TKK Dissertations in Information and Computer Science
Espoo 2008 TKK-ICS-D8

INFERRING RELEVANCE FROM EYE MOVEMENTS WITH WRONG MODELS

Jarkko Salojärvi

Dissertation for the degree of Doctor of Science in Technology to be presented with due permission of the Faculty of Information and Natural Sciences for public examination and debate in Auditorium T2 at Helsinki University of Technology (Espoo, Finland) on the 21st of November, 2008, at 12 o'clock noon.

Helsinki University of Technology
Faculty of Information and Natural Sciences
Department of Information and Computer Science

Teknillinen Korkeakoulu
Informaatio- ja luonnontieteiden tiedekunta
Tietojenkäsittelytieteen laitos

Distribution:

Helsinki University of Technology
Faculty of Information and Natural Sciences
Department of Information and Computer Science
P.O.Box 5400
FIN-02015 TKK
Finland

Tel. +358-9-451 3272
Fax +358-9-451 3277
Email: series@ics.tkk.fi

ISBN 978-951-22-9612-5 (Print)
ISBN 978-951-22-9613-2 (Online)
ISSN 1797-5050 (Print)
ISSN 1797-5069 (Online)
URL: `http://lib.tkk.fi/Diss/2008/isbn9789512296132/`

Multiprint Oy
Espoo 2008

ABSTRACT

Salojärvi, J. (2008): **Inferring Relevance from Eye Movements with Wrong Models.** Doctoral thesis, Helsinki University of Technology, Dissertations in Information and Computer Science, TKK-ICS-D8, Espoo, Finland.

Keywords: Conditional inference, incorrect model, probabilistic inference, eye movements, proactive information retrieval.

Statistical inference forms the backbone of modern science. It is often viewed as giving an objective validation for hypotheses or models. Perhaps for this reason the theory of statistical inference is often derived with the assumption that the "truth" is within the model family. However, in many real-world applications the applied statistical models are incorrect. A more appropriate probabilistic model may be computationally too complex, or the problem to be modelled may be so new that there is little prior information to be incorporated. However, in statistical theory the theoretical and practical implications of the incorrectness of the model family are to a large extent unexplored.

This thesis focusses on conditional statistical inference, that is, modeling of classes of future observations given observed data, under the assumption that the model is incorrect. Conditional inference or prediction is one of the main application areas of statistical models which is still lacking a conclusive theoretical justification of Bayesian inference. The main result of the thesis is an axiomatic derivation where, given an incorrect model and assuming that the utility is conditional likelihood, a *discriminative posterior* yields a distribution on model parameters which best agrees with the utility. The devised discriminative posterior outperforms the classical Bayesian joint likelihood-based approach in conditional inference. Additionally, a theoretically justified expectation maximization-type algorithm is presented for obtaining conditional maximum likelihood point estimates for conditional inference tasks. The convergence of the algorithm is shown to be more stable than in earlier partly heuristic variants.

The practical application field of the thesis is inference of relevance from eye movement signals in an information retrieval setup. It is shown that relevance can be predicted to some extent, and that this information can be exploited in a new kind of task, proactive information retrieval. Besides making it possible to design new kinds of engineering applications, statistical modeling of eye tracking data can also be applied in basic psychological research to make hypotheses of cognitive processes affecting eye movements, which is the second application area of the thesis.

TIIVISTELMÄ

Salojärvi, J. (2008): **Mielenkiinnonkohteiden päättely silmänliikkeistä vääriä malleja käyttäen.** Väitöskirja, Teknillinen Korkeakoulu, Dissertations in Information and Computer Science, TKK-ICS-D8, Espoo, Suomi.

Avainsanat: Ehdollinen mallitus, probabilistinen mallintaminen, väärä malli, silmänliikkeet, proaktiivinen tiedonhaku.

Tilastolliset menetelmät muodostavat modernin tieteen selkärangan. Menetelmien ajatellaan usein tarjoavan objektiivisen vahvistuksen hypoteeseille ja malleille, ja ehkä juuri tästä syystä tilastollisen päättelyn teoria onkin kehitetty lähtien oletuksesta, että "oikea" malli sisältyy käytettyyn malliperheeseen. Monissa käytännön sovelluksissa käytetyt mallit ovat kuitenkin vääriä. Tilannetta paremmin kuvaava malli voi esimerkiksi olla laskennallisesti liian raskas muodostaa, tai tutkimusta tehdään uudella alalla, jolloin paremman mallin kehittäminen ei onnistu puuttuvan tutkimustiedon takia. Tilastollisen mallituksen teoriassa väärien mallien käytön teoreettisia ja käytännön vaikutuksia on kuitenkin tutkittu vain vähän.

Väitöskirjassa tutkitaan ehdollista tilastollista päättelyä, eli havaintojen luokan ennustamista annettuna havaittu mittausaineisto, kun otetaan huomioon se, että käytetty malli on väärä. Vaikka ehdollinen päättely on yksi tilastollisten mallien pääsovellusaloista, puuttuu siltä yhä täysi teoreettinen perustelu Bayesilaisen mallituksen näkökulmasta. Väitöstyön päätulos on aksiomaattinen todistus sille, että kun käytössä on väärä malli ja tehtävänä on ennustaa ehdollinen todennäköisyys, niin sanottu ehdollinen likelihood, on *diskriminatiivinen posteriori* parhaiten tehtävään sopiva jakauma mallin parametriavaruudessa. Johdettu jakauma toimii ehdollisessa päättelyssä paremmin kuin klassinen Bayesilainen yhteisjakaumaan perustuva mallitus. Tämän lisäksi työssä esitellään teoreettisesti perusteltu expectation-maximization-tyyppinen algoritmi ehdollisen likelihoodin piste-estimaattien laskemiseen. Algoritmin osoitetaan suppenevan vakaammin kuin aiempien osittain heurististen menetelmien.

Väitöskirjan käytännön sovellus on mielenkiinnon kohteiden päättely silmänliikkeistä tiedonhakutehtävissä. Työssä osoitetaan, että kiinnostuksen päätteleminen on mahdollista, ja että kyseistä tietoa voidaan käyttää hyväksi täysin uudenlaisessa ennakoivan tiedonhaun sovelluksessa. Väitöstyön toisessa sovelluskohteessa silmänliikkeiden tilastollista mallitusta käytetään psykologiseen perustutkimukseen rakentamaan hypoteeseja silmänliikkeiden taustalla olevista kognitiivisista prosesseista.

Contents

Preface

This thesis work has been carried out in the Adaptive Informatics Research Centre of the Laboratory of Computer and Information Science and, since 2008, the Department of Information and Computer Science at the Helsinki University of Technology. I also have the pleasure of being a part of the Helsinki Institute for Information Technology. The work has been funded by the Academy of Finland through the PROACT programme and by the European Union under the PASCAL Network of Excellence, IST-2002-506778. Additionally, I have received funding from the Helsinki Graduate School in Computer Science and Engineering (HeCSE).

I wish to thank my supervisor Professor Samuel (Sami) Kaski for providing the optimal creative environment to work in, allowing me a great degree of scientific freedom in my research, while at the same time giving me the necessary amount of much-needed guidance. Without his attitude and understanding of life in general this work would not have been possible. I would also like to thank Doctor Kai Puolamäki for his deep mathematical intuition and relaxed attitude even in the most stressful of times before the paper deadlines.

I would also like to thank the reviewers of this thesis, Professor Petri Myllymäki and Doctor Guillaume Bouchard, for their valuable feedback.

I am also greatly indebted to my other co-authors besides Sami and Kai: Eerika Savia, Jaana Simola and Ilpo Kojo. Writing the papers with you and knowing you has been a pleasure.

I would also like to thank all the members of our MI research group, former and present (especially Jarkko Venna, Janne Nikkilä, Arto Klami, Janne Sinkkonen, Jaakko Peltonen, Merja Oja, Abhishek Tripathi) for valuable discussions on both science and life in general.

Additionally, I would like to express my gratitude to heavy metal music for giving me the correct attitude when dabbling with the fine points of probabilistic modeling — take only the essential and hit it with the largest hammer that you've got.

Finally, thanks to all my friends for their support and constant interest in the state of my thesis over the last years. I also wish to thank all the members of the Plant Stress group of University of Helsinki, headed by professor Jaakko Kangasjärvi. Getting to know each of you while learning a new research area has been a exhilarating experience.

I am grateful to my mother and Jouko for giving me support in the worst of times. Most of all, I wish to thank my children Timo and Anna for being there, and my love Reetta. With you three around I gladly face the prospect of getting a life after the thesis.

LIST OF PUBLICATIONS

The thesis consists of an introduction and the following publications:

1. Jarkko Salojärvi, Ilpo Kojo, Jaana Simola and Samuel Kaski. Can Relevance Be Inferred from Eye Movements in Information Retrieval? In: *Proceedings of the Workshop on Self-Organizing Maps (WSOM'03)*, pp 261–266, Hibikino, Japan, 2003.

2. Jarkko Salojärvi, Kai Puolamäki, and Samuel Kaski. Relevance Feedback from Eye Movements for Proactive Information Retrieval. In: *Heikkilä, Pietikäinen, Silvén (Eds.): Workshop on Processing Sensory Information for Proactive Systems (PSIPS 2004)*, pp. 37–42, Oulu, Finland, 2004.

3. Jarkko Salojärvi, Kai Puolamäki, and Samuel Kaski. Implicit Relevance Feedback from Eye Movements. In: W. Duch, J. Kacprzyk, E. Oja and S. Zadrozny, editors, *Artificial Neural Networks: Biological Inspirations, ICANN 2005*, Lecture Notes in Computer Science 3696, pp. 513–518. Springer-Verlag, Berlin, Germany, 2005.

4. Kai Puolamäki, Jarkko Salojärvi, Eerika Savia, Jaana Simola, and Samuel Kaski. Combining Eye Movements and Collaborative Filtering for Proactive Information Retrieval. In: G. Marchionini, A. Moffat, J. Tait, R. Baeza-Yates and N. Ziviani, editors. *SIGIR'05: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* pp. 146–153. ACM press, New York, USA, 2005.

5. Jarkko Salojärvi, Kai Puolamäki, and Samuel Kaski. Expectation Maximization Algorithms for Conditional Likelihoods. In: L- D. Raedt and S. Wrobel, editors. *Proceedings of the 22nd International Conference on Machine Learning (ICML-2005)*, pp. 753–760. ACM press, New York, USA, 2005.

6. Jarkko Salojärvi, Kai Puolamäki, and Samuel Kaski. On Discriminative Joint Density Modeling. In: J. Gama, R. Camacho, P. Brazdil, A. Jorge and L. Torgo, editors. *Machine Learning: ECML 2005.* Lecture Notes in Artificial Intelligence 3720, pp. 341–352. Springer-Verlag, Berlin, Germany, 2005.

7. Jarkko Salojärvi, Kai Puolamäki, Eerika Savia, and Samuel Kaski. Inference with Discriminative Posterior. Submitted to a journal.

8. Jaana Simola, Jarkko Salojärvi, and Ilpo Kojo. Using hidden Markov model to uncover processing states from eye movements in information search tasks. *Cognitive Systems Research* 9:237–251, 2008.

SUMMARY OF PUBLICATIONS AND THE AUTHOR'S CONTRIBUTION

Most of the ideas presented in the publications have been developed as a team work. Giving individual credit is therefore, to a large part, not meaningful.

Publication 1 is a first feasibility study of its kind on inferring relevance from eye movements in information retrieval tasks. Design of the experimental setup was a joint effort; data preprocessing and analysis was carried out by the author. The author wrote most of the paper.

Publication 2 applies hidden Markov models as well as discriminative hidden Markov models, optimized with Viterbi learning, for inferring relevance from eye movements for the first time. The author wrote most of the paper. The experimental setup was designed together. Experiments were carried out together with Kai Puolamäki.

Publication 3 reports results from a larger scale information retrieval experiment. The data collected here was used in a Pascal EU Network of Excellence challenge. The methods are further developments from Publication 2; the new learning algorithm developed in Publication 5 was applied for learning the discriminative hidden Markov model. The author wrote most of the paper. The experimental setup was designed together. Experiments were carried out by the author.

Publication 4 is a feasibility study on whether the relevance predictions from collaborative filtering can be complemented and improved by implicit feedback measured from eye movements. Design of the experimental setup was a joint effort. The discriminative Dirichlet mixture model was the author's idea. The author implemented most of the experiments. Writing of the paper was a collaborative effort.

Publication 5 develops a new algorithm for discriminative learning of generative models. The author wrote most of the paper. Theory was devised together with Kai Puolamäki. Experiments were designed and carried out by the author.

Publication 6 is a first attempt of explaining the theory of discriminative generative models. The author is responsible for most of the paper.

Publication 7 generalizes the theoretical developments of Publication 6 from point estimates into distributions, as well as presents the theory in a more elegant manner. Theoretical development as well as writing of the paper was a joint effort. The author designed and carried out the experiments.

Publication 8 applies the discriminative hidden Markov models developed in this thesis to psychological analysis of eye movements in information search tasks. Except for the analysis with linear mixed models, the author carried out all the data analysis. The paper was written together with Jaana Simola.

LIST OF ABBREVIATIONS AND SYMBOLS

$\boldsymbol{\beta}$	Parameter vector of a discriminative model
C	The number of classes
$\mathcal{D}, \mathbf{x}, c$	Denotes data
$D_\kappa(\cdot, \cdot)$	Bregman divergence for a convex function κ
$\kappa(\cdot)$	Log-partition function
$\mathcal{L}(X; \theta)$	Likelihood
$\mathcal{L}(C\|X; \theta)$	Conditional likelihood
μ	The dual parameters of θ for an exponential family distribution
N	The number of data items
S	The number of hidden states
$T(\mathbf{x})$	Sufficient statistics of $\mathbf{x}$
Θ	Used for denoting the parameter space of all models
$\overline{\Theta}$	Used for denoting the parameter space of realizable models
θ	Used for denoting the parameters of a given model, $\theta \in \overline{\Theta}$
$\tilde{\Theta}$	Used for denoting the parameters of a "true" model, $\theta \notin \overline{\Theta}$
$u(\cdot)$	Utility
CML	Conditional Maximum Likelihood
dEM	discriminative Expectation − Maximization
dMCMC	discriminative Markov chain Monte Carlo
EM	Expectation − Maximization
MCMC	Markov chain Monte Carlo
ML	Maximum Likelihood
int	Interior
dom	Domain

Chapter 1

Introduction

All models are wrong, some are useful.

(Box and Draper, 1987)

The goal of science is to develop useful models of reality. A model is information distilled from reality; it states an explicit hypothesis, with the aim of providing generalisation to other similar events. This forms the basis of science: the constructed models can be directly tested.

The focus of this thesis is on conditional inference with probabilistic models. The aim of probabilistic modeling is to take into account the randomness in a system. A deterministic system is a special case where something happens with probability one (or zero). In the other extreme, the system is uniformly random if random guessing, a coin flip, performs as well as any model developed for prediction. The use of probabilistic models is often referred to as statistical inference. Conditional inference is a special case where the probabilistic model is used for predicting unobserved quantities, given observations which are assumed to be relevant by the model. A practical example of statistical conditional inference is validation of a hypothesis by measuring a statistically significant difference between the predictions of the hypothesis and random behavior. This is a very crude example; the random behavior, the *null hypothesis*, can actually be a complex model in itself, if the most general framework, the so called Bayes factors (Kass and Raftery, 1995) are applied for determining the difference.

The main focus of this thesis within probabilistic modeling is on conditional inference in cases where the given model is known to be incorrect. In science, the model does not have to be the "truth" and, furthermore, it is possible that the "truth" cannot be modelled (this is why science considers only theories and hypotheses). As an example from physics, it has turned out that it is not possible to construct a "true" deterministic model describing the interactions of a system of particles. The reason lies partly in the computational complexity due to the multitude of possible interactions, but also at the fundamental level due to the inaccuracy of measurements; uncertainty in location or momentum according to Heisenberg's uncertainty principle. Another situation where the devised model may be incorrect occurs when a new research area is studied. When little prior information on the phenomenon is available, the models can naturally be quite far from being correct.

Probabilistic inference is often viewed as giving an objective validation of hypotheses. Perhaps for this reason the theory of probabilistic inference is derived under the assumption that the model family contains also the "truth" (Bernardo and Smith, 2000; Cox, 1946) — an assumption which is not true in most real-world cases. Probability theory acknowledges the case where inference is done with an incorrect model, but no conclusive theoretical analysis on its effects has been presented yet. In this thesis it is assumed that an incorrect probabilistic model is given, and the goal is to use it for discriminative inference — a special case of conditional inference where the quantity to be predicted is categorical. The incorrect model is known to be useful to some extent; a decent effort has been made in order to model the underlying phenomenon, but the given model is still known to be imperfect in some respects. The theoretical development under these assumptions leads to a theorem that proves the existence of a *discriminative posterior*; a posterior-type distribution in the parameter space of the model which is optimal for conditional inference. To be more precise, optimal conditional inference is obtained by integrating over this posterior in a Bayesian manner.

The intuition behind the discriminative posterior is that if the true model cannot be achieved, the best option is to apply the most useful "wrong" model. Publication 7 considers a general case where, at the beginning before any observations are made, a distribution describing the initial beliefs on the models is assumed to exist. Then observations are brought in one at a time, and the intent is to update the beliefs in such a manner that in light of the observations the most useful models will gain more weight. Publication 7 provides axioms, a set of basic assumptions, which are needed in order to update the beliefs in this manner.

What remains to be defined is the usefulness of a model. In case of the above example from physics, even if the "truth" cannot be modeled, it is still possible to construct a probabilistic model that describes the general behavior of the system, such as the distribution of the particles. The model developed for this task is, in the strict sense, incorrect but still useful since it can be used for predictions. As another example, it is still possible to gather understanding within new fields of research even with models that are known to be inadequate, given that they predict events better than random guessing. Already from these examples it can be said that usefulness in general is a subjective concept, and depends on the goal of the modeler.

In Publication 7 the problem is formalized by assuming that a hypothetical "true" model exists, but it does not necessarily belong to the set of models under consideration. Both of the above examples can then be interpreted as cases where the given model is close enough to the "true" model in order to be useful. Furthermore, the usefulness can then be expressed by using the hypothetical "true" model: the most useful model is the one that minimizes the error between the predictions by the model and the "true" model.

The error between the models and the "true" model is related to a measure called *expected utility*. The term comes from decision-theory, where the rationale of the decision maker is to maximize this quantity. That is, a rational decision-maker chooses an action which is expected to result in the best utility at the time of making the decision. In the case of probabilistic modeling, utility is a real number output by each of the models when an observation is given as their input — this is typically the probability of the observation, likelihood. Expected utility is obtained as a sum of the real numbers, weighed according to the beliefs in each of the models. Expected utility is therefore maximized by a belief distribution that weighs more those models that have a high utility value.

As said earlier, the beliefs are evolved from initial beliefs by updating them by the usefulness of the models in earlier similar tasks. In standard Bayesian inference, the real number output by the model is the joint probability of the observation, joint likelihood. Therefore the updated beliefs are in agreement with the performance of the model in modeling the full observation. However, the goal of conditional inference is different, to predict a certain unobserved quantity given some already observed quantities. The main message of the thesis is that with incorrect models, modeling the full distribution does not give beliefs which are best suited for conditional inference. The reason is that in conditional inference the utility is different. Beliefs should be thus updated according to the usefulness of the models in the task which they will eventually be used for. That is, the real number output by the models should be directly related to the quantity of interest instead of the probability of full observation. By assuming a "true" model, this intuition can be expressed more explicitly by requiring that a high utility value is associated with a small error between the predictions of the given model and the "true" model.

Discriminative posterior is the main contribution of Chapter 3 of the thesis. Additionally, the relations to two most famous earlier axiomatic proofs of Bayesian inference are discussed: the decision-theoretic framework which was originally presented by de Finetti, and the less rigorous but more intuitive axioms of Cox.

There are cases where, instead of computing a distribution over the different models, only one specific model is desired. For example, if the model is assumed to be close enough to the truth, inspection of the model parameters could possibly bring more understanding to the phenomenon to be modelled. The so-called conditional maximum likelihood estimate is a point estimate from the discriminative posterior (assuming a uniform prior). The second main contribution of this thesis is to present a first computationally feasible expectation–maximization (discriminative EM) algorithm for obtaining conditional maximum likelihood estimates. In addition to presenting the discriminative EM algorithm, Chapter 4 discusses various numerical methods for making inference within an exponential model family. Also Markov chain Monte Carlo methods for sampling from discriminative posterior are presented.

The third main contribution of this thesis is probabilistic modeling of eye movements in information retrieval tasks, discussed in Chapter 5. A controlled experimental setup is constructed to show that the relevance of text read by the user can be predicted to some extent. Then a prototype of a *proactive information retrieval* application is shown to be feasible. The application predicts the preferences of users by combining relevance predictions inferred from eye movements with relevance predictions from ratings given by other similar-minded users. The predictions can then be used in the background, for example for finding more relevant documents. Finally, probabilistic modelling is applied in psychological research for inferring cognitive states occurring during information retrieval tasks. Discriminatively optimized joint density models were applied throughout the work.

Before presenting the contributions of the thesis, the practice of probabilistic modeling is discussed in Chapter 2, namely the practical tools of Bayesian inference, the exponential family, and different ways of constructing probabilistic models using this model family.

1.1 The Main Contributions of the Thesis

Main contributions of the thesis are:

- The theoretical contribution is discriminative modeling when the given model family is known to be incorrect. It is shown that in this case, a discriminative posterior can be used for conditional inference that is consistent with conditional (marginal) likelihood.

- Markov Chain Monte Carlo methods are introduced for sampling from the discriminative posterior.

- The first computationally feasible Expectation-Maximization algorithm for finding a point estimate for conditional maximum likelihood is presented.

- Advanced machine learning methods are developed and applied for inferring relevance from eye movements in an information retrieval setup. A controlled experimental setup is designed where it is shown that relevance of read text can be inferred from eye movements. The measured eye movement data was used in a PASCAL Network of Excellence machine learning challenge during 2005.

- A prototype of a proactive information retrieval application is introduced which combines implicit relevance feedback inferred from eye movements with relevance information from other available sources.

Chapter 2

Probabilistic Modeling

> *Not to be absolutely certain is, I think, one of the essential things in rationality.*
>
> Bertrand Russell

This chapter describes the principles, tools, and practices of probabilistic modeling. The common assumptions and procedures that are usually followed are discussed first, namely the likelihood principle, point estimates, and the principle of marginalization. Next, the workhorse of probabilistic modeling is introduced — the exponential model family. Any distribution can be modeled arbitrarily closely with exponential family distributions, that is, they are *universal density estimators* (McLachlan and Peel, 2000). This fact can be seen as the justification for the choice of using the exponential model family. In principle, any other family of functions having the same property could be used; the exponential family is chosen because of mathematical convenience. The basic properties of the model family will be gone through with some detail, since they will be needed in Chapter 4 discussing different optimization methods for exponential family distributions. Then two alternative ways of constructing universal density estimators with exponential family models is discussed. In this Chapter, the new contribution of the thesis is the discussion on discriminative generative models in Section 2.6 from the point of view of exponential model families. The work was originally presented in Publication 6.

2.1 Modeling Principles

In general, there are two different philosophies to statistical modeling, frequentist and Bayesian. In a frequentist framework probabilities are defined as the proportion of successful events to the total amount of events. The setup requires the assumption that an experiment can be hypothetically repeated an infinite number of times in a similar setting. The task of the modeler is then to count the number of successes in order to determine the probabilities.

Bayesian framework, on the other hand, acknowledges that an event can occur only once. Probabilistic treatment requires the assumption of a *prior* — initial (subjective) beliefs on a successful event before observing data. After observing the event, the prior is then updated to *posterior* beliefs by using the Bayes formula.

Frequentist criticism is that because of prior beliefs, the framework cannot be regarded as fully objective. Subjective Bayesians accept the subjectivity, since all modeling can be regarded as subjective to some extent; for example the selection of the model family is a much more restrictive prior. Objective Bayesian answer is to construct more objective priors, for example by choosing a prior that maximizes the amount of information that can be gained from a future experiment (Bernardo and Smith, 2000), a *reference prior.* Further discussion on the subjectivity of Bayesian inference and its implications is left out of this thesis, since it is a deep research question considered in a multitude of publications over the last 80 years, and therefore a brief discussion would not do full justice to the topic. In practical data analysis, either one of the modelling approaches can be assumed. With large amounts of data frequentist and Bayesian inference give similar results, but with small data sets the results differ, often to the benefit of Bayesian modeling.

In this thesis the Bayesian point of view is assumed because of pragmatic considerations. Besides being better in practice for small data sets, the treatment of complex (hierarchical) models is more intuitive in Bayesian modeling; all variables are treated in the same manner instead of having to divide them to fixed or random effects.

According to Gelman et al. (2003), Bayesian data analysis proceeds in the following steps:

1. Construct a joint distribution for all observable and unobservable variables.

2. Condition on observed data: calculate and interpret the appropriate posterior distribution, the conditional distribution of the unobserved quantities of ultimate interest, given observed data.

3. Evaluate (the fit of) the model. If necessary, alter or expand the model and repeat the three steps.

In the following it is assumed that the modeling step 1 has been already carried out, and the aim is to do statistical inference given the model.

2.1.1 Setup and Notation

In the following sections, the general methodology of probabilistic modeling is discussed. Data are denoted by $D = \{\mathbf{x}_i\}_{i=1}^N$ or, in case of conditional inference, $D = \{c_i, \mathbf{x}_i\}_{i=1}^N$. Unless stated otherwise, discriminative inference is considered: the task is to predict the values c_i of a categorical variable C, given the associated variables $X = \mathbf{x}_i$.

The model parameters are denoted by θ, and they are assumed to belong to a modeling manifold $\overline{\Theta}$, the model family. The manifold itself belongs to a larger space Θ consisting of all possible models. This includes also the “true” model, denoted by $\tilde{\theta}$. The view is slightly controversial, since in reality no probabilistic model is perfectly true and all models are false to some extent. Therefore the “true” model $\tilde{\theta}$ cannot be attained. This can be taken into account by assuming that the data has not been generated by a model in our model family, $\tilde{\theta} \notin \overline{\Theta}$. The situation is illustrated also in Figure 2.1.

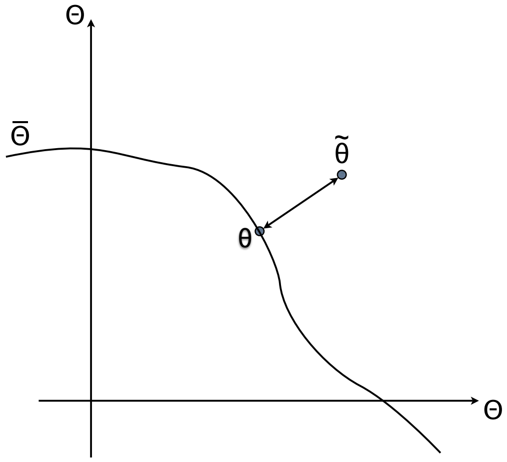

Figure 2.1: An illustration of the modeling setup. Θ denotes the set of all models. The set of realizable models, the selected model family, is a manifold plotted with gray colour and denoted by $\overline{\Theta}$. $\theta \in \overline{\Theta}$ defines a model with parameter values θ. The "true" model is denoted by $\tilde{\theta}$.

2.1.2 Objective Function

A central quantity in all probabilistic modeling is the likelihood function $\mathcal{L}(\mathcal{D};\theta)$ of data $\mathcal{D}$ and parameters θ. Frequentist axiomatizations, for example Birnbaum (1962), justify the likelihood function, whereas Bayesian axiomatic systems (Bernardo and Smith, 2000; Cox, 1946) justify the posterior

$$p(\theta \mid \mathcal{D}) = \frac{\mathcal{L}(\mathcal{D};\theta)p(\theta)}{\int d\theta \mathcal{L}(\mathcal{D};\theta)p(\theta)} \quad , \tag{2.1}$$

which is proportional to the likelihood multiplied by a prior $p(\theta)$.

Usually in modeling an assumption of exchangeability of data points is made at some level. In this case the likelihood factorizes, that is, it is expressed as a multiplication of the probabilities $p(\mathbf{x}_i|\theta)$ given by the model to each of the data points $D = \{\mathbf{x}_i\}_{i=1}^N$:

$$\mathcal{L}(\mathcal{D};\theta) = \prod_i p(\mathbf{x}_i|\theta) \quad . \tag{2.2}$$

Since multiplication is a commutative operation, the order of the operands can be changed without altering the end result.

The selection of the objective function is made by the modeler. In case of conditional inference where the interest is to predict the values of C given X, using conditional likelihood

$$\mathcal{L}(C|X;\theta) = \prod_i p(c_i|\mathbf{x}_i,\theta) \tag{2.3}$$

as an objective function has given good results.

2.1.3 Point Estimates

Once the objective function of a probabilistic model is defined, the model parameters can be optimized. One common approach is to compute a point estimate, that is, the parameter values that maximize the objective function. The parameter values maximizing the likelihood function are called a maximum likelihood (ML) estimate, that is

$$\hat{\theta}_{ML} = \arg\max_{\theta} \mathcal{L}(\mathcal{D};\theta) = \arg\max_{\theta} \prod_i p(\mathbf{x}_i|\theta) \quad . \tag{2.4}$$

The values maximizing the posterior are called maximum a posteriori (MAP) estimate. At the limit of infinite data the two estimates are the same, assuming that the support of the prior is the entire model space $\overline{\Theta}$.

In case of conditional likelihood, the obtained point estimate is conditional maximum likelihood (CML)

$$\hat{\theta}_{CML} = \arg\max_{\theta} \mathcal{L}(C|X;\theta) = \arg\max_{\theta} \prod_i p(c_i|\mathbf{x}_i,\theta) \quad . \tag{2.5}$$

2.1.4 Marginalization

In Bayesian inference, all random variables in the model have a probability distribution, making it possible to compute expectations over the random variables. Integration, or *marginalization*, is therefore the natural means of getting rid of "nuisance" variables which are not needed in predictions. For example, consider a case where the value of c^* needs to be predicted, given observations $\mathbf{x}^*$ and a model $p(c \mid \mathbf{x},\theta)$. The predictive distribution $p(c^*|\mathbf{x}^*,\mathcal{D})$ can be computed by carrying out the integration

$$p(c^*|\mathbf{x}^*,\mathcal{D}) = \int p(c^*|\mathbf{x}^*,\theta)p(\theta|\mathcal{D})d\theta \quad , \tag{2.6}$$

where

$$p(\theta|\mathcal{D}) = \frac{\mathcal{L}(\mathcal{D};\theta)p(\theta)}{\int \mathcal{L}(\mathcal{D};\theta)p(\theta)d\theta} \tag{2.7}$$

is the posterior obtained from earlier data $\mathcal{D} = \{\mathbf{x}_i, c_i\}_{i=1}^N$. In this case, the model parameters θ are "nuisance" variables that can be integrated out by computing an expectation over the parameter distribution $p(\theta|\mathcal{D})$.

The above Equation (2.7) provides also the means for sequential updates of beliefs in model parameters; the prior describes the initial beliefs and the posterior the beliefs after seeing the data. When new data is gathered, the posterior from earlier step is used as a new prior. Equation (2.7) is referred to as the Bayes formula.

Frequentist inference is sometimes seen as a special case of Bayesian inference where uniform priors and delta function posteriors are used. If the support of the prior covers the whole model parameter space, the two approaches are asymptotically (with infinite data) the same, since in this case the posterior reduces into a single peak at the maximum likelihood parameter values.

2.2 Exponential Model Family

Practical implementation of probabilistic inference requires that a certain parametric form of the model for $p(\mathbf{x}|\theta)$ or $p(c,\mathbf{x}|\theta)$ is assumed. In this thesis we assume that modeling is carried out with exponential family models. The motivation for using this specific model family is that it allows to form a universal density estimator; all distributions can be approximated arbitrarily closely with some specific type of an exponential family distribution (Altun et al., 2004; McLachlan and Peel, 2000). It is therefore safe to consider modeling within the exponential family without losing any generality. In this section a short review of theory of exponential families originally introduced by Barndorff-Nielsen (1978) is given. See (Barndorff-Nielsen, 1978; Buntine, 2002; Efron, 1978; Wainwright and Jordan, 2003; Jordan, 2004) for further references and discussion. The properties of the exponential family are discussed below with some detail. These properties will become useful in Section 4 when deriving different optimization methods for exponential family models.

The exponential family is formally defined as

Definition 2.2.1 (Canonical exponential family model) *A model $p(\mathbf{x} \mid \theta)$ belongs to the exponential family if the model can be expressed in the canonical form*

$$p(\mathbf{x} \mid \theta) = b(\mathbf{x})e^{\theta^T T(\mathbf{x}) - \kappa(\theta)} \quad , \tag{2.8}$$

where θ is a vector of natural parameters, and $T(\mathbf{x})$ the corresponding vector of sufficient statistics. The $b(\mathbf{x})$ is the base measure, and $\kappa(\theta)$ a convex normalization term.

In literature, $\kappa(.)$ is known as the log-partition function, moment generating function, or cumulant function. The selection of $\kappa(.)$ defines the exponential family.

For simplicity, the thesis concentrates on models that belong to *regular* exponential families

Definition 2.2.2 *If the parameter space θ is open, that is, if $\theta \in interior(\Theta)$, the model is called a regular exponential family (Wainwright and Jordan, 2003).*

and use *minimal* sufficient statistics, that is, the sufficient statistic has the smallest possible dimensionality (Bernardo and Smith, 2000).

Examples of exponential family distributions include Gaussian, multinomial and exponential models. Table 2.1 gives parametrizations of the Gaussian and multinomial distributions, the exponential family distributions applied in this thesis. Most of the distributions used in probabilistic modeling belong to the exponential family. Some exceptions of course occur such as uniform, Student-t, and Cauchy distributions. These can however be approximated with exponential family distributions[1].

2.2.1 Convexity

In exponential family models, the log-partition function $\kappa(\theta)$ is a convex function in the convex domain θ. Therefore the following results apply. See (Rockafellar, 1970; Boyd and Vandenberghe, 2004) for a detailed introduction to convex analysis.

First, the domain of θ, the subset of $\mathcal{R}^n$ where κ is defined, forms a *convex set*:

[1] For example the Student-t can be modeled as an infinite mixture of Gaussians all having the same mean but different standard deviations.

Table 2.1: Canonical representation of some exponential family distributions. The left column shows the notation in a canonical representation of an exponential family model. Notation in case of Gaussian: μ mean, σ standard deviation. Notation in case of Multinomial: It is assumed that the multinomial consists of $k \in 1 \ldots K$ slots. n_k is the count for a slot k, π_k the corresponding dual parameter, interpreted as the probability of picking the slot. N is the total count, $N = \sum_k n_k$.

	Gaussian	Multinomial
$T(x)$	$\begin{pmatrix} x^2 & x \end{pmatrix}$	n_k
θ	$\begin{pmatrix} -\frac{1}{2\sigma^2} & \frac{\mu}{\sigma^2} \end{pmatrix}$	$\log \pi_k$
$\kappa(\theta)$	$-\frac{\mu^2}{2\sigma^2} - \log \sigma$	$\log \left(\sum e^{\pi_k} \right)$
$\log b(x)$	$-\frac{1}{2} \log 2\pi$	$\log N! - \sum_k \log n_k!$

Definition 2.2.3 (Convex set (Boyd and Vandenberghe, 2004)) *A set S is convex if the line segment between any two points in S lies in S, that is, if for any $\mathbf{x}_1, \mathbf{x}_2 \in S$ and any ξ with $0 \leq \xi \leq 1$, the condition*

$$\xi \mathbf{x}_1 + (1 - \xi)\mathbf{x}_2 \in S$$

is fulfilled.

After defining a convex set, convexity itself can be defined.

Definition 2.2.4 (Convexity) *A function $f : \mathcal{R}^n \mapsto \mathcal{R}$ is convex if the domain $\mathbf{dom} f$ is a convex set and if for all $\mathbf{x}, \mathbf{y} \in \mathbf{dom} f$, and ξ with $0 \leq \xi \leq 1$, we have*

$$f(\xi \mathbf{x} + (1 - \xi)\mathbf{y}) \leq \xi f(x) + (1 - \xi) f(y).$$

In exponential families, the κ is twice differentiable, so the following two lemmas hold.

Lemma 2.2.5 (Convexity; differentiable function)
Assume a differentiable function $f : \mathcal{R}^n \mapsto \mathcal{R}$. Then f is convex if and only if the domain $\mathbf{dom} f$ is a convex set and if for all $\mathbf{x}, \mathbf{y} \in \mathbf{dom} f$

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T (\mathbf{y} - \mathbf{x})$$

holds.

Lemma 2.2.6 (Convexity; twice differentiable function)
Assume a twice differentiable function $f : \mathcal{R}^n \mapsto \mathcal{R}$. Then f is convex if and only if the domain $\mathbf{dom} f$ is a convex set and its Hessian is positive semidefinite, that is, for all $\mathbf{x} \in \mathbf{dom} f$

$$\nabla^2 f(\mathbf{x}) \succeq 0$$

holds.

A result of particular interest in convex algebra and in exponential families (Wainwright and Jordan, 2003) is the *Legendre-Fenchel transformation*

$$\kappa^*(\mu) = \sup_{\theta \in \Theta} \{\theta^T \mu - \kappa(\theta)\} \quad . \tag{2.9}$$

The $\kappa^*(\mu)$ is often referred to as the convex conjugate of $\kappa(\theta)$, and it is again a convex function.

The conjugate of a conjugate is again the function itself, that is,

$$\kappa^{**}(\theta) = \sup_{\mu \in \mathcal{M}} \{\theta^T \mu - \kappa^*(\mu)\} = \kappa(\theta) \quad . \tag{2.10}$$

2.2.2 Duality

The cumulant-generating function κ can also be used to obtain a dual parametrization of exponential models in terms of expected sufficient statistics μ and their covariance matrix Σ (Efron, 1978),

$$\mu = E_{p(\mathbf{x}|\theta)}\left\{T(\mathbf{x})\right\} = \tfrac{\partial\kappa}{\partial\theta}, \tag{2.11}$$

$$\Sigma = E_{p(\mathbf{x}|\theta)}\left\{\left(T(\mathbf{x})-\mu\right)\left(T(\mathbf{x})-\mu\right)^T\right\} = \tfrac{\partial^2\kappa}{\partial\theta\partial\theta} = \tfrac{\partial\mu}{\partial\theta}. \tag{2.12}$$

In cases where the exponential model family is regular, μ forms a dual parametrization of the exponential family model referred to as the mean value parametrization. The benefit of the dual parametrization is that natural parameters of the exponential families do not in general (with Gaussian being the exception) lie within the same space as the sufficient statistics (Efron, 1978), complicating their use and interpretation. However, the dual parameters lie in the same space as the mean of the sufficient statistics, which is why the dual parameters μ are the parameters we usually associate with exponential distributions (and sometimes they are referred to as expected sufficient statistics, for obvious reasons).

An exponential family model can also be expressed in a mixed parametrization consisting of natural parameters as well as mean value parameters (Wainwright and Jordan, 2003).

2.2.3 Bregman Divergence

Bregman divergence (Bregman, 1967; Banerjee et al., 2005) is a distance measure which is closely connected to exponential family models. The formal definition is

Definition 2.2.7 *The Bregman divergence for a convex function $\kappa(\theta)$ is:*

$$D_\kappa(\theta_1,\theta_2) = \kappa(\theta_1) - \kappa(\theta_2) - \nabla\kappa(\theta_2)(\theta_1-\theta_2). \tag{2.13}$$

The divergence is always non-negative due to convexity of κ, and $D_\kappa(\hat{\theta},\hat{\theta}) = \frac{\partial}{\partial\theta}\, D_\kappa(\theta,\hat{\theta})\Big|_{\theta=\hat{\theta}} = 0$.

The relation to exponential family models is shown in Banerjee et al. (2005), who prove that there exists a one-to-one mapping between exponential families and Bregman divergences:

Theorem 2.2.8 (Banerjee et al. (2005))
There is a bijection between regular exponential families and regular Bregman divergences.

The regularity of a Bregman divergence is defined in terms of an open input parameter space in a similar manner to the regularity of an exponential family, see Definition 2.2.2. For an exact definition, see Banerjee et al. (2005).

The mapping can be shown to exist for example in the following way. Consider a model $P_0(\mathbf{x} \mid \theta^*)$ within the exponential model family where the value θ^* is a value of the natural parameter corresponding to the sufficient statistics $T(x)$, found by solving a mapping $\frac{\partial}{\partial\theta}\kappa(\theta)\big|_{\theta=\theta^*} = \mu(\theta^*) = T(\mathbf{x})$. Then, by Equation (2.8) we have (Barndorff-Nielsen, 1978; Efron, 1978)

$$\frac{P(\mathbf{x} \mid \theta)}{P_0(\mathbf{x} \mid \theta^*)} = \exp\{-(\kappa(\theta) - \kappa(\theta^*) - T(\mathbf{x})^T(\theta-\theta^*))\} = e^{-D_\kappa(\theta,\theta^*)}. \tag{2.14}$$

Since κ is a convex function, the term in the exponential corresponds to a (minus) Bregman divergence (Bregman, 1967) between the θ and θ^*, denoted by $D_\kappa(\theta, \theta^*)$. This leads to an alternative representation of exponential family models (see Efron (1978), for example) by $p(\mathbf{x}|\theta) = e^{-D_\kappa(\theta,\theta^*)} P_0(\mathbf{x} \mid \theta^*)$. In this respect the Bregman divergence is the natural distance measure for the selected exponential family.

The second derivative of Bregman divergence is the Fisher information, denoted by Σ here.

Kullback-Leibler Divergence. The Kullback-Leibler divergence,

$$D_{KL}(p, q) = \sum_{i=1}^{n} p(i) \log \frac{p(i)}{q(i)} \quad ,$$

where p and q are n-dimensional vectors of probabilities and i an index over the components of the vectors, is an example of the Bregman divergence. This can be verified by setting $\theta_1 = \log q$, $\theta_2 = \log p$, $\kappa(\theta_1) = \log \sum_i e^{\theta_1(i)} = \log \sum_i q(i) = 0$, and, in a similar manner, $\kappa(\theta_2) = 0$. Additionally, $\nabla\kappa(\theta_2(i)) = \frac{e^{\theta_2(i)}}{\sum_i e^{\theta_2(i)}} = p(i)$.

2.2.4 Conjugate Families

A conjugate family for minimal canonical representations of the form in Equation (2.8) is defined to be

$$p(\theta \mid \varrho, \chi) = d(\varrho, \chi) e^{\chi^T \theta - \varrho\kappa(\theta)} \quad , \tag{2.15}$$

where $\varrho \in \mathcal{R}$, $\chi \in \mathcal{R}^k$ are parameters and $d(\varrho, \chi)$ a normalizing constant. The conjugate family obviously depends on the chosen minimal canonical representation.

2.2.5 Posterior

For an exponential family model, complemented with a conjugate prior distribution, the posterior is of the form

$$p(\theta|X, \varrho, \chi) = \frac{\prod_{i=1}^{N} p(\mathbf{x}_i|\theta) p(\theta|\varrho, \chi)}{\sum_\theta \prod_i p(\mathbf{x}_i|\theta) p(\theta|\varrho, \chi)}. \tag{2.16}$$

By noting the denominator by $Z(X)$, since it is constant with respect to θ, we may write the posterior in an exponential family notation by

$$p(\theta|X, \varrho, \chi) = \\ \frac{b(x) d(\varrho, \chi)}{Z(x)} \exp\left\{ \theta^T \left(\sum_{i=1}^{N} t(\mathbf{x}_i) + \chi \right) - (N + \varrho)\kappa(\theta) \right\}. \tag{2.17}$$

The equation gives a full description of the posterior and is of the same form as the conjugate prior distribution.

2.3 Models with Rich Sufficient Statistics

One way of modeling with exponential family distributions is to use rich sufficient statistics. A rich sufficient statistic is defined in a (possibly) infinite-dimensional feature space. Practical modeling in an infinite dimensional space can be made feasible by using the so-called kernel trick (Schökopf and Smola, 2002).

A theorem by Altun et al. (2004) states that an exponential family with rich sufficient statistics is a universal density estimator, that is, all distributions can be approximated arbitrarily closely with an exponential family model with rich sufficient statistics. The theory and implementation of such a model is beyond the scope of this thesis, an interested reader should see for example Schökopf and Smola (2002); Herbrich (2002).

2.4 Latent Variable Models

Instead of applying an exponential family model with rich sufficient statistics, an alternative method for obtaining a universal density estimator is to construct (possibly an infinite) mixture of simple exponential family distributions. Modeling with mixtures can be justified for example by the celebrated theorem of de Finetti (1931) (here we present the formulation of Bernardo and Smith (2000)):

Theorem 2.4.1 (Representation theorem for 0-1 random vectors)
If $x_1, x_2, \ldots$ is an infinitely exchangeable sequence of 0-1 random quantities with probability measure P, there exists a distribution function Q such that the joint mass function $p(x_1, \ldots, x_n)$ for $x_1, \ldots, x_n$ has the form

$$p(x_1, \ldots, x_n) = \int_0^1 \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} dQ(\theta) \quad ,$$

where $Q(\theta) = \lim_{n\to\infty} P[y_n/n \leq \theta]$, with $y_n = x_1 + \ldots + x_n$, and $\theta = \lim_{n\to\infty} y_n/n$.

In Bernardo and Smith (2000) a sequence is defined to be infinitely exchangeable if every finite subsequence is exchangeable. Exchangeability in turn means that under the given model (probability measure), all permutations of the subsequence are equally probable, that is, $p(x_1, \ldots, x_n) = p(x_{\pi(1)}, \ldots, x_{\pi(n)})$ for all permutations π defined on the set $\{1, \ldots, n\}$.

The above theorem states that all exchangeable binary sequences can be represented as a mixture of Bernoulli distributions (the mixture is infinite and thus denoted by an integral in Theorem 2.4.1). The assumption of exchangeability thus corresponds to modeling with a mixture: it corresponds to assuming that there is an underlying (latent) random variable, and the data are conditionally independent given the values of the random variable. Extensions to several different types of exchangeability assumptions exist (Bernardo and Smith, 2000). For example, if complete exchangeability of data items is assumed, a mixture of multinomials follows; on the other hand, if a centred spherical symmetry is assumed, a mixture of Gaussians follows. The components of the resulting mixtures belong to the exponential family, which motivates modeling with common exponential family models. Therefore, a proper mixture fulfilling the exchangeability assumptions made on the data will model the distribution arbitrarily closely. The Theorem 2.4.1 is also used as a general justification of the Bayesian approach of modeling with a prior distribution. For more discussion on the topic see for example (Bernardo and Smith, 2000; Koivisto, 2004).

When computing the likelihood of the data or making predictions, the latent variable is treated as a "nuisance" parameter and integrated out. Accurate modeling of data may require a large amount of mixture components. In practice a finite amount of mixture components is used, even in so-called infinite mixtures where the amount of mixture components is selected with a random process prior. Discussion on this topic is beyond the scope of the thesis; an interested reader should see for example Dirichlet process (Ferguson, 1973; Sethuraman, 1994) and Indian buffet process priors (Griffiths and Ghahramani, 2006).

The remainder of the thesis concentrates on modeling with finite mixture models. Generality is again not compromised, since for example all continuous distributions can be approximated arbitrarily well by a finite mixture of Gaussians with common variance or covariance matrix (McLachlan and Peel, 2000). Written in terms of minimal canonical exponential family models, a mixture model is expressed as:

$$p(\mathbf{x}_i|\Theta) = \sum_{j \in \mathcal{S}_\pi} \exp\{\theta_{\pi_j}^T T(\pi_j) - \kappa(\theta_\pi) + \theta_{x,j}^T T(\mathbf{x}_i) - \kappa(\theta_{x,j}) - \log b(\mathbf{x}_i, j)\} \quad , \tag{2.18}$$

where θ_{π_j} are the natural parameters of choosing the mixture component j, $T(\pi_j)$ is the sufficient statistics of picking a component j, $\kappa(\theta_\pi)$ is the partition function; $\theta_{x,j}$ are the natural parameters of the mixture component j that models $\mathbf{x}$, $T(\mathbf{x}_i)$ the sufficient statistics of data $\mathbf{x}_i$, $\kappa(\theta_{x,j})$ is the partition function, and $\log b(x_i, j)$ is the base measure. S_π denotes the set of mixture components, $j \in 1 \ldots J$.

Example: Hidden Markov Model. Hidden Markov models (see Figure 2.2 for an illustrative example) are a special type of mixture models, applied when modeling data which comes in the form of a sequence, for example in a case where the statistical properties of a signal change over time. The model explains these changes by a switching of a hidden (latent) state s_i $i \in \{1 \ldots S\}$ within the model, where S is the number of hidden states. Each of the states addresses an associated observation distribution $p(\mathbf{x}|\theta_i, s_i)$, from which the data is generated. While the general functional form of the observation distribution is assumed to remain the same, the parameters θ_i model a different distribution in each state. The changes in the distributions of the observations are thus associated with transitions between hidden states. The transitions are probabilistic and defined by a transition matrix $\mathbf{B}$ containing b_{ij} as its elements — the probabilities of making a transition from state i to state j. A full definition of HMMs requires one more set of parameters, $\pi(i)$, $i = 1 \ldots S$, which are the probabilities of initiating the observation sequence at state i.

Usually a first-order Markov property for the transitions is assumed, where the transition to the next state $s(t+1)$ depends only on the current state $s(t)$ at time t, that is, $p(s(t+1)|s(t), \ldots, s(0)) = p(s(t+1)|s(t))$. The main reason for making the assumption is to reduce the amount of parameters in the model; the model is still rather flexible even after this simplification.

For a time series $\mathbf{x}_{0 \ldots T} = \{\mathbf{x}(0), \ldots, \mathbf{x}(T)\}$ of observations, the full likelihood

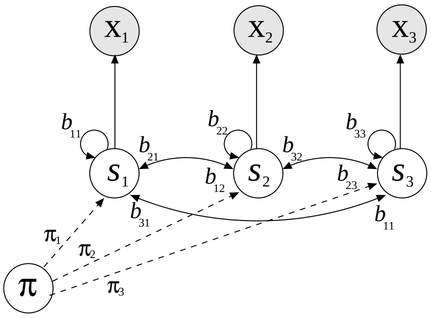

Figure 2.2: A graphical illustration of a hidden Markov model. Hidden states are denoted by $s_i,\ i \in \{1,2,3\}$. Transition probabilities from a hidden state s_i to a state s_j are denoted by b_{ij}. The entries b_{ij} form a transition matrix $\mathbf{B}$. The $\mathbf{x}_i$ are observations which are modelled according to the observation distribution associated with the hidden state s_i. The probabilities of beginning the observation sequence from state s_i are denoted by π_i.

of the HMM is then

$$p(\mathbf{x}_{0\ldots T}|\Theta) = \sum_{\mathcal{S}} \pi(s(0))p(\mathbf{x}(0)|s(0)) \prod_{t=1}^{T} p(\mathbf{x}(t)|s(t))p(s(t)|s(t-1)), \qquad (2.19)$$

where $\mathcal{S}$ denotes all "paths" through the model, that is, all S^{T+1} combinations of hidden states for a sequence of length $T+1$, and $\mathbf{x}(t)$ is the measured observation vector at time t. By inspecting an observation sequence $\mathbf{x}_{0\ldots T}$ as one data point, we can see that the HMM is a (complex) mixture model:

$$p(\mathbf{x}_{0\ldots T}|\Theta) = \sum_{s_{0\ldots T}\in\mathcal{S}} p(\mathbf{x}_{0\ldots T}|\mathbf{s}_{0\ldots T},\Theta)p(\mathbf{s}_{0\ldots T}|\Theta) \quad , \qquad (2.20)$$

where a path through the model is a mixture component. Usually all the probability distributions within the HMM are assumed to be of the exponential family. With the notation of the exponential families, a HMM can therefore be written as

$$\begin{aligned} p(\mathbf{x}_{0\ldots T}|\Theta) = \sum_{s_{0\ldots T}\in\mathcal{S}} & \exp\{\theta_\pi^T t(\pi) - \kappa(\theta_\pi) + \\ & + \sum_{t=1}^{T} \theta_{s_t}^T t(s_t; s_{t-1}) - \kappa(\theta_{s_t}) + \theta_{x_t}^T t(x_t) - \kappa(\theta_{x_t}) - \log b(x_t)\} \ . \end{aligned} \qquad (2.21)$$

The formula simplifies somewhat by taking into account that for multinomial models $\kappa(\theta) = 0$; in this case $\kappa(\theta_\pi) = \kappa(\theta_{s_t}) = 0$.

2.5 Discriminative Models

The above discussion has concentrated on *generative models*, a category of models which model the full generative process of the data. This is the approach recommended in standard Bayesian data analysis (Gelman et al., 2003). The reason for this type of modeling is that uncertainty in all variables can then be taken into account. Interestingly enough, there are many practical cases where Bayesian modeling does not proceed as recommended.

Instead of a joint model, *Bayesian regression* is known to give best results for prediction purposes. Such regression models are referred to as as *discriminative* models in the following (see e.g. Gelman et al. (2003); Hastie et al. (2001); McCullagh and Nelder (1990); Sutton and McCallum (2006)). Difference to joint density models is that discriminative models use some of the data as covariates, meaning that the distribution of some part of the data is not modelled; the associated data values are assumed to be given. In other words, instead of modelling the distribution of, say, random variables C and X jointly by $p(C, X)$, a discriminative model predicts the conditional probability $p(C|X)$, where $X = \mathbf{x}$ are covariates. Instead of likelihood, the objective function of discriminative models contains the conditional likelihood function, $\prod_{i=1}^{N} p(c_i \mid \mathbf{x}_i, \theta)$.

The predictions made by discriminative models are in practice often superior to generative modeling. Intuitively the reason for this is that the discriminative model is optimized to give best predictions of the dependent variable C, whereas a joint model is optimized to predict the joint distribution including also the margin $p(X)$.

In practice, with large amounts of data, generative models are inferior to discriminative models, since the assumed model is almost always incorrect, but with small sample sizes generative models may show better performance (Ng and Jordan, 2002). Section 2.6.1 provides more discussion on the topic.

An example class of discriminative models is the generalized linear model (McCullagh and Nelder, 1990), of which the logistic regression is a special case.

Example: Logistic Regression. Logistic regression is perhaps the best-known discriminative model. It computes the probability of a class $j \in 1 \ldots C$ for a data item $\mathbf{x}_i$ by

$$p(C = j|\mathbf{x}_i, \mathbf{B}) \equiv p_{ji} = \frac{e^{\boldsymbol{\beta}_{j,x}^T \mathbf{x}_i + \beta_{j,0}}}{1 + \sum_{j'=1}^{C-1} e^{\boldsymbol{\beta}_{j',x}^T \mathbf{x}_i + \beta_{j',0}}} \quad , \tag{2.22}$$

where $\mathbf{x}_i$ is the vector of independent variables, covariates, and $\boldsymbol{\beta}_j$ the vector of coefficients for a given class, separated into a vector $\boldsymbol{\beta}_{j,x}$ that multiplies the covariates, plus a constant $\beta_{j,0}$. Furthermore, we denote all β_j by matrix $\mathbf{B}$ consisting of β_j as its columns. To avoid overparametrization $\boldsymbol{\beta}_{C,x}$, $\beta_{C,0}$ are set to zero (Gelman et al., 2003; Hastie et al., 2001).

Each observation c_i can be considered as a draw from a multinomial, and hence the log-likelihood is

$$\log \mathcal{L}(C|X; \boldsymbol{B}) = \sum_{i=1}^{N} \sum_{j=1}^{C} \delta(c_i, j)(\boldsymbol{\beta}_{j,x}^T \mathbf{x}_i + \beta_{j,0}) - \log \left(1 + \sum_{j'=1}^{C-1} e^{\boldsymbol{\beta}_{j',x}^T \mathbf{x}_i + \beta_{j',0}} \right) \quad , \tag{2.23}$$

where $\delta(c_i, j)$ picks the class index j corresponding to the class of sample i, and $\boldsymbol{\beta}_{C,x}, \beta_{C,0}$ are set to zero.

The benefit of the logistic regression model is that its objective function is concave and therefore has one unique global maximum.

2.6 Discriminative Joint Density Models

There are two differences between generative and discriminative models:

1. Discriminative models optimize the conditional likelihood instead of joint likelihood.

2. Discriminative models do not form a model for the margin $p(X)$.

A joint density model can be converted into a discriminative model by changing the objective function from the joint likelihood $\prod_i p(c_i, \mathbf{x}_i|\theta)$ to the conditional likelihood $\prod_i p(c_i|\mathbf{x}_i, \theta)$ and transforming the model accordingly by applying the Bayes formula. In Publication 6 such models are referred to as discriminative joint density models. The conditional maximum likelihood estimate of a joint density model, a discriminative joint density model, has been used extensively for example in speech processing, where discriminative hidden Markov models (Schlüter and Macherey, 1998; Povey et al., 2003) are the current state-of-the-art.

In case of point estimates and given an incorrect model, the conditional likelihood and joint likelihood estimates differ asymptotically. The reason is that the objective functions are different (likelihood vs. conditional likelihood). The fact that the estimates differ can be shown for example by looking at the asymptotical behavior of the objective function of the joint likelihood. In this case the joint likelihood can be written as a Kullback-Leibler divergence between the model $p(c, \mathbf{x}|\theta)$ and the assumed "true" model $p(c, \mathbf{x}|\tilde{\theta})$,

$$
\begin{aligned}
\mathcal{D}_{KL}(\tilde{\theta}, \theta) = \sum_{c \in C, x \in X} p(c, \mathbf{x}|\tilde{\theta}) \log \frac{p(c, \mathbf{x}|\tilde{\theta})}{p(c, \mathbf{x}|\theta)} = \\
\sum_{C,X} p(c, \mathbf{x}|\tilde{\theta}) \log \frac{p(c|\mathbf{x}, \tilde{\theta})}{p(c|\mathbf{x}, \theta)} + \sum_{X} p(\mathbf{x}|\tilde{\theta}) \log \frac{p(\mathbf{x}|\tilde{\theta})}{p(\mathbf{x}|\theta)} \quad ,
\end{aligned}
\tag{2.24}
$$

where the first term is the conditional likelihood. If the true model is included in the model family, the latter term can be made to vanish through optimization, but otherwise, in case of an incorrect model, it is always larger than zero for joint likelihood models (this follows from the non-negativity of the Kullback-Leibler divergence). When the true model is not within the model family, the ML estimate is asymptotically always worse than the CML estimate (Nádas et al., 1988). On the other hand, if the model family is correct, ML estimate is better. The estimates are the same, but the asymptotic variance of ML estimate is lower than the asymptotic variance of a CML estimate (Nádas, 1983).

Compared to a fully discriminative model, a discriminative joint density model states a more specific hypothesis on the data, since a parametric shape of the margin $p(X)$ is assumed. Since the cost function is conditional on $\mathbf{x}$, the distribution of $\mathbf{x}$ is not modelled, but the assumption restricts the parameter space of the model (shown below in this section). Therefore a discriminative joint density model can work better with a small number of data items if the assumed model family is reasonably close to truth.

Model Specification. The general description of the discriminative joint density models considered in this thesis is presented in Publication 6.

The discriminative generative model is formed from a full generative model of the observed (categorical) variable C and the associated measurements X. The generative model is assumed to be a mixture; X is modelled using an exponential family distribution, given the mixture component l. In the generative process, a component l responsible for generating the data is picked first with probabilities $\pi(l),\ l \in 1 \ldots L$. For simplicity it is assumed that the mixture component contains a direct mapping to the observed class variable c_i. The set of components, values of latent variable, associated with the value c_i are denoted by $\mathcal{S}(c_i)$. Given the component, each measurement $\mathbf{x}_i$ consists of M different kinds of data sources indexed by m, each modelled with an appropriate exponential family distribution. Their parameters are denoted by $\theta_{l,m,x}$.

The model is made discriminative by optimizing the conditional likelihood

$$\mathcal{L}(C \mid X; \theta) = \prod_{i=1}^{N} p(c_i|\mathbf{x}_i, \theta) = \prod_i \frac{\prod_k \left(\sum_{l\in\mathcal{S}(c_k)} \pi(l) \prod_m p(\mathbf{x}_{i,m}|\theta_{l,m,x})\right)^{\delta(c_i,k)}}{\sum_{l'\in\mathcal{S}} \pi(l') \prod_m p(\mathbf{x}_{i,m}|\theta_{l',m,x})} \quad , \tag{2.25}$$

where the $\theta = \{\pi, \theta_{l,m,x}\}$ denote the parameters of the model. The model structure described above can be extended to more general cases, but this is left for further work. In principle, restricting the inspection to mixture models does not affect the generality, since a mixture model is a uniform density estimator, as discussed in Section 2.4.

Example: Discriminative Hidden Markov Model. A practical example of a discriminative joint density model is the discriminative hidden Markov model (dHMM), used in Publications 3, 5, and 8. The objective function of dHMM is conditional likelihood,

$$\mathcal{L}(C|X;\theta) = \prod_i p(c_i|\mathbf{x}^i_{0\ldots T}, \theta) = \prod_i \frac{p(c_i, \mathbf{x}^i_{0\ldots T}|\theta)}{p(\mathbf{x}^i_{0\ldots T}|\theta)} \quad , \tag{2.26}$$

where

$$\begin{aligned} p(c_i, \mathbf{x}^i_{0\ldots T}|\theta) = \sum_{s_{0\ldots T}\in\mathcal{S}(c_i)} & \exp\{\theta_\pi^T T(\pi) - \kappa(\theta_\pi) + \\ & + \sum_{t=1}^{T} \theta_{s_t}^T T(s_t; s_{t-1}) - \kappa(\theta_{s_t}) + \theta_{x_t}^T T(x_t) - \kappa(\theta_{x_t}) - \log b(x_t)\} \quad , \end{aligned} \tag{2.27}$$

and

$$\begin{aligned} p(\mathbf{x}_{0\ldots T}|\theta) = \sum_{s_{0\ldots T}\in\mathcal{S}} & \exp\{\theta_\pi^T T(\pi) - \kappa(\theta_\pi) + \\ & + \sum_{t=1}^{T} \theta_{s_t}^T T(s_t; s_{t-1}) - \kappa(\theta_{s_t}) + \theta_{x_t}^T T(x_t) - \kappa(\theta_{x_t}) - \log b(x_t)\} \ . \end{aligned} \tag{2.28}$$

2.6.1 Discriminative vs. Generative: Logistic Regression vs. Naive Bayes

In literature, for example in Ng and Jordan (2002), naive Bayes is referred to as the generative counterpart of logistic regression model; meaning that the difference between the models is the optimization criterion. Publication 6 inspects the topic more closely and shows that there is a subtle difference between the models: due to the distributional assumption of the naive Bayes model, its parameters actually form a subspace within the wider parameter space of logistic regression model. The fact has been known but the exact connection has not been explored, for example Sutton and McCallum (2006) refer to this as "added flexibility". The connection between logistic regression and naive Bayes models was discussed in Publication 6. In an independent work, Roos et al. (2005) reported similar results, but the application was different. The work concentrated on the correspondence between conditional Bayesian networks and logistic regression model, showing that in some cases a global optimum of a conditional Bayesian network can be found by mapping its parameters to those of the logistic regression model and solving the optimization problem there, by finding the unique global optimum of the logistic regression model.

Recently, the logistic regression model has been inspected in more detail by Banerjee (2007); the result is that by using a logistic regression model, we assume that the underlying generative distribution is a mixture of exponential family distributions where each mixture component has the same form. Logistic regression however does not fix the parametric form of the mixture components. Naive Bayes model makes a further assumption by fixing the parametric form of the mixture components, which restricts the parameter space. We discuss this in the following.

Naive Bayes and its Discriminative Version

The generative model behind naive Bayes models is a mixture model with one mixture component per class,

$$p(\mathbf{x}_i|\theta) = \sum_{j\in 1\ldots C} \exp\{\theta_{\pi_j}^T T(\pi_j) - \kappa(\theta_{\pi_j}) + \theta_{x,j}^T T(\mathbf{x}_i) - \kappa(\theta_{x,j}) - \log b(\mathbf{x}_i, j)\} \quad , \tag{2.29}$$

where $\pi(j)$ is the prior class probability, θ_{π_j} the corresponding natural parameters, and $\kappa(\theta_{\pi_j})$ the convex cumulant-generating function. Sufficient statistics are denoted by $T(\mathbf{x}_i)$, the natural parameters of the data generating mixture component j by $\theta_{x,j}$, and the respective cumulant-generating function by $\kappa(\theta_{x,j})$.

As an example, a special case of the naive Bayes is the Linear Discriminant Analysis (LDA; Sharma (1996)). In its classical form it assumes that data for each class is generated from a Gaussian distribution where all of the classes share the same within-class covariance S. LDA is optimized by computing a joint density maximum likelihood estimate: mean is estimated from class centroids, and variance by the within-class covariance.

After learning a classifier, classes for data samples are predicted with an a posteriori decision rule

$$p(C=j|\mathbf{x}_i) = \frac{\exp\{\theta_{\pi_j}^T T(\pi_j) + \theta_{x,j}^T T(\mathbf{x}_i) - \kappa(\theta_{x,j})\}}{\sum_{j'\in 1\ldots C} \exp\{\theta_{\pi_{j'}}^T T(\pi_{j'}) + \theta_{x,j'}^T T(\mathbf{x}_i) - \kappa(\theta_{x,j'})\}} \quad , \tag{2.30}$$

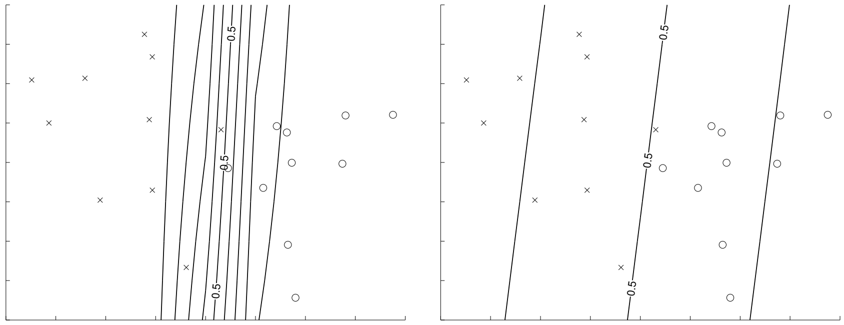

Figure 2.3: Difference of class distributions of LDA trained by maximizing conditional or joint likelihood. Discriminative modeling is optimal for predicting c (Left). In a joint likelihood model the class difference is optimized only implicitly, resulting in softer class borders (Right). In this toy example both models have the same covariance matrix, the within-class covariance, and only the class centroids are optimized. The contour plot shows the probability $p(c|x)$ in 0.1 intervals. "X" and "O" denote samples from the different classes.

where the common terms of the components cancel out; compare to Equation (2.29).

If the main use of a model is to predict the class c given $\mathbf{x}$, optimizing a model by using the objective function derived from Equation (2.30) seems more straightforward, and conditional likelihood the obvious choice. In case of point estimates, the solutions obtained by maximizing joint likelihood or conditional likelihood are asymptotically the same if the "true" data distribution follows the assumptions of the model. Otherwise the solutions differ (see Figure 2.3 for a toy example).

Connection

The relationship between discriminative naive Bayes and logistic regression models is shown next. A mapping between the parameters of naive Bayes and logistic regression is derived first. The derivation begins from the conditional log-likelihood of naive Bayes, written using Equation (2.30),

$$\begin{aligned}\log \mathcal{L}(C|X;\Theta) = &\sum_{i=1}^{N}\sum_{j=1}^{C}\delta(c_i,j)\{\theta_{\pi_j}^T T(\pi_j) + \theta_{x,j}^T T(\mathbf{x}_i) - \kappa(\theta_{x,j})\} \\ &- \log \sum_{j'\in 1\ldots C} \exp\{\theta_{\pi_{j'}}^T T(\pi_{j'}) + \theta_{x,j'}^T T(\mathbf{x}_i) - \kappa(\theta_{x,j'})\} \quad .\end{aligned} \tag{2.31}$$

To make the equations analogous to logistic regression models, the component C is taken as a common term, that is, $\exp\{\theta_{\pi_C}^T T(\pi_C) + \theta_{x,C}^T T(\mathbf{x}_i) - \kappa(\theta_{x,C})\}$. Skipping details, this results in

$$\begin{aligned}\log\mathcal{L}(C|X;\Theta) = &\sum_{i=1}^{N}\sum_{j=1}^{C-1}\delta(c_i,j)\{\log\frac{\pi_j}{\pi_C} + (\theta_{x,j} - \theta_{x,C})^T T(\mathbf{x}_i) - \kappa(\theta_{x,j}) + \kappa(\theta_{x,C})\} \\ &- \log\{1 + \sum_{j'\in 1\ldots C-1}\exp\{\log\frac{\pi_j}{\pi_C} + (\theta_{x,j} - \theta_{x,C})^T T(\mathbf{x}_i) - \kappa(\theta_{x,j}) + \kappa(\theta_{x,C})\}\} \quad .\end{aligned} \tag{2.32}$$

By comparing this to the corresponding log-likelihood of the logistic regression model in Equation (2.23), the following mapping can be found

$$\begin{aligned} \mathbf{x}_i &= T(\mathbf{x}_i) && (2.33)\\ \boldsymbol{\beta}_{j,x} &= \theta_{x,j} - \theta_{x,C} && (2.34)\\ \beta_{j,0} &= \log\frac{\pi_j}{\pi_C} - \kappa(\theta_{x,j}) + \kappa(\theta_{x,C}) \quad . && (2.35) \end{aligned}$$

The difference between the two models is that in logistic regression the components of the vector $\boldsymbol{\beta}_{j,x}$ are independent and can be freely set with no restrictions, whereas in naive Bayes the $\boldsymbol{\theta}_{x,j}$ are dependent through $\kappa(\boldsymbol{\theta}_{x,j})$.

For example, in a multinomial the dual parameters $\mu_{x,j}$ of $\theta_{x,j}$ must sum to one. The natural parameter space of the multinomial is thus restricted because of the mapping between the two parameter spaces; $\theta_{x,j} = \log \mu_{x,j}$.

Effectively, the condition $\sum_j \mu_{x,j} = 1$ removes one degree of freedom from the model. For example, setting one mixture component $\boldsymbol{\beta}_{j,x}(i)$ to $-\log \mu_{x,C}(i)$ (that is, the corresponding dual parameter is set to one) requires that others are set to $-\infty$ (the corresponding dual parameters are set to zero). Sketch of proof:

Lemma 2.6.1 *Multinomial mixture model has one degree of freedom less per mixture component than its logistic regression counterpart.*

Proof Assume a multinomial naive Bayes model with $\kappa(\theta_{j,x}) = \log \sum_{l=1}^{d} e^{\theta_{j,x}(l)}$, where d is the dimensionality of x. For simplicity, assume input data $\mathbf{x} = (0 \ldots 010 \ldots)^T$. Add a displacement λ to $\theta_{j,x}$. The effect on an individual mixture component j of the naive Bayes model in Equation (2.32) is

$$\log\frac{\pi_j}{\pi_C} + (\theta_{x,j} + \lambda\mathbf{1} - \theta_{x,C})^T T(\mathbf{x}) - \log\sum_{l=1}^{d} e^{\theta_{j,x}(l)+\lambda} + \kappa(\theta_{x,C}) \Leftrightarrow$$

$$\log\frac{\pi_j}{\pi_C} + (\theta_{x,j} - \theta_{x,C})^T T(\mathbf{x}) + \lambda - \log e^{\lambda} - \log\sum_{l=1}^{d} e^{\theta_{j,x}(l)} + \kappa(\theta_{x,C}) \Leftrightarrow$$

$$\log\frac{\pi_j}{\pi_C} + (\theta_{x,j} - \theta_{x,C})^T T(\mathbf{x}) - \log\sum_{l=1}^{d} e^{\theta_{j,x}(l)} + \kappa(\theta_{x,C}) \quad ,$$

where $\lambda\mathbf{1}^T T(\mathbf{x}) = \lambda\mathbf{1}^T(0 \ldots 010 \ldots)^T = \lambda$. Therefore λ does not affect predictions. Logistic regression does not have the same restriction, as can be seen by adding a similar displacement to $\beta_{k,x}$. Prediction for a class c_i is then

$$\sum_{j=1}^{C} \delta(c_i, j)(\beta_{j,x}^T \mathbf{x}_i + \beta_{j,0}) + \delta(c_i, k)\lambda - \log\left(1 + \sum_{j'=1}^{C-1} e^{\beta_{j',x}^T \mathbf{x}_i + \beta_{j',0} + \delta(j',k)\lambda}\right) \quad (2.36)$$

where the λ doesn't cancel out. ▮

The proof shows that there is a direction (an equipotential curve) in the space of natural parameters of multinomial mixture model where the probabilities given by the model do not change. Logistic regression does not have such equipotential curves and therefore has one additional degree of freedom.

2.6.2 Connection to Generalized Linear Models

In generalized linear models (GLM; McCullagh and Nelder (1990)), the dependent variable c is modelled with an exponential family distribution of the form $p(c_i|\mathbf{x}_i, \boldsymbol{B}) = \exp\{T(c_i)^T(\boldsymbol{B}^T\mathbf{x}_i) - F(\boldsymbol{B}^T\mathbf{x}_i) - \log b(c_i)\}$. The objective function to be optimized is the conditional likelihood,

$$\mathcal{L}(C|X, \boldsymbol{B}) = \prod_{i=1}^{N} p(c_i|\mathbf{x}_i, \boldsymbol{B}) = \prod_{i=1}^{N} \exp\{T(c_i)^T(\boldsymbol{B}^T\mathbf{x}_i) - F(\boldsymbol{B}^T\mathbf{x}_i) - \log b(c_i)\} \quad . \tag{2.37}$$

The GLM thus assumes a mapping $\theta = \boldsymbol{B}^T\mathbf{x}$ to natural parameters. Since the GLM is an exponential family distribution, the function $F(\cdot)$ is convex, and corresponds to the log-partition function $\kappa(\cdot)$ of ordinary exponential family models.

Analogously to exponential family models, the function $\mu = f(\theta) = \frac{\partial}{\partial\theta}F(\theta)$ provides a mapping to dual parameters. In GLM nomenclature, $f(\theta)$ is called the inverse of a *link function*. Most often used is the *canonical link* function which is obtained if the log-partition function is selected. That is, the inverse of the link is $f(\theta) = \frac{\partial}{\partial\theta}\kappa(\theta)$. The link $g(\theta)$ is solved from $g(\theta) = f^{-1}(\theta)$.

Generative Model for Generalized Linear Models

A parametric mapping can be established also between discriminative generative models and the GLM model class. The connection was shown in case of logistic regression model in Section 2.6.1, and now we continue by showing the mapping in a more general setting. For simplicity it is assumed below that exactly one mixture component k corresponds to each class label c. However, the theory applies to more general cases as well.

The conditional likelihood of discriminative joint density models in Equation (2.25) can be written as

$$\begin{aligned}\mathcal{L}(C \mid X; \theta) = \exp\Bigg\{ \sum_{i=1}^{N}\sum_{k=1}^{C} \delta(c_i, k) \log\Big(\pi(k) \prod_{m=1}^{M} p(\mathbf{x}_{i,m}|\theta_{k,m,x})\Big) \\ - \log \sum_{k'=1}^{K} \pi(k') \prod_{m} p(\mathbf{x}_{i,m}|\theta_{k',m,x})\Bigg\} \quad ,\end{aligned} \tag{2.38}$$

where the model in Equation (2.25) has been simplified such that the component k corresponds the class label c, that is, $k \in \{1 \ldots C\}$, where C is the number of classes. The N is the number of data items and M the number of exponential family distributions used for modelling $\mathbf{x}$.

By comparing Equations (2.38) and (2.2.1), we notice that Equation (2.38) corresponds to an exponential distribution with the mapping

$$\theta_{ik} = \log\left(\pi(k) \prod_{m} p(\mathbf{x}_{i,m}|\theta_{k,m,x})\right) \tag{2.39}$$

$$T(c_i) = \delta(c_i, k)$$

$$\kappa(\theta_i) = \log \sum_{k'} e^{\theta_{ik'}} \tag{2.40}$$

The $\log b(c_i)$ is omitted, since it can be considered a constant. By using the exponential family notation, Equation (2.39) can be written as

$$\theta_{ik} = \sum_m T(\mathbf{x}_{i,m})^T \theta_{k,m,x} - \kappa(\theta_{k,m,x}) - \log b(\mathbf{x}_{i,m}) + \theta_{\pi_k} - \kappa(\theta_\pi) \ . \tag{2.41}$$

The $\log b(\mathbf{x}_i)$-term and $\kappa(\theta_\pi)$ are cancelled out, since they are the same for all components k and can therefore be taken out of the summation in $\kappa(\theta_i)$ in Equation (2.40).

Inserting Eq. (2.41) into Eq. (2.38) results in

$$\mathcal{L}(C|X;\theta) = \prod_i p(c_i|\mathbf{x}_i,\theta) = \prod_i \exp\{T(c_i)^T \left(T(\mathbf{x}_i)^T \boldsymbol{\theta}_x - \kappa(\theta_{k,x}) + \theta_{\pi_k}\right) - \kappa(\theta_i)\} \ , \tag{2.42}$$

where the vector $T(\mathbf{x}_i)$ is formed by concatenation, $T(\mathbf{x}_i) = [T(\mathbf{x}_{i,1}) \ldots T(\mathbf{x}_{i,M})]$. The matrix $\boldsymbol{\theta}_x$ consists of vectors $\theta_{k,x} = [\theta_{k,1,x} \ldots \theta_{k,M,x}]$ as its columns. Finally, $\kappa(\theta_{k,x}) = [\sum_m \kappa(\theta_{1,m,x}) \ldots \sum_m \kappa(\theta_{K,m,x})]$, and the log-partition function $\kappa(\theta_i) = \log \sum_{k' \in \mathcal{S}} \pi(k') \prod_m p(\mathbf{x}_{i,m}|\theta_{k',m,x})$.

The similarity between the generative model and GLM can be seen by comparing Equations (2.42) and (2.37). First, the mapping between the parameters of GLM and discriminative joint density models is

$$\boldsymbol{B}^T \mathbf{x}_i \mapsto T(\mathbf{x}_i)^T \boldsymbol{\theta}_x - \kappa(\theta_{k,x}) + \theta_{\pi_k} \quad ,$$

corresponding to the log-probabilities of the mixture components of the discriminative joint density model.

The difference between the models is also visible in the mapping, the log-partition functions $\kappa(\theta_{k,x})$ in Equation (2.42). As shown earlier, in case of multinomial distribution, the $\kappa(.)$ removes one degree of freedom in the model. Analogously to the logistic regression case, the GLM, in contrast, does not have such a restriction. Since the parameter space is smaller, a discriminative joint density model should therefore perform better than a discriminative model with small amounts of data, assuming that the model family is approximately correct.

The difference between the models can be interpreted as introducing additional prior information into GLMs: it is assumed that the generative model for the margin of $\mathbf{x}$ comes from a mixture of exponential family distributions, where the parametric form of the distribution is known. This restricts the (effective) parameter space β of the generalized linear model making the model simpler to optimize.

2.6.3 Conclusion

Section 2.6 presents the characterization of discriminative joint density models, originally discussed in Publication 6. The models are obtained from joint density models $p(c, \mathbf{x}|\theta)$ by first forming a predictive distribution $p(c|\mathbf{x}, \theta)$ with the Bayes formula, and then optimizing the conditional likelihood. The general formalization of discriminative joint density models in Section 2.6 is a mixture where the identity of the mixture component is associated with the class c. However, the restricted model structure does not affect the generality, since mixture model (at the limit of infinite components) is a universal density estimator, as discussed in Section 2.4.

As discussed in Section 2.5, in case of conditional inference, discriminative models often perform better than joint density models in practice. Sections 2.6.1

and 2.6.2 show that the discriminative generative models are related to discriminative models. Section 2.6.1 shows the connection, the parametric mapping between two special cases, discriminative naive Bayes and logistic regression models. The more general case is then discussed in Section 2.6.2, where the connection between discriminative joint density models and generalized linear models is shown. It is shown that the correspondence between the mappings is close, but the parameter space of the discriminative generative models is more restricted. The restriction is due to the normalization term $\kappa(\theta)$ which makes the parameters θ dependent. In the corresponding discriminative model the parameters are independent. This is shown for the special case of multinomial naive Bayes in Section 2.6.1. Since the effective parameter space of discriminative joint density models is smaller, they are assumed to perform better with a small number of data items, if the assumed model is close to "true" model. Initial empirical evidence is shown in Publication 6.

Discriminative joint density modeling as a concept is not new. For example, in speech recognition the approach has been used for more than 25 years. The contribution of the thesis is to formalize the approach such that a connection to the discriminative models can be made, and to show that the parameter space of discriminative joint density models is more restricted. A related result is by Banerjee (2007), who shows the correspondence in an alternative manner — the model family of logistic regression incorporates the model family of any generative exponential family (mixture) model.

Chapter 3

Discriminative Inference with Incorrect Models

Statistics owes its central presence in science and life to the facts that (i) it is enormously useful for prediction; (ii) it is viewed as providing an objective validation for science.

(Berger, 2006)

One of the most fundamental questions arising from Section 2 is how does discriminative modeling, that is, modeling of $p(c \mid \mathbf{x})$ directly, fit into the framework of Bayesian data analysis. After all, the first step of Bayesian modeling, as recommended for example by Gelman et al. (2003) (see Section 2.1) is to construct a joint distribution $p(c, \mathbf{x})$ of the whole system to be modelled. The reason for doing this is certainly sound, since that way all uncertainty concerning a phenomenon can be taken into account. However, practice has shown that discriminative models often perform better in conditional inference. This may seem odd, since the models do not have a generative model of the covariates $\mathbf{x}$, and thus would seem directly applicable only in experimental setups where there is no noise associated with the covariates.

Discriminative modeling has been justified from the Bayesian generative modeling point of view by making the assumption that the generative models for $p(c \mid \mathbf{x})$ and $p(\mathbf{x})$ have different parameters which are independent also in the prior, that is, $p(c \mid \mathbf{x}, \theta)p(\mathbf{x} \mid \psi)p(\theta)p(\psi)$. Then also the posteriors separate, and the generative distribution of $p(\mathbf{x})$ can be left out from modeling $p(c \mid \mathbf{x})$ (Gelman et al., 2003). This still leaves open the more general case where the parameters are not independent.

A solution to the more general case is proposed in Publication 7. In the problem setup, a full generative model of $p(c, \mathbf{x})$ is assumed to be given, and the task is to do discriminative inference, that is, model the $p(c \mid \mathbf{x})$. If the given generative model is incorrect, it will turn out that a new kind of posterior, called *discriminative posterior*, is optimal for modeling $p(c \mid \mathbf{x})$. An intuitive explanation is that the utility (cost) function is different in discriminative inference, which results in preferring different models in the posterior than in standard joint density modeling. In order to prove the result, a new axiomatic derivation of probabilistic inference is given, bearing resemblance to the derivation of Cox (1946). The resulting posterior is a function of the conditional likelihood, and gives a justification for discriminative modeling as a special case. Before going more deeply into the theoretical results, the general setup and earlier derivations of probabilistic inference are discussed.

3.1 Bayesian Views to Inference

Probability is, in general, a measure of uncertainty. It rises in cases where our knowledge of a system is either incomplete, or the system itself in non-deterministic.

There exist two definitions of probability. The first one determines probabilities in terms of frequencies. A probability defines the proportion of successful events, given an infinite (or a very large) amount of repeats. This frequentist interpretation makes it simple to derive the rules of computing probabilities by simple arithmetics.

However, probabilities can be defined also in cases where an infinite amount of repeats is not available, not even hypothetically (Cox, 1946). The notion gives rise to the second definition of probabilities, which is that probability is a subjective evaluation of the amount of randomness in a system. The definition makes it possible to assign probabilities on different possible outcomes even before anything of the system is observed — a *prior*. This will bring up another problem: if the probabilities are subjective, how should they be updated when observations of the system are made?

In order to restrict the scope of the thesis, the discussion will be restricted to two most famous derivations of probabilistic inference. Both proofs are axiomatic — they try to define a minimal set of assumptions from which probabilistic inference can be derived. Before going deeper into the axiomatic systems of Bernardo and Smith (2000) or Cox (1946), we try to provide an overview of what needs to be defined in general for deriving probabilistic inference. Both of the aforementioned approaches define these, but the formalizations differ largely.

First of all, *the general setup* needs to be defined; the set of events that are under consideration and their properties. After this *a model* can be defined. The model provides a mapping from the set of events to real numbers, that is, a model is an entity that outputs a number that can be converted into a probability of a given event. A model can be for example a mathematical function (of the type considered in this thesis) or it can be an individual describing his beliefs on the probabilities of events.

Depending on the axiomatic system, the update of beliefs as such can be seen as the main objective of probabilistic inference (Cox, 1946), or a further goal can be defined. For example, in Bernardo and Smith (2000) the goal is defined by expected utility, an objective function to be maximized.

3.1.1 De Finetti: Rational Decisions

The general setup of Bayesian decision theory (Bernardo and Smith, 2000) considers situations where the decision maker chooses an *action* a from a set of possible actions $\mathcal{A}$, producing *events* $e \in \mathcal{E}$ and *consequences*. The action requires a (rational) decision, but the following events and consequences are beyond the control of the decision maker. There is no uncertainty associated with the consequences, that is, once an action is taken and an event has happened, a certain consequence will follow. In probabilistic modeling, this means that for a given event and model a certain real number will always result.

Uncertainty in the framework is manifested in (our subjective evaluation of) the probability of event-consequence pairs. In probabilistic modeling this corresponds to the beliefs in different models.

Bernardo and Smith (2000) show that Bayesian inference follows from the following five axioms:

1. An order of preference can be expressed for the possible consequences.
2. The actions are transitive. If action a_2 should be preferred over a_1, and a_3 over a_2, then a_3 should be preferable to a_1. That is, if $a_1 \leq a_2$ and $a_2 \leq a_3$, then $a_1 \leq a_3$.
3. The order of preference between consequences is fixed; it is invariant of the state of information concerning relevant events.
4. There exists a standard event $\mathcal{S}$ in event space and a function ϑ that maps the event into a real number between $[0, 1]$.
5. The set of standard events provides a scale against which every consequence can be precisely measured.

See Bernardo and Smith (2000) for more rigorous definitions and proofs.

The goal of the decision maker is to maximize the *utility* function assigning an order of preference for the actions, given event-consequence pairs. That is, utility u is a function that maps the actions and events $\mathcal{A} \times \mathcal{E}$ into real numbers. Bayesian inference can then be derived as maximizing the *expected utility* of the decision maker, defined as:

Definition 3.1.1 (Expected utility) *For a given action $a \in \mathcal{A}$ and events $e_j \in \mathcal{E}$, with $j \in J$, where J is a finite partition of an (un)certain event into event-consequence pairs,*

$$\bar{u}_p(a) = \sum_{j \in J} u(a, e_j) p(e_j)$$

is the expected utility of the action a.

Although the selection of utility function is subjective, two general requirements for utility function are usually acknowledged; the utility should be *proper*, meaning that the utility is maximized if "true" beliefs are reported, and it should be *local*, meaning that only the observed events affect the utility. These considerations motivate the selection of a utility function of form $A \log P(e_j) + B$, justifying the use of log-likelihood as utility, see Bernardo and Smith (2000).

3.1.2 Cox-Jaynes: Reasonable Expectation

One of the most widely known and perhaps most intuitive justifications of Bayesian inference is presented by Cox (1946). The general setup considers beliefs in events. The approach bypasses the complex definitions of Bernardo and Smith (2000) with the first axiom — beliefs in events can be described by real numbers. The second axiom considers how the beliefs should be related to each other; the complement of a belief should be able to be expressed as a function of the belief. The third required axiom is that the joint belief in two events V and V' can be expressed as a function of two beliefs; the belief in V and the belief in V', conditional on V. More formally, using the formulation of Halpern (1999)[1]:

1. Assume a (belief) function Bel that associates a real number with each pair (U, V) of subsets of a domain W such that $U \neq \emptyset$. That is, $Bel(V \mid U) \in \mathcal{R}$
2. There is a twice differentiable function S such that belief in complement $\bar{V}$ of V in W can be expressed as $Bel(\bar{V} \mid U)) = S(Bel(V \mid U))$ if $U \neq \emptyset$.
3. Given two events V' and V in W, there is a twice differentiable function F, with a continuous second derivative, such that the joint belief in the events can be computed as $Bel(V \cap V' \mid U) = F(Bel(V' \mid V \cap U), Bel(V \mid U))$, if $V \cap U \neq \emptyset$.

Under these assumptions Cox shows that Bel is isomorphic to a probability distribution, that is, there is a continuous one-to-one function $g : \mathcal{R} \to \mathcal{R}$ such that $g \circ Bel$ is a probability distribution on W, and

$$g(Bel(V \mid U)) \cdot g(Bel(U)) = g(Bel(V \cap U)) \text{ if } U \neq \emptyset \quad .$$

Convention is to choose g to be the identity function.

It is known that the set of Cox's axioms is not sufficient for justifying Bayesian inference in a more general case (e.g. in case of discrete variables), as a recent counter example (Halpern, 1999) shows. Despite serious effort on fixing the axioms, no satisfying solution has been found so far.

3.2 Using Incorrect Models

The reason why incorrect models have not been considered profoundly in Bayesian inference is perhaps the deep belief that given enough prior information, a model which is arbitrarily close to the "truth" can be obtained. Thus the sole reason why the model performs poorly is that it does not contain enough prior information, and therefore a natural solution in case of a poorly performing model is to design a better one. This is however not always possible or feasible. For example, no additional prior information to be used for improving a model may be available, or a more accurate model may be computationally infeasible.

Publication 7 studies discriminative modeling with incorrect models. As discussed in the introduction of this Chapter, conditional models often perform better than joint density models in discriminative tasks. Intuitively the reason for better performance is simple; a generative model of a joint distribution needs to divide its modeling power to all variables, whereas the conditional model can concentrate more on modeling the distributions of the variables of interest. The result has been proven already two decades ago for maximum likelihood and conditional maximum likelihood point estimates (Nádas, 1983; Nádas et al., 1988), but no theoretical proof in case of distributions has been shown.

As discussed earlier in Section 2.6, a joint density model can be learned either by conventional joint density modeling or by converting it to a discriminative joint density model. The interpretation in Publication 7 is that the selection corresponds to choosing the utility function. In joint density modeling the utility is to model the full distribution $p(c, \mathbf{x})$ as accurately as possible, which corresponds to computing

[1] This is because Cox doesn't state his axioms explicitly

the standard posterior incorporating the likelihood function. In discriminative learning the utility is to model the conditional distribution of $p(c \mid \mathbf{x})$, and the result is a *discriminative posterior* incorporating the conditional likelihood.

The key result in Publication 7 is the proof of discriminative posterior, presented below in Theorem 3.2.3. The posterior can be used to prove the following two claims:

Claim 3.2.1 (Well-known) *Given a discriminative model, a model $p(c \mid \mathbf{x}; \theta)$ for the conditional density, Bayesian regression results in consistent conditional inference.*

Claim 3.2.2 (New) *Given a joint density model $p(c, \mathbf{x} \mid \theta)$, discriminative posterior results in consistent conditional inference.*

Inference is said to be consistent if the utility is maximized with large data sets (Vapnik, 1995). Notice that although the first claim is well known, it has not been proven, aside from the special case where the priors for the margin $\mathbf{x}$ and $c \mid \mathbf{x}$ are independent, as discussed above and in Gelman et al. (2003). Claim 3.2.2 proves this more general case.

The discriminative posterior — the axiomatic proof that it works, and methods for computing the posterior are the contributions of Publication 7. Earlier, Grünwald et al. (2002) suggested the posterior in an extended abstract. However, no empirical or theoretical proof of the posterior was given. First empirical validation that the discriminative posterior works in practice has been presented by Cerquides and Mántaras (2005) in case of maximum a posteriori point estimates. In Publication 7 we show that similar results are obtained when considering the full posterior by applying Markov chain Monte Carlo sampling.

3.2.1 Point Estimates

The derivation of the discriminative posterior can be motivated by inspecting point estimates. At the limit of infinite amount of data, the joint posterior distribution $p(\theta \mid \mathcal{D}) \propto p(\theta) \prod_{(c,\mathbf{x}) \in \mathcal{D}} p(c, \mathbf{x} \mid \theta)$ becomes a point solution, $p(\theta \mid \mathcal{D}) = \delta(\theta - \hat{\theta})$. Therefore, a proper utility function minimizing the approximation error D_{JOINT} between the point estimate $\hat{\theta}_{JOINT}$ and the "true" model $\tilde{\theta}$ is

$$\begin{aligned} \hat{\theta}_{JOINT} &= \arg\min_{\theta \in \overline{\Theta}} D_{JOINT}(\tilde{\theta}, \theta) \quad \text{where} \\ D_{JOINT}(\tilde{\theta}, \theta) &= \sum_c \int p(c, \mathbf{x} \mid \tilde{\theta}) \log \frac{p(c, \mathbf{x} \mid \tilde{\theta})}{p(c, \mathbf{x} \mid \theta)} \, d\mathbf{x}. \end{aligned} \tag{3.1}$$

If the "true" model is in the model family, that is, $\tilde{\theta} \in \overline{\Theta}$, Equation (3.1) can be minimized to zero and the resulting point estimate is effectively the maximum a posteriori solution (with infinite data the effect of prior is insignificant). On the other hand, if the "true" model is not in the model family, $\tilde{\theta} \notin \overline{\Theta}$, the resulting point estimate is the most accurate description of the "true" joint distribution $p(c, \mathbf{x} \mid \tilde{\theta})$ with respect to D_{JOINT}.

In case of conditional inference, the problem is to find the best point estimate $\hat{\theta}_{COND}$ for the conditional distribution $p(c \mid \mathbf{x}, \theta)$. The average Kullback-Leibler

-divergence between the "true" conditional distribution at $\tilde{\theta}$ and its estimate at θ is given by

$$D_{COND}(\tilde{\theta}, \theta) = \int p(\mathbf{x} \mid \tilde{\theta}) \sum_{c} p(c \mid \mathbf{x}, \tilde{\theta}) \log \frac{p(c \mid \mathbf{x}, \tilde{\theta})}{p(c \mid \mathbf{x}, \theta)} \, d\mathbf{x} \; , \tag{3.2}$$

and the best point estimate with respect to D_{COND} is

$$\hat{\theta}_{COND} = \arg \min_{\theta \in \overline{\Theta}} D_{COND}(\tilde{\theta}, \theta) \; . \tag{3.3}$$

By Equations (3.1) and (3.2), we may write

$$D_{JOINT}(\tilde{\theta}, \theta) = D_{COND}(\tilde{\theta}, \theta) + \int p(\mathbf{x} \mid \tilde{\theta}) \log \frac{p(\mathbf{x} \mid \tilde{\theta})}{p(\mathbf{x} \mid \theta)} \, d\mathbf{x} \quad . \tag{3.4}$$

Therefore the point estimates $\hat{\theta}_{JOINT}$ and $\hat{\theta}_{COND}$ are different in general. If the model that has generated the data does not belong to the model family, that is $\tilde{\theta} \notin \overline{\Theta}$, then by Equation (3.4) the joint estimate is generally worse than the conditional estimate in conditional inference, measured in terms of conditional likelihood.

3.2.2 Discriminative Posterior

In Publication 7 the point estimate $\hat{\theta}_{COND}$ presented above is generalized to a *discriminative posterior distribution* over $\theta \in \overline{\Theta}$:

Theorem 3.2.3 (Discriminative posterior distribution)
It follows from axioms 1–6 listed below that, given data $\mathcal{D} = \{(c_i, \mathbf{x}_i)\}_{i=1}^{n}$, the discriminative posterior distribution $p_d(\theta \mid \mathcal{D})$ is of the form

$$p_d(\theta \mid \mathcal{D}) \propto p(\theta) \prod_{(c,\mathbf{x}) \in \mathcal{D}} p(c \mid \mathbf{x}, \theta) \quad .$$

The predictive distribution for new $\tilde{\mathbf{x}}$, obtained by integrating over this posterior, $p(c \mid \tilde{\mathbf{x}}, \mathcal{D}) = \int p_d(\theta \mid \mathcal{D}) p(c \mid \tilde{\mathbf{x}}, \theta) \, d\theta$, is consistent for conditional inference. That is, p_d is consistent for the utility of conditional likelihood.

Inference is consistent if the utility is maximized with large data sets (Vapnik, 1995).

The axioms leading to the discriminative posterior are

1. The posterior $p_d(\theta \mid \mathcal{D})$ can be represented by non-negative real numbers that satisfy $\int_{\overline{\Theta}} p_d(\theta \mid \mathcal{D}) d\theta = 1$.

2. A model $\theta \in \overline{\Theta}$ can be represented as a function $h((c, \mathbf{x}), \theta)$ that maps the observations $(c, \mathbf{x})$ to real numbers.

3. The posterior, after observing a data set D followed by an observation $(c, \mathbf{x})$, is given by $p_d(\theta \mid \mathcal{D} \cup (c, \mathbf{x})) = F(h((c, \mathbf{x}), \theta), p_d(\theta \mid \mathcal{D}))$, where F is a twice differentiable function in both of its parameters.

4. Exchangeability: The value of the posterior is independent of the ordering of the observations. That is, the posterior after two observations $(c, \mathbf{x})$ and $(c', \mathbf{x}')$ is the same irrespective of their ordering: $F(h((c', \mathbf{x}'), \theta), p_d(\theta \mid (c, \mathbf{x}) \cup \mathcal{D})) = F(h((c, \mathbf{x}), \theta), p_d(\theta \mid (c', \mathbf{x}') \cup \mathcal{D}))$.

5. The posterior must agree with the utility. For $\tilde{\theta} \in \Theta$, and $\theta_1, \theta_2 \in \overline{\Theta}$, the following condition is satisfied:

$$p_d(\theta_1 \mid \mathcal{D}_{\tilde{\theta}}) \leq p_d(\theta_2 \mid \mathcal{D}_{\tilde{\theta}}) \Leftrightarrow D_{COND}(\tilde{\theta}, \theta_1) \geq D_{COND}(\tilde{\theta}, \theta_2) \quad ,$$

   where $\mathcal{D}_{\tilde{\theta}}$ is a very large data set sampled from $p(c, \mathbf{x} \mid \tilde{\theta})$. We further assume that the discriminative posteriors p_d at θ_1 and θ_2 are equal only if the corresponding conditional KL-divergences D_{COND} are equal.

6. For fixed $\mathbf{x}$ the model reduces to the standard posterior. For the data set $\mathcal{D}_x = \{(c, \mathbf{x}') \in \mathcal{D} \mid \mathbf{x}' = \mathbf{x}\}$, the discriminative posterior $p_d(\theta \mid \mathcal{D}_x)$ matches the standard posterior, that is, for $p^x(c \mid \theta) \equiv p(c \mid \mathbf{x}, \theta)$, $p_d(\theta \mid \mathcal{D}) = p(\theta \mid \mathcal{D})$

The proof that the discriminative posterior results from these axioms is given in Publication 7. Here the emphasis is on the intuition behind each of the axioms, why each of them is needed.

Axioms 1 and 2 define the general setup (see Figure 2.1 for an illustration). The manifold $\overline{\Theta}$ consists of the set of all different models that can be realized by the model family. Each of the models is a mapping from the event space into a non-negative real number. The mapping does not have to be one-to-one; it can be injective or surjective as well. When an event occurs, the corresponding non-negative real number output by the model describes the belief in the model. By axiom 1 the real numbers output by the models are normalized into a distribution.

After introducing the general setup, what remains to be defined is how the distribution should be updated when more events are observed. Axiom 3 states a sequential update of the beliefs in models; the updated distribution should be a function of the earlier beliefs and the real numbers output by the models for the current event. The desired item here is that the updated distribution contains all the information conveyed by a new observation. Axiom 3 makes also technical assumptions which are needed in proofs (assumption that second derivatives exist). The assumption is similar to the one made by Cox (Cox, 1946), and therefore the proof of the discriminative posterior holds in the same scope; see Halpern (1999).

Axiom 4 is a further requirement; it states that the events should be exchangeable — the belief in the model should be the same after observing two events, irrespective of the order in which they were observed. That is, all events must contain the same amount of information. The axioms 1–4 can be used to derive probabilistic inference — how beliefs should be updated. The proofs bear resemblance to proofs of Cox, but the setup is slightly different, since in discriminative posterior updates beliefs in models instead of beliefs in events used by Cox.

Axiom 5 requires that the posterior should agree with the utility of the modeler; the models that result in high utility should have a larger value in the posterior. The utility inspected in Publication 7 is the Kullback-Leibler divergence between the predictions $p(c \mid \mathbf{x}, \tilde{\theta})$ of the "true" model and the given model $p(c \mid \mathbf{x}, \theta)$.

Finally, the axiom 6 effectively states that the choice of utility function coincides with the standard Bayesian inference for a fixed $\mathbf{x}$.

3.2.3 Connection to Bayesian Decision Theory

Discriminative modeling with incorrect models can be considered also from within the decision-theoretic framework of Bernardo and Smith (2000). The standard Bayesian approach needs only two modifications:

- The utility is changed from (log) likelihood into (log) conditional likelihood.
- The model family is considered to be an approximation of the "true" model.

The general setup of Bernardo and Smith (2000) is similar to the setup of discriminative posterior, that is, the framework of Bernardo and Smith (2000) allows also making the assumption that a "true" joint density model (true beliefs) exists.

Notation. The expected utility was defined in Section 3.1.1 as

$$\bar{u}_p(a) = \sum_{j \in J} u(a, e_j) p(e_j) \quad ,$$

where $\{e_j\}$ is a partition of event-consequence pairs. This can be expressed in a more familiar notation by noting that an event consists of generation of data $\{c_i, \mathbf{x}_i\}$ and the selection of a model, let us denote it here by θ_k. The beliefs in events $p(e_j)$ are then the probabilities of generating data and selecting a model θ_k, that is, $p(e_j) = p(c_i, \mathbf{x}_i, \theta_k)$. If the "true" model is assumed to exist, all probability is concentrated on $\tilde{\theta}$, that is, $p(\theta) = \delta(\tilde{\theta} - \theta)$. Therefore the utility reduces to summation over observed data, $\bar{u} = \sum_i u(a, c_i, \mathbf{x}_i, \tilde{\theta}) p(c_i, \mathbf{x}_i, \tilde{\theta})$. The "true" model is denoted by $p(c_i, \mathbf{x}_i | \tilde{\theta})$ in the following.

Utility. A discriminative task can be formulated in the decision-theoretic framework of Bernardo and Smith (2000) by describing the events in a sequential form where first the outcome of set X of events is observed, and then a set C of events. In this case the expected utility can be written as

$$\bar{u} = \sum_i u(a, c_i, \mathbf{x}_i, \tilde{\theta}) p(c_i | \mathbf{x}_i, \tilde{\theta}) p(\mathbf{x}_i | \tilde{\theta}).$$

The partitioning of the event-consequence space corresponds to the one in joint density modeling, it is only treated in a hierarchical manner.

As discussed in Section 3.1.1, the requirements of a proper, local utility justify the use of (log)likelihood $u(a, c_i, \mathbf{x}_i, \tilde{\theta}) = \log p(c_i, \mathbf{x}_i | \tilde{\theta})$ as utility. If X is always observed before the action needs to be chosen, it is reasonable to change the utility accordingly, from $u(a, \mathbf{x}_i, c_i, \tilde{\theta}) = \log p(c_i, \mathbf{x}_i | \tilde{\theta})$ to $u_d(a, \mathbf{x}_i, c_i, \tilde{\theta}) = \log p(c_i | \mathbf{x}_i, \tilde{\theta})$, since it is known that when predicting C there will be no uncertainty associated with X. The resulting expected utility is thus

$$\bar{u}_d = \sum_i p(c_i | \mathbf{x}_i, \tilde{\theta}) p(\mathbf{x}_i | \tilde{\theta}) \log p(c_i | \mathbf{x}_i, \tilde{\theta}) \quad .$$

Incorrect Model. When working with incorrect or approximate models of the "true" model, the suggested functional form for a proper, local utility function (see page 75 of Bernardo and Smith (2000)) is

$$\bar{u}_p(q) = -\sum_{j \in J} p_j \log \frac{p_j}{q_j} \quad . \tag{3.5}$$

The beliefs reported by the approximate model are denoted by q and the true beliefs by p. J is the partition of event-consequence pairs as earlier. The utility can be written also as the difference between the utilities of "true" model and the approximate model,

$$\bar{u}_p(q) = \sum_{j \in J} -p_j \log p_j + \sum_{j \in J} p_j \log q_j \quad . \tag{3.6}$$

In case of joint density modeling, the expected utility with a given incorrect model $q(c_i, \mathbf{x}_i|\theta)$ is therefore

$$\bar{u}(\theta) = -\sum_i p(c_i, \mathbf{x}_i|\tilde{\theta}) \log \frac{p(c_i, \mathbf{x}_i|\tilde{\theta})}{q(c_i, \mathbf{x}_i|\theta)} = const. + \sum_i p(c_i, \mathbf{x}_i|\tilde{\theta}) \log q(c_i, \mathbf{x}_i|\theta),$$

which justifies modeling with maximum likelihood.

In a similar manner, in discriminative tasks the utility function for a given incorrect model $q(c_i|\mathbf{x}_i, \theta)$ is

$$\bar{u}_d(\theta) = -\sum_i p(c_i, \mathbf{x}_i|\tilde{\theta}) \log \frac{p(c_i|\mathbf{x}_i, \tilde{\theta})}{q(c_i|\mathbf{x}_i, \theta)} = const. + \sum_i p(c_i, \mathbf{x}_i|\tilde{\theta}) \log q(c_i|\mathbf{x}_i, \theta) \quad ,$$

which justifies conditional maximum likelihood modeling.

Discriminative posterior is then a distribution on parameter space $\overline{\Theta}$ of the incorrect model $\theta_k \in \overline{\Theta}$ which maximizes the utility,

$$\bar{u}_d = \sum_k \bar{u}_d(\theta_k) q(\theta_k, c_i, \mathbf{x}_i),$$

This is the same as the expected approximation error of discriminative posterior, see Equation (3.7) below.

3.2.4 Relation to Cox's Axioms

Compared to discriminative posterior, the axioms presented by Cox consider only updating beliefs, and do not take the utility into account. Therefore the Cox's axioms should be contrasted to axioms 1–4 of the discriminative posterior. The axioms 1–2 together correspond to the axiom 1 by Cox. In discriminative posterior, axiom 2 by Cox is not stated explicitly. However, the requirement of axiom 1 that p_d is a probability distribution implicitly requires this. Axioms 3 and 4 together make up for axiom 3 of Cox. The difference is that instead of associativity, the discriminative posterior requires exchangeability. Proof goes in a very similar manner, and the end result is the same, multiplicativity.

There is additionally a slight difference in the general setups of the two axiomatizations. Instead of assuming beliefs in events, the discriminative posterior assumes beliefs in different models instead (Fine has presented similar assumptions earlier (Halpern, 1999)). The assumption of a continuous parameter space for the beliefs in models made by the discriminative posterior is perhaps more reasonable assumption than continuity of beliefs in events as in Cox's case.

3.2.5 Discriminative versus Generative Models

After introducing the discriminative posterior, the distinction between discriminative and generative models can be stated more clearly. The argument in Publication 7 is that the main difference between the two models is the utility function. The requirement of axiom 5 that the utility should agree with the posterior will eventually lead to the proper form of the posterior. In Publication 7 conditional inference was chosen as the utility, and a discriminative posterior was obtained. If the utility had been joint modeling of $(c, \mathbf{x})$, the standard posterior would have resulted. If the given model is incorrect the different utilities therefore lead to different learning schemes, discriminative or generative modeling.

The same problem can be formulated also in a different manner. It can be assumed that the utility of the modeler is always the same, joint density modeling. The resulting problem is then to find a proper way of extending a discriminative model into a joint density model, leading to the earlier justification of discriminative modeling as discussed in Gelman et al. (2003). Lasserre et al. (2006) extend this traditional generative view of discriminative modeling by extending a given generative model family (also (Kelm et al., 2006; Bouchard, 2007) use a similar approach). In their approach, the conditional model for $p(c \mid \mathbf{x}, \theta)$ is complemented with a model for $p(\mathbf{x} \mid \theta')$ to form a joint density model $p(c, \mathbf{x} \mid \theta, \theta') = p(c \mid \mathbf{x}, \theta)p(\mathbf{x} \mid \theta')$. The conditional density model $p(c \mid \mathbf{x}, \theta)$ is derived by Bayes rule from the given joint density model, $p(c, \mathbf{x} \mid \theta)$, and the model for the marginal $p(\mathbf{x} \mid \theta')$ is obtained by marginalizing it. In effect, a larger model family is postulated with additional parameters θ' for modeling the marginal $\mathbf{x}$. The approach of Lasserre et al. (2006) contains the discriminative posterior distribution as a special case in the limit where the priors are independent, that is, $p(\theta, \theta') = p(\theta)p(\theta')$ where the parameters θ and θ' can be treated independently.

Also Lasserre et al. (2006) can be viewed as giving a theoretical justification for Bayesian discriminative learning, based on generative modeling. The work introduces a method of extending a fully discriminative model into a generative model, making discriminative learning a special case of optimizing the likelihood (that is, the case where priors separate). However, since discriminative modeling is fit into the standard Bayesian framework of joint density modeling, the approach of Lasserre et al. (2006) is principled only if the "true" model belongs to the postulated larger model family $p(c, \mathbf{x} \mid \theta, \theta')$.

From a practical point of view it does not matter which philosophical view the modeler assumes, since both result in the same type of modeling. However, it is the view of the author that the discriminative posterior will eventually have more theoretical implications; first of all, in probabilistic modeling utility can be different from the one in joint density modeling, and secondly, the framework considers modeling also with incorrect models.

Performance Comparison. As discussed in Section 2.6, discriminative joint density models have a more restricted model structure than Bayesian regression models. Therefore, if the assumed generative model is close enough to the "truth", it is expected to perform better with small amounts of data. A full set of experiments validating this claim must be left for further work, but the initial experiments presented in Publication 7 seem to support the claim.

The claim can be backed up also by theoretical arguments. Since the discriminative posterior in Theorem 3.2.3 is normalized to unity, $\int_{\theta \in \overline{\Theta}} p_d(\theta \mid D) d\theta = 1$, the values of the posterior are generally smaller for larger model families (they define a larger manifold $\overline{\Theta}$). The expected approximation error of a model family can be formed by using Equation (3.2):

$$E_{p_d(\theta|\mathcal{D})}\left[K_{COND}(\tilde{\theta},\theta)\right] = -\int \sum_c p(\mathbf{x}, c \mid \tilde{\theta}) p_d(\theta \mid \mathcal{D}) \log p(c \mid \mathbf{x}, \theta) d\mathbf{x} d\theta + \text{const.} \tag{3.7}$$

The error is small when both the discriminative posterior distribution $p_d(\theta \mid \mathcal{D})$ and the conditional likelihood $p(c \mid \mathbf{x}, \theta)$ are large at the same time. For small amounts of data, if the model family is too large, the values of the posterior $p_d(\theta \mid \mathcal{D})$ are small. Since the discriminative joint density model has a more restricted model family than Bayesian regression, the values of the posterior are larger.

If the model is approximately correct, the discriminative joint density model will have a smaller approximation error than the Bayesian regression, that is, $p(c \mid \mathbf{x}, \theta)$ is large somewhere in $\overline{\Theta}$. This is analogous to selecting the model family that maximizes the evidence (in our case the expected conditional log-likelihood) in Bayesian inference; choosing a model family that is too complex leads to small evidence (see, e.g., Bishop (2006)). An interesting prospect of future work would be to study the extension of this view into model selection.

The frameworks of model selection have been divided in Bernardo and Smith (2000) into three categories; $\mathcal{M}$-closed, $\mathcal{M}$-completed, and $\mathcal{M}$-open. In the $\mathcal{M}$-closed case the "true" model is assumed to belong to the candidate models under consideration. In $\mathcal{M}$-completed case the models under consideration are approximations of the truth. The "true" model itself is known but it is not modelled because of for example computational reasons. The models under consideration are assumed to be derived with the knowledge of this "true" model. Finally, in the $\mathcal{M}$-open case the truth is not known. The setup considered in Publication 7 would correspond to model selection in the $\mathcal{M}$-open or in some special cases the $\mathcal{M}$-completed case.

Chapter 4

Discriminative Probabilistic Inference in Practice

In this section, the practical implementation of probabilistic modeling is discussed. The focus is on different optimization methods, mainly on the new optimization methods presented in Publications 5 and 7. The aim is to place the methods into a wider perspective with respect to other optimization methods for exponential models. For this means the general approach of presenting optimization methods in terms of Legendre-Fenchel transforms is assumed, as in Wainwright and Jordan (2003); Jordan (2004). See also Beal (2003) for related issues.

The section gives also a simple demonstration of the discriminative EM algorithm, which was left out from the original publications due to space limitations.

4.1 Methods for Probabilistic Inference

Before discussing the contributions of this thesis, the common steps of data analysis are discussed. The focus is on the first two steps of data analysis: model construction and optimization.

4.1.1 Selecting Model Family

The first task of the modeller in data analysis is to form a generative model of the data that incorporates as much prior information into the model as possible. For this means, the initial step is the selection of model families and modeling structures to be considered. This stage requires a considerable amount of subjective evaluations, and is often in practice done by *data exploration* – “looking at the data”. Model building begins by considering the properties of the data; what are the possible dependency structures in the data, that is, how can exchangeability be implemented within the model; do the consecutive items depend on each other, do the individual dimensions within the data item depend on each other, and within one dimension, can the values be considered exchangeable or do they share some underlying latent variable. An example of methods for data exploration are different kinds of visualizations, such as principal component analysis (PCA) (Hotelling, 1933) or self-organizing maps (SOM) (Kohonen, 2001). The selection of the used method affects heavily the end result of visualization (Venna and Kaski, 2007).

Once a set of model families are chosen, there are several ways of choosing the best one, such as Bayes factors (Kass and Raftery, 1995), Bayesian Deviance Criterion (Spiegelhalter et al., 2002), expected (predictive) utility (Vehtari and Lampinen, 2004), or older information-theoretic criteria such as Bayesian Information Criteria (Schwartz, 1978) or Akaike's Information Criteria (Akaike, 1974). The suitability and correctness of the criteria depends on the data at hand. In practice, the model is often validated by posterior predictive checking (Gelman et al., 2003), that is, subjective evaluation.

4.1.2 Point Estimates

Once a model family is chosen, the parameters of the model need to be optimized. An asymptotically consistent way is to find a point estimate, since with infinite data the posterior distribution reduces to a point estimate.

For example, in case of a single exponential family model with a conjugate prior, Equation (2.17) in Section 3 gives a full description of the posterior and is of the same form as the conjugate prior distribution, that is,

$$p(\theta|X,\varrho,\chi) = \frac{1}{Z}\exp\{\theta^T(\sum_{i=1}^{N} T(\mathbf{x}_i)+\chi) - (N+\varrho)\kappa(\theta) + \log b(\mathbf{x}_i) + \log d(\varrho,\chi)\}. \tag{4.1}$$

One possibility of optimizing a model is to solve the parameter values that maximize the posterior probability, that is, the maximum a posteriori (MAP) point estimate. The MAP estimate can be found by first taking a logarithm of Equation (4.1) and dropping out the terms not depending on θ

$$\log p(\theta|X,\varrho,\chi) \propto \theta^T(\sum_{i=1}^{N} T(\mathbf{x}_i)+\chi) - (N+\varrho)\kappa(\theta) \quad .$$

Since this is a concave function, the maximum occurs where the derivative with respect to θ is zero:

$$\begin{aligned}(\sum_{i=1}^{N} T(\mathbf{x}_i)+\chi) - (N+\varrho)\mu(\theta) = 0 & \qquad \Leftrightarrow \\ \mu(\theta) = (\sum_{i=1}^{N} T(\mathbf{x}_i)+\chi)/(N+\varrho) & \qquad .\end{aligned} \tag{4.2}$$

Here Equation (2.11) is applied to express the maximum in terms of dual parameters μ. For a single exponential family model the MAP solution is unique. Latent variable models usually have several local maxima.

Maximum Likelihood and Conditional Maximum Likelihood

From the point of optimization, computing a maximum likelihood estimate of a joint density model is far more simpler than computing a conditional maximum likelihood estimate.

As discussed in the earlier sections, the conditional (log) likelihood of the discriminative joint density model is

$$\sum_{i=1}^{N}\log p(c_i|\mathbf{x}_i,\theta) = \sum_{i=1}^{N}\log p(c_i,\mathbf{x}_i|\theta) - \sum_{i=1}^{N}\log\sum_{k=1}^{K} p(c=k,\mathbf{x}_i|\theta) \quad . \tag{4.3}$$

The first term is the (log) likelihood. The second term is the normalization, $p(\mathbf{x})$ obtained through marginalization. This term is the reason why maximum likelihood and conditional maximum likelihood have a different maxima, and also the reason why conditional maximum likelihood estimates are harder to compute.

4.1.3 Gradient Methods

The simplest method of optimizing a model is to use the so called batch gradient update; the derivative of the objective function is set to zero and the parameter values are solved. For example, in a mixture model the derivative will be

$$\frac{\partial}{\partial \theta_{\pi_k}} \quad : \quad \sum_i \sum_j p(j|\theta, x_i) \left\{ \delta(j,k) - \mu_{\pi_k} \right\} = \sum_i E_{p(j|\theta, x_i)} \left\{ \delta(j,k) - \mu_{\pi_k} \right\} \tag{4.4}$$

$$\frac{\partial}{\partial \theta_{\mathbf{x}_k}} \quad : \quad \sum_i p(j|\theta, x_i) \left\{ T(x_i) - \mu_{\theta_{x_j}} \right\} = \sum_i E_{p(j|\theta, x_i)} \left\{ T(x_i) - \mu_{\theta_{x_j}} \right\} \ , \tag{4.5}$$

where i indexes the data item and j the mixture component, and

$$p(j|\theta, x_i) = \frac{\exp\{\theta_\pi^T T(\pi_j) - \kappa(\theta_\pi) + \theta_{x_j} T(x_i) - \kappa(\theta_{x_j}) + \log b(x_i)\}}{\sum_k \exp\{\theta_\pi^T T(\pi_k) - \kappa(\theta_\pi) + \theta_{x_k} T(x_i) - \kappa(\theta_{x_k}) + \log b(x_i)\}} \quad .$$

If the parameters cannot be solved in a closed form, a gradient ascent update is used; update the parameters towards the gradient direction. The problem common to the simple gradient ascent-based algorithms is that the update step length needs to be selected empirically. Too optimistic step lengths may overshoot the local optimum resulting in a decrease of the objective function (and hence slower convergence), whereas too pessimistic step lengths lead to slow convergence. Line-search -based methods, such as conjugate gradient, try to cope with the problem by finding the maximum value of the cost function along the direction of the gradient (or some other direction, defined as a function of the gradient, as in conjugate gradient method). The optimal increase is found by carrying out for example a golden section search along the line. However, line search methods are not particularly suitable for cases where the computational cost of evaluating the objective function is high, since for example the conjugate gradient method requires an accurate line search, meaning several evaluations of the objective function.

So-called second order gradient ascent algorithms take a different approach. The methods try to take into account the curvature of the model and adjust the step length accordingly. This type of optimization methods are referred to as (quasi-)Newton methods. The classic Newton method constructs a (concave) quadratic approximation of the objective function and then jumps to the unique maximum of the approximation. This automatic step length selection may be costly in terms of computation since inversion of the Hessian matrix (second derivative of the objective function) of the model is required.

Quasi-Newton methods do not compute the Hessian, but rather maintain an approximation of the inverse of Hessian. A so-called BFGS algorithm, an acronym of Broyden-Fletcher-Goldfarb-Shanno (Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970), obtains a numerical approximation of the curvature using only gradient information, and uses this information to optimize the model parameters along the approximate natural gradient directions. The method carries out a line search along the natural gradient direction. A benefit of the Quasi-Newton methods is that the line search need not be as accurate.

4.1.4 Expectation – Maximization

Expectation-maximization (EM) algorithm is the most common algorithm used for optimizing latent variable models. It is a special case of the Quasi-Newton approach and operates by constructing a global lower bound for the objective function, likelihood where the latent variables have been marginalized out, with the aid of Jensen's inequality. The bound is tight (that is, the approximation and the true objective function have the same value) at the current values of the model parameters, $\hat{\theta}$, and its gradient equals that of the objective function at that point. The lower bound is global, and therefore it is guaranteed that optimizing the bound always increases the value of the objective function. Furthermore, the lower bound is concave, and can therefore be optimized with one (Newton) iteration. Since the gradients of the objective function and its lower bound are equal at $\hat{\theta}$, the EM can be interpreted as a kind of gradient ascent algorithm with automatic selection of step length. The main benefit of EM-type algorithms is therefore that the optimization can be carried out with less evaluations of the value of the objective function, which makes it suitable for optimizing complex models.

In (marginalized) joint likelihood latent variable models the cost function is

$$\sum_i \log p(\mathbf{x}_i|\theta) = \sum_i \log \sum_h p(\mathbf{x}_i, h|\theta) \quad , \tag{4.6}$$

where $x_i,\ i = 1, \ldots, N$ is the data, h the values of latent variable(s), and θ the model parameters. Below, $p(\mathbf{x}, h|\theta)$ is assumed to belong to the exponential family.

The objective function of the EM algorithm is a global lower bound for $\log p(\mathbf{x}|\theta)$, obtained with the aid of Jensen's inequality (Buntine, 2002):

$$\begin{aligned} \log p(\mathbf{x}|\theta) &= \log \sum_h p(\mathbf{x}, h|\theta) \geq \mathcal{F}(\theta) \\ &= \log p(\mathbf{x}|\theta) - D_{KL}(q(h), p(h|\mathbf{x}, \theta)) \qquad (4.7) \\ &= E_{q(h)}\{\log p(\mathbf{x}, h|\theta)\} + H(q(h)) \ , \qquad (4.8) \end{aligned}$$

where $D_{KL}(\cdot, \cdot)$ denotes the Kullback-Leibler divergence and $H(\cdot)$ the entropy. The globality of the lower bound is easy to see from Equation (4.7), since $D_{KL}(\cdot, \cdot)$ is always ≥ 0.

The EM algorithm operates by iteratively minimizing the Kullback-Leibler divergence of Equation (4.7) with respect to the distribution over the latent variable, $q(h)$, at $\theta = \hat{\theta}$, and then optimizing the likelihood, Equation (4.8), with respect to θ, keeping $q(h)$ fixed. In an EM algorithm there is always a distribution $q(h)$ such that the bound is tight (equality), making the bound tangential to the likelihood at $\theta = \hat{\theta}$ (Buntine, 2002).

If the distribution family $q(h)$ is not rich enough to include $p(h|\mathbf{x}, \hat{\theta})$, i.e. the Kullback-Leibler divergence of Equation (4.7) cannot generally be made to vanish, a variational algorithm results, where $q(h)$ is the variational approximation. EM algorithm is thus a special case of the variational method, see for example Beal (2003). The convergence guarantee of the EM algorithm can be shown to hold also when the update is applied only to subsets of the model parameters. The update scheme is then referred to as the generalized EM (Neal and Hinton, 1999).

Interpretation in Terms of Exponential Families

For simplicity, let us inspect the case of a simple mixture model

$$\prod_i \sum_j p(\mathbf{x}_i, h_j|\theta) = \prod_i \sum_j \exp\{\theta_\pi^T T(\pi_j) - \log \sum_{j'} e^{\theta_\pi(j)} + \theta_x(j) T(\mathbf{x}_i) - \kappa(\theta_x(j)) - \log b(\mathbf{x}_i)\} \quad , \tag{4.9}$$

where θ_π are the natural parameters of the probability of choosing a mixture component, the sufficient statistics $T(\pi_j)$ is a 0-1 vector which picks a mixture component j from θ_π, $\theta_x(j)$ denotes the natural parameters of a distribution generating $\mathbf{x}$ in a component j, and $\kappa(\theta_x(j))$ is the corresponding log-partition function. $b(\mathbf{x}_i)$ is again a constant.

Distribution. EM produces a global lower bound for the likelihood via Jensen's inequality for convex functions $\phi(\mathbf{x}_i)$,

$$\frac{\sum_i q_i \phi(\mathbf{x}_i)}{\sum q_i} \geq \phi\left(\sum_i \frac{q_i}{\sum q_{i'}} \mathbf{x}_i\right) \quad . \tag{4.10}$$

The inequality can be applied to Equation (4.9) by introducing a latent variable Q. For each data item i the realization of Q is a J-dimensional multinomial. The slots of the multinomial are denoted by $q(i,j)$. Equation (4.9) can therefore be written as

$$\begin{aligned} & \prod_i \sum_j e^{\theta_\pi^T T(\pi_j) - \log \sum_{j'} e^{\theta_\pi(j)} + \theta_x(j) T(\mathbf{x}_i) - \kappa(\theta_x(j)) - b(\mathbf{x}_i)} \frac{q(i,j)}{q(i,j)} \geq \\ & \prod_i e^{\sum_j q(i,j) \left(\theta_\pi^T T(\pi_j) - \log \sum_{j'} e^{\theta_\pi(j)} + \theta_x(j) T(\mathbf{x}_i) - \kappa(\theta_x(j)) - b(\mathbf{x}_i) - \log(q(i,j)\right)} \quad (4.11) \\ = \; & \prod_i e^{\sum_j q(i,j) \log p(\mathbf{x}_i, h_j|\theta) - q(i,j) \log q(i,j)} \quad . \quad (4.12) \end{aligned}$$

From Equation (4.12) the correspondence to Equation (4.8) is already easy to see. Equation (4.12) forms also an exponential family distribution, expressed in terms of dual parameters, $q = \mu$. The distribution can be expressed also in terms of natural parameters θ_q by computing the conjugate function via Legendre-Fenchel transformation,

$$\kappa(\theta)^* = \sup_\mu \{q(i,\cdot)^T \theta_{q_i} - q(i,\cdot)^T \log q(i,\cdot)\} = \log \sum_j e^{\theta_{q_i}(j)}. \tag{4.13}$$

The canonical representation of the EM lower bound is therefore

$$e^{\sum_i (\sum_j \theta_{q_i}(j) \log p(\mathbf{x}_i, h_j|\theta) - \log \sum_j e^{\theta_{q_i}(j)})} = e^{\theta_q^T T(\log p(\mathbf{x}, h|\theta)) - \kappa(\theta_q)} \quad , \tag{4.14}$$

which forms a distribution that resembles to a multinomial[1]. The lower bound thus corresponds to defining a single exponential family model for the whole likelihood with rich sufficient statistics.

[1] However, the support of sufficient statistics is not positive integers, but the complete axis of real values

E-step. Finding a distribution q such that Equation (4.12) is maximized forms the expectation step in the EM algorithm,

$$E : \kappa(\log p(\mathbf{x}, h_\cdot|\theta))^* = \sup_{q(i,\cdot)} \{q(i,\cdot) \log p(\mathbf{x}, h_\cdot|\theta) - q(i,\cdot) \log q(i,\cdot)\} \quad . \tag{4.15}$$

This is a Legendre-Fenchel transformation (see Wainwright and Jordan (2003) for additional discussion). Theory of convex functions (Rockafellar, 1970; Barndorff-Nielsen, 1978) can be used to show that the supremum is unique.

By solving the maximization problem and inserting q back into Equation (4.15), we get

$$\kappa(\log p(\mathbf{x}, h_\cdot|\theta))^* = \log \sum_j p(\mathbf{x}, h_j|\theta) \quad , \tag{4.16}$$

that is, the distribution q that maximizes Equation (4.15) gives the log likelihood of a mixture model in Equation (4.9), as it should.

M-step. After solving the distribution of latent variables in the E-step, model parameters can be optimized. This is done in the maximization step. To see how the updates look like, we continue from Equation (4.11), assuming that $q(i, j)$ are known and fixed. By taking a logarithm and omitting the terms that do not depend on parameters θ, we may write

$$\begin{aligned}
&\sum_i \sum_j q(i,j) \left(\theta_\pi^T T(\pi_j) - \log \sum_{j'} e^{\theta_\pi(j)} + \theta_x(j) T(\mathbf{x}_i) - \kappa(\theta_{x,j}) \right) = \\
&\theta_\pi^T \sum_i q(i,\cdot) - N \log \sum_{j'} e^{\theta_\pi(j)} + \sum_j \theta_x(j) \left(\sum_i q(i,j) T(\mathbf{x}_i) \right) \\
&\qquad\qquad - \sum_i q(i,j) \kappa(\theta_{x,j}) \quad .
\end{aligned}$$

Since a non-negative weighted sum of convex functions is convex (Boyd and Vandenberghe, 2004), the optimization problem can again be formalized as finding the Legendre-Fenchel conjugate, that is, for the case of parameters $\theta_{x,j}$

$$\sup_{\theta_{x,j}} \{\theta_{x,j} \sum_i q(i,j) T(\mathbf{x}_i) - (\sum_i q(i,j)) \kappa(\theta_{x,j})\} = \sup_{\theta_{x,j}} \{\theta_{x,j} \mu'_{x,j} - \kappa'(\theta_{x,j}, Q)\}$$

and similarly for θ_π. The solution is again unique as in the E step.

Update rules are obtained by requiring that the derivatives with respect to each of the variables are zero. Updates are expressed in terms of dual variables, that is,

$$\mu_{\pi,j} = \frac{1}{N} \sum_i q(i,j) \tag{4.17}$$

$$\mu_{x,j} = \frac{\sum_i q(i,j) T(\mathbf{x}_i)}{\sum_i q(i,j)} \tag{4.18}$$

Connection to Gradient Ascent

We next show the connection between the EM algorithm and gradient ascent. If we assume that all the probabilities within the model are expressed using exponential family distributions, the derivative of the model is

$$\frac{\partial}{\partial \theta} \sum_i \log \sum_h p(\mathbf{x}_i, h|\theta) = \sum_{i,h} p(h|\mathbf{x}_i, \theta)\{T(\mathbf{x}_i) - \mu(\theta)\}, \tag{4.19}$$

by using Equation (2.11), and denoting sufficient statistics associated with the current observation i by $T(\mathbf{x}_i)$.

Gradient ascent operates by iterating

$$\mu(\theta) = \mu(\hat{\theta}) + \gamma^{-1} \frac{\partial}{\partial \theta} \sum_i \log \sum_h p(\mathbf{x}_i, h|\theta),$$

where γ^{-1} is a small value. Inserting Equation (4.19) and evaluating the derivative at $\hat{\theta}$, we may solve for $\mu(\theta)$, resulting in

$$\mu(\theta) = \frac{\gamma \mu(\hat{\theta}) + \sum_{i,h} p(h|\mathbf{x}_i, \hat{\theta}) T(\mathbf{x}_i)}{\gamma + \sum_{i,h} p(h|\mathbf{x}_i, \hat{\theta})}. \tag{4.20}$$

This is similar to EM update rules. The difference is that in gradient ascent the update step length is regularized with γ.

4.2 Discriminative Expectation – Maximization

In case of discriminative joint density models the objective function to be optimized, the conditional likelihood, is a rational function,

$$\mathcal{L}(C|X;\theta) = \prod_{i=1}^{N} p(c_i|\mathbf{x}_i, \theta) = \prod_{i=1}^{N} \frac{p(c_i, \mathbf{x}_i|\theta)}{p(\mathbf{x}_i|\theta)} \ . \tag{4.21}$$

The computational complexity of conditional likelihood is higher than that of likelihood, since the denominator contains a marginalization over the variable C, that is, $p(\mathbf{x}_i|\theta) = \sum_k p(c = k, \mathbf{x}_i|\theta)$. Therefore line-search -based methods requiring several evaluations of the objective function can be infeasible in case of complex models. One solution is to incorporate more information of the model into the optimization by taking into account the curvature of the objective function.

The EM algorithm as such is not applicable, but an EM-type algorithm can be derived by forming a global lower bound as a combination of an EM-type lower bound for the numerator and an upper bound for the denominator. The approach, termed Conditional EM (CEM), was first presented and studied rigorously by Jebara and Pentland (2001); Jebara (2001). The resulting formulas, however, turned out to be very complicated, hindering their practical use, and obtaining even a conservative estimate of the bound was computationally demanding (Jebara, 2001). Moreover, the bounds allow only a very small step size which makes optimization slow and hence further increases computational demands. So far the only application of the CEM is Afify (2005), who applied the method for Gaussian mixture models.

In Publication 5, a faster EM-type algorithm (discriminative EM) for optimizing conditional likelihoods is proposed. The algorithm uses the same intuition as the CEM, but instead of guaranteed convergence, the DEM relies on local approximations which are considerably faster to compute. The trade-off is that the convergence is not always globally guaranteed (depending on the chosen approximation type). The computational complexity of the resulting DEM algorithm is somewhat bigger than the ordinary EM ($\mathcal{O}(S^3T^2)$ vs. $\mathcal{O}(S^2T)$, where S is the number of states and T the length of the data sequence), but it is already feasible for many model structures, such as hidden Markov models. The algorithm can be used in the same way as an ordinary EM.

We discuss the derivation of the discriminative EM algorithm in the following. As in the earlier CEM algorithm (Jebara and Pentland, 2001), the problematic part is the construction of an upper bound for the denominator in Equation (4.21). An insight to the problem can be obtained by looking at the second derivative of the log-likelihood.

Derivative and Curvature of Log-Likelihood

For a mixture model, the second order gradient, the curvature, of log-likelihood is

$$\begin{aligned}\frac{\partial^2}{\partial\theta_{\pi_k}\partial\theta_{\pi_j}}\mathcal{L}(X;\theta) =& \sum_i E_{p(j|\theta,x_i)}\{\delta(j,l)-\mu_{\pi_l}\}\{\delta(j,k)-\mu_{\pi_k}\}+\\ &-\left(E_{p(j|\theta,x_i)}\{\delta(j,l)-\mu_{\pi_l}\}\right)\left(E_{p(j|\theta,x_i)}\{\delta(j,k)-\mu_{\pi_k}\}\right)\\ &-E_{p(j|\theta,x_i)}\{\Sigma_{\pi_{kl}}\} = C(T(\mathbf{x}),\mu_{\pi_{kl}})-\Sigma_{\pi_{kl}}\end{aligned} \tag{4.22}$$

$$\begin{aligned}\frac{\partial^2}{\partial\theta_{x_k,1}\theta_{x_k,2}}\mathcal{L}(X;\theta) :& \sum_i E_{p(j|\theta,x_i)}\{T(\mathbf{x}_i)-\mu_{\theta_{x_k,1}}\}\{T(\mathbf{x}_i)-\mu_{\theta_{x_k,2}}\}+\\ &-\Big(E_{p(j|\theta,x_i)}\{T(\mathbf{x}_i)-\mu_{\theta_{x_k,1}}\}\Big)\Big(E_{p(j|\theta,x_i)}\{T(\mathbf{x}_i)-\mu_{\theta_{x_k,2}}\}\Big)\\ &-E_{p(j|\theta,x_i)}\{\Sigma_{\theta_{x_k,12}}\} = C(T(\mathbf{x}),\mu_{\theta_{x_k,12}})-\Sigma_{\theta_{x_k,12}}\end{aligned} \tag{4.23}$$

Here the $\Sigma_{\pi_{kl}} = \frac{\partial^2}{\partial\theta_{\pi_k}\partial\theta_{\pi_l}}\kappa(\theta_\pi)$ are elements of Fisher information matrix Σ which does not depend on data, and $C(T(\mathbf{x}),\mu_{\pi_{kl}})$ are elements of the expected covariance matrix $C(T(\mathbf{x}),\mu)$ of the data, computed over latent variables. Both matrices are positive definite and hence convex (Boyd and Vandenberghe, 2004), see also Section 2.2.1. The curvature of log-likelihood therefore alternates, with $C(T(\mathbf{x}),\mu)$ forming an upper bound and Σ a lower bound for the curvature. The curvature of the EM lower bound is Σ. The benefit of the lower bound thus is that it depends only on the chosen model, not data. The upper bound $C(T(\mathbf{x}),\mu)$ is more problematic, since it depends both on data and on parameters.

4.2.1 Derivation of DEM Lower Bound for Exponential Families

The upper and lower bounds of log-likelihood are next used for deriving the DEM algorithm in case of discrete latent variables. The derivation begins from the log-conditional likelihood,

$$\log\mathcal{L}(C|X;\theta) = \sum_i \log p(c_i,\mathbf{x}_i|\theta) - \log p(\mathbf{x}_i|\theta) = F(C,X;\theta) - G(X;\theta)\ . \tag{4.24}$$

Here we denote by $F(C, X; \theta)$ the joint log-likelihood and by $G(X; \theta)$ the marginalised log-likelihood. As mentioned earlier, in order to obtain a global lower bound for log-conditional likelihood, a lower bound for $F(C, X; \theta)$ and an upper bound for $G(X; \theta)$ is required.

A lower bound for $F(C, X; \theta)$ is simple to construct since it is a normal EM lower bound, obtained by Jensen's inequality,

$$\begin{aligned}
&\sum_i \log p(c_i, \mathbf{x}_i|\theta) = \\
&\sum_i \log \sum_{j \in J_{c,i}} \exp\{\theta_\pi^T T(\pi_j) - \log \sum_{j'} e^{\theta_\pi(j)} + \theta_x(j) T(\mathbf{x}_i) - \kappa(\theta_x(j)) - b(\mathbf{x}_i)\} \geq \\
&\sum_i \sum_j q(i,j) \big(\theta_\pi^T T(\pi_j) - \log \sum_{j'} e^{\theta_\pi(j)} + \theta_x(j) T(\mathbf{x}_i) \\
&\qquad\qquad - \kappa(\theta_x(j)) - b(\mathbf{x}_i) - \log(q(i,j))\big) \ .
\end{aligned} \tag{4.25}$$

As already mentioned in Section 4.1.4, the resulting EM lower bound has the following properties:

1. The bound is tight at current parameter values $\hat{\theta}$.
2. The bound has the same gradient as the original function at $\hat{\theta}$.
3. The curvature of the bound is a lower bound of the curvature of the likelihood – it is equal to Σ in Equation (4.23).

Since the curvature of the lower bound is always less (or equal) than the curvature of the log-likelihood, the bound is global.

The difficult part in forming an EM-type algorithm for the conditional likelihood is to construct an upper bound for $G(X; \theta)$. The requirements for the bound should naturally mirror those of the lower bound, that is:

1. The bound is tight at current parameter values $\hat{\theta}$.
2. The bound has the same gradient as the original function at $\hat{\theta}$.
3. The curvature of the bound is an upper bound of the curvature of the likelihood – it is greater than or equal to $C(T(\mathbf{x}), \mu)$ in Equation (4.23).

Additionally, the upper bound should preferably be of the same functional form as the EM lower bound in Equation (4.25), consisting of a a linear term plus a convex function $\theta T(\mathbf{x}_i) + \kappa(\theta)$, since the form is known to be simple to optimize and results in low computational complexity of the algorithm. In Publication 5, the upper bound is found by solving a trial function which is of the desired form. The resulting upper bound has the functional form

$$\begin{aligned}
\log p(\mathbf{x}|\theta) = G(X; \theta) \leq \log p(\mathbf{x}|\hat{\theta}) + \sum_h p(h|\mathbf{x}, \hat{\theta})(T(\mathbf{x}) - \mu(\hat{\theta}))(\theta - \hat{\theta}) \\
+ \Lambda D(\theta, \hat{\theta}) \ .
\end{aligned} \tag{4.26}$$

The first term makes the bound tight at $\hat{\theta}$, and the second term makes the gradient equal to that of the original function at $\hat{\theta}$. In the last term, the Λ is an appropriate constant and $D(\theta, \hat{\theta})$ a distance function. Since the first terms already match the function value and its gradient, it is naturally required that the distance function fulfills the conditions $D(\hat{\theta},\hat{\theta}) = \frac{\partial}{\partial\theta} D(\theta,\hat{\theta})\left.\right|_{\theta=\hat{\theta}} = 0$. Even though there exist also other function classes fulfilling this condition, the Bregman divergences are a form that result in simple update rules for the exponential family models. The second role of function $D(\theta,\hat{\theta})$ is to take care of the curvature of the bound outside of $\hat{\theta}$. The constant Λ is the curvature of the upper bound. The requirement is that it is an upper bound of the curvature of the log-likelihood at $\hat{\theta}$; $\Lambda \geq C(T(\mathbf{x}),\mu)$.

The DEM algorithm was applied to discriminative HMMs in Publication 5.

4.2.2 Relation to Other EM-type Approaches

Relation to Extended Baum-Welch

The current state-of-the-art optimization method in speech recognition with hidden Markov models is called extended Baum-Welch, EBW (Gopalakrishnan et al., 1991; Povey et al., 2003). The algorithm was first presented by Gopalakrishnan et al. (1991) for multinomial observation distributions, and extended by Normandin (1991) to Gaussian distributions.

The EBW can be easily extended to optimize mixture models, for example the Gaussian mixture model (Klautau, 2003), by assuming that instead of time series, the data consists of N samples of length one. The algorithm of Klautau (2003) uses the same heuristics as EBW.

The update rules of EBW resemble those of the EM algorithm, plus some heuristics, which have been developed experimentally during the last seventeen years. From our point of view, the algorithm can be interpreted to lower bound both $\log p(c,\mathbf{x}|\theta)$ and $\log p(\mathbf{x}|\theta)$ using Jensen's inequality. Using Equation (4.7), the result is

$$\begin{aligned}&\log p(c,x|\theta) - D_{KL}(q_C(h), p(h|c,\mathbf{x},\theta)) + \\ &\quad - \log p(\mathbf{x}|\theta) + D_{KL}(q_F(h), p(h|\mathbf{x},\theta)) \quad ,\end{aligned}$$

where $q_C(h), q_F(h)$ denote the latent variable distributions in cases where c is known ("clamped"), or marginalized out ("free"), respectively. Since the last term is positive, the globality of the lower bound cannot be guaranteed.

By looking at the derivatives of the EBW upper bound, the following can be said:

1. The bound is tight at current parameter values $\hat{\theta}$.
2. The bound has the same derivative as the original function at $\hat{\theta}$.
3. The curvature of the bound is assumed to be an upper bound of the curvature of log-likelihood, that is, $\Sigma + \gamma \geq C(T(c,\mathbf{x}),\mu) - \Sigma$ in Equation (4.23), where γ is heuristic.

Therefore, a regularization term γ is needed in the update formulas, resulting in a functional form similar to Equation (4.20). Gopalakrishnan et al. (1991) present a formula for computing a regularization value γ which is large enough such that convergence can be guaranteed. However, the resulting value is so large that even in the original publication approximations needed to be made for practical implementation of EBW. As noted in Section 4.1.4, a large regularization coefficient reduces the EBW to a gradient ascent-type optimization algorithm which is known to converge to a local optimum. Since the original publication, choosing the proper amount of regularization has been under considerable debate.

Assuming diagonal covariances, the update rules in EBW are of form

$$\mu = \frac{\sum(q^C - q^F)T(\mathbf{x}) + B\hat{\mu}}{\sum(q^C - q^F) + B} \tag{4.27}$$

$$\sigma^2 = \frac{\sum(q^C - q^F)T(\mathbf{x}^2) + B(\hat{\sigma}^2 + \hat{\mu}^2)}{\sum(q^C - q^F) + B} - \mu^2. \tag{4.28}$$

The update of σ^2 can be understood as an update by estimating the second moment, that is, $\sigma^2 = E[\mathbf{x}^2] - E[\mathbf{x}]^2$.

Currently the best heuristics for choosing B are reported in Woodland and Povey (2002), where B are set on a per-Gaussian level to the larger of i) twice the smallest value needed to ensure positive variances, or ii) $\sum q^F$ times a further constant E, which is generally set to 1 or 2.

EBW and TM Algorithms in Terms of a Log-Linear Upper Bound

Edwards and Lauritzen (2001) have presented a TM algorithm (possibly from Tilted Maximization) which constructs a linear approximation of the numerator, $G(X;\theta)$. In case of exponential family models, the algorithm is very close to EBW with a certain type of heuristics, as shown next.

The approximate bound can be explained alternatively as a method where first a log-linear upper bound is formed of the individual components in the joint likelihood $\log p(\mathbf{x}_i, h|\theta)$, and then a Jensen lower bound of the upper bound is computed.

Each of the components are first upper bounded by applying the following lemma:

Lemma 4.2.1 (Log-linear upper bound for exponential family models) *Due to the convexity of the log-partition function $\kappa(\theta)$, exponential family distributions have a log-linear upper bound of the form*

$$\begin{aligned}&\exp\{T(\mathbf{x})^T\theta - \kappa(\theta) - \log b(\mathbf{x})\} \leq \\ &\exp\{(T(\mathbf{x}) + \mu(\hat{\theta}))^T(\hat{\theta} - \theta) - \kappa(\hat{\theta}) - \log b(\mathbf{x})\}.\end{aligned} \tag{4.29}$$

Proof The term in the exponential is a Bregman divergence. Proof follows directly from the property that the Bregman divergence is always ≥ 0 ∎

This eventually yields update formulas of type:

$$\mu = \frac{\sum(q^C - q^F)T + q^F\hat{\mu}}{\sum q^C} = \frac{\sum(q^C - q^F)T + q^F\hat{\mu}}{\sum(q^C - q^F) + q^F}, \tag{4.30}$$

giving the EBW heuristic where $B = \sum q^F$.

This upper bound has two desirable properties:

1. The bound is tight at $\hat{\theta}$.

2. The bound has the same derivative as the original function at $\hat{\theta}$.

However, since the bound is a lower bound of an upper bound, nothing general can be said about its general behavior with respect to the original function. This is also acknowledged in Edwards and Lauritzen (2001).

The difference between EBW and the TM algorithm is that TM can be applied to more general models than exponential families, since it does not assume that $F(C, X; \theta)$ needs to be lower bounded. After computing the gradient using the approximation described above the algorithm proceeds by a line search along that direction. In this respect it is worse than DEM which tries to take into account the model curvature to compute the best step length.

Relation Between Discriminative EM and Conditional Expectation Maximization

In conditional EM, Jebara and Pentland (1999) bound the $\log p(\mathbf{x}|\theta)$ by the function itself (plus a constant), $p(\mathbf{x}|\theta) + 1$, and thus achieve a global lower bound. However, the resulting update rules are complicated, which hinders their practical use.

The next development in Jebara and Pentland (2001) is to solve the upper bound by taking a trial function which has the same functional form as the lower bound, that is, for arbitrary data item $\mathbf{y}$:

$$\sum_h q(h) \left[T(\mathbf{y})\theta - \log Z(\theta)\right] - \log b(\mathbf{y}),$$

and solve its coefficients $q(h)$, $T(\mathbf{y})$, $\log b(\mathbf{y})$ so that (i) the bound is tight at $\hat{\theta}$ (thus getting $\log b(\mathbf{y})$), and (ii) has the same derivative as the log-likelihood at $\hat{\theta}$ (getting $T(\mathbf{y})$).

Inserting $\log b(\mathbf{y})$ and $T(\mathbf{y})$, and regrouping the variables results in

$$\begin{aligned} \log p(\mathbf{x}|\hat{\theta}) &+ \sum_h p(h|\mathbf{x}, \hat{\theta})(T(\mathbf{x}) - \mu(\hat{\theta}))(\theta - \hat{\theta}) \\ &+ q(h)B(\theta, \hat{\theta}) \quad , \end{aligned} \tag{4.31}$$

where $T(\mathbf{x})$ and $\mu(\hat{\theta})$ are the observed and expected sufficient statistics, respectively. The term $B(\theta, \hat{\theta})$ is the Bregman divergence between θ and $\hat{\theta}$. Jebara and Pentland (2001) proceed by mapping the Bregman distances to a parabola (since every convex function has a diffeomorphic mapping to a parabola), and solve the remaining values, $q(h)$. It turns out that the mapping need not be solved explicitly. However, the mapping affects the resulting update rules by restricting the allowed values for $q(h)$, making optimisation difficult. Furthermore, although guaranteed, the speed of convergence and especially the computational demands reported in Jebara (2001) leave room for improvement.

The reason for the high computational complexity is that in order to obtain a global upper bound, the worst-case curvature of $C(T(\mathbf{x}), \mu)$ from Equation (4.23) needs to be found, resulting in another optimization problem. The curvature depends on the data and the current parameter values μ, and needs to be computed after each maximization step. Since curvature is a worst case estimate, it is necessarily large. As a result, the step length of the reverse Jensen approach is short, and the convergence of the algorithm thus slow. Figure 4.1 illustrates the fact.

Instead of finding the worst-case curvature, DEM estimates a local upper bound for the curvature. The estimated curvature is therefore not as pessimistic as in

CEM, and thus results in longer step lengths. The curvature can be computed directly using current parameter values, which results in lower computational complexity. The possible pitfall of the method is that convergence is not guaranteed, since the model could be (at least initially) close to the area of considerably higher curvature. However, as the optimization proceeds towards a local optimum, the overall curvature of the log-likelihood should be negative. Therefore the possibility of decreasing the likelihood because of choosing upper bounds with too low curvature should decrease as optimization is closer to convergence.

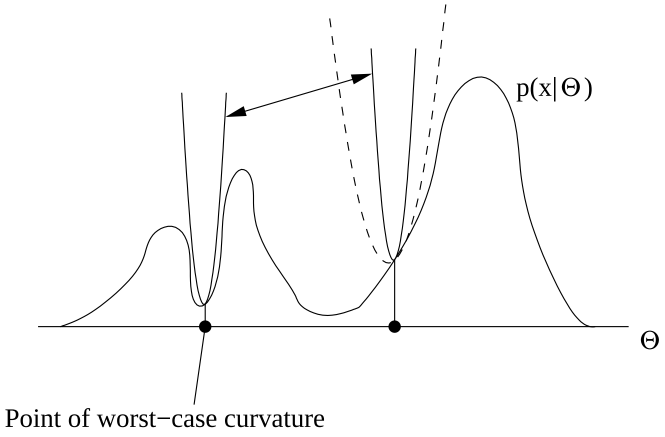

Figure 4.1: The reverse Jensen upperbound (Jebara and Pentland, 2001) searches a global upper bound for $p(\mathbf{x})$, shown in solid line. The DEM searches a local upper bound, plotted with dashed line. The curvature of the upper bound is directly related to the step length of an EM update, that is, how much the parameters are updated from their original values. Since the curvature of reverse Jensen is larger, the algorithm results in shorter step lengths and slower convergence.

4.2.3 Speedups and Alternative Methods

The benefits of the DEM over CEM are the simpler update rules and lower computational convergence. The downside is that the convergence is not globally guaranteed. However, as already discussed above, there is usually no need for guaranteed convergence in the worst conditions, since most likely they don't occur.

The computational complexity of the DEM is still somewhat higher that the corresponding EM algorithm. However, it is possible to make further approximations to make DEM faster. These are beyond the scope of this thesis and left for further work. One simple way is to compute the upper bound of the curvature, Λ, only once at the beginning of the optimization. Since the optimization proceeds to a maximum, the overall curvature of the log-likelihood should be negative when it is close to convergence, and therefore a more approximate upper bound suffices.

Recently Kim and Pavlovic (2007) provided a different approach; the problem is approached from the viewpoint of boosting (Schapire, 1990; Friedman et al., 2000). The algorithm finds a greedy solution by recursively adding mixture components to the earlier solution such that the conditional likelihood is maximized.

4.2.4 Example: DEM for Logistic Regression

The simplest example where discriminative EM algorithm can be applied is the logistic regression.

As discussed in Section 2.5, the log-likelihood of a logistic regression model is

$$\mathcal{L} = \sum_{i=1}^{N} \sum_{j=1}^{C} \delta(c_i, j) \beta_j^T \mathbf{x}_i - \log \left(\sum_{j'} e^{\beta_{j'}^T \mathbf{x}_i} \right) \quad . \tag{4.32}$$

Optimization with DEM

We next develop a discriminative EM-type algorithm for optimizing logistic regression models. We first note that the logistic regression is an exponential family model, that is,

$$p(C = c_i | \mathbf{x}_i, \beta) = \exp\{\sum_{j=1}^{C} \delta(c_i, j) \beta_j^T \mathbf{x}_i - \log \sum_{j'} e^{\beta_{j'}^T \mathbf{x}_i}\} \ .$$

By comparing this to Equation (2.2.1), we note that $T(\mathbf{x}) = \delta(c_i, j)$, $\theta_j = \beta_k^T \mathbf{x}_i$, and $\kappa(\theta) = \log \sum_{j'} e^{\beta_{j'}^T \mathbf{x}_i}$.

Now, by using Equation (4.29), an upper bound required for the DEM can be constructed,

$$\begin{aligned} -\log \sum_{j'} e^{\beta_{j'}^T \mathbf{x}_i} \quad &\leq \quad -\log \sum_{j'} e^{\hat{\beta}_{j'}^T \mathbf{x}_i} \\ &\quad - \left(\frac{\partial}{\partial \beta^T \mathbf{x}_i} - \log \sum_{j'} e^{\beta_{j'}^T \mathbf{x}_i} \Bigg|_{\beta = \hat{\beta}} \right)^T \left(\beta - \hat{\beta} \right)^T \mathbf{x}_i . \end{aligned}$$

This can be written in a more convenient form

$$-\log \sum_{j'} e^{\beta_{j'}^T \mathbf{x}_i} \leq -\log \sum_{j'} e^{\hat{\beta}_{j'}^T \mathbf{x}_i} + \sum_{k} \gamma(i,j) \left(\beta_j - \hat{\beta}_j \right)^T \mathbf{x}_i,$$

where $\gamma(i,j) = \frac{e^{\beta_j^T \mathbf{x}_i}}{\sum_{j'} e^{\beta_{j'}^T \mathbf{x}_i}}$. At this point we have a (log)linear upper bound for logistic regression model.

In order to have an EM-type optimization algorithm, we need to make the bound concave. This can be achieved by including a curvature term $\Lambda D(\beta, \hat{\beta})$, having the properties discussed in Section 4.2. The resulting upper bound is then of the form

$$\log \left(\sum_{j'} e^{\hat{\beta}_{j'}^T \mathbf{x}_i} \right) + \sum_{j} \gamma(i,j)(\beta_j - \hat{\beta}_j)\mathbf{x}_i + \Lambda D(\beta, \hat{\beta}), \tag{4.33}$$

where $\gamma(i,j)$ is the “latent” variable.

Choosing Λ

The curvature Λ can be computed from the trace of the Hessian. A local upper bound is

$$\Lambda(j) = \sum_{i,d} \gamma(j)(1-\gamma(j))\mathbf{x}_{id}^2 \quad , \tag{4.34}$$

where index d runs over the dimensionality of $\mathbf{x}$. In this case also a global upper bound is easy to obtain,

$$\Lambda^G = \sum_{i,d} \mathbf{x}_{id}^2. \tag{4.35}$$

Choosing $D(\beta, \hat{\beta})$

In the case of logistic regression we are free to choose either of the functional forms presented earlier. By using a parabola, the Bregman divergence for the Gaussian, as the functional form of the upper bound we get simple update formulas for β:

$$\beta_j = \hat{\beta}_j + \frac{1}{\Lambda}\sum_i \left(\delta(c_i, j) - \gamma(i,j)\right)\mathbf{x}_i \quad . \tag{4.36}$$

The similarity of the DEM to Newton methods is evident by comparing Equation (4.36) to Newton iteration formulas (see e.g. Bishop (1995)).

Experiment

An experiment was carried out with the data used in the logistic regression model of Publication 8. The performance of the discriminative EM algorithm with a global upper bound and local upper bound was compared to the performance of gradient ascent with a golden section line search. The performance measure was perplexity, measured against the required number of function value evaluations. As can be seen form Figure 4.2, the EM algorithm converges faster than the gradient ascent with line search. A conjugate gradient-based update scheme outperforms both. However, a similar type of boosting method for EM algorithm based on natural gradients is available (Honkela et al., 2007), which could improve the convergence speed of DEM.

4.3 Sampling from Posterior Distribution

In case of finite data, the posterior of the model parameters forms a distribution. Since this is the case in all real world applications, point estimates often result in suboptimal predictions. The reason for this is overfitting; a too complex model explains the learning data very well but does not generalize to new data as well as a simpler one (Hastie et al., 2001). Therefore a full posterior distribution should be used.

Solving the posterior distribution analytically is possible for a very small family of models, and thus approximate methods are used in general. There are two main approaches, either an approximation to a posterior (mode) is found by restricting the approximation to a model family which is tractable (Beal, 2003; Minka, 2001; Wainwright and Jordan, 2003), or a numerical approximation of the posterior is obtained by drawing samples from the posterior (Gelman et al., 2003; Carlin and Louis, 2000). In this thesis we concentrate on the latter.

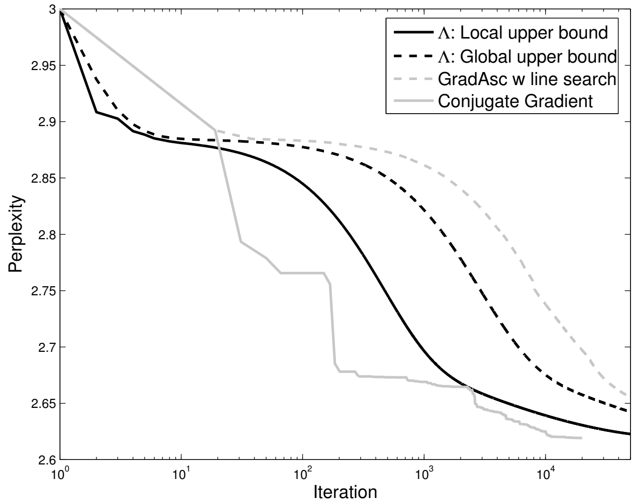

Figure 4.2: Alternative ways of choosing Λ result in different convergence. An example case where logistic regression is applied to eye movement data, plot of convergence, measured by perplexity, for the training data. Gray dashed line: The base line is formed by gradient ascent with (golden section) line search. Black dashed line: Global upper bound Λ results in slow but guaranteed convergence. Black solid line: A local approximation of λ can be computed by the trace of the Hessian, individually for each β_j. This results in faster, although not guaranteed convergence. Gray solid line: Conjugate gradient method is the state of the art method.

4.3.1 Markov Chain Monte Carlo

Many of the common approaches to obtain samples from the posterior fall under the general category of Markov chain Monte Carlo (MCMC) methods. The basis of the methods is the famous Metropolis-Hastings method (Metropolis and Ulam, 1949; Hastings, 1970). In the algorithm, samples are drawn from a jumping distribution J at time t, $J_t(\theta^* \mid \theta^{t-1})$. Samples from the distribution are accepted with probability

$$P_{acc} = \min\left\{1, \frac{p(\theta^* \mid \mathcal{D}) J_t(\theta^{t-1} \mid \theta^*)}{p(\theta^{t-1} \mid \mathcal{D}) J_t(\theta^* \mid \theta^{t-1})}\right\} \quad . \tag{4.37}$$

It is easy to prove that this type of sampling scheme converges to the target distribution $p(\theta \mid \mathcal{D})$ (Gelman et al., 2003).

Another commonly used MCMC method is Gibbs sampling (Geman and Geman, 1984; Gelfand and Smith, 1990). The method can be viewed as a special case of Metropolis-Hastings, in which the proposal distributions are defined in terms of conditional distributions solved analytically from the joint distribution.

4.3.2 Discriminative MCMC

Publication 7 presents also a method for MCMC sampling from the discriminative posterior. Another computational evidence is by Cerquides and Mántaras (2005) who show experimentally that maximum a posteriori point estimate is viable.

The discriminative posterior $p_d(\theta \mid \mathcal{D})$ can be sampled with an ordinary Metropolis-Hastings algorithm where the standard posterior $p(\theta \mid \mathcal{D})$ is simply replaced by the discriminative version. The difference to joint density MCMC sampling is that the likelihood is replaced by conditional likelihood when computing acceptance probabilities, that is,

$$p(\theta \mid \mathcal{D}) \propto \prod_{i=1}^{N} p(c_i, \mathbf{x}_i|\theta)p(\theta)$$

is replaced by

$$p_d(\theta \mid \mathcal{D}) \propto \prod_{i=1}^{N} p(c_i|\mathbf{x}_i\theta)p(\theta) \quad .$$

The proof that the Markov chain converges to the target distribution is straightforward and proceeds similarly to the joint density case.

Sampling from the discriminative posterior is computationally more demanding than joint density MCMC sampling, since each term in the conditional likelihood function is formed via Bayes formula, that is,

$$p(c_i|\mathbf{x}_i\theta) = \frac{p(c_i, \mathbf{x}_i|\theta)}{\sum_c p(c, \mathbf{x}_i\theta)} \quad . \tag{4.38}$$

Here the denominator term poses problems, since it involves a marginalization over the class variable c and latent variables h, that is,

$$p(\mathbf{x} \mid \theta) = \sum_c \int_{\mathrm{supp}(h)} p(\mathbf{x}, h, c \mid \theta) dh \quad .$$

In case of discrete latent variables, such as mixture models, the marginalization reduces to simple summations and can be computed exactly and efficiently. However, if the model contains continuous latent variables the integral needs to be evaluated numerically.

Treatment of Continuous Latent Variables

The practice of discriminative MCMC sampling from a discrete-valued latent variable model is covered in Publication 7. An item that still requires further research is sampling from a model with continuous latent variables. The problems that are encountered during sampling are illustrated next in case of a mixture of Latent Dirichlet Allocation models (see Publication 7).

The current state-of-the art generative models for text document collections are further developments of the Latent Dirichlet Allocation model (Blei et al., 2003). The data generation in the original model goes as follows: For each document, a K dimensional topic distribution π is first sampled from a Dirichlet with parameters α. Then, for each word in a document, a topic z is picked from a multinomial with the parameter values π. The word is then sampled from a topic-specific multinomial β_z, that is, the topic defines a row in a $K \times W$ dimensional matrix β consisting of multinomial parameter values, where W is the number of words in the vocabulary. The posterior of the model is thus

$$\begin{aligned} p(\theta \mid \mathcal{D}) &\propto \prod_{i=1}^{N} \int p(\pi \mid \alpha) \prod_{j=1}^{N_w^i} p(z_{ij} \mid \pi) p(w_{ij} \mid \beta, z_{ij}) d\pi \; p(\alpha, \beta) \quad , \\ &= \prod_{i=1}^{N} \int p(\pi \mid \alpha) \prod_{j=1}^{N_w^i} \sum_{k=1}^{K} \pi(k)\beta(k, j) d\pi \; p(\alpha, \beta) \quad , \end{aligned} \tag{4.39}$$

where N_w^i is the number of words in document i.

In Publication 7, another layer on top of the Latent Dirichlet Allocation model was added. A generating mixture component z_c is first sampled from π_c. For simplicity, the z_c is assumed to correspond directly the class of a document, c_i.The component indexes a row in the matrix α, and (for simplicity) contains a direct mapping to c. Now, given α, the generative model for words is an ordinary Latent Dirichlet Allocation model. For simplicity, four topic vectors were assumed. The resulting model structure was then similar the one used by Fei-Fei and Perona (2005). The posterior of the model is thus

$$\begin{aligned} p(\theta \mid \mathcal{D}) &\propto \prod_{i=1}^{N} p(z_i \mid \pi_c) \int p(\pi \mid \alpha, z_i) \prod_{j=1}^{N_w^i} p(z_{ij} \mid \pi) p(w_{ij} \mid \beta, z_{ij}) d\pi \; p(\alpha, \beta, \pi_c) \\ &= \prod_{i=1}^{N} \pi_c(c_i) \int p(\pi \mid \alpha, c_i) \prod_{j=1}^{N_w^i} \sum_{k=1}^{K} \pi(k) \beta(k, j) d\pi \; p(\alpha, \beta, \pi_c) \quad . \end{aligned} \tag{4.40}$$

In discriminative MCMC, the likelihood is replaced by conditional likelihood, that is,

$$\begin{aligned} & p_d(\theta \mid \mathcal{D}) \\ \propto & \prod_{i=1}^{N} \frac{p(z_c^i \mid \pi_c) \int p(\pi \mid \alpha, z_c^i) \prod_{j=1}^{N_w^i} p(z_{ij} \mid \pi) p(w_{ij} \mid \beta, z_{ij}) d\pi}{\sum_l p(z_i = l \mid \pi_c) \int p(\pi \mid \alpha, z_i = l) \prod_{j=1}^{N_w^i} p(z_{ij} \mid \pi) p(w_{ij} \mid \beta, z_{ij}) d\pi} \; p(\alpha, \beta, \pi_c) \\ = & \prod_{i=1}^{N} \frac{\pi_c(c_i) \int p(\pi \mid \alpha, c_i) \prod_{j=1}^{N_w^i} \sum_{k=1}^{K} \pi(k) \beta(k, j) d\pi}{\sum_l \pi_c(l) \int p(\pi \mid \alpha, l) \prod_{j=1}^{N_w^i} \sum_{k=1}^{K} \pi(k) \beta(k, j) d\pi} \; p(\alpha, \beta, \pi_c) \quad . \end{aligned} \tag{4.41}$$

Estimation of the discriminative posterior is therefore computationally rather intensive, since each term in the conditional likelihood requires an integration over the parameter π.

The computational problem in Equation (4.41) can be reduced to solving expectations of type $\mathrm{E}_{p(\pi \mid \alpha, z_c)}\{\prod_{j=1}^{N_w^i} \pi^T \beta(.,j)\}$. In Publication 7 the integrals were solved with Monte Carlo integration (Carlin and Louis, 2000).

Chapter 5

Probabilistic Modeling of Eye Movements in Information Retrieval Tasks

In addition to contributions to theory and algorithms, the third main contribution of this thesis is to apply probabilistic modeling for predicting relevance from eye movements in information retrieval tasks. Both research fields are well-established, but their combination is new.

Main body of research on eye movements has been published in psychology, where they have been studied as indicators of cognitive processes for more than a century (Rayner, 1998). Information search as such has not been studied, but there exists a considerable amount of research in visual search of images and reading.

Information retrieval (IR; Baeza-Yates and Ribeiro-Neto (1999); Singhal (2001)) is a research field that develops methods for finding relevant information from large (text) collections. The experimental setup of IR considers situations where the user types in an initial search, a *query*. The task of the search engine is then to return documents that best match the query. The field is a very fruitful area for application development, since the performance can be explicitly measured with well-established measures such as precision and recall. Precision is the proportion of retrieved documents that are relevant, and recall is the proportion of relevant documents retrieved by the system. Within information retrieval, the use of implicit feedback information in boosting queries has been studied to some extent (Kelly and Teevan, 2003), but the use of eye movements as the source of relevance information is a new concept.

Publications 1, 2, and 3 carry out a feasibility study on whether eye movements can be used as a source of implicit relevance feedback for predicting relevant items during information retrieval. The conclusion of the study is that relevance can be determined, to some extent. The data from the last stage of feasibility study in Publication 3 was used in a Pascal EU Network of Excellence machine learning challenge ”Inferring Relevance from Eye Movements” (Salojärvi et al., 2005).

After establishing that relevance can be determined from eye movements, the information is exploited in a second study reported in Publication 4, which implements an experiment where the relevance predictions from implicit feedback are combined with another source of available feedback information, collaborative filtering. The combined relevance prediction is better than either of the sources alone. The publication shows that a new type of proactive information retrieval application is feasible, and suggests an architecture for deriving and combining relevance predictions from several available background information sources.

Finally, in Publication 8, a discriminative hidden Markov model is applied in a psychological study where cognitive states of a user are inferred from eye movement data during different kinds of information search tasks.

Before discussing the contributions of the thesis from Section 5.3 onwards, the use of implicit feedback information in information retrieval is motivated in Section 5.1. Then the fundamental reasons why eye movements could be a good source of implicit feedback information are discussed in Section 5.2.

5.1 Explicit and Implicit Feedback in Information Retrieval

It has long been known that the performance of an information retrieval system can be boosted by adding a relevance feedback stage in the process (Rocchio, 1971). This type of feedback is *explicit*: the user interface explicitly asks for feedback from the user. A traditional method of using the feedback is query expansion, where the initial query is complemented with words from relevant documents. The technique has been shown to provide a remarkable improvement to search efficiency, which is why it was recommended to be incorporated into search engines already 18 years ago (Salton and Buckley, 1990). However, the feature is still not available in most of the current implementations. One of the main reasons for this is that giving relevance feedback requires additional effort from the user. Additionally, the simplest method of using explicit feedback is query-specific. Since there is no learning from search history, the feedback loop is needed in every search.

A natural next step is to reduce the workload of the user by determining the relevant documents automatically. There are several different solutions that can be pursued. The first possibility is to assume that the top documents that best match the initial query are automatically assumed to be relevant (Buckley et al., 1995), which can lead to very good or disastrous results, depending on the initial query.

The second option is to complement the query with implicit feedback information. Implicit feedback consists of data measured from the user without interrupting her natural work flow. Examples include measuring the reading time, or monitoring the selection, printing, and saving of documents. The implicit feedback is naturally less accurate than explicit feedback, but since the feedback stream is always available at no extra cost to the user, it is an attractive candidate. However, the improvement of implicit feedback on traditional searches has so far proved to be only marginal (Kelly and Teevan, 2003). For example, in case of reading time, the reason is that the measure is ambiguous; in addition to the possible effect on relevancy, the reading time is reported to be affected by at least task, topic, the expected duration of the task (endurance), the expected frequency of the task (frequency), the estimated progress in completing the task (stage), the estimated time the user is interested in the task (persistence) and knowledge about the topic (familiarity) (Kelly, 2004). Additionally, there is large individual variation, suggesting that an application using implicit feedback information should be user-specific. In summary, the implicit user information measured from traditional user interface is not accurate enough, and should be complemented with other available sources of information.

The third option is to complement the query with explicit relevance information which is already available by using earlier feedback from the user and other similar-minded users. This is the task of *collaborative filtering*, which operates on explicit ratings given by a large set of users. Traditional implementations of collaborative filtering have been memory-based; the ratings of all the users, *user profiles* are stored, and given a new set of ratings, the nearest neighboring user profiles are sought, measured for example in terms of Pearson correlation (Resnick et al., 1994; Shardanand and Maes, 1995).

However, the performance of memory-based collaborative filtering algorithms does not scale well as the number of users and rated documents increases. There are two main reasons: *scalability* – the computational complexity of nearest neighbor -finding algorithms increases linearly with the amount of users and documents, and (ii) *sparsity* – as the collection grows, even the most active users have ratings for only a small proportion of the available documents (Sarwar et al., 2001). One solution to the scalability problem is to restrict the set of candidates by constructing summaries of several similar users by model-based collaborative filtering. The summaries can be association rules (Sarwar et al., 2000), or they can be probabilistic models (Breese et al., 1998); the User Rating Profile model applied in Publication 4 belongs to this category.

Finally, better predictions can be achieved by combining the aforementioned approaches. Collaborative filtering has been combined at least with document contents (Melville et al., 2002), and with the traditional form of implicit feedback (reading time and features from mouse and keyboard usage, Zigoris and Zhang (2006)). At least according to Zigoris and Zhang (2006) the traditional form of implicit feedback has very limited, unstable predictive value by itself and only marginal value when combined with explicit feedback.

The conclusion is that new sources of implicit feedback information should be explored. In this thesis the use of eye movements as a new source is suggested. Before going into the experimental evaluation of whether this is feasible, the physiological and psychological reasons why the eye movements could be a good source of implicit feedback information are discussed.

5.2 Eye Movements as an Indicator of Interest

Most of the research on eye movements has been carried out in the field of psychology. The central assumption in psychological studies is the eye-mind link assumption (Just and Carpenter, 1976); in cognitively intensive tasks the attention lies where the eyes are fixated.

The fundamental reason why the gaze direction contains information on the focus of attention lies in the anatomy of the eye. Accurate viewing is possible only in the central *fovea* area (only 1–2 degrees of visual angle) where the density of photo-receptive cells is highly concentrated, see Figure 5.1. Functionally, this means that with a single glimpse the human vision can obtain information with high resolution only from a very small central area of the visual field. Therefore detailed inspection of a scene is carried out in a sequence of *saccades* (rapid eye movements) and *fixations* (the eye is fairly motionless). The trajectory is often referred to as a *scanpath*. Information on the environment is mostly gathered during fixations, and the duration of a fixation is correlated with the complexity of the object under inspection.

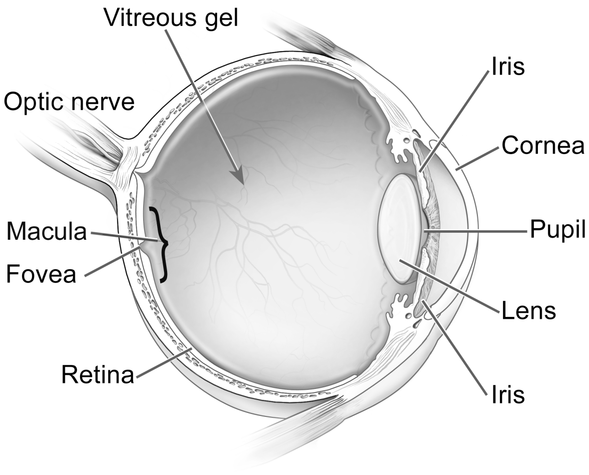

Figure 5.1: Anatomy of the eye. The amount of light entering the eye is controlled by opening or closing of the iris. The light then falls on retina, consisting of photoreceptor cells; there are two types of cells called rods and cones. Rods are sensitive to light in general, making up the vision in dim light conditions. Cones can separate different wavelengths of light, making up the color vision. However, they need more light which is why colors cannot be seen at night. The macula is an oval yellow spot of diameter 1.5 mm near the center of the retina. Within the macula are the fovea and foveola which contain a high density of cones. Near the center is the fovea, a small area that contains the largest concentration of cone cells in the eye and is responsible for accurate vision. The fovea area covers only 1–2 degrees of visual angle, and the cell density decreases rapidly towards the periphery. Image courtesy of National Eye Institute, National Institutes of Health.

Psychology

Inside the brain, the current hypothesis on the processes guiding eye movements is that they are at least partly automatic and triggered by rather low-level processes. Conscious control by top-down processes is possible, since fixations lasting several seconds can be maintained in cognitively intensive tasks. However, conscious control grows laborious over time, possibly because the top-down control requires active suppression of the low-level processes that usually guide eye movements. This characteristic must be taken into account for example in the design of computer interfaces that are (at least partly) controlled by eye movements. The effect is commonly known as the "Midas touch" (Jacob, 1993).

Two large research fields within psychology are related to information retrieval tasks, visual search and reading.

Visual Search. In case of viewing images, the eye movement pattern can be modeled computationally as a bottom-up process that uses a saliency map (Itti and Koch, 2001): fixations fall on the parts of images that have a high degree of saliency, that is, more detail. The basic eye movement pattern is then affected by top-down cognitive processes, for example the task affects the scan path, as reported already by Yarbus (1967). One well-defined cognitive task is *visual search* where an image is scanned for a particular object or feature (the target) among other objects or features (the distractors). There is some computational evidence that the visual search could be a purely top-down process (Zelinsky et al., 2006), suggesting that the visual system could possibly be primed in advance to find certain desired features.

Models of Reading. Research on eye movements during reading is a well-established field, see Rayner (1998) for an overview. Reading is highly controlled by top-down processes. For example, the durations of fixations have been noticed to depend on the frequency of occurrence of the words in general, and with how predictable the word is based on its context (Rayner, 1998). The interpretation of the result is that a lexical search from memory is performed – the more obscure the word, the longer the search takes.

Computational models on eye movement control during reading are aimed at explaining how different perceptual, cognitive or motor processes determine saccade initiation during reading. Current controversy is whether attention in reading is allocated serially to one word at a time, as assumed by the E-Z Reader model (Reichle et al., 2006; Pollatsek et al., 2006), or whether attention is spatially distributed so that several words are processed at the same time. This parallel hypothesis is implemented for example in the SWIFT (Richter et al., 2006), the Glenmore (Reilly and Radach, 2006) and the Competition/Interaction (Yang, 2006) models.

Usability Research

Outside of psychological research, eye movement studies have been widely used in usability studies (Jacob and Karn, 2003; Goldberg et al., 2002), where it is common to compute summary measures of eye movements on areas of interest, such as images or captions of text (see also poynter). The eye movements have also been used to give feedback on the subjective image quality (Vuori et al., 2004).

5.2.1 Applications Using Eye Movements

In addition to research, eye movements have been studied more recently in various engineering applications. During the last twenty years, eye movement measuring devices have become relatively accurate and progressively cheaper, allowing also free head movement of the user (see Morimoto and Mimica (2005) for an overview on eye tracking techniques). One of the most prominent application areas has been eye-typing (Majaranta and Räihä, 2002, 2007).

As discussed above in the introduction, user interfaces which are controlled with eye movements alone are laborious to use. The current research on the field has therefore concentrated more on using a combination of traditional input methods, complemented with information extracted from eye movements (Vertegaal, 2002; Hyrskykari et al., 2005).

Another line of application develoment is to enhance the traditional user interface by extracting implicit feedback information from eye movements. This information can be used *proactively* in the background in order to improve the performance of the interface, for example by trying to infer what the user would like to do next. The idea of using implicit feedback for enhancing traditional user interfaces, a *proactive user interface*, was originally suggested by Tennenhouse (2000). Historically, the first application where crude level user interest was inferred from eye movements was by Starker and Bolt (1990), who developed an interactive story teller, where the story told by the application concentrated more on items that the user was gazing at on a display. Rudimentary relevance determination is needed also in Hyrskykari et al. (2003), where a proactive translator is activated if the reader encounters a word which she has difficulties (these are inferred from eye movements) in understanding. The behavior of both of these applications is adjusted simply by measuring fixation durations which are longer than a predefined threshold. Another interesting proactive application is to exploit eye movements as one feedback channel in order to identify critical driving events in intelligent driver assistance systems (Miller et al., 2003; Torkkola et al., 2003).

Suitor (Simple User Interest Tracker) is a prototype attentive agent application by Maglio et al. (2000); Maglio and Campbell (2003). The application monitors eye movements during browsing of web pages in order to determine whether the user is reading or just browsing. If reading is detected, the document is defined relevant, and more information on the topic is sought and displayed. From a scientific perspective, the developed methods rely on heuristics, and no evaluation of the performance of the system is available. Publication 4 presents a controlled experimental setup to test the feasibility of a proactive information retrieval application.

5.3 Implicit Relevance Feedback from Eye Movements

Publications 1, 2, and 3 carry out the first study of eye movements in an information retrieval setup. The study follows the general procedures of data analysis, discussed in Sections 2.1 and 4.1.1. The initial step of data exploration is carried out in Publication 1, the next step, model construction, is then carried out in Publication 2. After verifying the experimental setup and choosing the appropriate models in initial experiments, a larger data set is gathered and the constructed models are applied to test whether the eye movements contain relevance information in Publication 3.

5.3.1 A Pilot Study

A pilot study is carried out in Publication 1. The publication presents the first controlled experimental setup for measuring eye movements in an information retrieval situation. In the experiment, the subject was shown a question and a list of twelve sentences, one of which contained the correct answer (C). Eight of the sentences were known to be irrelevant (I), and three relevant to the question (R). The experiment was carried out with three subjects, each completing a set of 20 question-answer tasks. The eye movements were measured with an iView-system manufactured by SensoMotoric Instruments GmbH, installed on a bicycle helmet.

The subject's head was stabilized with a chin rest to improve the accuracy of calibration. Even with the head rest, maintaining an accurate calibration throughout the experiment proved to be difficult, and it needed to be manually corrected before data analysis. Several eye movement features found in psychological literature were then computed from the raw fixation-saccade sequences. See Publication 1 for a full description of the preprocessing steps. The full description of computed features is available in Salojärvi et al. (2005).

The aim of the initial study was data exploration, discussed in Section 4.1.1. For this means, the data was visualized with principal component analysis, self-organizing maps, linear discriminant analysis, and self-organizing maps in learning metrics (Kaski et al., 2001). In summary, some of the features turn out to be redundant, showing strong correlations. Only a few of the measures were useful in predicting relevance, as can be seen from the eigenvectors of linear discriminant analysis presented in Figure 3 of Publication 1. The same set of features were obtained in a later analysis using a Bayesian multilayer perceptron with an automatic relevance determination prior (Kudjoi, 2004).

In addition to data exploration, relevance prediction from eye movement data was tested with linear discriminant analysis. Even with difficulties in calibration, a simple LDA was able to predict the class of the sentences at a level which was statistically significantly better than random guessing. This verifies that eye movements contain relevance information, and that the devised experimental setup is appropriate.

5.3.2 Model Building

The second stage after data exploration is model building, considered in Publication 2. The primary aim of modeling was to take the time series nature of the data into account. For this means, a time series representation of the data was constructed. In order to restrict the possible number of parameters in the model, the set of features was restricted to those that best predicted relevance in Publication 1. A time series representation of the features was then constructed.

Since it is a model for sequential data, the hidden Markov model (see Section 2.4) is a natural candidate for modeling the time series. Historically, perhaps the first publication to consider eye movements and HMMs is by Rimey and Brown (1991), who apply HMMs to generate fixation-saccade -type scanpaths with a camera. The first application of HMMs to real eye movements was by Salvucci and Anderson (2001), who modeled cognitive processes using eye movement data. The models however were not learned from data - manual tuning of the parameters was applied.

Publication 2 considers three alternative HMM types: (i) a separate two-state HMM for each of the three sentence types, (ii) a six-state discriminative HMM learned with a Viterbi algorithm, and (iii) a heuristically trained "global" HMM. All of the models performed significantly better than random classification. The separate HMMs (i) and discriminative HMM (ii) were chosen to be used for further analysis with a larger data set. The "global" HMM was left out mainly because of the heuristic training method.

5.3.3 Large-Scale Experiment

After initial confirmation that relevance can be predicted, and developing proper models for the task, a larger-scale experiment to validate the result was designed. The experimental setup was slightly revised and better measuring equipment was used for gathering the data.

Experimental Setup. The experimental setup used in Publication 3, and the Pascal challenge (Salojärvi et al., 2005) was a slight revision from the one used in the pilot study. The subject was first shown a question and then a list of ten sentences. One of them is the correct answer. Five of the sentences are known to be irrelevant, and four relevant to the question. The task of the subject was to identify the correct answer (while eye movements were measured), press ‘enter’, thus ending the measurement. Afterwards the associated number was typed in the following display. Each of the eleven test subjects carried out 50 assignments.

Equipment. The eye movements were measured with a Tobii 1750 eye tracker with a screen resolution of 1280x1024 pixels and a sampling rate of 50 Hz. The equipment allowed free head movement of the subjects within a 20x20x20cm area. The accuracy of the measurements was sufficient ($0.5°$) for mapping fixations to individual words. The calibration was accurate throughout the experiments.

Feature Extraction. The raw eye movement data, sequence of fixations (and saccades), was first mapped to the word which was closest to the fixation. Several features (21 in total) described in psychological research on reading were computed (Rayner, 1998; Calvo and Meseguer, 2002). Full description of the features is included in Salojärvi et al. (2005).

A time series representation of the set of features resulting in best classification accuracy was then constructed, see Publications 2 and 3. The features were computed for each word along the eye movement trajectory. The whole trajectory was segmented to sequences occurring on the same title, and a label was assigned to each sequence according to the class of the title.

Results. The hidden Markov model applied in Publication 3 consisted of two hidden states per class $c \in \{C, I, R\}$, thus having six hidden states in total. Both joint density and a discriminative HMMs were trained. As a result, the HMM shows significantly better performance than a simple linear discriminant analysis using averaged features and, furthermore, a discriminative HMM is significantly better than joint density HMM. The conclusions are that

1. Relevance can be predicted from eye movements
2. The time series contains information on relevance
3. Discriminative training helps.

5.3.4 Related Research

Since the initial data analysis presented in Publication 1, research on applying eye movements as a source of implicit feedback in information retrieval has attracted more attention. For example, Granka et al. (2004); Joachims et al. (2005, 2007) approach the same problem from a different direction by investigating with quantitative measures how users behave in a real, less-controlled information retrieval task.

In a similar vein, Moe et al. (2007) carries out a visual qualitative analysis of eye movement trajectories in information retrieval tasks to test three possible measures that could indicate relevance: thorough reading, total viewing time, and number of regressions. The measure best corresponding with relevance was *thorough reading*, defined as consisting of mainly horizontal eye movements with many fixations per line and at least half a line read.

A similar result has been reported in an on-going work by Buscher et al. (2008), who develop a new measure of relevancy of a document, computed as the ratio of read text to the text that was either read or skimmed. According to Buscher et al. (2008), there seems to be a fair dependency between the developed measure and the (subjective) relevancy evaluations of the subjects. The applied reading detection algorithm is further development of the heuristic method by Campbell and Maglio (2001). The measure agrees nicely with earlier results and intuition – relevant documents attract more attention.

5.4 Proactive Information Retrieval

Publication 4 presents a new task, *proactive information retrieval.* In the task the user interest is inferred by observing user behavior and merged with other information available on the topic. The closest earlier work is by (Melville et al., 2002), discussed in Section 5.1, who combine collaborative filtering with document contents in order to boost the performance of a search engine. The work is similar in the sense that collaborative filtering is combined with other relevant information, the document contents. However, user behavior was not modelled.

Prototype Application. In a prototype application presented in Publication 4, implicit relevance feedback from eye movements is complemented by bringing in information from a collaborative filtering model. Relevance is predicted from the two different sources using probabilistic modeling. Implicit feedback is inferred from eye movements with discriminative hidden Markov models as described above in Section 5.3 and Publications 2 and 3. Relevance from collaborative filtering is modelled with the User Rating Profile model (Marlin, 2004), a probabilistic model which is a variant of latent Dirichlet allocation model discussed in Section 4.3, computed using Markov Chain Monte Carlo techniques.

Experiment. In the experimental setup the subjects are seeking interesting scientific articles by browsing their titles. The subjects were shown a set of six titles of scientific articles at a time, and they were instructed to choose two of the most interesting ones. Each subject performed 80 evaluation tasks. The background data for collaborative filtering was gathered from 22 subjects who performed the experiment via a web form. Three subjects performed the same task while their eye movements were measured, see Publication 4 for full details. The experimental setup differs from the earlier ones presented in Section 5.3 in two respects; the task language was English, whereas the subjects' native language was Finnish. The second difference was that instead of defining the correct answer by task instruction, the experiment required subjective evaluation of interest. Both of these could possibly have effects that affect the accuracy of relevance prediction from eye movements. However, a discriminative HMM was still able to predict relevance significantly better than chance.

Combining Relevance Predictions. A modular approach for combining the relevance predictions was assumed, because the architecture can thus be easily extended to several sources of information. A Dirichlet mixture model was designed for combining the predictions. The model was optimized in a discriminative manner with the conditional maximum likelihood criterion. For new document titles the prediction accuracy with the combined relevance predictions from eye movements and collaborative filtering was significantly better than by chance. The best prediction accuracy still leaves room for improvement but shows that proactive information retrieval and combination of many sources of relevance feedback is feasible.

Further Related Work. In a further work on the subject, Hardoon et al. (2007) study a case where the query given to the search engine is constructed from eye movements alone. The result is that the query can be learned from a small set of read documents, such that relevance predictions for a large set of unseen documents are ranked significantly better than by random guessing.

Additionally it is also studied whether the query can be further improved by including also explicit feedback information. The result depends on the topic, but the overall precision is improved.

5.5 Inferring Cognitive States

Traditional psychological research on eye movements concentrates on observing how the eyes behave when the subject is in a certain cognitive state. The state is induced on the subject by task instruction or experiment design. As a result, there exists a considerable amount of eye movement features that are observed given a certain cognitive state. For example, the fixation duration is longer when a subject is reading words that she has problems in understanding (Rayner, 1998) (compared to normal reading).

The problem can be approached also in a reverse manner – to infer from eye movements what cognitive processes are possibly active in the mind. In Publication 8, a reverse inference approach was assumed, with the aim of making hypotheses on hidden cognitive states in an experiment where the tasks resemble everyday information search tasks.

Experiment. The experimental setup was an extension of the one in Section 5.3; the presentation of the tasks was similar, but in addition to finding the correct answer, the experiment included two more conditions, a key word search and true interest. The idea was that if the eye movements contain information of different cognitive states, the tree different conditions can be predicted from eye movements.

Modeling. Discriminative hidden Markov models were applied in modeling. Instead of word-level features used earlier, the HMM was estimated from low-level eye movement features, the fixation durations, saccade lengths and directions (see

Publication 8 for a detailed description). The reason is that the aim of the study was to associate the distributions of these low-level features to what is known from earlier research where the cognitive state is known.

Results. The data analysis carried out with the model shows evidence that the subjects shift their eye movement behavior while they proceeded in tasks. The assignment is typically began from a set of HMM states reflecting a *scanning* type of behavior; long saccades with no preference on direction, accompanied with rather short fixations. The task completion then moves to a second state, *reading*, characterized by frequent forward saccades of the length of an average word. Also the mean fixation durations and the amount of regressions agree with previous research results on reading (Rayner, 1998). The third state suggests a more careful analysis of sentences, possibly of *deciding* whether the sentence is the correct answer to a given task. The saccades landed almost always on the previously seen lines and were directed either forward or backward. The distance covered by saccades was about the length of an average word. The participants ended the assignments while they were in this state.

Discussion. The general setup of the work in Publication 8 bears resemblance to Liechty et al. (2003), who inferred cognitive states of visual attention in an advertisement viewing task with a Bayesian HMM using Markov chain Monte Carlo sampling. Also Liechty et al. (2003) interpreted the hidden states to correspond to a cognitive process, termed local and global attention. However, there are two important differences: first of all, the experimental setup is different, and second, the work of Liechty et al. (2003) considers only HMMs with two hidden states. The main difference between the two experiments is that in Liechty et al. (2003) the subjects carry out browsing of advertisements at their own pace without any task instruction, whereas in the experimental setup of Publication 8 a task instruction is given.

The additional decision of relevance in the experimental setup of Publication 8 could possibly cause additional cognitive top-down processes which are visible from the eye movement patterns. The situation could be seen as analogous to visual search, discussed in Section 5.2. Initially, when there is no information on the location of the correct answer, the subject engages in browsing, corresponding to altering between the states of global (scanning) and local (reading) attention as in Liechty et al. (2003). At some point, when the correct answer is thought to be found, a top-down process similar to visual search is triggered, and the eyes key in on the relevant information to check the correctness. In order to prove that the brain could operate in this manner, it is important to infer the proper amount of hidden states as objectively as possible. Publication 8 used the criteria of Robertson et al. (2004); Miloslavsky and van der Laan (2002) to decide the number of hidden states from data. The fact that three hidden states for each task type is the proper amount suggested by the data does support this hypothesis. However, the data cannot give any evidence of the actual origins of the third hidden state.

Chapter 6

Conclusions

When solving a given problem one should avoid solving a more general problem as an intermediate step.

(Vapnik, 1995)

This thesis presents a framework for making conditional statistical inference with models that are known to be wrong. An intuitive way of acting in this case is that if our model is known to be wrong, its performance must be optimized such that the prediction error in the variable of interest is minimized, with the possible cost of modeling the distributions of other variables. The intuition leads to the key result of the thesis; in case of an incorrect model, there are different "posteriors" – distributions in the parameter space of the model – that best agree with each inference problem, that is, utility function. Publication 7 studies the special case of classification and shows that discriminative posterior is the optimal posterior distribution for this case. In this thesis we show that the same distribution can be derived from decision theory as in Bernardo and Smith (2000); Publication 7 derives the distribution from axioms that bear resemblance to the famous Cox axioms (Cox, 1946). The work gives a justification to Bayesian regression and discriminative modeling in general. The applicability of the result is wide-ranging, since our setup considers decision-making in real life; we do not have a true/full model of the world but we nevertheless desire to be rational. It can be seen as another application of the no-free-lunch theorem (Wolpert and Macready, 1997); the sampling of conventional posterior cannot give a solution which is optimal for all inference. That is, problem-specific information must be brought in for obtaining best inference. From another point of view, the result that the utility of the decision maker must be taken into account when optimizing the models – also when sampling from the posterior – is supported also by Kleinberg et al. (1998) in a data mining setup.

Our result goes against a view to conventional Bayesian inference where the posterior from a joint density model is computed, and then all inference is carried out with it. The reason why this approach has been adopted is that the theory of Bayesian inference has earlier been derived under the assumption that the "true" model is in the model family, which is the only case where the standard posterior is optimal for all inference. The results of this thesis do not make standard Bayesian modeling invalid; it can still be applied in cases where the model family is general enough. For example, in Gaussian process modeling the model family is infinite, and can thus be thought of as containing also the “true” model.

In addition to the theoretical contribution, we also present methods for computing the discriminative posterior. We show that standard Markov chain Monte Carlo sampling is applicable – the only thing that needs to be done is to replace the likelihood function by conditional likelihood. The MCMC methods presented in this thesis are first implementations of the sampling; a lot of further work is required in order to find the most efficient way of sampling. In practical applications, point estimates are often computed for computational reasons in case where the model is so complex that sampling from the posterior would be inefficient. For this means, we present an EM algorithm for computing conditional maximum likelihood estimates for discriminative joint density models. The computational complexity of the algorithm presented in Publication 5 drops considerably from the earlier proposed algorithm (Jebara, 2001), thus making the algorithm feasible, but is still higher than comparable methods for maximum likelihood estimation. However, it is possible to make further approximations. This is left for further work, since a good trade-off is required; if the approximations are too loose, they result in EM steps where the value of likelihood can drop. On the other hand, if more strict approximations based on global upper bounds are made, they will result in EM steps which are shorter, and thus increase the amount of iterations required for convergence.

The methods developed in this thesis are applied to modeling of eye movements in information retrieval tasks. The application area is novel; Publication 1 was the first publication to consider this topic. Modeling of eye movements requires probabilistic modeling, since they contain a considerable amount of noise. As discussed in Section 5.2, the current hypothesis is that eye movements are partially guided by low-level processes which can be regarded as automatic. An example of this is that when viewing images, gaze concentrates more on areas with high degree of saliency. This kind of behavior thus does not tell us anything of the ongoing high-level cognitive process, for example the task given to the subject. Probabilistic modeling allows us to develop and test hypotheses in spite of these noise processes, by either taking the processes into account in modeling, or by averaging the noise out by handling several repeats of an experiment. In the former case, enough of the process must be known in order to be able to construct a model for the noise process (such as a saliency map in the example of viewing images). In the latter case the experimental setup must be designed such that systematic effects are canceled out by randomized experimental design.

Since the application area is new, the models applied to analyzing eye movements cannot be assumed to be close to “truth” – the case where discriminative posterior will perform better than joint density modeling. In this thesis, most of the modeling was carried out with conditional maximum likelihood estimates, since the development of the discriminative posterior was the final result of the thesis. One line of further research could thus be to continue the work by applying the sampling-based approach to eye movement data. Nevertheless, also the point estimates show the main result; relevance can be inferred from eye movements, as shown in Publication 3. Improving the models will only make the predictions better, but also with our initial methods this makes it possible to design new engineering applications, such as a proactive information retrieval system studied in Publication 4.

Besides engineering applications, the applied models can be also used for making hypotheses for psychological research. For example, the evidence presented in Publication 8 suggests that a new cognitive state, coined decision making, could be predicted from eye movements in information retrieval tasks. One interesting line of research would be to combine the eye movement experiment with brain imaging to determine the areas of the brain that are active during the new state. The predicted duration of the decision making state is one second, which should be long enough for measuring the brain activity with functional magnetic resonance imaging (fMRI).

Bibliography

M. Afify. Extended Baum-Welch reestimation of Gaussian mixture models based on reverse Jensen inequality. In *INTERSPEECH-2005*, pages 1113–1116, 2005.

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

Y. Altun, A. J. Smola, and T. Hofmann. Exponential families for conditional random fields. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence (UAI-2004)*, pages 2–9, Arlington, Virginia, 2004. AUAI Press.

R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval.* Addison-Wesley, Wokingham, UK, 1999.

A. Banerjee. An analysis of logistic models: Exponential family connections and online performance. In *Proceedings of the 2007 SIAM International Conference on Data Mining*, pages 204–215, 2007.

A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.

O. Barndorff-Nielsen. *Information and exponential families in statistical theory.* John Wiley & Sons, 1978.

M. J. Beal. *Variational Algorithms for Approximate Bayesian Inference.* PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

J. Berger. The case for objective Bayesian analysis. *Bayesian Analysis*, 1(3):385–402, 2006.

J. M. Bernardo and A. F. M. Smith. *Bayesian Theory.* John Wiley & Sons Ltd, Chichester, England, 2000.

A. Birnbaum. On the foundations of statistical inference. *Journal of the American Statistical Association*, 57(298):269–306, 1962.

C. M. Bishop. *Pattern recognition and machine learning.* Springer, 2006.

C. M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, Oxford, 1995.

D. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

G. Bouchard. Bias-variance tradeoff in hybrid generative-discriminative models. In *ICMLA '07: Proceedings of the Sixth International Conference on Machine Learning and Applications*, pages 124–129, Washington, DC, USA, 2007. IEEE Computer Society.

G. E. P. Box and N. R. Draper. *Empirical Model-Building and Response Surfaces.* John Wiley & Sons, 1987. p. 424.

S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, Cambridge, UK, 2004.

J. S. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI-98)*, pages 43–52, San Francisco, 1998. Morgan Kaufmann.

L. M. Bregman. The relaxation method for finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7:191–204, 1967.

C. G. Broyden. The convergence of a class of double-rank minimization algorithms, II: The new algorithm. *IMA Journal of Applied Mathematics*, 6(3):222–231, 1970.

C. Buckley, J. Allan, G. Salton, and A. Singhal. Automatic query expansion using SMART: TREC 3. In *Proceedings of the Third Text Retrieval Conference (TREC-3)*, pages 69–80. NIST Special Publication 500-225, 1995.

W. Buntine. Variational extensions to EM and multinomial PCA. In T. Elomaa, H. Mannila, and H. Toivonen, editors, *Proceedings of the 13th European Conference on Machine Learning*, volume 2430 of *Lecture Notes in Artificial Intelligence*, pages 23–34. Springer-Verlag, 2002.

G. Buscher, A. Dengel, and L. van Elst. Eye movements as implicit relevance feedback. In *CHI '08: CHI '08 extended abstracts on Human factors in computing systems*, pages 2991–2996, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-012-X. doi: http://doi.acm.org/10.1145/1358628.1358796.

M. G. Calvo and E. Meseguer. Eye movements and processing stages in reading: Relative contribution of visual, lexical and contextual factors. *The Spanish Journal of Psychology*, 5(1):66–77, 2002.

C. Campbell and P. Maglio. A robust algorithm for reading detection. In *Workshop on Perceptive User Interfaces (PUI '01).* ACM Digital Library, November 2001. ISBN 1-58113-448-7.

B. P. Carlin and T. A. Louis. *Bayes and empirical Bayes methods for data analysis (2nd edition).* Chapman & Hall / CRC, Boca Raton, FL, 2000.

J. Cerquides and R. L. Mántaras. Robust Bayesian linear classifier ensembles. In J. Gama, R. Camacho, P. Brazdil, A. Jorge, and L. Torgo, editors, *Machine Learning: ECML 2005*, pages 72–83, Berlin, Germany, 2005. Springer-Verlag.

R. T. Cox. Probability, frequency and reasonable expectation. *American Journal of Physics*, 17:1–13, 1946.

B. de Finetti. Funzione caratteristica di un fenomeno aleatorio. *Atti della R. Academia Nazionale dei Lincei, Ser. 6*, 4:251–299, 1931. Memorie, Classe di Scienze Fisiche, Mathematiche e Naturali.

D. Edwards and S. L. Lauritzen. The TM algorithm for maximising a conditional likelihood function. *Biometrika*, 88(4):961–972, 2001.

B. Efron. The geometry of exponential families. *The Annals of Statistics*, 6(2):362–376, March 1978.

L. Fei-Fei and P. Perona. A Bayesian hierarchical model for learning natural scene categories. In *Proceedings 2005 IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 524–531, Los Alamitos, CA, USA, 2005. IEEE Computer Society. doi: http://dx.doi.org/10.1109/CVPR.2005.16.

T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1(2):209–230, 1973.

R. Fletcher. A new approach to variable metric algorithms. *The Computer Journal*, 13 (3):317–322, 1970.

J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *The Annals of Statistics*, 38(2):337–374, 2000.

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

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis (2nd edition)*. Chapman & Hall/CRC, Boca Raton, FL, 2003.

S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intell.*, 6(6):721–741, 1984.

J. H. Goldberg, M. J. Stimson, M. Lewenstein, N. Scott, and A. M. Wichansky. Eye tracking in web search tasks: design implications. In *ETRA '02: Proceedings of the symposium on Eye tracking research & applications*, pages 51–58. ACM Press, 2002.

D. Goldfarb. A family of variable-metric methods derived by variational means. *Mathematics of Computation*, 24(109):23–26, January 1970.

P. Gopalakrishnan, D. Kanevsky, A. Nádas, and D. Nahamoo. An inequality for rational functions with applications to some statistical estimation problems. *IEEE Transactions on Information Theory*, 37(1):107–113, 1991.

L. A. Granka, T. Joachims, and G. Gay. Eye-tracking analysis of user behavior in WWW search. In *Proceedings of SIGIR'04*, pages 478–479. ACM Press, 2004.

T. Griffiths and Z. Ghahramani. Infinite latent feature models and the Indian buffet process. In *Advances in Neural Information Processing Systems 18*, pages 475–482, Cambridge, MA, 2006. MIT Press.

P. Grünwald, P. Kontkanen, P. Myllymäki, T. Roos, H. Tirri, and H. Wettig. Supervised posterior distributions. Presentation at the Seventh Valencia International Meeting on Bayesian Statistics, Tenerife, Spain, 2002. `http://homepages.cwi.nl/ pdg/presentationpage.html`.

J. Halpern. A counter example to theorems of Cox and Fine. *Journal of Artificial Intelligence Research*, 10:67–85, 1999.

D. R. Hardoon, J. Shawe-Taylor, A. Ajanki, K. Puolamäki, and S. Kaski. Information retrieval by inferring implicit queries from eye movements. In *Eleventh International Conference on Artificial Intelligence and Statistics*, 2007.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, 2001.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

R. Herbrich. *Learning Kernel Classifiers*. MIT Press, Cambridge, Massachusetts, 2002.

A. Honkela, M. Tornio, T. Raiko, and J. Karhunen. Natural conjugate gradient in variational inference. In *Proceedings of the 14th International Conference on Neural Information Processing (ICONIP 2007)*, Kitakyushu, Japan, November 13-16 2007. to appear.

H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417–41,498–520, 1933.

A. Hyrskykari, P. Majaranta, and K.-J. Räihä. Proactive response to eye movements. In G. W. M. Rauterberg, M. Menozzi, and J. Wesson, editors, *INTERACT'03*. IOS press, 2003.

A. Hyrskykari, P. Majaranta, and K.-J. Räihä. From gaze control to attentive interfaces. In *Proceedings of HCII 2005*, Las Vegas, NV, July 2005. Lawrence Erlbaum Associates, Inc.

L. Itti and C. Koch. Computational modeling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203, 2001.

R. Jacob and K. Karn. Eye tracking in human-computer interaction and usability research: Ready to deliver the promises (section commentary). In J. Hyona, R. Radach, and H. Deubel, editors, *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research*, pages 573–605. Elsevier Science, Amsterdam, 2003.

R. J. K. Jacob. Eye movement-based human-computer interaction techniques: Toward non-command interfaces. In H. R. Hartson and D. Hix, editors, *Advances in Human-Computer Interaction*, volume 4, pages 151–190. Ablex Publishing Co., Norwood, NJ, 1993.

T. Jebara. *Discriminative, Generative and Imitative Learning*. PhD thesis, Media laboratory, MIT, 2001.

T. Jebara and A. Pentland. On reversing Jensen's inequality. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 231–237, Cambridge, MA, April 2001. MIT Press.

T. Jebara and A. Pentland. Maximum conditional likelihood via bound maximization and the CEM algorithm. In D. A. C. Michael J. Kearns, Sara A. Solla, editor, *Advances in Neural Information Processing Systems 11*, pages 494–500, Cambridge,MA, 1999. The MIT Press.

T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 154–161, New York, NY, USA, 2005. ACM Press. ISBN 1-59593-034-5. doi: http://doi.acm.org/10.1145/1076034.1076063.

T. Joachims, L. Granka, B. Pan, H. Hembrooke, F. Radlinski, and G. Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Trans. Inf. Syst.*, 25(2):7, 2007. ISSN 1046-8188. doi: http://doi.acm.org/10.1145/1229179.1229181.

M. I. Jordan. Graphical models. *Statistical Science: Special Issue on Bayesian Statistics*, 19:140–155, 2004.

M. A. Just and P. A. Carpenter. Eye fixations and cognitive processes. *Cognitive Psychology*, 8:441–480, 1976.

S. Kaski, J. Sinkkonen, and J. Peltonen. Bankruptcy analysis with self-organizing maps in learning metrics. *IEEE Transactions on Neural Networks*, 12:936–947, 2001.

R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90:773–795, 1995.

D. Kelly. *Understanding implicit feedback and document preference: A naturalistic user study.* PhD thesis, Rutgers University, 2004.

D. Kelly and J. Teevan. Implicit feedback for inferring user preference: a bibliography. *SIGIR Forum*, 37(2):18–28, 2003. ISSN 0163-5840. doi: http://doi.acm.org/10.1145/959258.959260.

B. M. Kelm, C. Pal, and A. McCallum. Combining generative and discriminative methods for pixel classification with multi-conditional learning. In *Proceedings of the 18th International Conference on Pattern Recognition (ICPR 2006)*, volume 2, pages 828–832. IEEE, 2006.

M. Kim and V. Pavlovic. A recursive method for discriminative mixture learning. In *Proceedings of the 24th International Conference on Machine Learning (ICML-2007)*, Corvallis, OR, 2007. To appear.

A. Klautau. Discriminative Gaussian mixture models: A comparison with kernel classifiers. In T. Fawcett and N. Mishra, editors, *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)*, pages 353–360. AAAI Press, 2003.

J. Kleinberg, C. Papadimitriou, and P. Raghavan. A micro-economic view of data mining. *Data Mining and Knowledge Discovery, 2(4), 1998*, 2(4):311–324, 1998.

T. Kohonen. *Self-Organizing Maps.* Springer, Berlin, 3rd edition, 2001.

M. Koivisto. *Sum-Product Algorithms for the Analysis of Genetic Risks.* PhD thesis, Department of Computer Science, University of Helsinki, 2004. Report A 2004-1.

I. Kudjoi. Feature set selection for inferring relevance from eye movements. Special assignment, information technology, Helsinki University of Technology, July 2004.

J. A. Lasserre, C. M. Bishop, and T. P. Minka. Principled hybrids of generative and discriminative models. In *Proceedings 2006 IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 87–94, Los Alamitos, CA, USA, 2006. IEEE Computer Society. doi: http://dx.doi.org/10.1109/CVPR.2006.227.

J. Liechty, R. Pieters, and M. Wedel. Global and local covert visual attention: Evidence from a Bayesian hidden Markov model. *Psychometrika*, 68:519–541, 2003.

P. P. Maglio and C. S. Campbell. Attentive agents. *Commun. ACM*, 46(3):47–51, 2003. ISSN 0001-0782. doi: http://doi.acm.org/10.1145/636772.636797.

P. P. Maglio, R. Barrett, C. S. Campbell, and T. Selker. Suitor: an attentive information system. In *Proceedings of the 5th international conference on Intelligent user interfaces*, pages 169–176. ACM Press, 2000. ISBN 1-58113-134-8. doi: http://doi.acm.org/10.1145/325737.325821.

P. Majaranta and K.-J. Räihä. Twenty years of eye typing: Systems and design issues. In *Proceedings of ETRA 2002, Eye Tracking Research and Applications Symposium*, pages 15–22, New Orleans, LA, USA, 2002. ACM Press.

P. Majaranta and K.-J. Räihä. Chapter 9: Text entry by gaze: Utilizing eye-tracking. In I. MacKenzie and K. Tanaka-Ishii, editors, *Text entry systems: Mobility, accessibility, universality.* Morgan Kaufmann, San Francisco, 2007.

B. Marlin. Modeling user rating profiles for collaborative filtering. In *Advances in Neural Information Processing Systems 16*, pages 627–634, Cambridge, MA, 2004. MIT Press.

P. McCullagh and J. A. Nelder. *Generalized Linear Models.* CRC Press, 2nd edition, 1990.

G. J. McLachlan and D. Peel. *Finite Mixture Models.* John Wiley & Sons, New York, NY, 2000.

P. Melville, R. J. Mooney, and R. Nagarajan. Content-boosted collaborative filtering for improved recommendations. In *Proceedings of the 18th National Conference on Artificial Intelligence (AAAI-2002)*, pages 187–192. AAAI Press, 2002.

N. Metropolis and S. Ulam. The Monte Carlo Method. *Journal of the American Statistical Association*, 44(247):335–341, 1949.

B. Miller, C. H. Hwang, K. Torkkola, and N. Masseya. An architecture for an intelligent driver assistance system. In *Proceedings of IEEE Intelligent Vehicles Symposium*, pages 639–644, June 2003.

M. Miloslavsky and M. J. van der Laan. Fitting of mixtures with unspecified number of components using cross validation distance estimate. *Computational Statistics and Data analysis*, 41:413–428, 2002.

T. Minka. Expectation Propagation for approximative Bayesian inference. In J. S. Breese and D. Koller, editors, *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, pages 362–369, 2001.

K. K. Moe, J. M. Jensen, and B. Larsen. A qualitative look at eye-tracking for implicit relevance feedback. In B.-L. Doan, J. Jose, and M. Melucci, editors, *Proceedings of the Workshop on Context-Based Information Retrieval (CIR'07)*, volume 326, Roskilde, Denmark, August 2007. CEUR-WS.org.

C. H. Morimoto and M. R. M. Mimica. Eye gaze tracking techniques for interactive applications. *Comput. Vis. Image Underst.*, 98(1):4–24, 2005. ISSN 1077-3142. doi: http://dx.doi.org/10.1016/j.cviu.2004.07.010.

A. Nádas. A decision theoretic formulation of a training problem in speech recognition and a comparison of training by unconditional versus conditional maximum likelihood. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 31(4):814–817, 1983.

A. Nádas, D. Nahamoo, and M. A. Picheny. On a model-robust training method for speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 39(9):1432–1436, 1988.

R. Neal and G. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In M. I. Jordan, editor, *Learning in Graphical Models*, pages 355–368. MIT Press, Cambridge, MA, 1999.

A. Y. Ng and M. I. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 841–848. MIT Press, Cambridge, MA, 2002.

Y. Normandin. *Hidden Markov models, maximum mutual information estimation and the speech recognition problem.* PhD thesis, McGill University, 1991.

A. Pollatsek, E. D. Reichle, and K. Rayner. Tests of the E-Z reader model: Exploring the interface between cognition and eye-movement control. *Cognitive Psychology*, 52: 1–56, 2006.

D. Povey, P. Woodland, and M. Gales. Discriminative MAP for acoustic model adaptation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP'03)*, volume 1, pages 312–315, 2003.

poynter. URL `http://www.poynter.org/eyetrack2000/`.

K. Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372–422, 1998.

E. D. Reichle, A. Pollatsek, and K. Rayner. E-Z reader: A cognitive-control, serial-attention model of eye-movement behavior during reading. *Cognitive Systems Research*, 7:4–22, 2006.

R. G. Reilly and R. Radach. Some empirical tests of an interactive activation model of eye movement control in reading. *Cognitive Systems Research*, 7:34–55, 2006.

P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. Grouplens: an open architecture for collaborative filtering of netnews. In *CSCW '94: Proceedings of the 1994 ACM conference on Computer supported cooperative work*, pages 175–186, New York, NY, USA, 1994. ACM. ISBN 0-89791-689-1. doi: http://doi.acm.org/10.1145/192844.192905.

E. Richter, R. Engbert, and R. Kliegl. Current advances in swift. *Cognitive Systems Research*, 7:23–33, 2006.

R. S. Rimey and C. M. Brown. Controlling eye movements with hidden Markov models. *International Journal of Computer Vision*, 7:47–65, 1991.

A. W. Robertson, S. Kirshner, and P. Smyth. Downscaling of daily rainfall occurrence over northeast Brazil using a hidden Markov model. *Journal of Climate*, 17(22):4407–4424, 2004.

J. J. Rocchio. Relevance feedback in information retrieval. In G. Salton, editor, *The SMART Retrieval System – Experiments in Automatic Document Processing*, pages 313–323. Prentice Hall, Inc., Englewood Cliffs, NJ, 1971.

R. T. Rockafellar. *Convex analysis*. Princeton University Press, Princeton, New Jersey, 1970.

T. Roos, H. Wettig, P. Grünwald, P. Myllymäki, and H. Tirri. On discriminative Bayesian network classifiers and logistic regression. *Machine Learning*, 59(3):267–296, June 2005.

J. Salojärvi, K. Puolamäki, J. Simola, L. Kovanen, I. Kojo, and S. Kaski. Inferring relevance from eye movements: Feature extraction. Technical Report A82, Helsinki University of Technology, Publications in Computer and Information Science, March 2005. http://www.cis.hut.fi/eyechallenge2005/.

G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41(4):288–297, 1990.

D. D. Salvucci and J. R. Anderson. Automated eye-movement protocol analysis. *Human-Computer Interaction*, 16:39–86, 2001.

B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Analysis of recommendation algorithms for e-commerce. In *EC'00: Proceedings of the 2nd ACM conference on Electronic commerce*, pages 158–167, New York, NY, USA, 2000. ACM. ISBN 1-58113-272-7. doi: http://doi.acm.org/10.1145/352871.352887.

B. Sarwar, G. Karypis, J. Konstan, and J. Reidl. Item-based collaborative filtering recommendation algorithms. In *WWW '01: Proceedings of the 10th international conference on World Wide Web*, pages 285–295, New York, NY, USA, 2001. ACM. ISBN 1-58113-348-0. doi: http://doi.acm.org/10.1145/371920.372071.

R. E. Schapire. The strength of weak learnability. *Machine Learning*, 5(2):197–227, 1990. doi: http://dx.doi.org/10.1023/A:1022648800760.

R. Schlüter and W. Macherey. Comparison of discriminative training criteria. *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, pages 493–496, 1998.

B. Schökopf and A. J. Smola. *Learning with Kernels.* MIT Press, Cambridge. Massachusetts, 2002.

G. Schwartz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

D. F. Shanno. Conditioning of quasi-Newton methods for function minimization. *Mathematics of Computation*, 24:647–656, 1970.

U. Shardanand and P. Maes. Social information filtering: Algorithms for automating 'word of mouth'. In *Proceeding of Computer Human Interaction*, pages 210–217, 1995.

S. Sharma. *Applied Multivariate Techniques.* John Wiley & Sons, Inc., 1996.

A. Singhal. Modern information retrieval: A brief overview. *IEEE Data Engineering Bulletin*, 24(4):35–43, 2001.

D. Spiegelhalter, N. Best, B. Carlin, and A. van der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, 64:583–639, 2002.

I. Starker and R. A. Bolt. A gaze-responsive self-disclosing display. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 3–10. ACM Press, 1990. ISBN 0-201-50932-6. doi: http://doi.acm.org/10.1145/97243.97245.

C. Sutton and A. McCallum. An introduction to conditional random fields for relational learning. In L. Getoor and B. Taskar, editors, *Introduction to Statistical Relational Learning*, chapter 4. MIT Press, 2006.

D. Tennenhouse. Proactive computing. *Commun. ACM*, 43(5):43–50, 2000. ISSN 0001-0782. doi: http://doi.acm.org/10.1145/332833.332837.

K. Torkkola, N. Masseya, B. Leivian, C. Wood, J. Summers, and S. Kundalkar. An architecture for an intelligent driver assistance system. In *Proceedings of the International Conference on Machine Learning and Applications (ICMLA)*, pages 81–85, June 2003.

V. Vapnik. *The Nature of Statistical Learning Theory.* Springer, 1995.

A. Vehtari and J. Lampinen. Model selection via predictive explanatory power. Technical Report B38, Laboratory of Computational Engineering, Helsinki University of Technology, 2004.

J. Venna and S. Kaski. Comparison of visualization methods for an atlas of gene expression data sets. *Information Visualization*, 6:139–54, 2007.

R. Vertegaal. Designing attentive interfaces. In *ETRA '02: Proceedings of the 2002 symposium on Eye tracking research & applications*, pages 23–30, New York, NY, USA, 2002. ACM. ISBN 1-58113-467-3. doi: http://doi.acm.org/10.1145/507072.507077.

T. Vuori, M. Olkkonen, M. Pölönen, A. Siren, and J. Häkkinen. Can eye movements be quantitatively applied to image quality studies? In *Proceedings of the third Nordic conference on Human-computer interaction*, pages 335–338, 2004.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. Technical Report 649, Department of statistics, University of California, Berkeley, September 2003.

D. Wolpert and W. Macready. No free lunch theorems for optimization. *Evolutionary Computation, IEEE Transactions on*, 1(1):67–82, 1997.

P. C. Woodland and D. Povey. Large scale discriminative training of hidden Markov models for speech recognition. *Computer Speech and Language*, 16:25–47, 2002.

S.-N. Yang. An oculomotor-based model of eye movements in reading: The competition/interaction model. *Cognitive Systems Research*, 7:56–69, 2006.

A. Yarbus. *Eye Movements and Vision.* Plenum Press, New York, 1967.

G. Zelinsky, W. Zhang, B. Yu, X. Chen, and D. Samaras. The role of top-down and bottom-up processes in guiding eye movements during visual search. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems*, volume 18, pages 1569–1576, Cambridge, MA, 2006. MIT Press.

P. Zigoris and Y. Zhang. Bayesian adaptive user profiling with explicit & implicit feedback. In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 397–404, New York, NY, USA, 2006. ACM. ISBN 1-59593-433-2. doi: http://doi.acm.org/10.1145/1183614.1183672.