its variance by $(\Sigma_{j|\setminus j})_{k,k}$.

We obtain $p(b_{k,j}|b_{k-1,j})$, the conditional distribution of boundary position $b_{k,j}$ given the position of its direct neighbor $k-1$ in column j by the same formula:

$$\begin{gathered}p(b_{k,j}|b_{k-1,j}) = \mathcal{N}(b_{k,j};\mu_{k|k-1,j},\sigma^2_{k|k-1,j}),\\ \mu_{k|k-1,j} = \mu_{k,j} - \sigma^{-2}_{k|k-1,j}(K_{jj})_{k,k-1}(b_{k-1,j}-\mu_{k-1,j}),\quad \sigma^2_{k|k-1,j} = (K_{jj})_{k,k}\,.\end{gathered} \tag{3.2.12}$$

We now can express the probabilities $p(c_{1,j}|b)$ and $p(c_{k,j}|c_{k-1,j},b)$ introduced in (3.1.9) in terms of the integrals

$$\begin{aligned}p(c_{1,j}=n|b) &= \int_{n-\frac{1}{2}}^{n+\frac{1}{2}} p(b_{1,j}=\tau|b_{\setminus j})d\tau,\\ p(c_{k,j}=n|c_{k-1,j}=m,b) &= \\ &\int_{n-\frac{1}{2}}^{n+\frac{1}{2}}\int_{m-\frac{1}{2}}^{m+\frac{1}{2}} p(b_{k,j}=\tau|b_{\setminus j})p(b_{k,j}=\tau|b_{k-1,j}=\nu)d\tau d\nu.\end{aligned}$$

Turning our attention back to the problem of calculating $E_{q_b}[\log p(c|b)]$, we notice that the terms of $p(c|b)$ depend on $b_{\setminus j}$ via $(\mu_{j|\setminus j})_k$, hence on q_b too. It suffices to adopt the most crude numerical integration formula (integrand = step function) in order to make this dependency explicit $\int_{a-1/2}^{a+1/2} f(x)dx \approx f(a)$.

By applying the logarithm to $p(c|b)$, we obtain a representation that is convenient for the evaluation of $\int_b \cdots q_b db$. Recall that q_b is normally distributed with moments $\bar{\mu}$ and $\bar{\Sigma}$. Therefore the moments of $b_{\setminus j}$ with respect to q_b are given by

$$E_{q_b}[b_{\setminus j}] = \bar{\mu}_{\setminus j}, \qquad E_{q_b}[b_{\setminus j} b_{\setminus j}^T] = \bar{\Sigma}_{\setminus j,\setminus j} + \bar{\mu}_{\setminus j}\bar{\mu}_{\setminus j}^T . \tag{3.2.13}$$

We now established all necessary prerequisites to write the terms $E_{q_b}[\log p(c_{1,j}|b)]$ and $E_{q_b}[\log p(c_{k,j}|c_{k-1,j}, b)]$ in an explicit form, that is suitable for an optimization with respect to $\bar{\mu}$ and $\bar{\Sigma}$. For the first term we have:

$$\begin{aligned} &E_{q_b}[\log p(c_{1,j} = n|b)] \\ &\qquad = E_{q_b}[\log p(b_{1,j} = n|b_{\setminus j})] \\ &\qquad = C - \frac{1}{2(\Sigma_{j|\setminus j})_{1,1}} \left(n^2 - 2n E_{q_b}\left[(\mu_{j|\setminus j})_1\right] + E_{q_b}\left[\left((\mu_{j|\setminus j})_1\right)^2\right] \right). \end{aligned}$$

Recall the definition (3.2.11) of $(\mu_{j|\setminus j})_1$. Abbreviating the kth row of the matrix $\Sigma_{j|\setminus j} K_{j,\setminus j}$ with $(a_k^j)^T$ and moving terms independent of $\bar{\mu}$ and $\bar{\Sigma}$ to C, we obtain

$$= C - \frac{1}{2(\Sigma_{j|\setminus j})_{1,1}} \left(2(n - \mu_{1,j})(a_1^j)^T E_{q_b}[b_{\setminus j}] + (a_1^j)^T (E_{q_b}[b_{\setminus j} b_{\setminus j}^T] - 2\mu_{\setminus j} E_{q_b}[b_{\setminus j}]) a_1^j \right).$$

Finally, by replacing the expectations with the respective moments in (3.2.13) yields

$$\begin{aligned} &E_{q_b}[\log p(c_{1,j} = n|b)] = \\ &\qquad C - \frac{1}{2(\Sigma_{j|\setminus j})_{1,1}} \left(2(n - \mu_{1,j})(a_1^j)^T \bar{\mu}_{\setminus j} + (a_1^j)^T B a_1^j \right). \end{aligned} \tag{3.2.14}$$

with $B = \bar{\Sigma}_{\setminus j,\setminus j} + \bar{\mu}_{\setminus j}\bar{\mu}_{\setminus j}^T - 2\mu_{\setminus j}\bar{\mu}_{\setminus j}^T$.

The probability inside the second term

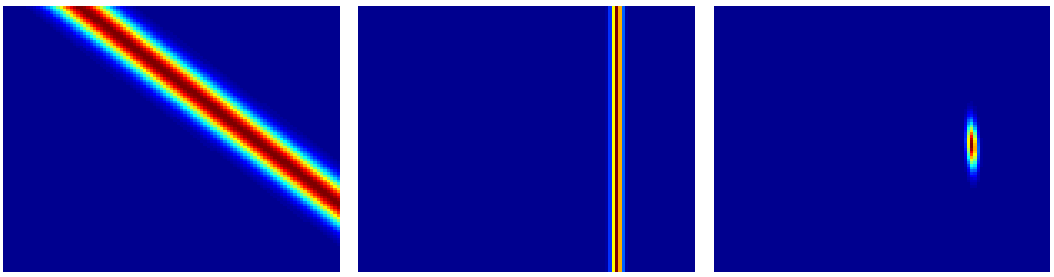
$$E_{q_b}\left[\log p(c_{k,j} = n|c_{k-1,j} = m, b)\right] = E_{q_b}[\log(p(b_{k,j} = n|b_{\setminus j}) p(b_{k,j} = n|b_{k-1,j} = m))],$$

is a product of two Gaussians and therefore again Gaussian, modulo normalization. Furthermore, the dependency on $b_{\setminus j}$ is the same. We can therefore, by using the formula for the product of two Gaussians, e.g. [RW06], show that

$$\begin{aligned} &E_{q_b}\left[\log p(c_{k,j} = n|c_{k-1,j} = m, b)\right] = \\ &\qquad \tilde{C} - \frac{1}{2(\Sigma_{j|\setminus j})_{k,j}} \left(2(n - \mu_{k,j})(a_k^j)^T \bar{\mu}_{\setminus j} + (a_k^j)^T B a_k^j \right), \end{aligned} \tag{3.2.15}$$

with B as above, indices 1 replaced by k and a different constant $\tilde{C}$. Note that the dependency on m, and thereby on $c_{k-1,j}$ is hidden inside $\tilde{C}$. We now made the expectation with respect to q_b explicit.

Expectation with respect to q_c. Similar arguments as for $p(c|b)$ hold for $p(c|y)$ too: We can split the sum over $c_{\bullet,j}$ into parts depending (at most) on two neigh-



(a) Local: $b_{k,j}|b_{k-1,j}$ (b) Global: $b_{k,j}|b_{\setminus j}$ (c) Both terms combined

Figure 3.5 - Illustration of a transition matrix $\Omega_{k,j}$ (c) and the local (a) and global (b) shape information it is composed of. The plots show the exponential version that is used during optimization the optimization of q_c (see Section 3.3.1), in order to illustrate the inherent sparsity that we utilize to speed up the calculation of q_c.

boring boundaries $c_{k-1,j}$ and $c_{k,j}$. See the previous section for details. We define matrices $\Omega_{k,j}$ and vectors $\omega_{1,j}$ as

$$\begin{aligned}(\Omega_{k,j})_{m,n} &= E_{q_b}[\log p(c_{k,j} = n | c_{k-1,j} = m, b)],\\ (\omega_{1,j})_n &= E_{q_b}[\log p(c_{1,j} = n | b)],\end{aligned}$$

for $k = 2, \ldots, N_b$, $j = 1, \ldots, M$ and $1 \leq m \leq n \leq N$. Finally, we can write the expectation of the second term in vectorized form as

$$-\sum_{j=1}^{M}\left((q_{c;1,j})^T\omega_{1,j} + \sum_{k=2}^{N_b}\langle q_{c;k\wedge k-1,j}, \Omega_{k,j}\rangle\right). \tag{3.2.16}$$

Figure 3.5 shows the two components (left and center panel) of a transition matrix $\Omega_{k,j}$ (right panel) for $m, n = 101, \ldots, 200$. For a better illustration of the inherent sparsity of $\Omega_{k,j}$, we applied the exponential function to each term. We see how $\Omega_{k,j}$ is composed by combining prior information about the relative distance between $b_{k,j}$ and $b_{k-1,j}$ (a) with the distribution of $b_{k,j}$ conditioned on information from all other columns via $\mathbb{E}_{q_b}[b_{\setminus j}] = \bar{\mu}_{\setminus j}$ (b).

3.2.4 Third Summand $\log p(b)$ of $J(q_b, q_c)$

Since the third term is independent of q_c the expectation with respect to q_c vanishes. Before we continue, note that

$$a^T B a = \sum_{i,j} a_i a_j B_{ij} = \operatorname{tr}(aa^T B) = \langle aa^T, B\rangle, \tag{3.2.17}$$

for symmetric $n \times n$ matrices B, since the trace can be seen as the inner product. Furthermore with $E_{q_b}[b] = \bar{\mu}$ and $E_{q_b}[bb^T] = \bar{\Sigma} + \bar{\mu}\bar{\mu}^T$ we can write for the expectation

of $p(b)$

$$\begin{aligned} -\int_b q_b(b)\log p(b)db &= C + \frac{1}{2}E_{q_b}[b^T K b - 2b^T K\mu + \mu^T K\mu], \\ &= C + \frac{1}{2}\langle K, \overline{\Sigma} + \bar{\mu}\bar{\mu}^T - 2\bar{\mu}\mu^T + \mu\mu^T\rangle. \end{aligned} \tag{3.2.18}$$

3.2.5 Entropy Terms of $J(q_b, q_c)$

Finally, we make explicit the entropies of q_b and q_c. For the normal distribution q_b we have that

$$-H[q_b] = \int_b q_b(b)\log q_b(b)db = C - \frac{1}{2}\log|\overline{\Sigma}|, \tag{3.2.19}$$

see for example [PP12, Eq. (389)]. For the negative entropy $-H[q_c]$ we can use the fact (c.f. (2.4.26)) that for tree-structered Markov random field, their entropy is given by the sum over singleton entropies and mutual information of unary and pairwise marginals. Adapting the notation in (2.4.26) to that used for q_c, we have

$$\begin{aligned} -H[q_c] &= -\sum_{j=1}^{M}\sum_{k=1}^{N_b} H_{k,j}[q_{c;k,j}] + \sum_{j=1}^{M}\sum_{k=2}^{N_b} I_{k\wedge k-1;j}[q_{c;k\wedge k-1;j}], \\ &= \sum_{j=1}^{M}\Bigg(\sum_{k=1}^{N_b}\sum_{c_{k,j}} q_c(c_{k,j})\log q_c(c_{k,j}) \\ &\qquad + \sum_{k=2}^{N_b}\sum_{c_{k-1,j}}\sum_{c_{k,j}} q_c(c_{k,j}, c_{k-1,j})\log\frac{q_c(c_{k,j}, c_{k-1,j})}{q_c(c_{k-1,j})q_c(c_{k,j})}\Bigg). \end{aligned}$$

3.2.6 Explicit Formulation of the Objective Function $J(q_b, q_c)$

In the previous sections we derived explicit formulations of all expectations in the functional $J(q_b, q_c)$ (3.2.4). We now combine all terms to obtain the reformulated minimization problem

$$\begin{aligned} \min_{q_c,\bar{\mu},\overline{\Sigma}} \; & -\sum_{j=1}^{M}\Big((q_{c;1,j})^T\theta_{1,j} + \sum_{k=2}^{N_b}\langle q_{c;k\wedge k-1,j}, \Theta_{k,j}\rangle + (q_{c;N_b,j})^T\theta_{N_b,j}\Big) \\ & + \frac{1}{2}\langle K, \overline{\Sigma} + \bar{\mu}\bar{\mu}^T - 2\bar{\mu}\mu^T\rangle - \frac{1}{2}\log\det\overline{\Sigma} - H[q_c] + C, \end{aligned} \tag{3.2.20}$$

subject to normalization (2.4.24) and marginalization constraints (2.4.25) for q_c. Note that we merged the terms of (3.2.9) and (3.2.16) into $\theta_{k,j}$ and $\Theta_{k,j}$, that is $\Theta_{k,j} = \Omega_{k,j} + \Psi_{k,j}$, $\theta_{1,j} = \omega_{1,j} + \psi_{1,j}$ and $\theta_{N_b,j} = \psi_{N_b,j}$. Furthermore, note that $\log\det\overline{\Sigma}$ automatically enforces $\overline{\Sigma} \succ 0$ (see Example 2.10).

We derived the above optimization problem by minimizing the Kullback-Leibler divergence between $p(b, c|y)$ and $q(b, c)$, see (3.2.3). Recall that we discussed this type of variational inference, called mean field method, in Section 2.4.2.5. There we also showed that one can view mean field methods from a second perspective, that

of maximizing a lower bound on $A(\theta)$:

$$\max_{\mu \in \mathcal{M}_F(\mathcal{G})} \{\langle \theta, \mu \rangle - A_F^*(\mu)\}. \tag{3.2.21}$$

We can rearrange (3.2.20) to bring it into the form (3.2.21). First note that

$$\mu = \left\{ q_c, \bar{\mu}, \overline{\Sigma} + \bar{\mu}\bar{\mu}^T \right\},$$

and

$$\theta = \left\{ \{\theta_{1,j}\}_{j=1}^{M}, \{\Theta_{k,j}\}_{k=1,j=1}^{N_b,M}, \{\theta_{N_b,j}\}_{j=1}^{M}, K\mu, K \right\}.$$

Confer Examples 2.49 and 2.50 for more information about canonical and mean parameters of MRFs and GMRFs. Finally, by observing that $A_F^*(\mu)$ denotes the negative entropies of q_c and q_b, we have shown the correspondence between (3.2.20) and (3.2.21).

3.3 Optimization

We alternatingly optimize the objective function (3.2.20) with respect to the sufficient statistics of q_c and q_b. The former corresponds to inference on trees and can be accomplished by the sum-product algorithm presented earlier. The optimization with respect to $\bar{\mu}$ and $\overline{\Sigma}$ is given in closed form. Both subproblems are convex, thus by alternatingly optimizing with respect to q_b and q_c, the functional $J(q_b, q_c)$, being bounded from below over the feasible set of variables, is guaranteed to converge to some minimum.

3.3.1 Optimization of q_c

Taking into account only terms in (3.2.20) that depend on parameters of q_c, we obtain an optimization problem that can be split into column-wise *convex* subproblems that are instances of the Bethe variational problem (2.4.27). As discussed in Section 2.4.2.4, this can be solved to global optimality with the sum-product algorithm.

3.3.2 Optimization of q_b

We now turn to the problem of optimizing (3.2.20) with respect to $\bar{\mu}$ and $\overline{\Sigma}$. Recall that in Section 3.2.3 we defined vectors $\omega_{1,j}$ (3.2.14) and matrices $\Omega_{k,j}$ (3.2.15), whose entries were all dependent on subsets of $\bar{\mu}$ and $\overline{\Sigma}$. We furthermore have terms originating from the entropy of q_b and the expectation of $p(b)$ with respect to q_b. Inspecting all these terms, we see that $\bar{\mu}$ and $\overline{\Sigma}$ are independent from each other, so we can optimize them separately.

3.3.2.1 Optimization With Respect to $\bar{\Sigma}$

We begin by stating the final result, namely that the optimization with respect to $\bar{\Sigma}$ is given by

$$\min_{\bar{\Sigma}} -\frac{1}{2}\log|\bar{\Sigma}| + \frac{1}{2}\langle K + \tilde{P}, \bar{\Sigma}\rangle, \tag{3.3.1}$$

which has the closed-form solution: $\bar{\Sigma} = (K + \tilde{P})^{-1}$. The newly introduced matrix $\tilde{P}$ contains the dependencies of terms $\omega_{1,j}$ and $\Omega_{k,j}$ on $\bar{\Sigma}$.

Derivation of $\tilde{P}$. Only considering terms in the nth entry of $(\omega_{1,j})$ that depend on $\bar{\Sigma}$, we obtain

$$(\omega_{1,j})_n(\bar{\Sigma}) = -\frac{1}{2(E_{j|\setminus j})_{1,1}}(a_1^j)^T\bar{\Sigma}_{\setminus j,\setminus j}a_1^j,$$

and accordingly for $(\Omega_{k,j})_{m,n}(\bar{\Sigma})$ with indices 1 replaced by k. We defined $(a_k^j)^T$ in Section 3.2.3 as the kth row of $\Sigma_{j|\setminus j}K_{j,\setminus j}$, hence as a column vector of length $N_bM - N_b$. We introduce the extended version $\tilde{a}_k^j$ of length N_bM, padded with zero entries such that

$$(\tilde{a}_k^j)^T\bar{\Sigma}\,\tilde{a}_k^j = (a_k^j)^T\bar{\Sigma}_{\setminus j,\setminus j}a_k^j. \tag{3.3.2}$$

Note that entries of $(\Omega_{k,j})(\bar{\Sigma})$ and $(\omega_{1,j})(\bar{\Sigma})$ are independent of m and n and therefore independent of q_c. Thus

$$(q_{c;1,j})^T\omega_{1,j}(\bar{\Sigma}) = 1\cdot\omega_{1,j}(\bar{\Sigma}), \qquad \langle q_{c;k\wedge k-1,j}, \Omega_{k,j}(\bar{\Sigma})\rangle = 1\cdot\Omega_{k,j}(\bar{\Sigma}).$$

Using $b^TBb = \langle bb^T, B\rangle$, we obtain for (3.2.16)

$$\begin{aligned}
&-\sum_{j=1}^{M}\Big((q_{c;1,j})^T\omega_{1,j}(\bar{\Sigma}) + \sum_{k=2}^{N_b}\langle q_{c;k\wedge k-1,j}, \Omega_{k,j}(\bar{\Sigma})\rangle\Big)\\
&\qquad = \frac{1}{2}\sum_{j=1}^{M}\sum_{k=1}^{N_b}\left\langle \frac{1}{(E_{j|\setminus j})_{k,k}}\tilde{a}_k^j(\tilde{a}_k^j)^T, \bar{\Sigma}\right\rangle\\
&\qquad = \frac{1}{2}\langle\tilde{P}, \bar{\Sigma}\rangle.
\end{aligned}$$

Since $\tilde{P}$ is independent of q_c and depends only on the sufficient statistics of $p(b)$, we do not have to update it while optimizing $J(q_b, q_c)$. Furthermore, since it is composed of linear combinations of submatrices of K, it can be expressed solely in terms of W and $\sigma^2 I$ (c.f. Section 2.2.5.3).

3.3.2.2 Optimization With Respect to $\bar{\mu}$

Again we first state the final result. Taking into account all terms in $J(q_b, q_c)$ that depend on $\bar{\mu}$, we get

$$\min_{\bar{\mu}} \frac{1}{2}\langle K + \tilde{P}, \bar{\mu}(\bar{\mu} - 2\mu)^T\rangle + \tilde{p}^T\bar{\mu}, \tag{3.3.3}$$

with the solution $\bar{\mu} = \mu - (K + \tilde{P})^{-1}\tilde{p}$. Again $\tilde{p}$ captures the dependencies of $\omega_{1,j}$ and $\Omega_{k,j}$, this time on $\bar{\mu}$, and is derived below. $\tilde{P}$ is the same as above. To minimize (3.3.3), we use conjugate gradient descent which enables us to calculate $\bar{\mu}$ using $(K + \tilde{P})$ instead of $(K + \tilde{P})^{-1}$.

Derivation of $\tilde{p}$. Only considering terms in $\omega_{1,j}$ depending on $\bar{\mu}$, we obtain

$$(\omega_{1,j})_n(\bar{\mu}) = -\frac{1}{2(E_{j|\setminus j})_{1,1}}(2(n - \mu_{1,j})(a_1^j)^T\bar{\mu}_{\setminus j} + (a_1^j)^T(\bar{\mu}_{\setminus j}\bar{\mu}_{\setminus j}^T - 2\mu_{\setminus j}\bar{\mu}_{\setminus j}^T)a_1^j)$$

and accordingly for $(\Omega_{k,j})_{m,n}(\bar{\mu})$. The first term is dependent on n and thus on q_c, whereas the remaining terms are again independent and q_c marginalizes out as above. Using again $(\tilde{a}_1^j)^T$ as the extended version of $(a_1^j)^T$ (see (3.3.2)) we obtain

$$\begin{aligned}
&-\sum_{j=1}^{M}\Big((q_{c;1,j})^T\omega_{1,j}(\bar{\mu}) + \sum_{k=2}^{N_b}\left\langle (q_{c;k\wedge k-1,j})^T, \Omega_{k,j}(\bar{\mu})\right\rangle\Big) \\
&\quad= \frac{1}{2}\sum_{j=1}^{M}\sum_{k=1}^{N_b}\frac{1}{(E_{j|\setminus j})_{k,k}}\left(2\left(E_{q_c}[c_{k,j}] - \mu_{k,j}\right)(\tilde{a}_k^j)^T\bar{\mu} + \left\langle \tilde{a}_k^j(\tilde{a}_k^j)^T, \bar{\mu}(\bar{\mu} - 2\mu)^T\right\rangle\right) \\
&\quad= \frac{1}{2}\sum_{j=1}^{M}\sum_{k=1}^{N_b}2\tilde{p}_{k,j}^T\bar{\mu} + \left\langle \tilde{P}_{k,j}, \bar{\mu}(\bar{\mu} - 2\mu)^T\right\rangle \\
&\quad= \tilde{p}^T\bar{\mu} + \frac{1}{2}\left\langle \tilde{P}, \bar{\mu}(\bar{\mu} - 2\mu)^T\right\rangle.
\end{aligned}$$

Since $\tilde{p}$ is dependent on q_c, its gets updated at every iteration.

3.3.3 Initialization

We start the optimization of $J(q_b, q_c)$ by initializing the distribution q_c and setting $E_{q_b}[p(b_{k,j}|b_{\setminus j})]$ to a uniformity, since we yet lack the distribution q_b. Afterwards, we initialize q_b via (3.3.1) and (3.3.3). Subsequently, we iterate both optimizations alternatingly until $J(q_b, q_c)$ converges.

4 Evaluation

4.1 Experiments

4.1.1 Data Acquisition

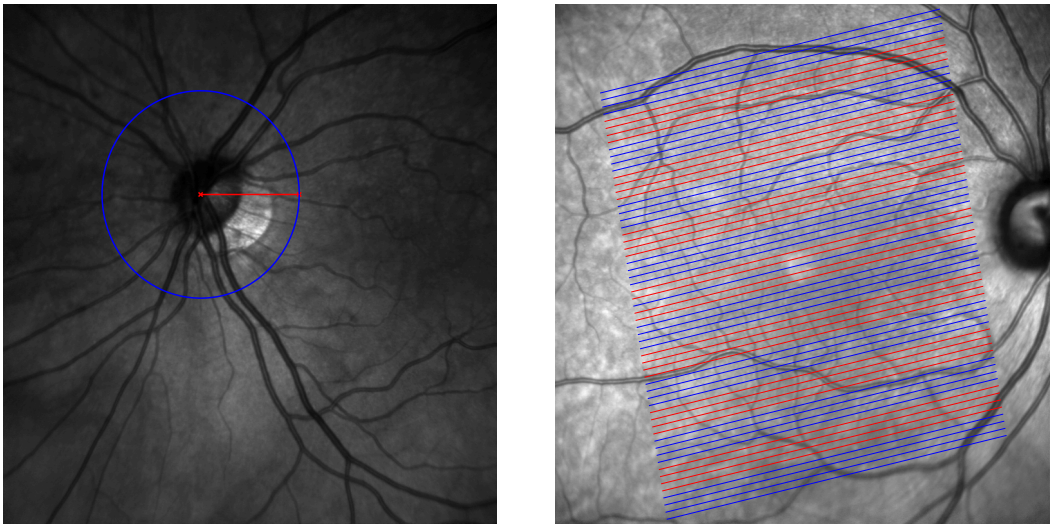
Circular B-scans *measured around the optic nerve head* were acquired from 80 healthy as well as from 66 glaucomatous subjects using a Spectralis HRA+OCT device (Heidelberg Engineering, Germany). Each scan had a diameter of 12°, corresponding to approximately $3.4\,mm$, and consisted of $M = 768$ A-scans of depth resolution $3.87\mu m$/pixel ($N = 496$ pixels), see Figure 4.1 (a). A medical expert provided ground truth for the boundary separating NFL and GCL, crucial for the diagnosis of glaucoma, as well as a grading for the pathological scans: *pre-perimetric* glaucoma (PPG), meaning the eye is exhibiting structural symptoms of the disease but the visual field and sight are not impaired yet, as well as *early*, *moderate* and *advanced* primary open-angle glaucoma (PGE, PGM and PGA). Ground truth for the remaining eight boundaries was produced by the author of this thesis. To measure interobserver variability, a second set of labels for the healthy B-scans was obtained.

The second data set consisted of *fovea-centered* 3-D volumes, acquired from 35 healthy subjects using the same device as above. Each volume was composed of 61 B-Scans of dimension 500×496, covering an area of approximately $5.7 \times 7.3\,mm$. Ground truth was obtained as follows: Each volume was divided into 17 regions, and a B-scan randomly drawn from each region was labeled with the previously introduced nine boundaries, see Figure 1.2. Figure 4.1 (b) depicts the location of all 61 B-Scans and their partition into regions indicated by color.

4.1.2 Various Configurations of Appearance Terms

In Equation (3.1.5) we introduced switches β^l and β^t to enable or disable layer and boundary appearance terms, respectively. In the discussion following (3.2.3) we described discriminative appearance models as an alternative to generative ones and pointed out that they yielded better performance. In this section we will report how different combinations of appearance terms performed.

Using the set of healthy circular scans, we tested the model with generative layer as well as boundary terms, i.e. $\beta^t = \beta^l = 1$. This configuration turned out to be sensitive to texture distortions caused for example by blood vessels. This resulted in initializations above the retinal layers, since the model misinterpreted the dark area as part of the choroid, as shown in Figure 4.2 (a). We then disabled the layer appearance terms, i.e. set $\beta^l = 0$. This solved the previous issue, but spuriously led to some columns being initialized below the retina, due to very high probabilities for



(a) 2-D circular scan (b) 3-D volume

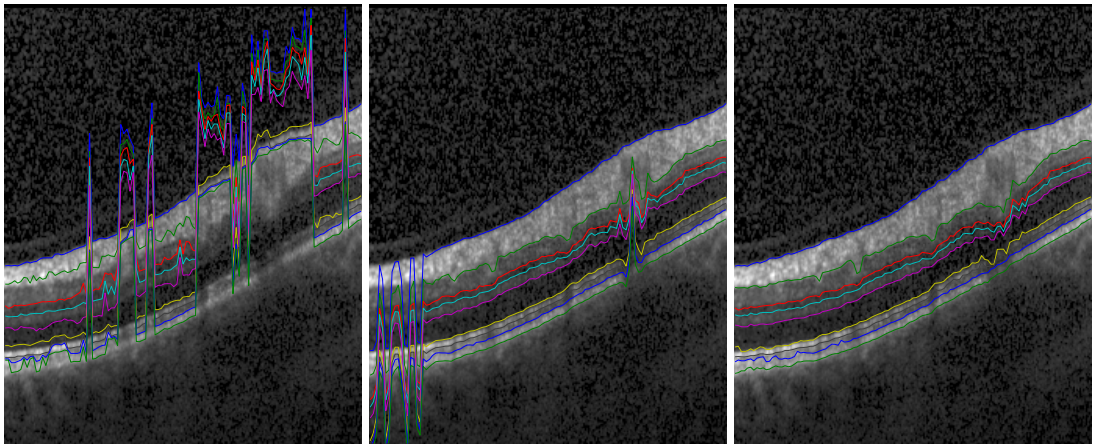
Figure 4.1 - SLO fundus images that exemplarily depict (a) the trajectory and radius of a 2-D circular scan centered around the *optic nerve head* and (b) the area covered by a 3-D volume consisting of 61 B-Scans centered at the *foveola*, the light spot in the center, c.f. (I) in Figure 2.9 (a). Alternating coloring illustrates the partitioning into 17 different regions, from which one scan respectively was randomly selected for manual labeling. Confer Figures 2.8 and 2.9 for a physiological classification of the depicted regions.

some boundary classes caused by relatively small class model variances, i.e. narrow and steep normal distributions. For patches close to the mean, the probabilities for those classes happened to be up to 100 times larger than for other classes. This caused false positive class responses in the choroid to displace the entire initialization for some columns, as displayed in Figure 4.2 (b).

Switching to discriminative probabilities solved this issue as well, since the local normalization limits all probabilities to 1 and gives each appearance class the same influence. Thus false-positives did not possess the probability mass any more to displace the segmentation of a complete column, see Figure 4.2 (c). Notice that the layer terms, although switched off by setting $\beta^l = 0$, are utilized indirectly, since they contribute to the normalization of terms $p(x_{i,j}(c)|y_{i,j})$, see (3.1.6). Thus large layer appearance terms disfavor layer boundaries and thus can rule out certain parts of the OCT scan for segmentation.

4.1.3 Model Parameters

Table 4.1 summarizes the model parameters and the values they were set to. For the appearance models we set α_{glasso} to 0.01, which resulted in sparse covariance matrices $\Sigma_{x_{i,j}}$ that speed up computations significantly. A patch-size of 15×15 and the projection onto the first $q_{\text{pca}} = 20$ eigenvectors resulted in smooth segmentation boundaries. Similar, we used $q_{\text{ppca}} = 20$ eigenvectors to build the shape prior model, after examining the eigenvalue spectrum of the empirical covariance matrix S.



(a) $\beta^l, \beta^t = 1$, $p(y|c)$ (b) $\beta^l = 0$, $p(y|c)$ (c) $\beta^l = 0$, $p(c|y)$

Figure 4.2 - (a)-(c) Close-up view of initialization results for different configurations of appearance terms. Switches β^l and β^t include or exclude layer and transition appearance terms.

Table 4.1 - Set of model parameter values used throughout all experiments.

	Appearance			Shape	Inference
Parameters	α_{glasso}	q_{pca}	Patch-Size	q_{ppca}	Variance of $p(b_{k,j}\|b_{\setminus j})$
Value	0.01	20	15×15	20	10

An important parameter during the inference is the variance of $p(b_{k,j}|b_{\setminus j})$, which balances the influence of appearance and shape. Artificially increasing this parameter results in broader normal distributions (that is wider stripes in Figure 3.5 (b)), which allows q_c to take into account more observations around the mean of $p(b_{k,j}|b_{\setminus j})$ during the discrete inference part. At the same time the influence of the appearance terms on q_b is reduced, which results in smoother estimates for the mean $\bar{\mu}$ of q_b. Thus increasing the variance loosens the coupling between q_c and q_b and vice versa. A ten-fold increased variance turned out to provide a good balance between local appearance terms and shape regularization as well as between run-time and prediction accuracy.

We used the very same set of parameter values for all our experiments and performed no fine tuning separately for each data set. Hence it is plausible to assume that these values perform well on a broad range of data sets.

4.1.4 Error Measures and Test Framework

For each boundary as well as the entire scan we computed the unsigned distance in μm ($1px = 3.87\mu m$) between estimates $\hat{c}_{k,j} = \mathbb{E}_{q_c}[c_{k,j}]$ and manual segmentations $\tilde{c}_{k,j}$, that is

$$E^k_{\text{unsgn}} = \frac{1}{M}\sum_{j=1}^{M} |\hat{c}_{k,j} - \tilde{c}_{k,j}|, \qquad E_{\text{unsgn}} = \frac{1}{N_b}\sum_{k=1}^{N_b} E^k_{\text{unsgn}}\,.$$

For volumes we additionally averaged over all 17 scans in the volume. For each data set we provide the unsigned error averaged over all scans and it's standard deviation (SD).

Results were obtained via cross-validation: After splitting each data set into a number of subsets, each subset in turn is used as a test set, while the remaining subsets are used for training. This provided an estimate of the ability to segment new (unseen) test scans. We used 10-fold cross-validation for the set of non-pathological circular scans and leave-one-out cross-validation for the volumes data set, to maximize the number of training examples in each split. For the set of glaucomatous scans, we used no cross-validation but used a single model trained on all healthy scans as predictor.

4.1.5 Implementation and Running Time

We implemented our approach in MATLAB. The main bottle-neck, the sum-product algorithm used to find an optimal solution for $q_c(c)$, was implemented in C and incorporated into MATLAB via the Mex-interface. To further decrease running ti.e. we exploi.e. the inherent sparsity of the transition matri.e. $\Omega_{k,j}$, as illustrated in Figure 4.1. Also, wherever possible we transferred expensive matrix-vector multiplications to the GPU, using a wrapper for MATLAB called GPUmat [MMG08]. Segmenting all 61 B-Scans of a 3-D volume took around 60 s, with memory requirements of about 2 GB, measured on a Core i7-2600K 3.40GHz.

4.2 Results

4.2.1 Circular Scans

Average boundary-wise results are summarized in Table 4.2. In general, boundaries 1 and 6 to 9 turned out to be easier to segment than boundaries 2 to 5. For boundary 1 this stems from easily detectable textures, whereas boundaries 6-9 with their regular shape profit disproportionately from regularization by the shape prior. Boundaries 2-5 on the other hand pose a harder challenge with their high variability of texture and shape. The upper row in Figure 4.3 shows an example close to the average segmentation performance.

For the pathological scans segmentation performance was comparable to that of the normal scans, but naturally decreased with the progression of the disease. This may have had several causes: Since glaucoma is known to cause a thinning of the nerve fiber layer (NFL) [SHP+95, BZB+01], the shape prior trained on healthy scans may encounter difficulties adapting to very abnormal glaucomatous shapes. Furthermore, we observed a reduced scan quality for glaucomatous scans, also reported by others [ISW+05, SIH+06, MHMT10], which in turn reduced the meaningfulness of the appearance terms. For advanced primary open-angle glaucoma, the NFL can even vanish at some locations. The appearance model for this layer, trained on healthy data, is not able to detect these extreme anomalies, which resulted in a comparatively

Table 4.2 - Results in $\mu m \pm$SD ($3.87 \mu m \mathrel{\widehat{=}} 1 px$) for 2-D circular scans (separately for healthy eyes as well as the different degrees of glaucoma, pre-perimetric, early, moderate and advanced) and 3-D scans of healthy subjects. Numbers within brackets denote the respective data set size.

	2-D Healthy	2-D Glaucoma				3-D Healthy
k	*All (80)*	*PPG (22)*	*PGE (22)*	*PGM (13)*	*PGA (9)*	*All (35)*
1	2.06±0.57	2.60±0.85	3.76±1.42	4.51±1.18	6.53±2.76	1.36±0.18
2	4.68±1.13	6.66±2.41	5.65±1.66	6.74±1.64	9.95±4.74	3.32±0.37
3	3.67±0.84	4.57±1.18	5.37±1.33	5.49±1.00	8.80±3.03	3.17±0.44
4	3.31±0.78	4.43±1.09	5.78±1.48	5.44±1.19	8.30±2.21	3.23±0.56
5	3.30±0.75	4.34±1.63	4.40±1.14	4.15±0.68	5.05±0.92	3.27±0.66
6	2.10±0.76	2.67±1.37	2.76±0.97	2.88±1.62	2.99±1.92	1.61±0.23
7	2.34±1.05	2.59±1.11	2.95±1.27	2.21±0.68	2.42±0.44	1.86±0.32
8	2.81±1.42	2.82±1.00	3.40±1.22	2.94±1.40	4.19±1.97	2.27±0.40
9	2.01±1.14	2.06±0.65	1.63±0.48	1.64±0.25	2.36±1.18	2.07±0.48
∅	**2.92±0.53**	**3.64±0.68**	**3.97±0.73**	**4.00±0.53**	**5.62±1.25**	**2.46±0.22**

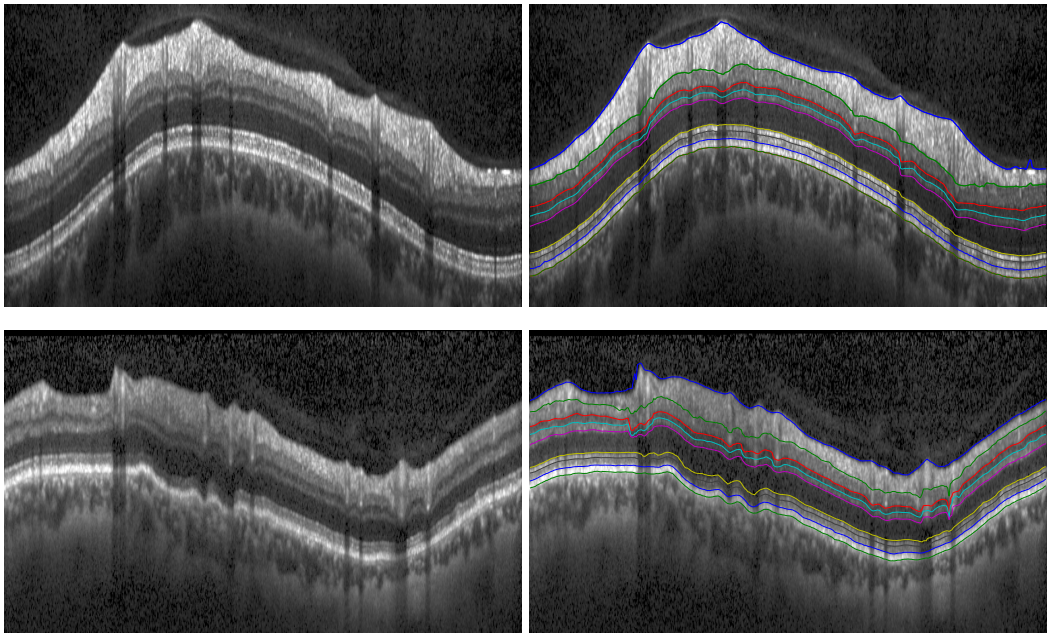


Figure 4.3 - Top: Segmentation ($E_{\text{unsgn}} = 2.97\,\mu m$) of a non-pathological circular scan. Bottom: Segmentation ($E_{\text{unsgn}} = 5.09\,\mu m$) of an advanced glaucomatous scan.

Table 4.3 - Interobserver variability as well as prediction performance for the set of 80 healthy circular scans, for which we obtained a second set of labels. The algorithm was trained on the average ground truth. Errors in $\mu m \pm$SD ($3.87\mu m \mathrel{\widehat{=}} 1px$). Interestingly the performance improves over to the case of only one set of labels, compare column four with column one of Table 4.2.

	Obs.1 vs. Obs.2	Algo. vs. Obs.1	Algo. vs. Obs.2	Algo. vs. Avg. Obs.
1	2.86±0.46	2.35±0.62	3.69±0.76	2.74±0.66
2	7.57±1.06	5.51±1.30	6.15±1.35	4.56±1.00
3	4.62±1.13	3.74±0.91	4.26±0.85	3.25±0.74
4	3.63±0.65	3.31±0.75	3.35±0.74	2.74±0.73
5	3.39±0.66	3.31±0.75	3.36±0.75	2.83±0.70
6	1.87±0.59	2.09±0.73	2.05±0.73	1.82±0.71
7	2.36±1.14	2.33±0.99	2.55±1.03	2.08±0.92
8	3.54±1.78	3.23±1.44	2.51±1.33	2.21±1.15
9	1.37±0.51	1.94±1.03	2.17±1.02	1.91±1.01
⌀	**3.47±0.37**	**3.09±0.50**	**3.34±0.52**	**2.68±0.50**

bad performance for some scans. We discuss possible modifications to overcome this problem in Section 4.3.

The bottom panels in Figure 4.3 show an example of a PGA-type scan and its segmentation. The scan exhibits the discussed reduced scan quality. Furthermore, the segmentation proves that the shape model can generalize well to pathological shapes as well as scan artifacts.

4.2.1.1 Interobserver Variability

A second set of labels was created for the healthy circular B-scan data set by a colleague. For training and testing we utilized the same set-up as described earlier (10-fold cross-validation, parameters as reported in Table 4.1), but used the mean of both labelings for training. In Table 4.3 we compare the predicted segmentations separately with each of the two labelings and with the mean labeling. Furthermore, we report the average absolute distance between both observers, the *interobserver variability*.

We see, that the resulting prediction errors are well within the range of the interobserver variability. The performance for the mean labels even improves compared to set-up with only one set of labels, c.f. the first column of Table 4.2. This suggests an increased robustness of the averaged ground truth towards scan artifacts, ambiguous image regions and labeling bias.

4.2.1.2 Qualitative Evaluation

A key property of our model is the *inference of full probability distributions over segmentations* q_c and q_b, instead of only modes thereof. This allowed us to rate the

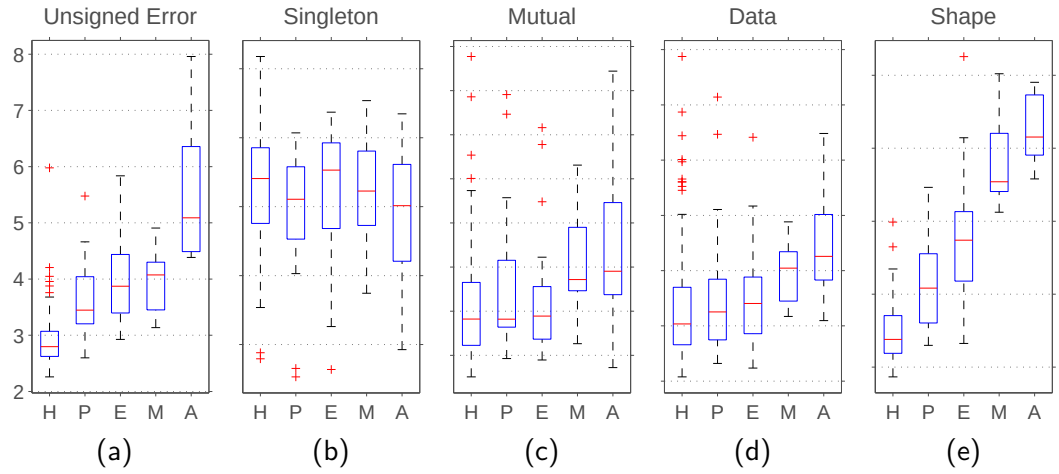


Figure 4.4 - Different terms (b-e) of the objective function $J(q_b, q_c)$ and the unsigned error (a) for healthy (**H**) as well as glaucomatous scans (**P**PG, PG**E**, PG**M** and PG**A**). While "Shape" is very discriminative for glaucomatous scans, "Mutual" and "Data" correlate well with the unsigned error.

quality of the prediction as a whole as well as point out areas with low certainty, and classify a scan as normal or potentially pathological. To this end, we evaluated the different terms of the objective function $J(q_b, q_c)$. Figure 4.4 shows boxblots of four terms (b-e) and compares them to the unsigned error (a). Singleton entropy (b) and mutual information (c) are the two summands of the negative entropy of q_c, see (3.2.20). The data (d) and shape (e) terms represent the first two summands of (3.2.4), introduced in Section 3.2.2 and 3.2.3.

The shape term, which measures how much the data-driven distribution q_c differs from the shape-driven expectation $E_{q_b}[\log p(c|b)]$, is highly discriminative between healthy and pathological scans. The mutual information on the other hand exhibit a good correlation with the unsigned error. It measures the dependence between neighboring random variables $c_{k,j}$ and $c_{k-1,j}$. Imaging $c_{k,j}$ and $c_{k-1,j}$ each having a single strong peak in $q_{c;k,j}$ and $q_{c;k-1,j}$. Their joint probability $q_{c;k\wedge k-1,j}$ will reveal almost no dependency. On the other hand, if we have several possibilities for each variable caused for example by poor data terms, then their dependency increases and thereby the mutual information. We will use these two terms in the forthcoming evaluation.

Classification. A state-of-the-art method for the clinical diagnosis of glaucoma is based on NFL thickness, averaged for example over the whole scan or one of its four quadrants [BZB+01, LCC+05, CKFB09, LRZ+11]. Estimates of the NFL thickness for all circular scans were obtained using the software of the Spectralis OCT device, version 5.6. We compared this established method against the prediction of the shape term discussed above. Using the same setup as in [BZB+01], we report sensitivities for specificities of 70% and 90%, as well as the area under the curve

Table 4.4 - Sensitivities for the detection of glaucoma. We compare NFL-based features that measure average thickness in different parts of the scan (indicated by their name), and our global shape based feature. Bold numbers indicate the highest detection rate for the respective specificity and glaucoma class.

Specificity	**70%**			**90%**			**AUC**		
Type	*PPG*	*PGE*	*PGM*	*PPG*	*PGE*	*PGM*	*PPG*	*PGE*	*PGM*
Average	68.2	90.9	**100.0**	36.4	86.4	**100.0**	0.72	0.93	**1.00**
Superior	63.6	81.8	92.3	45.5	77.3	76.9	0.78	0.84	0.90
Inferior	45.5	72.7	92.3	13.6	31.8	53.8	0.69	0.77	0.89
Temporal	63.6	**95.5**	**100.0**	54.5	90.9	**100.0**	0.74	**0.95**	0.99
Nasal	36.4	63.6	92.3	18.2	45.5	61.5	0.51	0.74	0.89
Shape	**77.3**	**95.5**	**100.0**	**63.6**	**95.5**	**100.0**	**0.84**	**0.95**	**1.00**

(AUC) of the receiver operating characteristic (ROC)[1], see Table 4.4. In all cases, our shape-based discriminator performs at least as good as the best thickness-based one. Especially for pre-perimetric scans, which feature only subtle structural changes, our approach improves diagnostic accuracies significantly. For this most interesting group, Figure 4.5 (a) provides ROC curves of the two *overall* best performing NFL measures and our shape-based measure.

Global Quality. We obtained a *global* quality measure, by combining the mutual information and the shape term. Given the values for all scans, we re-weighted both terms into the ranges $[0, 1]$ and took their sum. Thereby we could establish a quality index that had a very good correlation of 0.82 with the unsigned segmentation error. See Figure 4.5 (b) for a plot of all quality index/error pairs and a linear fit thereof. The estimate of this fit and the true segmentation error differs on average by only $0.51\,\mu m$. This shows that the model is able to additionally deliver the quality of its segmentation.

Local Quality. Finally, we determined a way to distinguish *locally* between regions of high and low model confidence. This could for example point out regions where a manual (or potentially automatic) correction is necessary. To this end we examined the local correlation (that is on a column-wise level) of the mutual information terms with the unsigned error. We calculated its mean for instances with segmentation errors smaller than 0.5 and bigger than 2 pixels. This yielded three ranges of confidence in the quality of the segmentation. For each image we fine-tuned these ranges by dividing by $\max(\text{Quality Index}(\text{CurrentImage}), 1)$.

Figure 4.6 (a) shows a PGA-type scan with annotated segmentation, whose error is $6.83\,\mu m$. The advanced thinning of the NFL and the partly blurred appearance caused the segmentation to fail in some parts of the scan. Close-ups (b) and (c) show that the model correctly identified those erroneously segmented regions. The average

[1] The AUC can be interpreted as the probability, that a random pathological scan gets assigned a higher score than a random healthy scan.

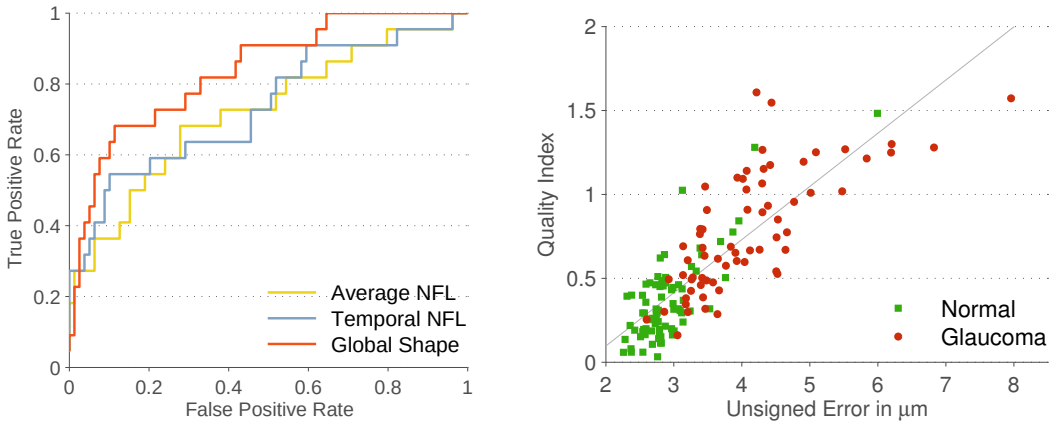


Figure 4.5 - (left) ROC curves for the two overall best performing NFL-based classifiers and our shape prior based approach for the least advanced and therefore hardest to detect glaucoma class. (right) High correlation of our quality index, obtained by combining terms (c) and (e) from Figure 4.4, with the actual unsigned error. On average, the estimated error (linear fit) differed by only $0.51\,\mu m$ from the true segmentation error.

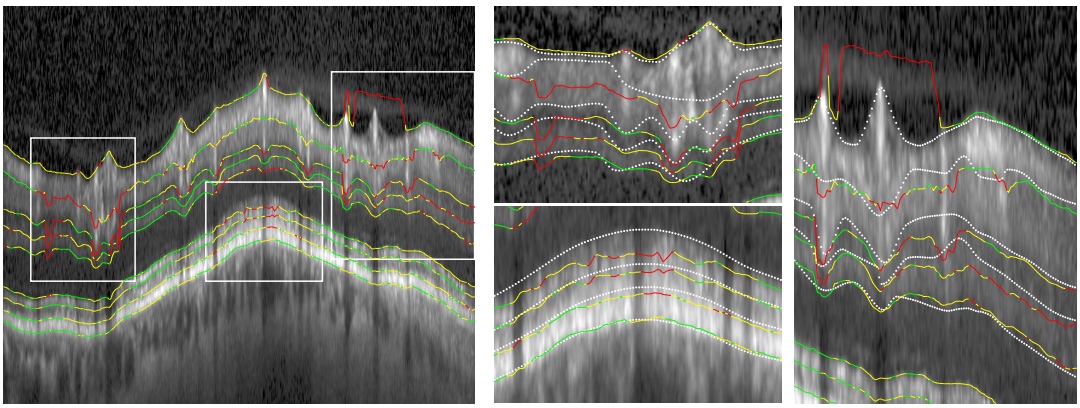


Figure 4.6 - (a) An advanced primary open-angle glaucoma scan and the segmentation thereof ($E_{\text{unsgn}} = 6.81\,\mu m$), augmented by the local quality estimates of the model, with red representing highest uncertainty. (b and c) Close-ups of the three areas, the model is (correctly) most uncertain about. White dotted lines represent ground truth.

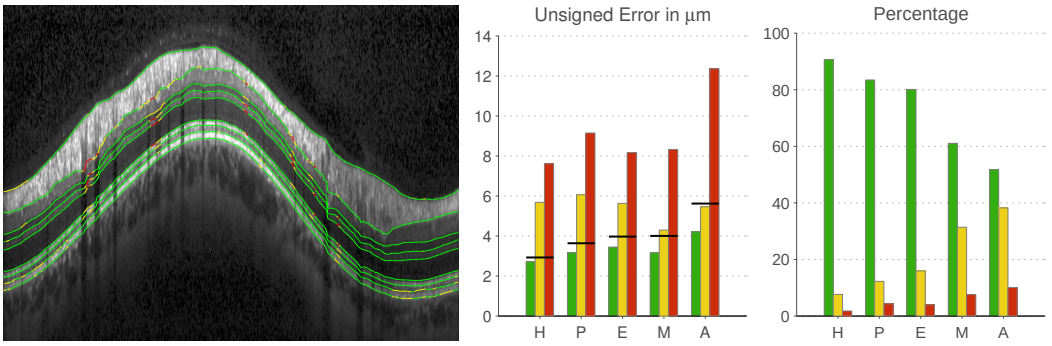


Figure 4.7 - (a) A healthy scan ($E_{\text{unsgn}} = 3.05\,\mu m$) accompanied by a positive quality assessment. (b) Average segmentation errors broken down for each type of grading and healthy as well as pathological scans. Black lines donate the average segmentation errors as reported in Table 4.2. (c) Average amount of assignment to each class for healthy and pathological scans.

errors of the three categories are 4.67, 5.43 and 18.36 μm respectively. Figure 4.7 (a), on the other hand, shows a scan from a healthy eye with a segmentation error of 2.83 μm, that is accompanied by a throughout positive quality rating.

We examined the accuracy of the local quality index numerically for all scans. Figure 4.7 (b) reports the average unsigned error for normal (H) and glaucomatous scans (P, E, M and A) as well as all three grades of certainty, and compares it to the average segmentation error for each data set, given as black lines. As for the global quality index, also locally the model is able to distinguish between correct and erroneous regions. Figure 4.7 (c) shows the average percentage of each segmentation that was assigned to one of the quality ratings. While normal scans are mostly labeled as OK, this changes gradually as the disease progresses.

4.2.2 Volumetric Scans

In contrast to 2-D scans, the labeling of OCT volumes is very time consuming, hence our data set only consisted of 35 samples. Thus we were left with less data points to train a shape model of much higher dimension. Consequently, we observed a reduced ability of $p(b)$ respectively $q_b(b)$ to generalize well to unseen scans. We tackled this problem by reducing the dimensionality of $p(b)$ and by interpolating it for intermediate columns, which fixed the problem only to some extent.

We further pursued this i.e. and suppressed the connectivity between different B-scans inside the volume, which corresponds to a block-diagonal covariance matrix Σ, where each block is obtained separately using PPCA. This significantly reduced the amount of parameters that had to be determined, and improved accuracy significantly. The last column in Table 4.2 reports results for all boundaries.

The average segmentation error of 2.46 μm is significantly smaller than for circular scans, as well as its standard deviation of 0.22 μm. Reasons are smoother boundary shapes and less severe texture artifacts caused by blood vessels. Representative for

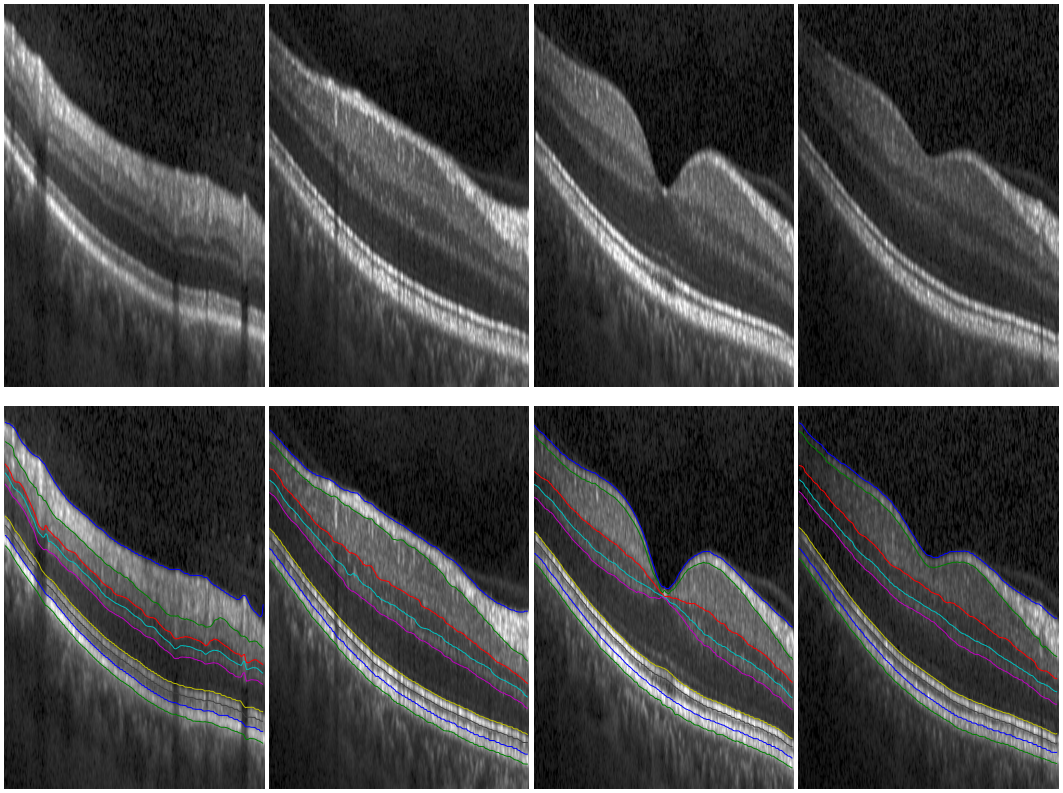


Figure 4.8 - Four segmented B-Scans from regions 2, 6, 9 and 11 of the same volume ($E_{\text{unsgn}} = 2.53\,\mu m$).

the average segmentation performance, Figure 4.8 shows B-scans of the same volume from four different regions, with an error of 2.53 μm averaged over all scans in the volume.

4.2.2.1 Noise Robustness

Speckle noise is a known issue when dealing with OCT scans. It is caused by random interference between reflected laser beams that are mutually coherent [SXY99]. In this section we investigate the robustness of our segmentation approach to speckle noise. To this end we artificially added speckle noise to our OCT scans, using the Matlab function `imnoise(A,'speckle',var)`, which modifies the image A via

$$A_{\text{new}} = A + \epsilon A,$$

where ϵ is uniformly distributed with mean 0 and variance `var`. Note that this was done for test scans only, whereas the model was still trained on the unmodified data. We examined parameter values `var` $= \{0, 0.1, 0.2, \ldots, 1\}$. Figure 4.9 shows close-ups of the fovea-region with (a) no added speckle noise and (b)-(d) with added speckle noise of variances 0.2, 0.6 and 1. Again the examples where picked to be representative for the average segmentation error.

It turned out that for very high levels of speckle noise, the column-wise *initialization* became more and more error-prone, with single columns being way off. This caused the subsequent iterative inference to produce very inaccurate configurations. We therefore added a simple routine, that detected those outliers, and in case of detection solved a MRF for boundary one spanning all columns. The resulting marginal densities were then used to clamp the appearance terms of the first boundary during the initialization step. This worked well in almost all cases except for some cases where boundary 6 was mistaken for boundary 1. But since this happened in only 0.5% of all cases, most of them in the same volume, we did not persuaded this any further, but dropped the corresponding results (since the mean is very sensitive to outliers).

Figure 4.10 illustrates how the segmentation error evolves with increasing speckle noise. Error bars indicate standard deviation. We added results reported by [DCA+13], as their setup is almost identical to ours (same noise parameters, very similar data obtained from the same OCT device), and their model constitutes a state-of-the-art retina segmentation approach based on graph-cuts with *local* probabilistic shape terms. We see that the segmentation performance remains stable much longer for our model, while [DCA+13] report a significant drop in performance already for the lowest amount of speckle noise. This underlines the usefulness of a *global* shape prior, which helps to deal with situations of poor data terms

4.3 Discussion

A novel probabilistic approach for the segmentation of retina layers in OCT scans was presented. It incorporates *global* shape information, which distinguishes it from

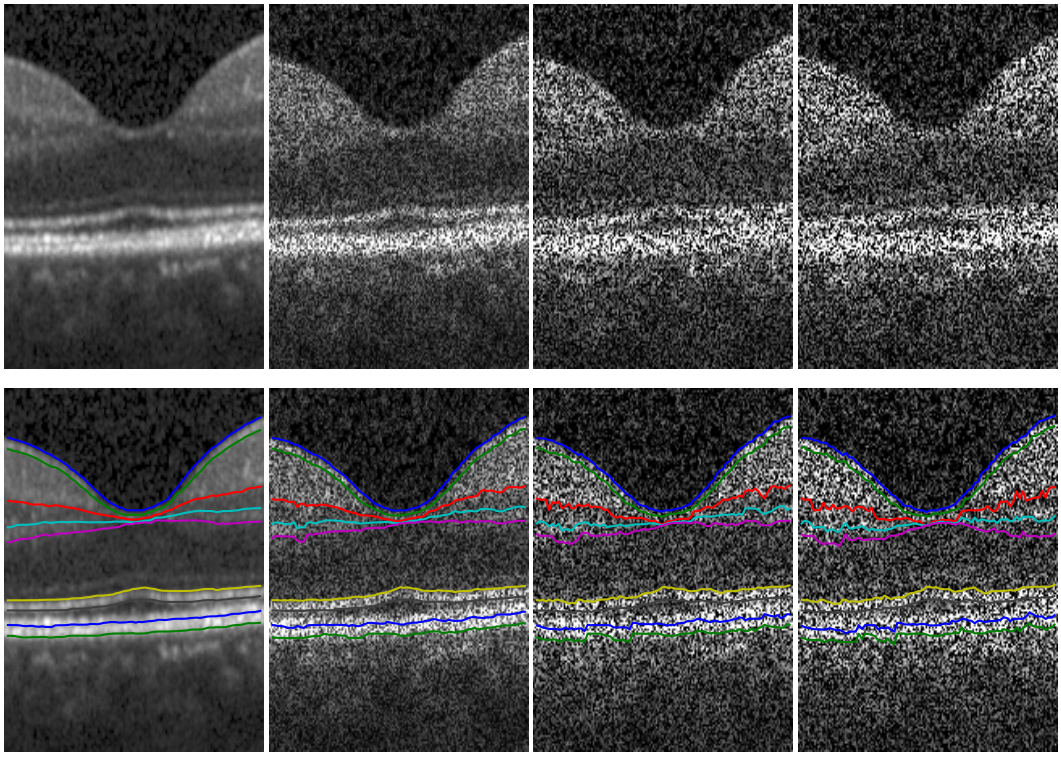


Figure 4.9 - Close-ups of the fovea region showing the unmodified scan in (a), and with added speckle noise for noise parameter set to 0.2 (b), 0.6 (c) and 1.0 (d). The chosen example is representative for the mean error reported in Figure 4.10.

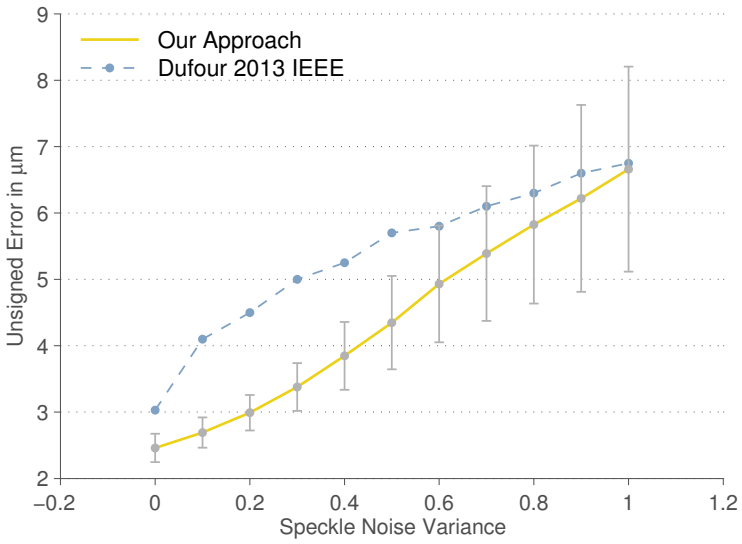


Figure 4.10 - Response of our segmentation model when confronted with increase amounts of artificially added speckle noise. State-of-the-art results of [DCA+13] are given for comparison, as they use an almost identical setup but their model only encompasses local shape information. Error bars show standard deviation.

other existing approaches, that rely solely on local shape information or at best on sparse global information. To obtain a deterministic approximate of the *full* posterior distribution $p(c, b|y)$, we employed variational methods, which entail efficiently solvable optimization problems. We demonstrated the applicability of our approach for a variety of different OCT scans as well as showed the benefit of inferring full probability distributions over segmentations rather than point estimates thereof.

3-D Segmentation Performance. Especially for 3-D OCT volumes, our segmentation performance was significantly better than results recently reported from approaches that use no shape information [VvdSLdB11, YRW+10], local hard-constrained shape information [GAW+09], local probabilistic shape information [DCA+13, SBG+13] or sparse global shape information [KPH+10]. Taking into account, for better comparability, only publications that used data sets obtained from the same OCT device used in this publication, the following trend evolves:

- While *no shape information* lead to only mediocre results: $6.20\,\mu m$ and $5.28\,\mu m$ for healthy and moderate glaucomatous data respectively [VvdSLdB11],
- adding *local shape information via hard constraints* yielded improved segmentation performance: $3.54 \pm 0.56\,\mu m$ as evaluated by [DCA+13] but comparable to the model proposed by [GAW+09].
- Additionally using *probabilistic local constraints*, [DCA+13] recently could again boost performance to $3.03 \pm 0.54\,\mu m$.
- Finally, by adding *global shape information*, we could in turn increase the segmentation performance to $2.46 \pm 0.22\,\mu m$.

We also showed the increased robustness of our model when faced with speckle noise compared to the model of [DCA+13]. Although this clearly seems to support the use of global shape information for regularization, keep in mind that a concluding comparison can only be carried out using the same data set. Nevertheless, we believe that these results highlight the usefulness of global shape regularization for the segmentation of retinal layers in OCT images.

Reported time requirements vary greatly, and our running time of 60 s is slower than the 18 s and 15 s reported by [DCA+13] and [YRW+10], but faster than the remaining approaches cited above, ranging from 900 s [VvdSLdB11] to over two hours [SBG+13]. Since the bottleneck of our approach, optimizing q_c, is done separately for each image column j, further speed-ups could be achieved by parallelizing the existing C implementation or transferring the calculation of q_c to the GPU.

2-D Segmentation Performance. We also evaluated the performance for healthy and pathological 2-D circular scans and, in both cases, obtained good results. The only exception was the group of advanced glaucomatous scans, which was mainly caused by the appearance models. Patches with near-zero layer thickness are too far away from the mean to be detected from the Gaussian distribution trained on healthy data.

Apart from including patches from glaucomatous scans into the training set, a useful extension could be to define a mixture of Gaussians for each appearance class, adding patches centered below or above pixel (i, j), which model its surrounding but not the layer/boundary itself. Also fixing the issues reported in Section 4.1.2 and utilize all appearance terms, those of boundaries *and* layers, should help in that case. Finally, given more pathological examples especially for PGM and PGA, one could learn a pathological shape prior and let the model choose the more probable shape prior based on the initialization.

Pathology detection. We investigated different ways to utilize the inferred distributions q_c and q_b. Experiments showed, that the model is quite sensitive to abnormal shapes and thus can act as a detector of glaucoma, with a higher sensitivity than established methods solely based on NFL thickness. This could relate to recent findings, that glaucoma causes a thinning of *all* inner retinal layers: NFL, GCL, IPL and (to a lesser extent) INL [TLL+08]. To confirm these promising results, further studies with more patients enrolled will be needed.

Quality Assessment. Another benefit of our approach is the ability to assess the quality of the segmentation, altogether for the whole scan or for each boundary position separately. In the context of screening large patient databases, the former could be a valuable tool to minimize the effort of a physician in reassessing the results. The latter could facilitate a automatic or manual post-processing, targeted specifically at regions with a high error probability. A thorough investigation of these regions could reveal a suitable approach.

To facilitate further research in the area of OCT segmentation and related areas, we published our source code together with a documentation on our project page: `http://graphmod.iwr.uni-heidelberg.de/Project-Details.132.0.html`.

5 Shape Prior Obeying Ordering-Constraints

5.1 Introduction

5.1.1 Overview, Motivation

In many real-world scenarios we are given a random vector $X \in \mathbb{R}^d$, distributed according to some underlying density function p, whose components satisfy the *ordering constraint*:

$$X_1 \leq X_2 \leq \ldots \leq X_d. \tag{5.1.1}$$

This set of linear constraints defines the cone K, given by

$$K = \left\{ x \in \mathbb{R}^d \colon x_1 \leq \cdots \leq x_d \right\}. \tag{5.1.2}$$

Many instances arise in physiology, where boundaries of cell tissue naturally satisfy such an ordering constraint. Examples from the literature are the segmentation of the left heart ventricle [EKKN13, DHDP13], cross-section images of the artery [TKK+11, UAO+12] or the segmentation of retinal layers in OCT images [DCA+13, RSS14]. Figure 1.1 illustrates the arterial wall segmentation and the non-medical application of tree ring segmentation.

In the application presented in this thesis, the ordering constraint arises during the column-wise discrete inference of boundary positions, see Section 3.1.3. There we require for all image columns j, that for pairs of neighboring boundaries k and $k-1$ the following holds:

$$p(c_{k,j} = n, c_{k-1,j} = m | b) = 0, \quad \forall\, m < n,$$

where $c_{k,j}$ are the respective discrete boundary positions.

We now assume a set of i.i.d realizations of X denoted by $\mathcal{D} = (x^i)_{i=1}^N, x^i \in K$. Given $\mathcal{D}$ we wish to infer an estimate of the underlying distribution, whose support is restricted to the cone K. This constraint immediately rules out all classical parametric density estimators, as they lack the ability to restrict their support to such a linear subspace. For example, consider the Gaussian density estimate for two neighboring boundaries close to the foveola. In that region the upper retina layers vanish in order to allow as much light as possible to reach the photo-receptor layers located below, c.f. Figure 2.8 (b). This leads to estimates of f, where a significant amount of probability mass is located outside of K as illustrated in Figure 5.1.

We therefore turn our attention to non-parametric density estimation. Since maximizing the likelihood of a non-parametric density without any regularization

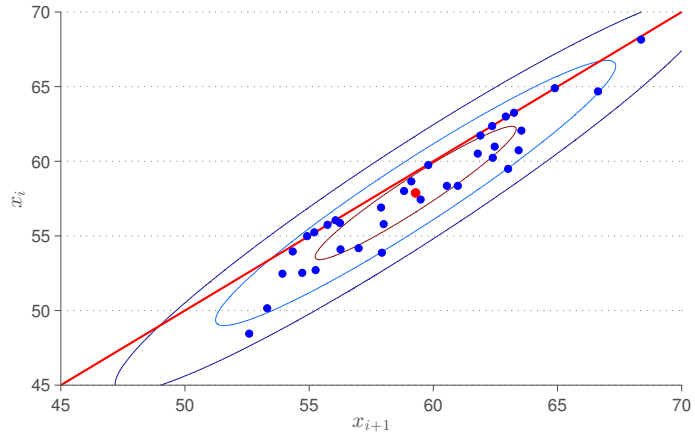


Figure 5.1 - Samples of two neighboring layer boundaries close to the foveola (c.f. Figure 2.8 (b)), that all satisfy the constraint $X_{i+1} \geq X_i$. Contour lines show one, two and three standard deviations of the normal distribution. The red line denotes the boundary of the cone constraint (5.1.2). A significant amount of probability mass is located outside that cone.

results in a density f of Dirac deltas at the sample points x^i, all approaches perform some sort of regularized maximum likelihood estimation.

5.1.2 Related Work

A class of approaches that received a great deal of interest lately is that of *shape-constraint* density estimators, which dates back to the seminal work of [Gre56]. Recent approaches proposed the estimation of a *log-concave* density function f in $\mathbb{R}$ [Ruf07, DR09] and more recently in $\mathbb{R}^d, d \geq 2$ [CSS10, KM10, SW10], such that

$$f(x) = \exp(-g(x)), \tag{5.1.3}$$

where $g(x) : \mathbb{R}^d \to (-\infty, \infty]$ is a convex function. The class of log-concave density functions is an immensely rich one, including many parametric density functions as special cases: all (non-degenerate) normal distributions, Wishart distributions, gamma distributions with shape-parameter bigger than one, beta(α, β) distributions with $\alpha, \beta \geq 1$ and many more.

Log-concave densities enjoy many desirable properties: All its level sets are convex and bounded. Log-concave density estimators do not rely on a tuning parameter contrary to for example kernel density estimators, a popular class of non-parametric density estimators [Par62]. The property of log-concavity itself is preserved under marginalization and conditioning and the product of log-concave densities is log-concave again. Finally, in case a limiting density exists the weak limits of log-concave densities are also log-concave [DJD88]. Sampling is especially efficient, with applications to MCMC and related sampling algorithms, see for example [FKP94] and [Bro98].

Log-concave density estimation can be seen from two perspectives: The first one is that of maximizing the log-likelihood of the sample set $\mathcal{D}$ (Definition 2.27) subject

to the log-concavity constraint on the estimated density f. Dual to that perspective is that of maximizing the Shannon entropy of f (Definition 2.23) subject to absolute continuity of the corresponding measure F with respect to the Lebesgue measure [KM10].

Outline. In the next section we will introduce the log-likelihood perspective and from that derive its dual formulation (similar to the treatment in [KM10]). In Section 5.3 we describe the optimization procedure, that is based on a discretization, and consists of solving a differentiable convex optimization problem with an interior-point method. Section 5.4 presents some numerical evaluations. An alternative approach is discussed in Section 5.5.

5.2 Log-Concave Density Estimator

5.2.1 Primal Formulation

Given a data set $\mathcal{D} = \{x^i\}_{i=1}^N$, the primal objective function is given by the following constrained maximum likelihood problem:

$$\min_g \Phi_0(g) = \frac{1}{N}\sum_{i=1}^N g(x^i), \quad \text{such that } g \text{ is convex} \quad \text{and} \quad \int e^{-g(x)}dx = 1, \quad (5.2.1)$$

where $g = -\log f$. However, the non-convexity of the set $\{g \,|\, \int e^{-g} = 1\}$ renders the whole problem intractable. But, as pointed out by [Sil82], a convex formulation can be obtained by moving the integral constraint into the objective function:

$$\min_g \Phi_1(g) = \frac{1}{N}\sum_{i=1}^N g(x^i) + \int e^{-g(x)}dx, \qquad \text{such that } g \text{ is convex.} \qquad (5.2.2)$$

The following theorem ensures equality between both problem formulations:

Theorem 5.1 *([Sil82]).* The convex function $\hat{g}$ minimizes Φ_0 over g subject to $\int e^{-g(x)}dx = 1$ if and only if $\hat{g}$ minimizes Φ_1.

Proof. Let g be any convex function and define $g^* = g + \log \int e^{-g}$. We have $\int e^{-g^*} = \int \frac{1}{\int e^{-g}} e^{-g} = 1$. Plugging g^* into Φ_1 yields

$$\Phi_1(g^*) = \Phi_1(g) + 1 - \int e^{-g} + \log \int e^{-g}.$$

The maximum of $f(x) = 1 - x + \log x$ is obtained for $f(x = 1) = 0$ which induces $\Phi_1(g^*) \leq \Phi_1(g)$. Thus g minimizes $\Phi_1(g)$ if and only if $\int e^{-g} = 1$. But for any g with $\int e^{-g} = 1$, $\Phi_1(g)$ and $\Phi_0(g) + 1$ are identical, which proofs the theorem. □

One can show that functions g minimizing Φ_1 are *finitely generated*: For every collection (X, Y)[1] of points $x^i \in \mathbb{R}^d$ and $y^i \in \mathbb{R}$, we define the function $g_{(X,Y)}$ finitely

[1] To remain consistent with [KM10], we denote from hereon by X any set of points and the set $\mathcal{D}$ in particular. This overrides the meaning of X as a random variable.

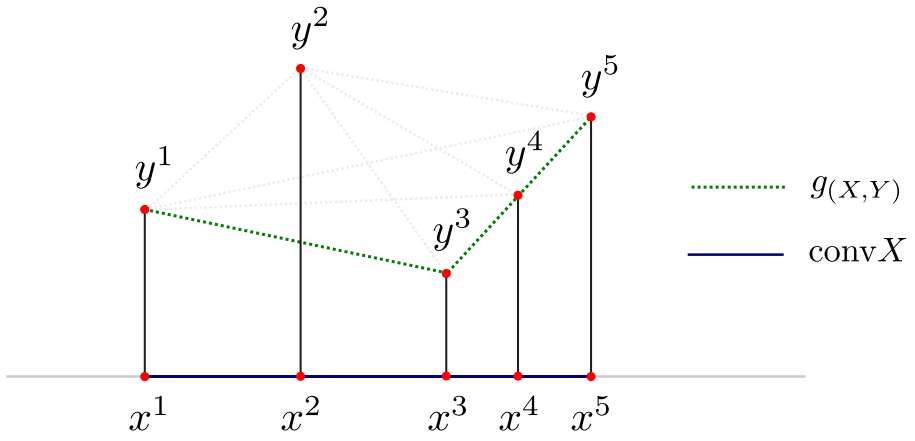


Figure 5.2 - Dotted lines visualize convex combinations of points (x^i, y^i) and (x^j, y^j). As defined in (5.2.3), $g_{(X,Y)}$ denotes the lower convex hull of (X, Y).

generated by (X, Y) to be

$$g_{(X,Y)}(x) = \inf\left\{\sum_{i=1}^{N} \lambda_i y^i \,|\, x = \sum_{i=1}^{N} \lambda_i x^i,\ \sum_{i=1}^{N} \lambda_i = 1,\ \lambda_i \geq 0\right\}. \tag{5.2.3}$$

Hence, by [Roc70, Corol. 19.1.2], $g_{(X,Y)}$ is a *polyhedral convex* function which has a polyhedral epigraph, being the *lower convex hull* of the points (X, Y) (see Figure 5.2 for an example in $\mathbb{R}$). That is equivalent to saying that $g_{(X,Y)}$ is the supremum of a finite number of affine functions, hence convex [RWW09, Prop. 2.9(b)]. Clearly, for any convex function h with $h(x^i) \leq y^i$ for all i, it holds that $h(x) \leq g_{(X,Y)}(x)$ for all x [RWW09, Prop. 2.31]. From the convention $\inf\{\emptyset\} = +\infty$ follows $\mathrm{dom}\, g_{(X,Y)} = \mathrm{conv}\, X$.

Let us denote, for a fixed set of points X, the collection of finitely generated functions $g_{(X,Y)}$ by $\mathcal{G}(X)$. The following theorem enables us to restrict the mimimization of $\Phi_1(g)$ to the set $\mathcal{G}(X)$:

Theorem 5.2 *([KM10]).* For any convex function h, we can find a function $g \in \mathcal{G}(X)$ such that $\Phi_1(g) \leq \Phi_1(h)$. Strict inequality holds whenever $h \notin \mathcal{G}(X)$ and $\mathrm{conv}\, X$ has nonempty interior.

For the sake of mathematical clarity in the subsequent development, we reformulate the primal optimization problem (5.2.2) once more. Let $\mathcal{C}(X)$ denote the space of functions continuous on $\mathrm{conv}\, X$ and $\mathcal{K}(X)$ the cone of closed (lower semicontinuous) convex functions on $\mathrm{conv}\, X$. Clearly $\mathcal{G}(X) \subset \mathcal{K}(X) \subset \mathcal{C}(X)$. In view of Theorem 5.2, any solution of (5.2.2) is also a solution of

$$\min_{g \in \mathcal{C}(X)} \frac{1}{N} \sum_{i=1}^{N} g(x^i) + \int e^{-g(x)} dx, \qquad \text{such that } \ g \in \mathcal{K}(X), \tag{5.2.4}$$

and vice versa.

5.2.2 Dual Formulation

This section will outline the derivation of the problem *dual* to (5.2.4) as given in [KM10]. To do that we require further prerequisites: Let $\mathcal{C}^*(X)$ denote the *dual space* of $\mathcal{C}(X)$, which is the space of signed, finite, regular Borel measures on conv X. The *bilinear form* associated with $\mathcal{C}(X)$ and $\mathcal{C}^*(X)$ is given by

$$\langle g, G\rangle = \int g\, dG \qquad g \in \mathcal{C}(X),\, G \in \mathcal{C}^*(X). \tag{5.2.5}$$

The *polar cone* to $\mathcal{K}(X)$ is defined as

$$\mathcal{K}^*(X) = \left\{ G \in \mathcal{C}^*(X) \,|\, \langle g, G\rangle \leq 0 \;\text{ for all } g \in \mathcal{K}(X) \right\}. \tag{5.2.6}$$

Moreover, $P_n \in \mathcal{C}^*(X)$ denotes the *empirical measure* of the set X. Finally, the conjugate f^* of a function f was introduced in Definition 2.8.

Let $\Upsilon(g)$ be the indicator function of the cone $\mathcal{K}(X)$ (c.f. Definition 2.2). Then we can equivalently express (5.2.4) as

$$\inf_{g \in \mathcal{C}(X)} \Phi(g) + \Upsilon(g). \tag{5.2.7}$$

Fenchel's duality theorem[2] [Roc70, Thm. 31.1] states that (under certain conditions, that are satisfied for the problem at hand [KM10]) strong duality holds between (5.2.7) and

$$\sup_{G \in \mathcal{C}^*(X)} -\Upsilon^*(-G) - \Phi^*(G), \tag{5.2.8}$$

where $-\Upsilon^*(-G)$ expresses the constraint $G \in \mathcal{K}^*(X)$ [KM10].

The conjugate function of $\Phi(g)$ is given by

$$\begin{aligned}\Phi^*(G) &= \sup_{g \in \mathcal{C}(X)} \left\{ \langle G, g\rangle - \frac{1}{N}\sum_{i=1}^{N} g(x^i) - \int e^{-g}\, dx \right\} \\ &= \sup_{g \in \mathcal{C}(X)} \left\{ \langle G - P_n, g\rangle - \int e^{-g}\, dx \right\} = \Psi^*(G - P_n),\end{aligned}$$

where we defined $\Psi^*(H)$ as the conjugate function of $\Psi(g) = \int \psi(g) = \int e^{-g}$. [R+71, Corol. 4A] states that for measures H that are absolutely continuous with respect to the Lebesgue measure, the functional $\Psi^*(H)$ is given by

$$\Psi^*(H) = \int \psi^*\left(\frac{dH}{dx}\right) dx, \tag{5.2.9}$$

and $\Psi^*(H) = +\infty$ otherwise. The argument of ψ^* denotes the *Radon-Nikodym derivative* and defines an integrable function $h(x)$ such that $H(A) = \int_A h(x)\, dx$. The

[2] Together with the fact that for a convex function $h(x) = -g(x)$ its conjugate is given by $h^*(x^*) = -g^*(-x^*)$ [Roc70, p. 308].

explicit form of $\psi^*(h)$ is given by

$$\psi^*(h) = -h \log -h + h, \qquad \operatorname{dom} \psi^* = \{h \,|\, h \leq 0\}. \tag{5.2.10}$$

Finally, let $f = -h$. We now obtained all ingredients to give the final statement:

Theorem 5.3. The problem

$$\sup_f \; -\int f(x) \log f(x)\, dx, \qquad \text{such that } \; f = \frac{d(P_n - G)}{dx}, \quad G \in \mathcal{K}^*(X), \tag{5.2.11}$$

is the strong dual problem of (5.2.4). Furthermore, for any dually feasible f, it holds that $f \geq 0$ and $\int f dx = 1$, hence that f is a probability density.

Indeed, the non-negativity of f is induced by the domain of $\psi^*(-f)$. From the definition of the polar cone $\mathcal{K}^*(X)$ (5.2.6) it follows that $\langle G, g \rangle \leq 0$ for any $g \in \mathcal{K}(X)$. Therefore

$$0 \geq \langle G, 1 \rangle = -\langle G, -1 \rangle \geq 0 \Longrightarrow \langle G, 1 \rangle = 0,$$

and for every dual feasible f it follows that

$$\int f(x)\, dx = \langle P_n - G, 1 \rangle = \langle P_n, 1 \rangle - \langle G, 1 \rangle = \int 1\, dP_n = 1.$$

5.3 Discretization and Optimization

There exist two popular approaches in the literature, that propose numerical procedures for solving the optimization problems outlined in the previous section. The first one by [CSS10] solves a reformulated version of the primal problem (5.2.2) which is convex but non-differentiable. Therefore, they have to resort to subgradient-based methods with no theoretical proof of convergence (although in practice convergence is no issue). Furthermore, their approach calculates in each step the convex hull of the point set (X, Y) mentioned in the previous section, and therefore run-time scales strongly with the number of samples.

The second approach by [KM10] solves the dual problem formulation (5.2.11) on a discrete grid where individual points are subsumed by the grid points that enclose them. The resulting optimization problem is smooth and differentiable, and can be solved by interior-point methods. Their approach trades the dependency on sample size for that on grid density, although this is no issue for 2-D.

Outline. As a prerequisite for the discretization, we derive an approximation of the Hessian of g based on finite differences in Section 5.3.1. Section 5.3.2 will establish the discritization of (5.2.2) based on a regular grid and derive the discrete version of the primal function. While [KM10] go on to derive the discrete dual, we show in Section 5.3.3 how to directly optimize the discrete primal formulation. In Section 5.3.4 we extend their approach and derive a general formulation for the case of three or more dimensions.

5.3.1 Finite Difference Approximation of Derivatives

A twice-differentiable function $g(x)$ is convex if

$$\nabla^2 g(x) \succeq 0, \qquad \forall\, x \in \operatorname{dom} g(x). \tag{5.3.1}$$

The Hessian $\nabla^2 g(x)$ is defined in terms of the second partial derivatives of g, which are in turn defined as limits of difference quotients for some $h \to 0$. The next section will deal with a discrete approximation of $g(x)$ on a regular grid, where such limits can not be taken.

We therefore resort to *finite difference approximations* of derivatives. Using multi-index notation

$$|\alpha| = \alpha_1 + \ldots + \alpha_n, \quad a! = a_1! \ldots a_n!, \quad x^\alpha = x_1^{\alpha_1} \ldots x_n^{\alpha_n}, \tag{5.3.2}$$

the following Theorem states that one can approximate a k times differentiable function g around a given point by a kth order Taylor polynomial:

Theorem 5.4 *(Taylor's theorem [JJ99]).* Let $g : \mathbb{R}^n \to \mathbb{R}$ be a k times differentiable function at the point $a \in \mathbb{R}^n$. Then there exists $h_\alpha : \mathbb{R}^\alpha \to \mathbb{R}$ such that

$$\begin{aligned} g(x) &= P_k(x) + \sum_{|\alpha|=k} h_\alpha(x)(x-a)^\alpha \\ P_k(x) &= \sum_{|\alpha|\le k} \frac{D^\alpha g(a)}{\alpha!}(x-a)^\alpha \end{aligned} \tag{5.3.3}$$

and $\lim_{x\to a} h_\alpha(x) = 0$.

The first term constitutes the Taylor polynomial $P_k(x)$ of degree k, while the term $h_\alpha(x)(x-a)^\alpha$ denotes the approximation error $g(x) - P_k(x)$.

Using a Taylor polynomial $P_2(x)$ around the point $x = (x_1, x_2)$, one can approximate the function g by

$$\begin{aligned} &g(x_1 + h_1, x_2 + h_2) \approx \\ &\quad g(x) + \frac{\partial g(x)}{\partial x_1} h_1 + \frac{\partial g(x)}{\partial x_2} h_2 + \frac{\partial^2 g(x)}{\partial x_1^2}\frac{h_1^2}{2!} + \frac{\partial^2 g(x)}{\partial x_2^2}\frac{h_2^2}{2!} + \frac{\partial^2 g(x)}{\partial x_1 \partial x_2}\frac{h_1 h_2}{2!}. \end{aligned}$$

which becomes more accurate as $\|h\| \to 0$.

We can now derive approximations for the terms in $\nabla^2 g(x)$. Adding the approximations for $g(x_1 + h_1, x_2)$ and $g(x_1 - h_1, x_2)$, the terms $-\frac{\partial g(x)}{\partial x_1} h_1$ and $\frac{\partial g(x)}{\partial x_1} h_1$ cancel and one obtains

$$\frac{\partial^2 g(x)}{\partial x_1^2} \approx \frac{g(x_1 + h_1, x_2) + g(x_1 - h_1, x_2) - 2g(x)}{h_1^2},$$

and similar for $\frac{\partial^2 g(x)}{\partial x_2^2}$. The approximation for the off-diagonal terms,

$$\frac{\partial^2 g(x)}{\partial x_1 \partial x_2} \approx \frac{g(x-h) - g(x_1+h_1, x_2-h_2) - g(x_1-h_1, x_2+h_2) + g(x+h)}{4h_1h_2},$$

is obtained by adding the Taylor expansions of the terms in the numerator. We will make use of these approximations in the next section, to calculate Hessian matrix evaluations at the grid points.

5.3.2 Discretization of $\Phi_1(g)$

Rectangular Grid. Let us define a regular rectangular grid consisting of grid points $\xi_\alpha \in \mathbb{R}^n$ where we use the multi-index notation $\alpha = (\alpha_1, \ldots, \alpha_n)$ to denote the grid position $(\xi_{\alpha_1}, \ldots, \xi_{\alpha_n})$. The distance along each coordinate is δ. Finally, we denote function evaluations at grid point ξ_α by $\gamma_\alpha = g(\xi_\alpha)$.

Convexity Constraint. In the previous section we derived an approximation for Hessian matrices H^α at grid points ξ_α based on finite differences. For a grid point $\xi_{ij} = (\xi_i, \xi_j) \in \mathbb{R}^2$, its entries are given by

$$\begin{aligned}
H_{11}^{ij} &= [g(\xi_i+\delta, \xi_j) - 2g(\xi_i, \xi_j) + g(\xi_i-\delta, \xi_j)]/\delta^2, \\
H_{22}^{ij} &= [g(\xi_i, \xi_j+\delta) - 2g(\xi_i, \xi_j) + g(\xi_i, \xi_j-\delta)]/\delta^2, \\
H_{12}^{ij} &= [g(\xi_i+\delta, \xi_j+\delta) - g(\xi_i+\delta, \xi_j-\delta) \\
&\qquad - g(\xi_i-\delta, \xi_j+\delta) + g(\xi_i-\delta, \xi_j-\delta)]/4\delta^2, \\
H_{21}^{ij} &= H_{12}^{ij}.
\end{aligned} \tag{5.3.4}$$

Convexity of g is then *approximately* enforced by imposing positive semi-definiteness of H^{ij} at all grid points ξ_{ij} except those on the grid boundary. In $\mathbb{R}^2$ positive semi-definiteness of H is enforced by

$$H_{11}^{ij}, H_{22}^{ij} \geq 0, \qquad \det H^{ij} = H_{11}^{ij} H_{22}^{ij} - (H_{12}^{ij})^2 \geq 0, \qquad \forall ij. \tag{5.3.5}$$

These constraints correspond to *rotated quadratic cone* constraints:

$$\mathcal{Q}_{\text{rot}} := \{(x, y, z) \in \mathbb{R}^{1+2} \mid x^2 \leq yz, y \geq 0, z \geq 0\}. \tag{5.3.6}$$

Thus, we specifically have $(H_{12}^{ij}, H_{11}^{ij}, H_{22}^{ij}) \in \mathcal{Q}_{\text{rot}}$ for all grid points ξ_{ij} which more explicitly means

$$\left\| \begin{pmatrix} 2H_{12}^{ij} \\ H_{11}^{ij} - H_{22}^{ij} \end{pmatrix} \right\| \leq H_{11}^{ij} + H_{22}^{ij} \qquad \Leftrightarrow \qquad \underbrace{\begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & 1 & 1 \end{pmatrix}}_{:=R_{\mathcal{Q}}} \begin{pmatrix} H_{12}^{ij} \\ H_{11}^{ij} \\ H_{22}^{ij} \end{pmatrix} \in \mathcal{L}^2. \tag{5.3.7}$$

Denoting (5.3.4) as

$$H^{ij} := (H_{12}^{ij} H_{11}^{ij} H_{22}^{ij})^\top =: G^{ij} \gamma^{ij} \tag{5.3.8}$$

with

$$G^{ij}\gamma^{ij} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1/4 & -1/4 & -1/4 & 1/4 \\ 1 & 1 & 0 & 0 & -2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & -2 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \gamma_{i-1,j} \\ \gamma_{i+1,j} \\ \gamma_{i,j-1} \\ \gamma_{i,j+1} \\ \gamma_{i,j} \\ \gamma_{i-1,j-1} \\ \gamma_{i-1,j+1} \\ \gamma_{i+1,j-1} \\ \gamma_{i+1,j+1} \end{pmatrix}, \tag{5.3.9}$$

and setting

$$A^{ij} := R_{\mathcal{Q}} G^{ij}, \tag{5.3.10}$$

the constraints (5.3.7) read

$$A^{ij}\gamma^{ij} \in \mathcal{L}^2, \qquad \forall ij. \tag{5.3.11}$$

Assembling them into a global system yields

$$A\gamma \in \mathcal{K}. \tag{5.3.12}$$

Later on, it will be convenient to make explicit the connection between (5.3.11) and (5.3.12). Each index ij in (5.3.11) refers to an internal grid node $\xi_{ij} = (\xi_i, \xi_j)$. The local vectors H^{ij} defined by (5.3.8) and (5.3.4) may not correspond to consecutive components of the vector $A\gamma$ in (5.3.12). We therefore introduce matrices I^{ij} such that

$$I^{ij} A\gamma = A^{ij}\gamma^{ij}, \qquad \forall ij. \tag{5.3.13}$$

Accordingly, the cone $\mathcal{K}$ in (5.3.12) is defined by

$$A\gamma \in \mathcal{K} \qquad \Leftrightarrow \qquad I^{ij} A\gamma \in \mathcal{L}^2, \qquad \forall ij. \tag{5.3.14}$$

Integral Constraint. For the integral constraint, [KM10] suggest performing straightforward Riemann sums on the rectangular grid

$$\int_{\operatorname{conv} X} \exp(-g(x))\, dx \approx \sum_{\alpha} s_\alpha \exp(-\gamma_\alpha) = s^T \Psi(\gamma), \tag{5.3.15}$$

where the index runs over all points in the grid and weights $s_\alpha = \delta^2$ denote the area of the approximating squares. Naturally, the accuracy of such an approximation depends heavily on the density of the grid.

Log-Likelihood-Term. Since sample points x^i most probably won't lie on the rectangular grid, evaluation of $g(x^i)$ is done by linear interpolation of values γ_α at grid points ξ_α directly enclosing x^i. This is sufficiently accurate given a fine enough

grid. We define

$$\frac{1}{N}\sum_{i=1}^{N} g(x^i) \approx \omega^T L\gamma \tag{5.3.16}$$

where L is a interpolation operator, selecting the appropriate grid points such that $(L\gamma)_i \approx g(x^i)$ and ω is a weighting vector of the observations, typically $\omega_i = 1/N$.

5.3.3 Discrete Objective Function and Numerical Optimization

Using the formulations (5.3.14), (5.3.15) and (5.3.16), the discrete equivalent of the primal problem (5.2.2) reads

$$\inf_{\gamma}\ \omega^\top L\gamma + s^\top \Psi(\gamma), \qquad \text{such that}\ \ A\gamma \in \mathcal{K}, \tag{5.3.17}$$

which is a convex optimization problem with *generalized inequality constraints* (c.f. Definition 2.9).

As pointed out in Section 2.1, this type of problem can be transformed into an unconstrained convex problem using logarithmic barrier functions. To every grid point ξ_α in the interior of the grid corresponds a generalized inequality

$$-I^\alpha A\gamma \preceq_{\mathcal{L}^2} 0 \quad \Longleftrightarrow \quad I^\alpha A\gamma \in \mathcal{L}^2.$$

Recall, that we defined $I^\alpha A\gamma$ in (5.3.13) as the column vector $A^\alpha\gamma^\alpha$ consisting of three elements. Identifying $\psi(-f_i(x))$ with

$$\psi(I^\alpha A\gamma) = \log\left((A_3^\alpha\gamma^\alpha)^2 - ((A_1^\alpha\gamma^\alpha)^2 + (A_2^\alpha\gamma^\alpha)^2)\right), \tag{5.3.18}$$

we can collect all log-barrier terms in the function

$$\phi(A\gamma) = -\sum_{\alpha}\psi(I^\alpha A\gamma). \tag{5.3.19}$$

Following (2.1.13) we can formulate the unconstrained version of (5.3.17):

$$\inf_{\gamma}\ t\left(\omega^T L\gamma + s^T\Psi(\gamma)\right) + \phi(A\gamma) := \pi(\gamma), \tag{5.3.20}$$

where we introduced the weighting term t. As pointed out in Section 2.1, a valid approach for that type of problem is the *barrier method*, that iteratively solves the unconstrained convex problem $\pi(\gamma)$ to optimality and sets $t := \mu t$ with $\mu > 1$ after each step. The optimization of $\pi(\gamma)$ is performed by *Newtons method*, that requires the calculation of the gradient and the Hessian of $\pi(\gamma)$ for each descent step $\Delta\gamma^k$, c.f. (2.1.14) and (2.1.15).

The gradient and Hessian of $\pi(\gamma)$ are

$$\begin{aligned} \nabla\pi &= t(\omega^T L + \theta) + \nabla\phi, & \theta_\alpha &= -s_\alpha\exp(-\gamma_\alpha),\\ \nabla^2\pi &= t\Theta + \nabla^2\phi, & \Theta_{\alpha\alpha} &= s_\alpha\exp(-\gamma_\alpha). \end{aligned}$$

The derivation of $\nabla\phi$ and $\nabla^2\phi$ is given below. Note that, although the problem is high-dimensional with hundreds of thousands of grid points, the Hessian $\nabla^2\pi$ is a sparse banded matrix. While Θ is diagonal, the off-diagonal entries of $\nabla^2\phi$ reflect the neighborhood structure imposed by the local Hessian matrices H^α, which access only the direct neighbors of each ξ_α.

We therefore were able to utilize fast versions of Cholesky decomposition designed for sparse matrices, to calculate the Newton step (2.1.15) in an efficient way. Furthermore, we utilized Matlabs sparse library to speed up several other computations.

Gradient and Hessian of $\phi(A\gamma)$. Let us denote the argument of the logarithm in (5.3.18) by c_α, such that

$$\phi(A\gamma) = -\sum_\alpha \log(c_\alpha).$$

Then the gradient of ϕ is given by

$$\nabla\phi = -\sum_\alpha (c_\alpha)^{-1}\nabla c_\alpha,$$
$$\nabla c_\alpha = 2\left((A_3^\alpha\gamma^\alpha)(A_3^\alpha)^T - (A_1^\alpha\gamma^\alpha)(A_1^\alpha)^T - (A_2^\alpha\gamma^\alpha)(A_2^\alpha)^T\right).$$

Again differentiating with respect to γ yields the Hessian

$$\nabla^2\phi = -\sum_\alpha -(c_\alpha)^{-2}\nabla c_\alpha(\nabla c_\alpha)^T - (c_\alpha)^{-1}\nabla^2 c_\alpha,$$

with $\nabla^2 c_\alpha$ given by

$$\nabla^2 c_\alpha = 2\left(A_3^\alpha(A_3^\alpha)^T - A_1^\alpha(A_1^\alpha)^T - A_2^\alpha(A_2^\alpha)^T\right).$$

5.3.4 N-D case

The extension of both the log-likelihood term (5.3.16) and the integral constraint (5.3.15) to a grid with three or more dimensions is straightforward. However, the positive semidefiniteness constraints of local Hessian matrices H^α (5.3.5) require a different log-barrier function. As pointed out in Example 2.10, the barrier function for the constraint $H^\alpha \in \mathcal{S}_+^n$ is

$$\psi(H^\alpha) = \log\det H^\alpha, \qquad \text{dom } \psi(H^\alpha) = \text{int}\,\mathcal{S}_+^n, \tag{5.3.21}$$

such that

$$\phi(\gamma) = -\sum_\alpha \log\det H^\alpha. \tag{5.3.22}$$

Note that we dropped the selection and rotation matrix A, but directly relate entries of H^α to entries in γ. But these are merely notational issues. Using Laplace expansion, the determinant of a symmetric 3×3 matrix can be calculated using the formula

$$\det H^\alpha = H_{11}^\alpha H_{22}^\alpha H_{33}^\alpha + 2H_{12}^\alpha H_{23}^\alpha H_{13}^\alpha - H_{11}^\alpha(H_{23}^\alpha)^2 - H_{22}^\alpha(_{13}^\alpha)^2 - H_{33}^\alpha(H_{12}^\alpha)^2. \tag{5.3.23}$$

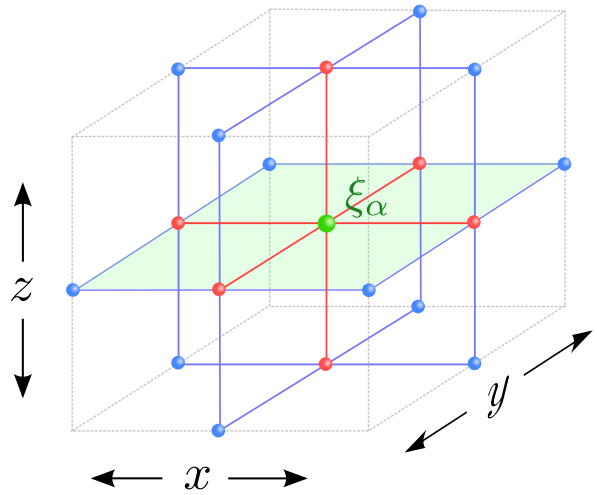


Figure 5.3 - Denotes, for a three-dimensional grid, the neighbors of ξ_α that are required to calculate all entries of H^α. Blue squares denote those neighbors that are required for off-diagonal terms H_{12}, H_{23} and H_{13}, red lines those (including ξ_α itself) that are required for H_{11}, H_{22} and H_{33}. The green layer denotes the 2-D case.

The entries of H are determined from γ using the same scheme as in (5.3.4) adapted to the third dimension and visualized in Figure 5.3. Entries H_{11}, H_{22} and H_{33} are denoted by red lines and entries H_{12}, H_{13} and H_{23} by blue squares. The green layer denotes the 2-dimensional case.

Instead of 8 neighboring grid points, each grid point ξ_α now is related to 18 other grid points. This increases the bandwidth in $\nabla^2\phi$. Moreover, the number of grid points now grows cubically with the number of points per dimension. This significantly increases the complexity of calculating the Newton step. We examine this issue in Section 5.4.4.

Gradient and Hessian of $\phi(\gamma)$. Applying the chain rule, we obtain

$$\nabla\phi = -\sum_\alpha \frac{1}{\det H^\alpha}\nabla \det H^\alpha$$

$$\nabla^2\phi = -\sum_\alpha -\frac{1}{(\det H^\alpha)^2}\nabla \det H^\alpha(\nabla \det H^\alpha)^T + \frac{1}{\det H^\alpha}\nabla^2 \det H^\alpha$$

Calculating $\nabla \det H^\alpha$ and $\nabla^2 \det H^\alpha$ amounts to writing each entry H^α_{ij} as a linear combination of entries in γ as visualized in Figure 5.3 and differentiating $\det H^\alpha$ with respect to γ. Since these calculations are straightforward but rather tedious we omit them here.

5.4 Experiments

The experimental section is concerned with two things: The first one is to demonstrate soundness of our implementation for 2-D and 3-D. Secondly, it will outline numerical properties of our implementation and compare them with results obtained from the

R package `LogConcDead` of [CSS10]. Finally, we will obtain a log-concave density estimate for the motivating example shown in Figure 5.1 that obeys the ordering constraint.

5.4.1 Setup

If not otherwise stated, we choose a grid of size 300×300 in accordance with [KM10], which yielded satisfactory results without artifacts and very close to the results of [CSS10] (see Section 5.4.3 for more details). We used values $t = \{1, 10, \ldots, 10^8\}$ as parameters of the barrier method.

5.4.2 Student's Criminals Data Set

In order to examine the validity of our implementation, we choose the Student's criminal data set [Stu08], that was also used in [KM10]. This bivariate data set contains the heights and left middle finger lengths of 3000 British criminals. Figure 5.4 a) shows the output of our implementation whereas panel b) shows the output of `LogConcDead`. Contour lines denote the same level sets. The plots show that the two different approaches yield almost identical density estimates.

The latter more clearly exposes the polyhedral character of the solution as predicted by Theorem 5.2. This originates from the fact, that the numerical approach proposed by [CSS10] explicitly models $g(x)$ as a polyhedral function, while the approach proposed by [KM10] and implemented here does not rely on an explicit polyhedral parametrisation of g and solves the dual problem using an interior point method. The corresponding log-barrier function enforces *strict* convexity of the grid function as defined above, and hence may introduce slight approximation errors in case of solutions that are actually located at the boundary of the feasible set.

5.4.3 Influence of Grid and Sample Size

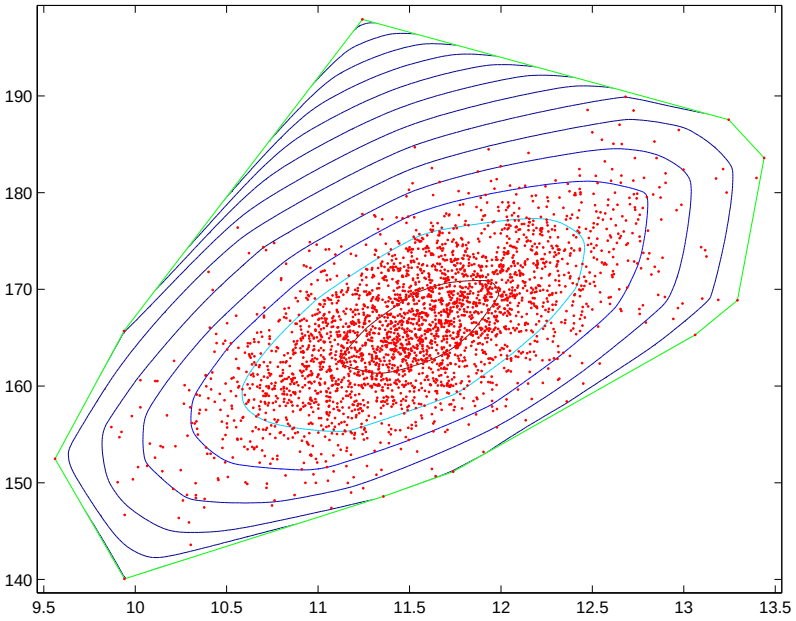
We pointed out before the different characteristics of the optimization approaches of [CSS10] and [KM10]: While the former is sensitive to sample size the latter's performance is governed by the grid density. We will examine both aspects in this section.

To test the influence of the grid size, we sampled once 250 data points from a normal distribution with parameters

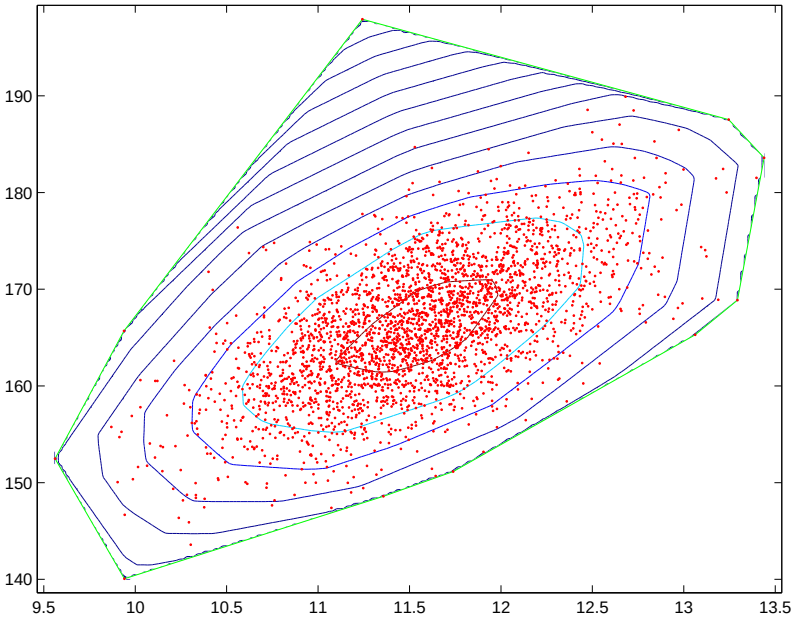
$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} 2 & 0.5 \\ 0.5 & 1 \end{pmatrix}.$$

We examined different grid sizes and run times are reported in Figure 5.5 (a)[3]. For comparison, the run time of Cule's approach for 250 samples is 3.5 s, indicated by the red line. The computational bottleneck in our implementation is (naturally) the Cholesky decomposition necessary for the Newton step, which accounts for roughly 85% of total run time. Furthermore, we observed that with growing grid density

[3] Grid size denotes number of grid points along each dimension.

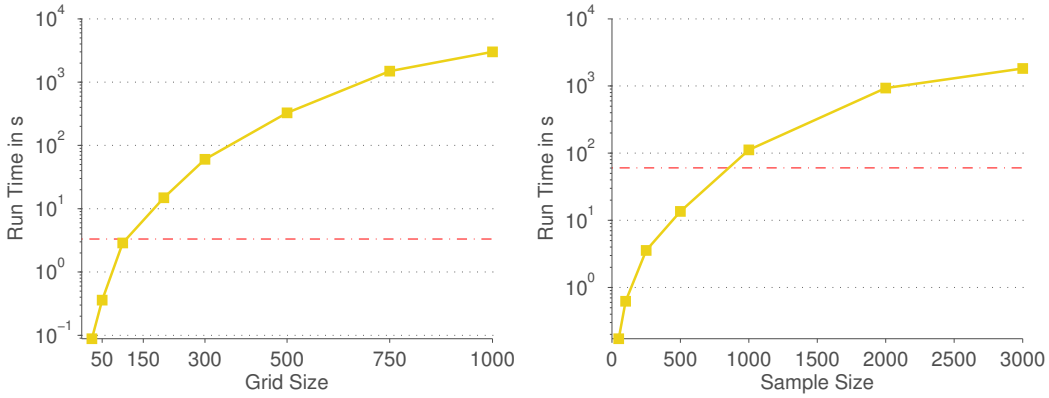


(a)



(b)

Figure 5.4 - Log-concave density estimations for the Student's criminal data set. Comparison of (a) our implementation (interior-point-approach applied to the primal formulation (5.3.20)) with the results (b) obtained from the R package `LogConcDead` of Cule et al. The estimated densities are almost the same.



(a) Approach by Koenker et al. (b) Approach by Cule et al.

Figure 5.5 - The influence of grid size and sample size on the run time. While the approach by [KM10] is sensitive to the grid density (a), the one of [CSS10] is sensitive to sample size (b). 250 samples where used in the left plot. Red lines denote run times of the respective other approach.

Newton's method required more iterations. Our implementation needed 60 s for a 300×300 grid , and 15 s for a 200×200 grid. On the other hand, the grid-based approach is insensitive to sample size, while the run time of Cule's approach increases, as shown in Panel (b) of Figure 5.5.

Figure 5.6 (a) shows that very low and very high grid sizes yields estimates with reduced sample log-likelihood. For low densities this is mainly caused by boundary effects and inaccurate interpolation of sample points. For very high densities this is caused by ill-conditioned Hessian matrices, which in turn result in poor Newton steps. Nonetheless, even for the "optimal" grid size of 300×300, there remains a small gap to the log-likelihood of the estimate of [CSS10], caused by the distinct ansatz of [KM10]. Confer also the discussion in Section 5.4.2.

5.4.4 Density Estimation in 3-D

In Section 5.3.4 we discussed the extension of the approach of [KM10] to the n-dimensional case. Since there the number of grid points grows cubically, running the approach with 300 grid points along each dimensions turned out too expansive. Therefore, as a proof of concept, we sampled 250 data points from a 3-dimensional normal distribution and obtained the log-concave density estimate using a $50 \times 50 \times 50$ grid.

The result is shown in Figure 5.6 (b). Recall, that for the 3-D problem the bandwidth of the Hessian $\nabla^2 \pi$ grows since the calculation of local Hessian matrices H^α requires more neighbors of grid point ξ_α. This in turn makes the Cholesky decomposition more expansive. Total run time was about 600 s while the run time of a 2-D example with the same amount of grid points only requires 100 s. From the viewpoint of our retina segmentation model this is not important though, since the shape prior distributions are calculated offline.

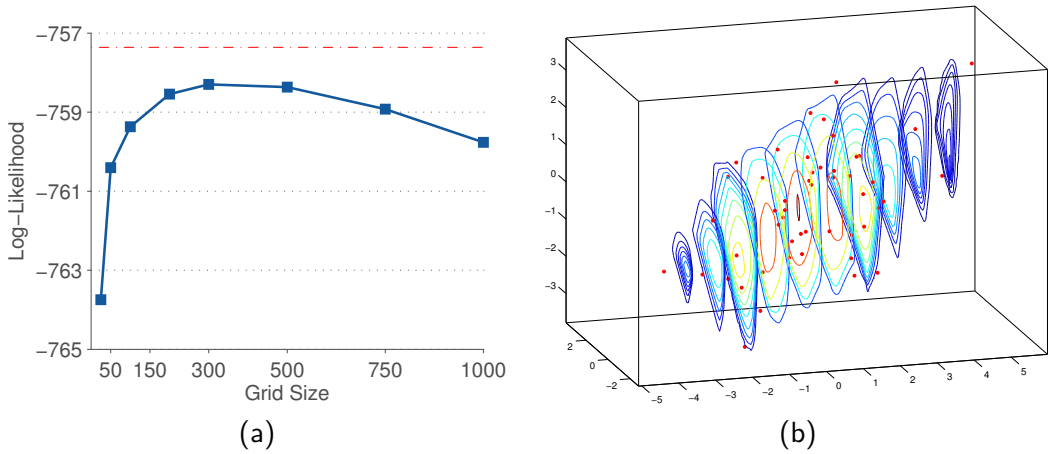


Figure 5.6 - (a) Sample log-likelihood for different grid sizes and the one returned by Cule's approach as a red line. See Section 5.4.3 for a discussion of that plot. (b) Log-concave density estimate in 3-D on a $50 \times 50 \times 50$ grid.

5.4.5 Log-Concave Shape Prior

In Figure 5.1 we showed a Gaussian density estimate for a set of data points that represent boundary positions very close to the foveola, a region where the upper retina layers can become very thin or even vanish. Figure 5.7 (a) again shows that data set together with its Gaussian density estimate. In Panel (b) the log-concave density estimate for the same data set can be seen. Contour lines in both plots denote the same level sets. Since the log-concave density concentrates all probability mass inside the convex hull of the data, it naturally has a higher density at almost all data points. Furthermore, the log-concave density captures the skewed characteristic of the data set in contrast to the normal density.

One problem remains: How to handle test samples which lie outside the convex hull of the training samples. In the literature exist smoothed versions of the log-concave estimator [CS13], but this would eliminate the inherent ordering property. A better alternative would be to project any sample onto the polyhedral convex set $\operatorname{conv} X$, which can be solved as a convex quadratic program, see [BV04, Sec. 8.1.1].

5.5 Discussion

We investigated non-parametric log-concave density estimation as a way of obtaining density estimates that obey the *ordering constraint* inherent in many problems. We examined the approach of [KM10] and outlined its optimization in detail. We also discussed its extension to more than two dimensions. We yet did not integrate it into our segmentation model, but that is rather straightforward and we discuss that point in the next chapter.

While the approach seems applicable in the 2-D and 3-D case, four or more dimensions don't seem feasible. On the other hand the approach by [CSS10] becomes

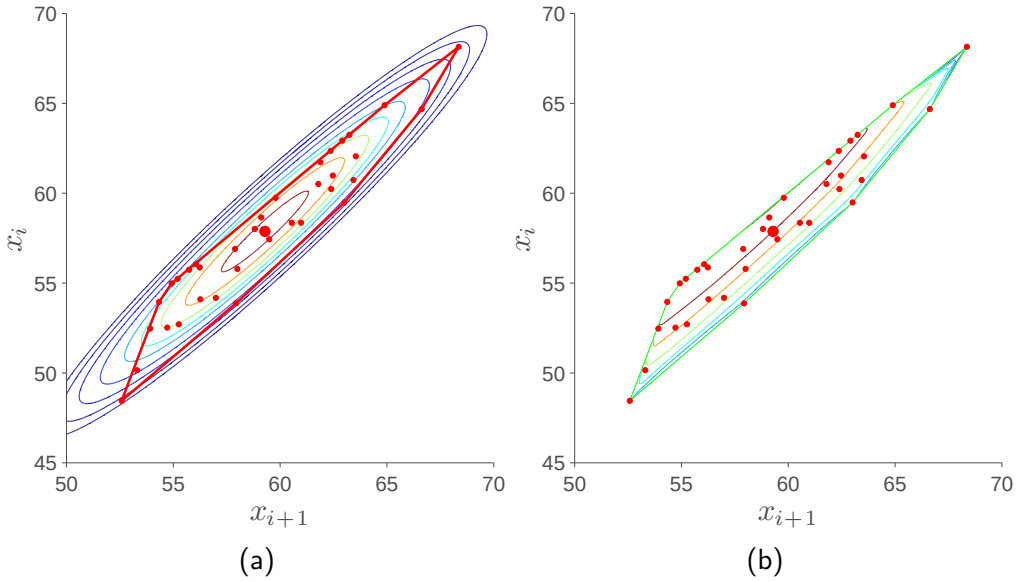


Figure 5.7 - Normal density versus a log-concave density estimate for positions of two neighboring boundaries close to the fovea (same sample as in Figure 5.1). Colors in both plots denote the same level sets. The right density obeys the ordering constraint inherent in the data set.

quickly intractable with increasing sample size. As a quick recap, Cule's approach relies on the "internal" representation of the epigraph of g, that is on a set if points (X, Y) whose lower convex hull defines the graph of g. To obtain a polyhedral representation of g they need to conduct a triangulation of (X, Y) at each step, whose simplicies then form the polyhedral function $g \in \mathcal{G}(X)$ (c.f. Equation (5.2.3)).

A first analysis of results produced by `LogConcDead` [CSS10] showed that only a small amount of points in (X, Y) are *extreme points* in the sense of [Roc70, Chap. 18], i.e. zero-dimensional faces of $\text{conv}(X, Y)$. One example is given in Figure 5.8 for 250 sample points, out of which 20 are extreme points, colored blue. Hyperplanes, corresponding to the 1-dimensional simplices connecting extreme points, are given by dashed lines for better visualization. Only four of the 20 simplicies generate most of $g(x)$.

This suggests an approach relying on the "external" representation of $\text{epi}\, g$ in terms of its *faces*, corresponding to the alternate definition of convexity in terms of a set of hyperplanes supporting g

$$g(x_1) \geq g(x_2) + p^T(x_2 - x_1), \qquad p \in \partial g(x_1), \ \ \forall x_1, x_2 \in \text{conv}\, X, \tag{5.5.1}$$

where the subdifferential $\partial g(x)$ denotes the convex set of *subgradients* [Roc70, Chap. 23] of g at x. For polyhedral functions a finite set of such hyperplanes suffices to accurately represents g.

A class of functions that automatically satisfy these constraints are *max-affine* functions, defined as

$$g(x) = \max_k \ a_k + \beta_k^T x, \tag{5.5.2}$$

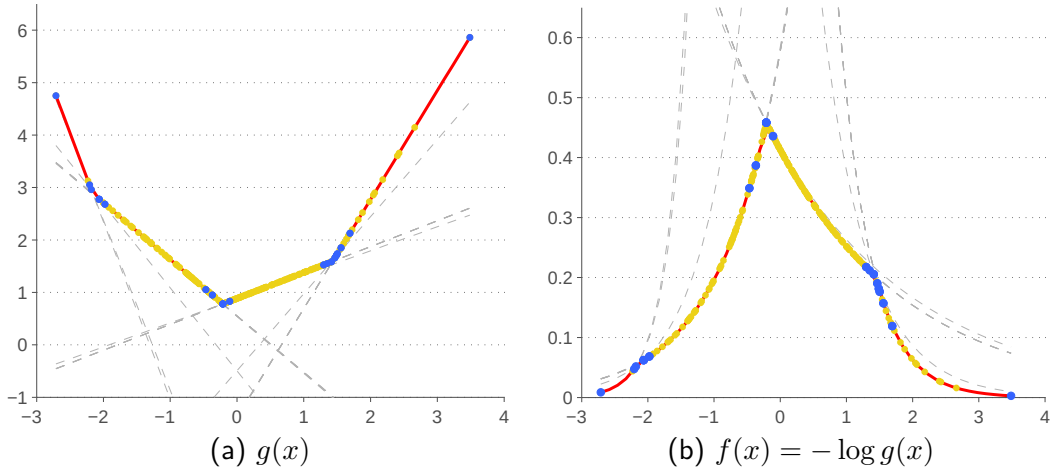


Figure 5.8 - Output of `LogConcDead`, the R package of [CSS10] for a 1-D sample of 250 points. Blue points denote *extreme points* of the lower convex hull, which constitutes $g(x)$. The dashed lines denote hyperplanes that correspond to the 1-dimensional faces connecting those extreme points. Although there exist 20 different hyperplanes, only four of them contribute significantly to the shape of $g(x)$ and $f(x)$ respectively.

for a set of hyperplanes $\alpha = \{a_1, \beta_1, \ldots, a_K, \beta_K\}$. Some approaches for convex regression exist [MB09, HD13], that fit max-affine functions by minimizing the least-squares error

$$J(\alpha) = \sum_{i=1}^{N} \left(g(x^i) - y^i \right)^2 = \sum_i \left(\max_k \left(a_k + \beta_k^T x^i \right) - y^i \right)^2. \tag{5.5.3}$$

Although convex for each polyhedral region where the kth hyperplane is active, globally this problem is non-convex. The same holds true for the log-concave density estimation problem based on the same function class, that is

$$\begin{aligned} J(\alpha) &= \frac{1}{N} \sum_{i=1}^{N} g(x^i) + \int e^{-g(x)} dx \\ &= \frac{1}{N} \sum_{i=1}^{N} \max_k \left(a_k + \beta_k^T x^i \right) + \int e^{-\max_k \left(a_k + \beta_k^T x \right)} dx. \end{aligned} \tag{5.5.4}$$

Working out a dedicated optimization method might provide a further promising direction of research in order to exploit sparse representations of g in dimensions $d > 3$.

6 Conclusion

Summary. In this thesis we proposed and evaluated a segmentation approach for the retina segmentation task. By the application to various data sets we could demonstrate its state-of-the-art performance. We furthermore showed how to make use of the fact that we infer an approximation of the full posterior distribution. This made it possible to establish an estimator for the quality of the predicted segmentations as well as a detector for pathological scans which both showed promising results. Finally, we published our revised code along with a documentation to serve as a possible baseline for future segmentation approaches.

In the last chapter we outlined a possible extension of our approach, that is the inclusion of local shape distributions that are adapted to the task of representing ordering constraints. We demonstrated their applicability in the 2-D and 3-D case. Their integration into the segmentation approach remains future work though (see the discussion below).

Future work. We already outlined some possible directions for future work before. Here we will express some additional thoughts.

- *(Connectivity of discrete graphical models)* Currently our segmentation model uses a separate discrete graph for each image column, where only direct neighbors are connected. Communication along image columns is governed by the shape prior. Other forms of more complex discrete graphs are conceivable, by either increasing the intra-column connectivity and/or by introducing connectivity between image columns. An adaptive approach could be to increase connectivity in regions with poor appearance information based e.g. on the values in the precision matrix of the shape prior distribution, since these values reflect conditional dependencies of boundary positions.

- *(Utilizing log-concave shape prior information)* Following Equation (3.1.9), we defined that the prior terms of our discrete graphical models consist of a local and a global component, both emerging from the Gaussian shape prior distribution. We could alter that for the local terms over neighboring boundaries, such that they would be given by the non-parametric log-concave densities discussed in Chapter 5. This would induce the ordering constraint in a more natural way. The modification would result in altered transition matrices $\Omega_{k,j}$ of the discrete graph q_c, as the normalizing constant $\tilde{C}$ would change, c.f. (3.2.15). It would not change however the calculation of the sufficient statistics of q_b and thus not alter the interplay between q_c and q_b.

Index

Bibliography

[ABK12] B. Andres, T. Beier, and J. Kappes. OpenGM: A C++ library for discrete graphical models. *ArXiv e-prints*, 2012.

[ADD00] R. B. Ash and C. Doleans-Dade. *Probability and Measure Theory*. Academic Press, 2000.

[Alf06] D. V. Alfaro. *Age-Related Macular Degeneration: A Comprehensive Textbook*. Lippincott Williams & Wilkins, 2006.

[AN85] P. Airaksinen and H. Nieminen. Retinal nerve fiber layer photography in glaucoma. *Ophthalmology*, 92(7):877–879, 1985.

[AO11] B. Antic and B. Ommer. Video parsing for abnormality detection. In *International Conference on Computer Vision (ICCV 2011)*, pages 2415–2422, 2011.

[ASG+08] C. Ahlers, C. Simader, W. Geitzenauer, G. Stock, P. Stetson, S. Dastmalchi, and U. Schmidt-Erfurth. Automatic segmentation in three-dimensional analysis of fibrovascular pigmentepithelial detachment using high-definition optical coherence tomography. *Brit. J. Ophthalmol.*, 92(2):197–203, 2008.

[BEGd08] O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learning Res.*, 9:485–516, 2008.

[BFT07] M. Baroni, P. Fortunato, and A. L. Torre. Towards quantitative analysis of retinal features in optical coherence tomography. *Med. Eng. Phys.*, 29(4):432–441, 2007.

[BGF13] F. E. Bachl, C. S. Garbe, and P. W. Fieguth. Bayesian inference on integrated continuity fluid flows and their application to dust aerosols. In *International Geoscience and Remote Sensing Symposium (IGARSS 2013)*, pages 2246–2249, 2013.

[Bis06] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer New York, 2006.

[Bro98] S. P. Brooks. MCMC convergence diagnosis via multivariate bounds on log-concave densities. *Ann. Stat.*, 26(1):398–433, 1998.

[BV04] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[BVZ01] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(11):1222–1239, 2001.

[BZB+01] C. Bowd, L. M. Zangwill, C. C. Berry, E. Z. Blumenthal, C. Vasile, C. Sanchez-Galeana, C. F. Bosworth, P. A. Sample, and R. N. Weinreb. Detecting early glaucoma by assessment of retinal nerve fiber layer thickness and visual function. *Invest. Ophthalmol. Vis. Sci.*, 42(9):1993–2003, 2001.

[CET98] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In *European Conference on Computer Vision (ECCV 1998)*, pages 484–498. Springer, 1998.

[CHKM07] M. Cerda, N. Hitschfeld-Kahler, and D. Mery. Robust tree-ring detection. In *Advances in Image and Video Technology (PSIVT 2007)*, pages 575–585. Springer, 2007.

[CKFB09] R. T. Chang, O. Knight, W. J. Feuer, and D. L. Budenz. Sensitivity and specificity of time-domain versus spectral-domain optical coherence tomography in diagnosing early to moderate glaucoma. *Ophthalmology*, 116(12):2294–2299, 2009.

[Cli90] P. Clifford. Markov random fields in statistics. In *Disorder in physical systems*, pages 19–32. Oxford University Press, 1990.

[CRD07] D. Cremers, M. Rousson, and R. Deriche. A review of statistical approaches to level set segmentation: integrating color, texture, motion and shape. *Int. J. Comput. Vision*, 72(2):195–215, 2007.

[CS13] Y. Chen and R. J. Samworth. Smoothed log-concave maximum likelihood estimation with applications. *Statist. Sinica*, 23, 2013.

[CSS10] M. Cule, R. Samworth, and M. Stewart. Maximum likelihood estimation of a multi-dimensional log-concave density. *J. R. Stat. Soc. Series B Stat. Methodol.*, 72(5):545–607, 2010.

[CSYI03] M. A. Choma, M. V. Sarunic, C. Yang, and J. A. Izatt. Sensitivity advantage of swept source and fourier domain optical coherence tomography. *Opt. Express*, 11(18):2183–2189, 2003.

[CT06] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2006.

[CTCG95] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models-their training and application. *Comput. Vis. Image Und.*, 61(1):38–59, 1995.

[dBCP+03] J. F. de Boer, B. Cense, B. H. Park, M. C. Pierce, G. J. Tearney, and B. E. Bouma. Improved signal-to-noise ratio in spectral-domain compared with time-domain optical coherence tomography. *Opt. Lett.*, 28(21):2067–2069, 2003.

[DCA+13] P. Dufour, L. Ceklic, H. Abdillahi, S. Schroder, S. De Dzanet, U. Wolf-Schnurrbusch, and J. Kowal. Graph-based multi-surface segmentation of OCT data using trained hard and soft constraints. *IEEE Trans. Med. Imaging*, 32(3):531–543, 2013.

[DF08] W. Drexler and J. Fujimoto. State-of-the-art retinal optical coherence tomography. *Prog. Retin. Eye. Res.*, 27(1):45–88, 2008.

[DFJ54] G. Dantzig, R. Fulkerson, and S. Johnson. Solution of a large-scale traveling-salesman problem. *Oper. Res.*, 2(4):393–410, 1954.

[DHDP13] J. F. Dreijer, B. M. Herbst, and J. A. Du Preez. Left ventricular segmentation from MRI datasets with edge modelling conditional random fields. *BMC Med. Imaging*, 13(1):24, 2013.

[DJD88] S. W. Dharmadhikari and K. Joag-Dev. *Unimodality, Convexity and Applications*, volume 8. Academic Press New York, 1988.

[DR09] L. Dümbgen and K. Rufibach. Maximum likelihood estimation of a log-concave density and its distribution function: Basic properties and uniform consistency. *Bernoulli*, 15(1):40–68, 2009.

[EKKN13] A. Eslami, A. Karamalis, A. Katouzian, and N. Navab. Segmentation by retrieval with guided random walks: Application to left ventricle segmentation in MRI. *Med. Image Anal.*, 17(2):236–253, 2013.

[FHD+93] A. Fercher, C. Hitzenberger, W. Drexler, G. Kamp, H. Sattmann, et al. In vivo optical coherence tomography. *Am. J. Ophthalmol.*, 116(1):113, 1993.

[FHT08] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[FKP94] A. Frieze, R. Kannan, and N. Polson. Sampling from log-concave distributions. *Ann. Appl. Probab.*, 4(3):812–837, 1994.

[FSP05] D. C. Fernández, H. M. Salinas, and C. A. Puliafito. Automated detection of retinal layer structures on optical coherence tomography images. *Opt. Express*, 13(25):10200–10216, 2005.

[GAW+09] M. K. Garvin, M. D. AbrÃămoff, X. Wu, S. R. Russell, T. L. Burns, and M. Sonka. Automated 3-d intraretinal layer segmentation of macular spectral-domain optical coherence tomography images. *IEEE Trans. Med. Imaging*, 28(9):1436–1447, Sep 2009.

[Gre56] U. Grenander. On the theory of mortality measurement: Part II. *Scand. Actuar. J.*, 1956(2):125–153, 1956.

[HD13] L. A. Hannah and D. B. Dunson. Multivariate convex regression with adaptive partitioning. *JMRL*, 14(1):3261–3294, 2013.

[Hec87] E. Hecht. *Optics*. Addison-Wesley, 2nd edition, 1987.

[HFN73] W. F. Hoyt, L. Frisen, and N. M. Newman. Fundoscopy of nerve fiber layer defects in glaucoma. *Invest. Ophthalmol. Vis. Sci.*, 12(11):814–829, 1973.

[Hot33] H. Hotelling. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.*, 24(6):417, 1933.

[HSL+91] D. Huang, E. A. Swanson, C. P. Lin, J. S. Schuman, W. G. Stinson, W. Chang, M. R. Hee, T. Flotte, K. Gregory, C. A. Puliafito, et al. Optical coherence tomography. *Science*, 254(5035):1178–1181, 1991.

[Hub95] D. H. Hubel. *Eye, Brain, and Vision*. Scientific American Library New York, 1995.

[ISW+05] H. Ishikawa, D. M. Stein, G. Wollstein, S. Beaton, J. G. Fujimoto, and J. S. Schuman. Macular segmentation with optical coherence tomography. *Invest. Ophthalmol. Vis. Sci.*, 46(6):2012–2017, 2005.

[JJ99] H. Jeffreys and B. Jeffreys. *Methods of Mathematical Physics*. Cambridge University Press, 1999.

[KAH+13] J. H. Kappes, B. Andres, F. A. Hamprecht, C. Schnörr, S. Nowozin, D. Batra, S. Kim, B. X. Kausler, J. Lellmann, N. Komodakis, et al. A comparative study of modern inference techniques for discrete energy minimization problems. In *Conference on Computer Vision and Pattern Recognition (CVPR 2013)*, 2013.

[KF09] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[KFKA09] Y. H. Kwon, J. H. Fingert, M. H. Kuehn, and W. L. Alward. Primary open-angle glaucoma. *New Engl. J. Med.*, 360(11):1113–1124, 2009.

[KHH+02] M. A. Kass, D. K. Heuer, E. J. Higginbotham, C. A. Johnson, J. L. Keltner, J. P. Miller, I. Parrish, K. Richard, M. R. Wilson, M. O. Gordon, et al. The ocular hypertension treatment study: a randomized trial determines that topical ocular hypotensive medication delays or prevents the onset of primary open-angle glaucoma. *Arch. Ophthalmol.*, 120(6):701, 2002.

[KHH+11] Y. Kotera, M. Hangai, F. Hirose, S. Mori, and N. Yoshimura. Three-dimensional imaging of macular inner structures in glaucoma by using spectral-domain optical coherence tomography. *Invest. Ophthalmol. Vis. Sci.*, 52(3):1412–1421, 2011.

[KM10] R. Koenker and I. Mizera. Quasi-concave density estimation. *Ann. Stat.*, 38(5):2998–3027, 2010.

[Kol06] V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(10):1568–1583, 2006.

[KPH+10] V. Kajić, B. Povazay, B. Hermann, B. Hofer, D. Marshall, P. L. Rosin, and W. Drexler. Robust segmentation of intraretinal layers in the normal human fovea using a novel statistical model based on texture and shape analysis. *Opt. Express*, 18(14):14730–14744, 2010.

[KVGH+99] D. Kamal, A. Viswanathan, D. Garway-Heath, R. Hitchings, D. Poinoosawmy, and C. Bunce. Detection of optic disc change with the heidelberg retina tomograph before confirmed visual field change in ocular hypertensives converting to early glaucoma. *Br. J. Ophthalmol.*, 83(3):290–294, 1999.

[KWT88] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *Int. J. Comput. Vision*, 1(4):321–331, 1988.

[Lau96] S. L. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

[LCC+05] C. K.-s. Leung, K. K.-L. Chong, W.-m. Chan, C. K.-F. Yiu, M.-y. Tso, J. Woo, M.-K. Tsang, K.-k. Tse, and W.-h. Yung. Comparative study of retinal nerve fiber layer measurement by StratusOCT and GDx VCC, II: structure/function regression analysis in glaucoma. *Invest. Ophthalmol. Vis. Sci.*, 46(10):3702–3711, 2005.

[LGF00] M. E. Leventon, W. E. L. Grimson, and O. Faugeras. Statistical shape influence in geodesic active contours. In *Computer Vision and Pattern Recognition (CVPR 2000)*, volume 1, pages 316–323, 2000.

[LHF+03] R. Leitgeb, C. Hitzenberger, A. F. Fercher, et al. Performance of fourier domain vs. time domain optical coherence tomography. *Opt. Express*, 11(8):889–894, 2003.

[LRZ+11] M. T. Leite, H. L. Rao, L. M. Zangwill, R. N. Weinreb, and F. A. Medeiros. Comparison of the diagnostic accuracies of the Spectralis, Cirrus, and RTVue optical coherence tomography devices in glaucoma. *Ophthalmology*, 118(7):1334–1339, 2011.

[LS88] S. L. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *J. R. Statist. Soc. B*, 50(2):157–224, 1988.

[MB06] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *Ann. Stat.*, 34(3):1436–1462, 2006.

[MB09] A. Magnani and S. P. Boyd. Convex piecewise-linear fitting. *Optim. Eng.*, 10(1):1–17, 2009.

[MGF08] M. Mozaffarieh, M. C. Grieshaber, and J. Flammer. Oxygen and blood flow: players in the pathogenesis of glaucoma. *Mol. Vis.*, 14:224, 2008.

[MHMT10] M. A. Mayer, J. Hornegger, C. Y. Mardin, and R. P. Tornow. Retinal nerve fiber layer segmentation on FD-OCT scans of normal subjects and glaucoma patients. *Biomed. Opt. Express*, 1(5):1358–1383, 2010.

[MMG08] P. Messmer, P. J. Mullowney, and B. E. Granger. GPULib: GPU computing in high-level languages. *Comput. Sci. Eng.*, 10(5):70–73, 2008.

[MTRP09] C. McGrory, D. Titterington, R. Reeves, and A. Pettitt. Variational Bayes for estimating the parameters of a hidden Potts model. *Stat. Comput.*, 19:329–340, 2009.

[MWBC09] A. Mishra, A. Wong, K. Bizheva, and D. A. Clausi. Intra-retinal layer segmentation in optical coherence tomography images. *Opt. Express*, 17(26):23719–23728, 2009.

[MWJ99] K. P. Murphy, Y. Weiss, and M. I. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *Uncertainty in Artificial Intelligence (UAI 1999)*, pages 467–475, 1999.

[Nea93] R. M. Neal. Probabilistic inference using Markov Chain Monte Carlo methods, 1993.

[NVT+13] J. Novosel, K. A. Vermeer, G. Thepass, H. G. Lemij, and L. J. van Vliet. Loosely coupled level sets for retinal layer segmentation in optical coherence tomography. In *International Symposium on Biomedical Imaging (ISBI 2013)*, pages 1010–1013, 2013.

[OS88] S. Osher and J. A. Sethian. Fronts propagating with curvature-dependent speed: algorithms based on Hamilton-Jacobi formulations. *J. Comput. Phys.*, 79(1):12–49, 1988.

[Par62] E. Parzen. On estimation of a probability density function and mode. *Ann. of Math. Stat.*, pages 1065–1076, 1962.

[Pea88] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[PP12] K. B. Petersen and M. S. Pedersen. The matrix cookbook. Technical report, Technical University of Denmark, 2012.

[QA+82] H. A. Quigley, E. M. Addicks, et al. Quantitative studies of retinal nerve fiber layer defects. *Arch. Ophthalmol.*, 100(5):807–814, 1982.

[QKD+92] H. Quigley, J. Katz, R. Derick, D. Gilbert, A. Sommer, et al. An evaluation of optic disc and nerve fiber layer examinations in monitoring progression of early glaucoma damage. *Ophthalmology*, 99(1):19–28, 1992.

[Qui99] H. A. Quigley. Neuronal death in glaucoma. *Prog. Retin. Eye Res.*, 18(1):39–57, 1999.

[R+71] R. Rockafellar et al. Integrals which are convex functionals, II. *Pacific J. Math*, 39(2):439–469, 1971.

[RH05] H. Rue and L. Held. *Gaussian Markov Random Fields: Theory and Applications*. Chapman & Hall/CRC, 2005.

[Roc70] R. T. Rockafellar. *Convex Analysis*, volume 28. Princeton University Press, 1970.

[RSS11] F. Rathke, S. Schmidt, and C. Schnörr. Order preserving and shape prior constrained intra-retinal layer segmentation in optical coherence tomography. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2011)*, volume 6893, pages 370–377. Springer, 2011.

[RSS14] F. Rathke, S. Schmidt, and C. Schnörr. Probabilistic intra-retinal layer segmentation in 3-D OCT images using global shape regularization. *Med. Image Anal.*, 18(5):781–794, 2014.

[Ruf07] K. Rufibach. Computing maximum likelihood estimators of a log-concave density function. *J. Statist. Comput. Simulation*, 77(7):561–574, 2007.

[RW06] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[RWW09] R. T. Rockafellar, R. J.-B. Wets, and M. Wets. *Variational Analysis*, volume 317. Springer, 3rd edition, 2009.

[RyC11] S. Ramòn y Cajal. *Histologie du systéme nerveux de l'homme & des vertébrés*. Paris, Maloine, 1911.

[SBG+13] Q. Song, J. Bai, M. Garvin, M. Sonka, J. Buatti, X. Wu, et al. Optimal multiple surface segmentation with shape and context priors. *IEEE Trans. Med. Imag.*, 32(2):376–386, 2013.

[Sch99] H. Schubert. Anatomy and physiology: structure and function of the neural retina. In M. Yanoff and J. Duker, editors, *Ophthalmology*. London: Mosby, 1999.

[SHB14] M. Sonka, V. Hlavac, and R. Boyle. *Image Processing, Analysis, and Machine Vision*. Cengage Learning, 2014.

[SHP+95] J. S. Schuman, M. R. Hee, C. A. Puliafito, C. Wong, T. Pedut-Kloizman, C. P. Lin, E. Hertzmark, J. A. Izatt, E. A. Swanson, J. G. Fujimoto, et al. Quantification of nerve fiber layer thickness in normal and glaucomatous eyes using optical coherence tomography. *Arch. Ophthalmol.*, 113(5):586–596, 1995.

[SIH+93] E. A. Swanson, J. Izatt, M. R. Hee, D. Huang, C. Lin, J. Schuman, C. Puliafito, and J. G. Fujimoto. In vivo retinal imaging by optical coherence tomography. *Opt. Lett.*, 18(21):1864–1866, 1993.

[SIH+06] D. Stein, H. Ishikawa, R. Hariprasad, G. Wollstein, R. Noecker, J. Fujimoto, and J. Schuman. A new quality assessment parameter for optical coherence tomography. *Br. J. Ophthalmol.*, 90(2):186–190, 2006.

[Sil82] B. W. Silverman. On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Stat.*, 10(3):795–810, 1982.

[SMP+77] A. Sommer, N. R. Miller, I. Pollack, A. E. Maumenee, T. George, et al. The nerve fiber layer in the diagnosis of glaucoma. *Arch. Ophthalmol.*, 95(12):2149–2156, 1977.

[SPF04] J. Schuman, C. Puliafito, and J. Fujimoto. *Optical Coherence Tomography of Ocular Diseases*. Slack Incorporated, 2004.

[Stu08] B. Student. The probable error of a mean. *Biometrika*, 6(1):1–25, 1908.

[SW10] A. Seregin and J. A. Wellner. Nonparametric estimation of multivariate convex-transformed densities. *Ann. Stat.*, 38(6):3751–3781, 2010.

[SXY99] J. M. Schmitt, S. Xiang, and K. M. Yung. Speckle in optical coherence tomography. *J. Biomed. Opt.*, 4(1):95–105, 1999.

[TB99] M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *J. R. Stat. Soc.*, 61(3):611–622, 1999.

[TCL+09] O. Tan, V. Chopra, A. T.-H. Lu, J. S. Schuman, H. Ishikawa, G. Wollstein, R. Varma, and D. Huang. Detection of macular ganglion cell loss in glaucoma by fourier-domain optical coherence tomography. *Ophthalmology*, 116(12):2305–2314, 2009.

[Tib96] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B Stat. Methodol.*, 58(1):267–288, 1996.

[TKK+11] S. Tsantis, G. C. Kagadis, K. Katsanos, D. Karnabatidis, G. Bourantas, and G. C. Nikiforidis. Automatic vessel lumen segmentation and stent strut detection in intravascular optical coherence tomography. *Med. Phys.*, 39(1):503–513, 2011.

[TLL+08] O. Tan, G. Li, A. T.-H. Lu, R. Varma, D. Huang, et al. Mapping of macular substructures with optical coherence tomography for glaucoma diagnosis. *Ophthalmology*, 115(6):949–956, 2008.

[UAO+12] G. J. Ughi, T. Adriaenssens, K. Onsea, P. Kayaert, C. Dubois, P. Sinnaeve, M. Coosemans, W. Desmet, and J. D'hooge. Automatic segmentation of in-vivo intra-coronary optical coherence tomography images to assess stent strut apposition and coverage. *Int. J. Cardiovas. Imag.*, 28(2):229–241, 2012.

[VvdSLdB11] K. A. Vermeer, J. van der Schoot, H. G. Lemij, and J. F. de Boer. Automated segmentation by pixel classification of retinal layers in ophthalmic OCT images. *Biomed. Opt. Express*, 2(6):1743–1756, 2011.

[WGHH+98] G. Wollstein, D. F. Garway-Heath, R. A. Hitchings, et al. Identification of early glaucoma cases with the scanning laser ophthalmoscope. *Ophthalmology*, 105(8):1557–1563, 1998.

[WJ08] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1(1-2):1–305, 2008.

[WJW03] M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky. Tree-reweighted belief propagation algorithms and approximate ML estimation by pseudo-moment matching. In *Workshop on Artificial Intelligence and Statistics (AISTATS 2003)*, volume 21, pages 97–105, 2003.

[WK04] R. N. Weinreb and P. T. Khaw. Primary open-angle glaucoma. *The Lancet*, 363(9422):1711–1720, 2004.

[WLK+02] M. Wojtkowski, R. Leitgeb, A. Kowalczyk, T. Bajraszewski, and A. F. Fercher. In vivo human retinal imaging by Fourier domain optical coherence tomography. *J. Biomed. Opt.*, 7(3):457–463, 2002.

[YFW+01] J. S. Yedidia, W. T. Freeman, Y. Weiss, et al. Generalized belief propagation. In *Neural Information Processing Systems (NIPS 2001)*, volume 13, pages 689–695, 2001.

[YHSB+03] F. Yang, G. Holzapfel, C. Schulze-Bauer, R. Stollberger, D. Thedens, L. Bolinger, A. Stolpen, and M. Sonka. Segmentation of wall and plaque in in vitro vascular MR images. *Int. J. Cardiovas. Imag.*, 19(5):419–428, 2003.

[YHSS11] A. Yazdanpanah, G. Hamarneh, B. R. Smith, and M. V. Sarunic. Segmentation of intra-retinal layers from optical coherence tomography images using an active contour approach. *IEEE Trans. Med. Imaging*, 30(2):484–496, 2011.

[YRW+10] Q. Yang, C. A. Reisman, Z. Wang, Y. Fukuma, M. Hangai, N. Yoshimura, A. Tomidokoro, M. Araie, A. S. Raza, D. C. Hood, and K. Chan. Automated layer segmentation of macular OCT images using dual-scale gradient information. *Opt. Express*, 18(20):21293–21307, 2010.

[ZNO+07] A. M. Zysk, F. T. Nguyen, A. L. Oldenburg, D. L. Marks, and A. S. Boppart. Optical coherence tomography: a review of clinical development from bench to bedside. *J. Biomed. Opt.*, 12(5):1–21, 2007.