Department of Biomedical Engineering and Computational Science

Approximate Bayesian Inference Methods for Regression and Classification with Gaussian Processes and Neural Networks

Pasi Jylänki

Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 152/2013

Approximate Bayesian Inference Methods for Regression and Classification with Gaussian Processes and Neural Networks

Pasi Jylänki

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Science, at a public examination held at the auditorium F239a of the school on 17 October 2013 at 12.

Aalto University
School of Science
Department of Biomedical Engineering and Computational Science
Bayesian Methodology

Supervising professor
Prof. Jouko Lampinen

Thesis advisor
Dr. Aki Vehtari

Preliminary examiners
Assoc. Prof. Ole Winther, Technical University of Denmark, Denmark
Dr. Hannes Nickisch, Philips Research, Hamburg, Germany

Opponent
Prof. Manfred Opper, TU Berlin, Germany

Aalto University publication series
DOCTORAL DISSERTATIONS 152/2013

ISBN 978-952-60-5354-7
ISBN 978-952-60-5355-4 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-5355-4

Unigrafia Oy
Helsinki 2013

Finland

Aalto University

Abstract

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Author
Pasi Jylänki

Name of the doctoral dissertation
Approximate Bayesian Inference Methods for Regression and Classification with Gaussian Processes and Neural Networks

Publisher School of Science

Unit Department of Biomedical Engineering and Computational Science

Series Aalto University publication series DOCTORAL DISSERTATIONS 152/2013

Field of research Computational Science

Manuscript submitted 11 June 2013 **Date of the defence** 17 October 2013

Permission to publish granted (date) 19 August 2013 **Language** English

☐ **Monograph** ☒ **Article dissertation (summary + original articles)**

Abstract

During the recent decades much research has been done on a very general approximate Bayesian inference framework known as expectation propagation (EP), which has been found to be a fast and very accurate method in many experimental comparisons. A challenge with the practical application of EP is that a numerically robust and computationally efficient implementation is not straightforward with many model specifications, and that there is no guarantee for the convergence of the standard EP algorithm. This thesis considers robust and efficient application of EP using Gaussian approximating families in three challenging inference problems. In addition, various experimental results are presented to compare the accuracy of EP with several alternative methods for approximate Bayesian inference.

The first inference problem considers Gaussian process (GP) regression with the Student-t observation model, where standard EP may run into convergence problems, because the posterior distribution may contain multiple modes. This thesis illustrates the situations where standard EP fails to converge, reviews different modifications and alternative algorithms for improving the convergence, and presents a robust EP implementation that relies primarily on parallel EP updates and uses a provably convergent double-loop algorithm with adaptively selected step size in difficult cases.

The second inference problem considers multi-class GP classification with the multinomial probit model, where a straightforward EP implementation requires either multi-dimensional numerical integrations or a factored posterior approximation for the latent values related to the different classes. This thesis describes a novel nested EP approach that does not require numerical integrations and approximates accurately all between-class posterior dependencies of the latent values, but still scales linearly in the number of classes.

The third inference problem considers nonlinear regression using two-layer neural networks (NNs) with sparsity-promoting hierarchical priors on the inputs, where the challenge is to construct sufficiently accurate and computationally efficient approximations for the likelihood terms that depend in a non-linear manner on the network weights. This thesis describes a novel computationally efficient EP approach for simultaneous approximate integration over the posterior distribution of the weights, the hierarchical scale parameters of the priors, and the residual scale. The approach enables flexible definition of weight priors with different sparseness properties, and it can be extended beyond standard activation functions and NN model structures to form flexible nonlinear predictors from multiple sparse linear models.

Keywords approximate Bayesian inference, expectation propagation, Gaussian processes, neural networks

ISBN (printed) 978-952-60-5354-7 **ISBN (pdf)** 978-952-60-5355-4

ISSN-L 1799-4934 **ISSN (printed)** 1799-4934 **ISSN (pdf)** 1799-4942

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2013

Pages 204 **urn** http://urn.fi/URN:ISBN:978-952-60-5355-4

Aalto-yliopisto

Tiivistelmä

Aalto-yliopisto, PL 11000, 00076 Aalto www.aalto.fi

Tekijä
Pasi Jylänki

Väitöskirjan nimi
Approksimatiivisia bayesilaisia päättelymenetelmiä regressioon ja luokitteluun gaussisilla prosesseilla ja neuroverkoilla

Julkaisija Perustieteiden korkeakoulu

Yksikkö Lääketieteellisen tekniikan ja laskennallisen tieteen laitos

Sarja Aalto University publication series DOCTORAL DISSERTATIONS 152/2013

Tutkimusala Laskennallinen tiede

Käsikirjoituksen pvm 11.06.2013 **Väitöspäivä** 17.10.2013

Julkaisuluvan myöntämispäivä 19.08.2013 **Kieli** Englanti

☐ **Monografia** ☒ **Yhdistelmäväitöskirja (yhteenveto-osa + erillisartikkelit)**

Tiivistelmä

Viimeisien vuosikymmenien aikana on tutkittu paljon bayesilaiseen approksimatiiviseen päättelyyn soveltuvaa expectation-propagation-menetelmää (EP), joka on osoittaunut nopeaksi ja erittäin tarkaksi useissa kokeellisissa vertailuissa. Haasteena EP:n soveltamisessa on se, että numeerisesti robusti ja laskennallisesti tehokas käytännön toteutus ei ole suoraviivaista useilla mallimäärittelyillä, ja että normaalimuotoisen EP-algoritmin konvergoituminen ei ole taattu kaikissa tilanteissa. Tämä työ käsittelee robustia ja laskennallisesti tehokasta EP:n toteuttamista gaussisilla approksimaatioilla kolmessa vaativassa mallinnusongelmassa. Lisäksi työssä esitellään useita kokeellisia tuloksia, joissa EP:n tarkkuutta verrataan keskeisiin vaihtoehtoisiin approksimaatiomenetelmiin.

Ensimmäinen mallinnusongelma käsittelee regressiota gaussisilla prosesseilla ja Student-*t*-havaintomallilla, missä EP-algoritmi voi ajautua konvergenssiongelmiin johtuen posteriorijakauman mahdollisesta monimoodisuudesta. Tässä työssä havainnollistetaan tilanteita, joissa normaalimuotoinen EP-algoritmi ei konvergoidu ja käydään läpi erilaisia algoritmimuunnoksia konvergenssin parantamiseksi. Lisäksi esitellään uudentyyppinen algoritmitoteutus, jossa hyödynnetään ensisijaisesti rinnakkaisia EP-päivityksia ja vaikeissa tilanteissa todistettavasti konvergoituvaa kaksoissilmukka-algoritmia mukautuvalla askelpituudella.

Toinen mallinnusongelma käsittelee monen luokan luokittelua multinomiprobitmallilla, missä suoraviivainen EP-toteutus edellyttää joko moniulotteisia numeerisia integrointeja tai riippumattomia posterioriapproksimaatioita. Tässä työssä esitellään uudenlainen sisäkkäisiä EP-approksimaatioita hyödyntävä algoritmi, joka ei vaadi numeerisia integrointeja ja approksimoi tarkasti luokkien väliset posterioririippuvuudet mutta skaalautuu tästä huolimatta lineaarisesti luokkien lukumäärän suhteen.

Kolmas mallinnusongelma käsittelee epälineaarista regressiota kaksikerroksisella neuroverkolla, jossa on harvoja ratkaisuja suosiva hierarkkinen priori sisäänmenoille. EP-toteutuksessa haasteena on riittävän tarkkojen ja laskennallisesti tehokkaiden approksimaatioiden muodostaminen havaintomallin termeille, jotka riippuvat epälineaarisesti kaikista verkon parametreista. Tässä työssä esitellään laskennallisesti tehokas EP-toteutus, jossa integroidaan sekä verkon kertoimien, hierarkkisten skaalaparametrien että kohinaparametrin yli. Toteutus mahdollistaa monipuolisten harvojen priorien määrittelyn ja se voidaan laajentaa yleisille aktivaatiofunktioille ja monipuolisille mallirakenteille. Tämä mahdollistaa monipuolisten epälineaaristen ennustemallien toteuttamisen harvoja lineaarimalleja yhdistelemällä.

Avainsanat approksimatiivinen bayesilainen päättely, expectation propagation, gaussiset prosessit, neuroverkot

ISBN (painettu) 978-952-60-5354-7 **ISBN (pdf)** 978-952-60-5355-4

ISSN-L 1799-4934 **ISSN (painettu)** 1799-4934 **ISSN (pdf)** 1799-4942

Julkaisupaikka Helsinki **Painopaikka** Helsinki **Vuosi** 2013

Sivumäärä 204 **urn** http://urn.fi/URN:ISBN:978-952-60-5355-4

Preface

The research work presented in this thesis has been carried out in the Bayesian Methodology group at the Department of Biomedical Engineering and Computational Science (BECS) at the Aalto University during the years 2009-2013.

I can say that my journey towards completing this thesis has been eventful, to say the least. The first research project I started to work in ended, the second one relocated, the Laboratory of Computational Engineering (LCE) where I started my doctoral studies merged with the Laboratory of Biomedical Engineering to form the present BECS, and later also the Helsinki University of Technology merged with two other universities from the Helsinki area to form the present Aalto University. Despite all these changes BECS, in all of its forms, has provided me a safe and instructive environment for learning and research. My doctoral studies have been supported financially by the Academy of Finland (grants 129230, 129670, and 218248), the EU FP6 project MAIA, and the Finnish Foundation for Technology Promotion. I thank all the parties for their support.

During these eventful years I have been fortunate to have been surrounded by talented people who have helped me in various ways and have made this work possible. First of all, I want to thank my instructor Dr. Aki Vehtari, who took me into the Bayesian Methodology Group, helped and supported me at all times throughout these years, and most importantly, gave me the freedom to explore new ideas. I am also thankful to my present supervisor Prof. Jouko Lampinen and to my former supervisor Prof. Kimmo Kaski for the chance to work broadly on different interesting topics and for providing an excellent working environment. I thank warmly Dr. Jarno Vanhatalo, Dr. Aapo Nummenmaa and soon-to-be Dr. Jaakko Riihimäki for instructive and productive collaboration. I am also

thankful to Prof. Ole Winther and Dr. Hannes Nickisch for providing a thorough preliminary examination of this thesis and helpful comments for improving it.

I also want to thank all my former colleagues at BECS and LCE for scientific collaboration, numerous helpful discussions on work-related matters and also relaxing chats on everyday topics. Out of all these people I would like to mention especially Janne Ojanen, Ville Mäntynen, Dr. Jouni Hartikainen, Dr. Simo Särkkä, Dr. Pekka Marttinen, Dr. Tommi Mononen, Dr. Mari Myllymäki, Dr. Eli Parviainen, Tomi Peltola, Arno Solin, Ville Tolvanen, Dr. Laura Kauhanen, Dr. Toni Tamminen, Dr. Ilkka Kalliomäki, Tommi Nykopp, Janne Lehtonen, Aatu Kaapro, Dr. Toni Auranen, Dr. Sebastian von Alfthan, Dr. Taru Tukiainen, Dr. Linda Kumpula, Dr. Ville-Petteri Mäkinen, Jaakko Niemi and Antti Kangas. Not forgetting the ever-so-important administrative people at BECS who have helped me on numerous occasions; out of them I want to thank especially Mikko Hakala, Timo Aarnio, Jarkko Salmi, and Jari Siven for providing a great computing environment for scientific work.

Beyond work, I want to thank my parents Mauno and Elma, and my brother Ari and his son Joona for support, encouragement and company whenever I have needed it. I also want to thank all my friends for having taken my mind off this work on various occasions, such as band rehearsals that often resulted in high quality musical experiences, or just afterworks and other events that at times resulted in intellectual discussions and at times in not so intellectual but definitely very relaxing ones. Last but most of all, I wish to express my deepest gratitude to my beloved Sofia for having had faith in me and having kept me going during all these, at times also stressful, years. I would not be writing this if you wouldn't have been there.

Espoo, August 30, 2013,

Pasi Jylänki

Contents

List of Publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

I Jarno Vanhatalo, Pasi Jylänki and Aki Vehtari. Gaussian process regression with Student-t likelihood. In *Advances in Neural Information Processing Systems 22*, edited by Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, 1910–1918, 2009.

II Pasi Jylänki, Jarno Vanhatalo and Aki Vehtari. Robust Gaussian Process Regression with a Student-t Likelihood. *Journal of Machine Learning Research*, 12, 3227–3257, Nov 2011.

III Jaakko Riihimäki, Pasi Jylänki and Aki Vehtari. Nested Expectation Propagation for Gaussian Process Classification with a Multinomial Probit Likelihood. *Journal of Machine Learning Research*, 14, 75–109, Jan 2013.

IV Pasi Jylänki, Aapo Nummenmaa and Aki Vehtari. Expectation Propagation for Neural Networks with Sparsity-promoting Priors. *Journal of Machine Learning Research*, Accepted for publication conditioned on minor revisions, preprint: arXiv:1303.6938 [stat.ML], 2013.

Author's Contribution

In all publications the co-authors contributed to writing by revising the text and suggesting modifications. In addition, Vehtari contributed to the initiation of the conducted research and participated in the methodological developments through discussions.

Publication I: "Gaussian process regression with Student-t likelihood"

Vanhatalo had the main contribution in writing the article as well as designing and running the experiments. Jylänki contributed in developing the methodology and implemented the variational methods used for comparisons.

Publication II: "Robust Gaussian Process Regression with a Student-t Likelihood"

Jylänki had the main responsibility for designing and writing the article, developing the methodology as well as implementing and running the experiments. Vanhatalo contributed in developing the methodology, implementing the experiments, and writing the manuscript.

Publication III: "Nested Expectation Propagation for Gaussian Process Classification with a Multinomial Probit Likelihood"

Riihimäki and Jylänki developed the methodology and wrote the article jointly. Jylänki contributed more in the theoretical derivations and Riihimäki had the main responsibility for implementing and running the experiments.

Publication IV: "Expectation Propagation for Neural Networks with Sparsity-promoting Priors"

Jylänki had the main responsibility for deriving the methodology, designing and implementing the experiments, and writing the manuscript. Nummenmaa contributed in designing and writing the manuscript.

1. Introduction

In the recent years, ever increasing amounts of data are being produced in different functions of the modern society. This is being accelerated by the huge amount content uploaded in the internet, for example, through user accounts in various webservices, but the interest in utilizing the available databases using more elaborate methods is increasing also in science and more traditional areas of business and public service. For example, customer data is collected by retailers via user accounts or membership cards that reward for customer loyalty; measurements are made during production, use, or maintenance of various products; commercial measurement devices require efficient data analysis software for operation; data including the medical conditions and attributes from thousands of patients is collected by health care organizations; genome-wide association studies aiming to find risk factors for specific traits or illnesses can consist of millions of nucleotid variations from thousands of subjects; scientific brain measurements can contain signals from thousands of spatial locations recorded at several time instant for several subjects.

Utilization of these databases requires well-defined models as well as efficient computational methods for estimating the unknown model parameters and computing predictions together with other model summaries of interest. This thesis focuses on a supervised predictive modeling approach where a probabilistic model is assumed on the target variables of interest conditioned on known explanatory (input) variables and a set of unknown model parameters. The model definition encompasses the modeler's prior beliefs on the unknown functional relationship between the input variables and the target variables together with the assumptions on the process that generates the observations given the latent function values. The properties of the model are controlled by the adjustable model parameters, and the goal of the inference problem is to learn an estimate of these

parameters that accurately represents the underlying model properties using a set of observed input-output pairs. Once the parameters estimates are determined, they can be used to construct a predictive model, which can be utilized for making predictions on the target variables associated with future inputs, or alternatively to study the predictive relevances of the different input variables or to gain evidence on possible latent phenomena in the data, which are quantified by certain subsets of the model parameters. To achieve these goals, this thesis adopts the Bayesian inference approach from a practical machine learning perspective.

The main challenge in the Bayesian modeling approach is that the inference on the unknown model parameters and the future target variables is analytically intractable with many practically relevant model specifications. Usually analytical solutions are available only for the most trivial models such as a linear predictor with a Gaussian observation model and suitable conjugate priors for the coefficients and the observation noise [see, e.g., Gelman et al., 2004, Bishop, 2006]. Therefore many approximation methods have been proposed to facilitate the Bayesian inference with non-analytical models. Perhaps the most commonly used approach is to draw samples from the posterior distribution using Markov chain Monte Carlo (MCMC) methods, because the wide selection of different sampling algorithms allows a straightforward implementation for virtually any kind of model [see, e.g., Robert and Casella, 2004, Gelman et al., 2004, Bishop, 2006]. The drawback with the MCMC methods is their computational cost, which is manifested especially when the number of unknown model parameters is large and their posterior dependencies are strong causing slow converge of the MCMC chains. Therefore, many computationally cheaper analytical approximation methods are used extensively including methods such as the Laplace approximation (LA) [see, e.g., Laplace, 1774, Tierney and Kadane, 1986, Bishop, 2006, Rasmussen and Williams, 2006, Rue et al., 2009], different variational Bayes (VB) methods based on the variational mean field (VMF) and local variational bound (LVB) approximations [see, e.g., Jordan et al., 1999, Jaakkola and Jordan, 2000, Attias, 2000, Beal, 2003, Bishop, 2006, Murphy, 2012], VB methods based on direct variational minimization of the Kullback-Leibler (KL) divergence [Nickisch and Rasmussen, 2008, Opper and Archambeau, 2009], and expectation propagation (EP) [see, e.g., Minka, 2001a,b, 2005, Opper and Winther, 2005, Heskes et al., 2005]. EP has been found to be a relatively fast and very accurate method in many experimental com-

parisons [Minka, 2001a,b, Kuss, 2006, Nickisch and Rasmussen, 2008, Hernández-Lobato et al., 2010, Cseke and Heskes, 2011]. A challenge with the practical application of EP is that a numerical robust and computationally efficient implementation is not straightforward with many model specifications, and that there is no theoretical guarantee for the convergence of the standard EP algorithm [Minka, 2001b, Minka and Lafferty, 2002, Heskes and Zoeter, 2002, Seeger, 2005, 2008].

The main focus of this thesis is numerically robust and computationally efficient application of EP using Gaussian approximating families in different kinds of non-analytical inference problems. In addition, various experimental results are presented to compare the accuracy of EP with LA, VB, and MCMC approximations.

Publications I and II consider GP regression with the outlier-robust Student-t observation model. The challenge with the Student-t model is that the conditional posterior distribution of the latent function values may contain multiple modes and that the potential outlying observations result in local increases in the approximate posterior uncertainty on the corresponding latent function values with the LA and EP approximations. The latter property can be seen as negative precision contributions in the approximate posterior covariances contrary to the always non-negative contributions with log-concave models such as the logit and probit used in binary GP-classification [Nickisch and Rasmussen, 2008]. This requires some additional care when implementing the LA and EP approximations following the standard algorithms described by Rasmussen and Williams [2006] and can also result in clearly different behavior between the approximate methods. The LA approximation requires a robust and efficient method for determining the conditional mode of the latent function values given the hyperparameters, and a robust way for determining the approximate marginal likelihood in case the Hessian of the conditional posterior is close to singular at the local mode. Publication I describes a robust implementation of LA using the EM algorithm [see, e.g., Gelman et al., 2004] for determining the conditional mode and computational modifications that enable robust evaluation of the marginal likelihood approximation. By experimental comparisons with MCMC and the commonly used VMF approximation [Tipping and Lawrence, 2003, Kuss, 2006] it is also shown that LA provides a good alternative for VMF in terms of speed and accuracy. Applying EP for the Student-t model is theoretically straightforward, because each likelihood term depends only on a single latent

value, but the practical implementation requires that the posterior representation is kept numerically stable during the EP updates, and that the convergence of the algorithm can be verified also in difficult cases. Publication II discusses the converges problems of EP caused by the potential multimodalities in the conditional posterior distribution, describes a robust EP implementation based on parallel EP updates [van Gerven et al., 2009] and a provably convergent double-loop algorithm [Minka, 2001b, Opper and Winther, 2005], and provides more extensive comparisons with alternative approximate methods including VMF, LVB [Gibbs and MacKay, 2000, Nickisch and Rasmussen, 2008], and MCMC [Neal, 1997, Gelman et al., 2004, Vanhatalo and Vehtari, 2007].

Publication III describes a novel nested EP approach for multi-class classification with the multinomial probit model and GP priors [see, e.g., Girolami and Rogers, 2006]. The challenge with the multinomial probit model is that the tilted distributions related to each likelihood term depend on multiple latent values (one for each output class), which is why a straightforward EP implementation requires either multi-dimensional numerical integrals [Seeger and Jordan, 2004] or a factorized approximation for the latent values related to the different classes [Seeger et al., 2006, Girolami and Zhong, 2007]. Furthermore, a straightforward implementation that takes account of the posterior dependencies between the latent values from different classes would result in posterior computations that scale cubically with respect to number of classes c, which may become prohibitive with larger c. The proposed approach applies inner EP approximations within an outer main EP loop on a well-known integral representation of the multinomial probit model to approximate the multivariate integrals required for determining the mean vector, covariance matrix, and normalization term associated with each tilted distribution. The resulting algorithm does not require numerical quadrature integrations and the intrinsic parametric structure of the inner EP approximations results in a posterior representation that scales linearly with c and achieves therefore similar complexity with the LA approach described by Williams and Barber [1998]. Additional computational speed-up is achieved by introducing an incremental update scheme where damped updates [Minka and Lafferty, 2002, Heskes and Zoeter, 2002] are done on the scalar site parameters of the inner EP approximations instead of the related multivariate site parameters associated with the outer EP. The accuracy of the proposed nested EP approach is assessed by comparisons

with LA, VMF, and MCMC approximations.

Publication IV proposes a novel EP approach for nonlinear regression with two-layer neural networks (NNs) and sparsity-promoting hierarchial priors on the inputs. From a practical modeling perspective, GPs with neural network covariance function allow convenient integration over the uncertainty on the unknown latent function resulting from a similar two-layer NN with infinitely many hidden units [Williams, 1998, Rasmussen and Williams, 2006]. However, with infinite GP network the inherent complexity of the posterior computations scale cubically with respect to the number of observations n, which is why additional sparse approximations are required with large data sets [Quiñonero-Candela and Rasmussen, 2005]. Furthermore, the inference on the covariance function hyperparameters, which control, e.g., the nonlinearity of the latent function with respect to each input-dimension, is analytically intractable. Therefore, additional (e.g, MCMC) approximations are required, if the marginal maximum a posterior (MAP) estimates of the hyperparameters are not sufficient in problems with a large number of input features. One aim of Publication IV is to study whether computationally efficient nonlinear predictors with flexible input priors could be constructed by adapting the existing EP methodology presented for sparse linear models [Seeger, 2008, Hernández-Lobato et al., 2008, van Gerven et al., 2009] to finite-parametric NNs with a linear input-layer. Compared with the GP models considered in Publications I–III, a key technical difference in this NN approach is that EP approximations are formed, in addition to the non-Gaussian likelihood terms, also for the prior terms of the network weights. In this respect, the inference problem resembles the existing EP approaches for generalized linear models (GLMs), where both the likelihood and the prior terms are intractable [see, e.g., Seeger et al., 2007, Hernández-Lobato et al., 2008]. The challenge in the EP implementation is to construct a sufficiently accurate and computationally efficient Gaussian approximations for the likelihood terms that depend in a non-linear manner from all the network weights. Similarly to Publication III, this requires determining the moments of the multivariate tilted distributions associated with each likelihood term. Once such likelihood term approximations are obtained, adapting the existing EP methodology for sparse linear models is rather straightforward.

Publication IV describes a novel approach for approximating the moments of the tilted distributions, which is based on utilizing a suitable

factorized structure for the posterior approximation, and a combination of the approximate linear filtering paradigm used in the unscented Kalman filter [Wan and van der Merwe, 2000] and a similar Gaussian approximation that has been used by Ribeiro and Opper [2011] to form factorizing EP approximation for linear perceptrons. The proposed approach requires only one-dimensional numerical quadratures for determining the means and variances of the potentially multimodal tilted distributions and results in a computationally efficient algorithm, whose complexity scales linearly with respect to both n and the number of hidden units. The complexity scales similarly to an ensemble of independent sparse linear models and also the resulting approximate predictive model can be interpreted as a nonlinear combination of independent sparse linear models associated with each hidden unit. Compared with the inference approaches considered in Publications I–III that rely on marginal MAP estimates of the hyperparameters, in Publication IV, EP is used to approximately integrate over the posterior uncertainty of all the model parameters including the network weights, the hierarchical scale parameters of the weight priors, and the parameter controlling the observation noise magnitude. In addition to the existing EP methodology for sparse linear models, Publication IV proposes a flexible EP-based hierarchical prior framework that enables flexible definition of weight priors with different sparseness properties such as independent Laplace priors with a common scale parameter [Seeger, 2008] or Gaussian automatic relevance determination (ARD) priors with different relevance parameters for all inputs [Neal, 1996]. The computational efficiency and predictive accuracy of the approach is assessed by comparisons with two other models with ARD priors: an infinite GP network with MAP estimates of the hyperparameters [Rasmussen and Williams, 2006], and a finite NN with MCMC integration over all the model parameters [Neal, 1996].

The rest of the thesis is organized as follows. Section 2 contains an introduction to and a literature review on approximate Bayesian inference. Section 2.1 gives first a general introduction to the Bayesian predictive modeling approach and 2.2 defines next a general predictive model structure used in the upcoming discussion on various approximate inference methods. Section 2.3 gives a detailed description of EP with both Gaussian and general exponential family approximations and summarizes alternative provably convergent double-loop algorithms. After that shorter descriptions are given on other approximate inference methods including

VMF in Section 2.4, LVB in Section 2.5, and Laplace's method in Section 2.6. To conclude the general overview of the approximate Bayesian inference methods, Section 2.7 discusses shortly various approaches proposed for improving the approximate Gaussian marginal distributions based on successive use of LA and EP. Section 3 introduces the different case studies considered in Publications I–IV and links the associated approximate inference methods to the general discussion of Section 2. Finally Section 4 gives a final discussion on some of the key aspects considered in this thesis and discusses different possibilities for future research.

2. Approximate Bayesian Inference Methods

This thesis focuses on a supervised statistical modeling approach where a probabilistic model $p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$ is assumed on a vector (or collection) of target variables $\mathbf{y}$ conditioned on a vector of input variables $\mathbf{x}$ and a vector of unknown model parameters denoted by $\boldsymbol{\theta}$. A common approach is to choose a physically motivated model or a sufficiently flexible general purpose model for the unobserved (latent) functional relationship between the inputs $\mathbf{x}$ and the outputs $\mathbf{y}$, denoted here by $\mathbf{f}(\mathbf{x}, \boldsymbol{\theta})$, and to model the uncertainty on the observations (for example, the observation noise) given the latent function values $\mathbf{f}(\mathbf{x}, \boldsymbol{\theta})$ using a suitable probabilistic observation model $p(\mathbf{y}|\mathbf{f}(\mathbf{x}, \boldsymbol{\theta}), \boldsymbol{\theta})$. In this definition, $\boldsymbol{\theta}$ contains both the parameters of the unknown function and the observation model. Commonly used general purpose approaches for modeling the latent functional relationship $\mathbf{f}(\mathbf{x}, \boldsymbol{\theta})$ are linear models, finite-parametric nonlinear models such as neural networks and spline models, and infinite-parametric kernel models such as Gaussian processes (GP) [see, e.g., Bishop, 2006, Rasmussen and Williams, 2006]. Common observation models are, for example, regression using the Gaussian model and more robust alternatives such as the Student-t model, classification with the binary logit and probit models together with their generalizations for multiple target classes, modeling of count observations with the Poisson and the binomial model, and survival analysis using the Weibull model [Gelman et al., 2004, Bishop, 2006].

The main goal of the modeling approach is to learn the unobserved model parameters $\boldsymbol{\theta}$ from a set of n observations (input-output pairs), denoted by $\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n$, and to construct a predictive model $p(\mathbf{y}_*|\mathcal{D}, \mathbf{x}_*)$, which can be used to make predictions for the target variables $\mathbf{y}_*$ associated with future inputs vectors $\mathbf{x}_*$. Another common objective is to obtain a reliable estimate for the unknown parameters $\boldsymbol{\theta}$, which can be used, for example, to study the predictive relevances of the different input features

contained in $\mathbf{x}$, or to gain evidence on possible latent phenomena in the data which are quantified based on certain components of $\boldsymbol{\theta}$.

2.1 The Bayesian Modeling Approach

To determine reliable estimates for the unknown parameters $\boldsymbol{\theta}$ and to obtain the predictive model $p(\mathbf{y}_*|\mathcal{D}, \mathbf{x}_*)$, this thesis adopts the Bayesian approach from a practical machine learning perspective. The fundamental idea of Bayesian inference was considered independently already by Bayes [1763] and [Laplace, 1774] but only in the last few decades Bayesian methods have been utilized widely in various modeling applications mainly because of the rapid development of computational resources and inference methodology. A thorough theoretical description of the Bayesian approach is given by Bernardo and Smith [2000] and various statistical modeling applications are presented by Gelman et al. [2004] and O'Hagan and Forster [2004]. Many useful Bayesian methods for practical machine learning applications are summarized by Bishop [2006] and Murphy [2012].

The fundamental idea behind the Bayesian modeling approach can be summarized by Bayes' theorem, which in case of our supervised modeling approach can be written as

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{Y}|\mathbf{X})}, \tag{2.1}$$

where all the input and output variables related to the n observations are denoted by $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$ and $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^n$, respectively. In equation (2.1), $p(\boldsymbol{\theta})$ is the prior probability distribution assigned to the unknown parameters $\boldsymbol{\theta}$ before observing the data $\mathcal{D}$, $p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}) = \prod_{i=1}^n p(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\theta})$ is the likelihood function of $\boldsymbol{\theta}$ resulting from the observations $\mathcal{D}$ made according to the chosen model, $p(\boldsymbol{\theta}|\mathcal{D})$ is the posterior probability of $\boldsymbol{\theta}$ after $\mathcal{D}$ has been observed, and $p(\mathbf{Y}|\mathbf{X}) = \int p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$ is the marginal probability of the observations conditioned on the current model and prior assumption. Bayes' theorem states that the revised probability of $\boldsymbol{\theta}$ after observing the data $\mathcal{D}$ is obtained by multiplying the prior probability $p(\boldsymbol{\theta})$ with the joint conditional probability of the observations $p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta})$ according to the law of conditional probability and normalizing the result by $p(\mathbf{Y}|\mathbf{X})$.

The posterior distribution $p(\boldsymbol{\theta}|\mathcal{D})$ contains all information on $\boldsymbol{\theta}$ provided by the observed data combined with the prior beliefs, and quantities of interest associated with certain components of $\boldsymbol{\theta}$ can be obtained by sum-

marizing the corresponding marginal distributions of $p(\boldsymbol{\theta}|\mathcal{D})$. For example, inference on element θ_j can be done by reporting point estimates such as the mode, mean, and median of $p(\theta_i|\mathcal{D})$, or by reporting the interval that contains θ_i with 95% posterior probability, that is, the Bayesian credible interval. Testing a hypothesis such as $\theta_j \geq 0$ can be done by simply calculating the marginal posterior probability $P(\theta_j \geq |\mathcal{D})$.

The main difference with the classical frequentist modeling approach is that a Bayesian is willing to treat $\boldsymbol{\theta}$ as a random variable by assigning a prior uncertainty $p(\boldsymbol{\theta})$ to it, and to quantify the posterior uncertainty with $p(\boldsymbol{\theta}|\mathcal{D})$. In the classical frequentist approach, the unknown parameters take unique values and it is not permitted to treat them as random variables [see, e.g., O'Hagan and Forster, 2004]. In practice, the classical approach relies on maximum likelihood (ML) point estimates of $\boldsymbol{\theta}$, and the level of confidence on the estimate is summarized by a single realization of a confidence interval computed using the observed data $\mathcal{D}$. For example, assuming a 5% confidence level, the interval contains the true value of $\boldsymbol{\theta}$ in a 95% proportion of all samples of size n generated from the model. The classical approach can result in counter-intuitive and erratic behavior in some cases, because it does not follow the so-called likelihood principle, which states that all inference on $\boldsymbol{\theta}$ should be based on the likelihood provided by the observed data $\mathcal{D}$, not on what could have been observed [O'Hagan and Forster, 2004, Murphy, 2012]. For example, the result of hypothesis testing may depend on the decisions made on when to stop the collection of data even though the actual observations $\mathcal{D}$ remain the same.

Using the posterior distribution (2.1), the predictive model for future target variables $\mathbf{y}_*$ given inputs $\mathbf{x}_*$ can be written as

$$p(\mathbf{y}_*|\mathcal{D}, \mathbf{x}_*) = \int p(\mathbf{y}_*|\mathbf{x}_*, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta}, \tag{2.2}$$

where the posterior expectation of the chosen predictive model $p(\mathbf{y}_*|\mathbf{x}_*, \boldsymbol{\theta})$ is computed with respect to the posterior uncertainty of $\boldsymbol{\theta}$ given the observed data. From a theoretical perspective, the Bayesian solution to the prediction problem $p(\mathbf{y}_*|\mathbf{x}_*, \mathcal{D})$ can be written in the form of (2.2), with rather general conditions without defining first a specific parametric model structure $p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$ with prior $p(\boldsymbol{\theta})$ and computing the posterior distribution according to (2.1). Assuming $\mathbf{y}_1, ..., \mathbf{y}_n, \mathbf{y}_*$ conditioned on the respective inputs $\mathbf{x}_1, ..., \mathbf{x}_n, \mathbf{x}_*$ to be an exchangeable sequence of real-valued random vectors equipped with a probability measure P and a corresponding existing density function p, it follows that the probability of the sequence can be expressed independently as $P(\mathbf{Y}, \mathbf{y}_*|\mathbf{X}, \mathbf{x}_*, F) =$

$F(\mathbf{y}_*|\mathbf{x}_*)\prod_{i=1}^{n}F(\mathbf{y}_i|\mathbf{x}_i)$ conditioned on an unknown random distribution function F. Assuming further that the distribution function of F exist and that it can be expressed using a finite dimensional parameter $\boldsymbol{\theta}$ with a probability distribution function $p(\boldsymbol{\theta})$, the equation (2.2) follows directly by integration over the uncertainty of F [Bernardo and Smith, 2000]. In practice, the Bayes prediction is obtained by choosing a model that can sufficiently well express the modeler's beliefs about the unknown distribution function $F(\mathbf{y}|\mathbf{x})$ by defining $p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})$ and $p(\boldsymbol{\theta})$, and integrating over all the nuisance parameters $\boldsymbol{\theta}$. In the classical frequentist approach or other widely-used point-estimate based methods, the predictive model (2.2) would be summarized using only a point-estimate of the unknown model parameters denoted by $\hat{\boldsymbol{\theta}}$, that is, $p(\mathbf{y}_*|\mathcal{D},\mathbf{x}_*)=p(\mathbf{y}_*|\mathbf{x}_*,\hat{\boldsymbol{\theta}})$.

One benefit of the Bayesian approach is that it constitutes a flexible framework for the modeling process: Prior knowledge on the modeling problem can be incorporated in a principled way using the prior distribution $p(\boldsymbol{\theta})$ as a proxy, and consequently more general model specifications $p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})$ can be adapted for a variety of different modeling problems. In addition, intricate latent dependencies can be modeled using hierarchical models, where the prior distributions of $\boldsymbol{\theta}$ are defined conditional on higher-level hyperparameters $\boldsymbol{\phi}$, that is, $p(\boldsymbol{\theta}|\boldsymbol{\phi})$, and hyperpriors $p(\boldsymbol{\phi})$ are assigned to $\boldsymbol{\phi}$ [for more on hierarchical models see, e.g., Gelman et al., 2004]. For example, with generalized linear models scalar observations y can be modeled using a suitable observation model $p(y|f(\mathbf{x},\boldsymbol{\theta}),\boldsymbol{\phi}_1)$, where the latent function is defined as $\mathrm{f}(\mathbf{x},\boldsymbol{\theta})=\boldsymbol{\theta}^{\mathrm{T}}\mathbf{x}$, and $\boldsymbol{\phi}_1$ contains the hyperparameters related to the observation model. The priors of the coefficients $\boldsymbol{\theta}$ can be conditioned on higher-level hyperparameters according to $p(\theta_j|\boldsymbol{\phi}_{l_j})$, where the hyperparameters $\boldsymbol{\phi}_2=\{\boldsymbol{\phi}_l\}_{l=1}^{L}$ control the scale of the coefficients belonging to certain predefined groups indexed by $l=1,...,L$, and the group membership of θ_j is defined by l_j. This hierarchical prior definition can be used to suppress the harmful effects of potentially irrelevant input features within the automatic relevance determination (ARD) framework [Mackay, 1995, Neal, 1996], or to couple the magnitudes of the weights belonging to some known category in modeling problems such as multi-class classification or multitask learning [Bishop, 2006, Murphy, 2012]. In many commonly used models, the intermediate-level parameters $\boldsymbol{\theta}$ are often called latent variables, because they are not directly observed but they are used to model the latent dependencies between the observations and subsequently inferred or integrated over using the ob-

served data. For example, in GP models (assuming scalar observations) the unobserved functional relationship $f(\mathbf{x}, \boldsymbol{\theta})$ is modeled by assigning a multivariate Gaussian prior $p(\mathbf{f}|\boldsymbol{\phi}_2)$ to a finite set of function values $\mathbf{f} = [f_1, ..., f_n]^{\mathrm{T}}$ related to the observed input-output pairs according to $f_i = f(\mathbf{x}_i)$ and $p(y_i|f(\mathbf{x}_i, \boldsymbol{\theta}), \boldsymbol{\phi}_1) = p(y_i|f_i, \boldsymbol{\phi}_1)$ [Rasmussen and Williams, 2006]. Here the components of $\mathbf{f} \equiv \boldsymbol{\theta}$ are often called latent values, and the scale and smoothness properties of their prior are controlled by hyperparameters $\boldsymbol{\phi}_2$. An additional example of the beneficial properties of the Bayesian approach is that it provides a principled framework for sequential estimation [see e.g., Särkkä, 2006] and model selection based on expected utilities [Bernardo and Smith, 2000, Vehtari and Ojanen, 2012].

An essential property of the Bayesian approach is the integration over the unknown parameters $\boldsymbol{\theta}$ in equations (2.1) and (2.2). In general, when the number of observations n becomes large enough compared to the dimension of $\boldsymbol{\theta}$, the posterior distribution $p(\boldsymbol{\theta}|\mathcal{D})$ becomes sharply peaked around the MAP estimate (2.3). In such cases, integration over the posterior uncertainty of $\boldsymbol{\theta}$ is not necessary for good predictive accuracy but deriving posterior summaries such as credible intervals may not be feasible without quantification of the posterior uncertainty. For example, if certain components of $\boldsymbol{\theta}$ are weakly determined by the observations through the likelihood $p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$, their marginal posterior uncertainties may be significant.

When the number of unknown model parameters (the dimension of $\boldsymbol{\theta}$) increases compared with n, using point estimates of $\boldsymbol{\theta}$ to make predictions with $p(\mathbf{y}_*|\mathbf{x}_*, \hat{\boldsymbol{\theta}})$ can worsen the predictive accuracy of the model significantly, because the model can overfit to the finite set of observations. Typically, this can be seen as an almost perfect predictive accuracy with the training data set but with an independent validation set the performance can be worse than with a simple baseline prediction with the mean values of $\mathbf{Y}$. The problem is exacerbated with very flexible nonlinear models such as neural networks[1], but overfitting can be problematic also with linear models if the number of coefficients is larger than n, which is typical for linear inverse problems, where no unique ML-solution exists [Kaipio and Somersalo, 2005]. The overfitting effects associated with ML-estimates can often be reduced by assigning a suitable regularizing prior to $\boldsymbol{\theta}$ and

[1]The recent work of Hinton et al. [2012] gives good examples of overfitting with extremely complex multi-layer neural network models (>100000 parameters) with large data sets and simple non-Bayesian ways to reduce these effects.

making predictions with the maximum a posterior (MAP) estimate

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathcal{D}) = \arg\max_{\boldsymbol{\theta}} p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}) p(\boldsymbol{\theta}), \tag{2.3}$$

where in the last step the normalization term of (2.1) is neglected, because it does not depend on $\boldsymbol{\theta}$. For example, a commonly used approach with linear models is to use the Lasso (or $\mathcal{L}_1$-norm) regularization that yields unique MAP solutions in underdetermined problems [Tibshirani, 1994]. $\mathcal{L}_1$ regularization results in truly sparse estimates, where many of the linear coefficients get exactly zero values, which means that the corresponding input features are effectively pruned out of the predictive model reducing the potentially harmful effects of irrelevant features.

Taking the Bayesian treatment of the unknown parameters further, it is common to utilize hierarchical model structures that enable analytical integration over $\boldsymbol{\theta}$ conditioned on the hyperparameters $\boldsymbol{\phi}$. If the dimension of $\boldsymbol{\phi}$ is not too large compared to n and the values of $\boldsymbol{\phi}$ identify well from the data, good predictive performance is often obtained by making predictions with $p(\mathbf{y}_*|\mathbf{x}_*, \mathcal{D}, \hat{\boldsymbol{\phi}}) = \int p(\mathbf{y}_*|\mathbf{x}_*, \boldsymbol{\theta}, \hat{\boldsymbol{\phi}}) p(\boldsymbol{\theta}|\mathcal{D}, \hat{\boldsymbol{\phi}}) d\boldsymbol{\theta}$ using the marginal MAP estimate given by

$$\hat{\boldsymbol{\phi}} = \arg\max_{\boldsymbol{\phi}} p(\boldsymbol{\phi}|\mathcal{D}) = \arg\max_{\boldsymbol{\phi}} p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\phi}) p(\boldsymbol{\phi}), \tag{2.4}$$

where it is assumed that a closed-form expression can be computed for the marginal likelihood $p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\phi}) = \int p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\phi}) p(\boldsymbol{\theta}|\boldsymbol{\phi}) d\boldsymbol{\theta}$. This is the standard approach with GP models, where integration is done over the latent function values $\mathbf{f} = [f(\mathbf{x}_1), ..., f(\mathbf{x}_n)]^{\mathrm{T}}$ conditioned on marginal MAP estimates of the hyperparameters $\boldsymbol{\phi}_2$ and $\boldsymbol{\phi}_1$ that control the prior $p(\mathbf{f}|\boldsymbol{\phi}_2)$ and the observation model $p(\mathbf{y}|\mathbf{f}, \boldsymbol{\phi}_1)$ [Rasmussen and Williams, 2006]. Continuing the previous $\mathcal{L}_1$ regularization example concerning linear models, the overfitting effects can be mitigated also by introducing hierarchical priors $p(\theta_j|\phi_j)$, where hyperparameters ϕ_j control the prior scales of the respective linear coefficients θ_j. Choosing a suitable family for $p(\theta_j|\phi_j)$ and determining the marginal MAP estimates of $\boldsymbol{\phi}$ according to (2.4), can result in sparse conditional mean estimates $\mathrm{E}(\boldsymbol{\theta}|\mathcal{D}, \hat{\boldsymbol{\phi}})$, because the scale parameters ϕ_j related to the potentially irrelevant or unnecessary input features are driven to zero. This approach implements the ARD framework (also known as sparse Bayesian learning) for linear models [Tipping, 2001, Qi et al., 2004]. With linear models, an interesting connection has been found between the MAP solution (2.3) and the marginal MAP solution (2.4): The ARD solution is exactly equivalent to a MAP estimate

of the coefficients obtained using a particular class of nonfactorial coefficient prior distributions [Wipf and Nagarajan, 2008, Wipf et al., 2011]. This class of priors includes models that have desirable advantages such as fewer local minima compared with the regular MAP estimates. This is an interesting example of the theoretical benefits of integration over the intermediate parameters $\boldsymbol{\theta}$ in the model hierarchy even though the actual number of unknowns in the point-estimate based inference stays the same.

From a theoretical perspective, the best approach would be to integrate over all uncertain (nuisance) parameters (both $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$) and to make predictions with $p(\mathbf{y}_*|\mathbf{x}_*, \mathcal{D}) = \int p(\mathbf{y}_*|\mathbf{x}_*, \boldsymbol{\theta}, \boldsymbol{\phi}) p(\boldsymbol{\theta}, \boldsymbol{\phi}|\mathcal{D}) d\boldsymbol{\theta} d\boldsymbol{\phi}$ [Bernardo and Smith, 2000]. Unfortunately, the integrals involved in both the posterior distribution (2.1) and the predictive distribution (2.2) are analytically intractable with many practically relevant and interesting model specifications and closed-form solutions are available only with the simplest model definitions with suitable conjugate priors, such as the linear model with Gaussian observation model and inverse-gamma or inverse-Wishart prior on the residual variance parameter [see, e.g., Minka, 2000, Gelman et al., 2004]. Full integration over both $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ can in many cases be approximated only with MCMC methods, and the practical benefits of the full integration can often be seen as robust and consistent performance of MCMC methods compared with analytical approximations such as LA, VB, and EP in many experimental comparisons (see, e.g., Publications II and III, Lampinen and Vehtari [2001], Nickisch and Rasmussen [2008], Rue et al. [2009], Vanhatalo et al. [2010]). A drawback with the MCMC methods is that they can be computational expensive in many cases [see, e.g., Nickisch and Rasmussen, 2008], which is why much research has been done to improve the faster analytical alternatives. The upcoming sections give a detailed discussion of EP, after which shorter summaries are presented on VMF, LVB, and LA.

2.2 The Predictive Model Structure

In the following discussion it is assumed that the posterior distribution can be written as

$$p(\boldsymbol{\theta}, \boldsymbol{\phi}|\mathcal{D}) = Z^{-1} \prod_{i=1}^{n} p(\mathbf{y}_i|\mathbf{x}_i, \mathbf{U}_i^{\mathrm{T}}\boldsymbol{\theta}, \boldsymbol{\phi}) p(\boldsymbol{\theta}|\mathbf{X}, \{\mathbf{U}_j\}_{j=n+1}^{n+m}, \boldsymbol{\phi}) p(\boldsymbol{\phi}), \qquad (2.5)$$

where $Z = p(\mathbf{Y}|\mathbf{X})$ and the prior for $\boldsymbol{\theta} \in \mathbb{R}^d$ can be factored into m terms according to

$$p(\boldsymbol{\theta}|\mathbf{X}, \{\mathbf{U}_j\}_{j=n+1}^{n+m}, \boldsymbol{\phi}) = \prod_{j=n+1}^{n+m} t_j(\mathbf{U}_j^{\mathrm{T}}\boldsymbol{\theta}, \boldsymbol{\phi}). \tag{2.6}$$

In this notation each likelihood term $p(\mathbf{y}_i|\mathbf{x}_i, \mathbf{U}_i^{\mathrm{T}}\boldsymbol{\theta}, \boldsymbol{\phi})$ depends on the parameters $\boldsymbol{\theta}$ through a linear transformation $\mathbf{z}_i = \mathbf{U}_i^{\mathrm{T}}\boldsymbol{\theta}$, where $\mathbf{z}_i$ is a latent random variable associated with each term and $\mathbf{U}_i$ is a known transformation matrix that can depend on the inputs $\mathbf{X}$. Similarly, each prior factor $t_j(\mathbf{U}_j^{\mathrm{T}}\boldsymbol{\theta}, \boldsymbol{\phi})$ depends on $\boldsymbol{\theta}$ only through a transformed random variable $\mathbf{z}_j = \mathbf{U}_j^{\mathrm{T}}\boldsymbol{\theta}$. Although it is assumed that $\boldsymbol{\theta}$ is a real-valued random vector, certain components of $\boldsymbol{\theta}$ can be constrained to some bounded or half-bounded intervals by a suitable prior definition $p(\boldsymbol{\theta}|\mathbf{X}, \boldsymbol{\phi})$. In (2.5), $\boldsymbol{\phi}$ contains the hyperparameters associated with both the observation model $p(\mathbf{y}_i|\mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\phi})$ and the prior $p(\boldsymbol{\theta}|\mathbf{X}, \{\mathbf{U}_j\}_{j=n+1}^{m}, \boldsymbol{\phi})$, and these hyperparameters can be either discrete or continuous random variables. In case the type-II MAP estimate of $\boldsymbol{\phi}$ is used for predictions, the parameterization and the prior $p(\boldsymbol{\phi})$ are chosen so that $\boldsymbol{\phi}$ can be conveniently optimized using an approximation for the conditional marginal likelihood $p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\phi})$. Otherwise if the posterior uncertainty on $\boldsymbol{\phi}$ is approximated simultaneously within the approximate inference framework for $\boldsymbol{\theta}$, it is assumed that the hyperparameters are divided into L a priori independent groups according to $\boldsymbol{\phi} = \{\boldsymbol{\phi}_1, ..., \boldsymbol{\phi}_L\}$ and that the prior distribution of each group is some suitable member of the exponential family of distributions:

$$\begin{aligned} p(\boldsymbol{\phi}) &= \prod_{l=1}^{L} p(\boldsymbol{\phi}_l) = \prod_{l=1}^{L} Z(\boldsymbol{\lambda}_{0,l})^{-1} \exp\left(\boldsymbol{\lambda}_{0,l}^{\mathrm{T}} \mathbf{g}_l(\boldsymbol{\phi}_l)\right) \\ &= Z(\boldsymbol{\lambda}_0)^{-1} \exp\left(\boldsymbol{\lambda}_0^{\mathrm{T}} \mathbf{g}(\boldsymbol{\phi})\right), \end{aligned} \tag{2.7}$$

where $\boldsymbol{\lambda}_{0,l}$ are the natural parameters and $\mathbf{g}(\boldsymbol{\phi}_l)$ the sufficient statistics specific to the chosen prior distribution of group l in its canonical form, and $Z(\boldsymbol{\lambda}_{0,l}) = \int \exp\left(\boldsymbol{\lambda}_{0,l}^{\mathrm{T}} \mathbf{g}(\boldsymbol{\phi}_l)\right) d\boldsymbol{\phi}_l$ is the normalization factor (the logarithm of the normalizing factor $\log Z(\boldsymbol{\lambda}_{0,l})$ is also known as the log-partition function). The natural parameters and sufficient statistics of the combined prior $p(\boldsymbol{\phi})$ can be written as $\boldsymbol{\lambda}_0 = [\boldsymbol{\lambda}_{0,1}^{\mathrm{T}}, ..., \boldsymbol{\lambda}_{0,L}^{\mathrm{T}}]^{\mathrm{T}}$, $\mathbf{g}(\boldsymbol{\phi}) = [\mathbf{g}_1(\boldsymbol{\phi}_1)^{\mathrm{T}}, ..., \mathbf{g}_L(\boldsymbol{\phi}_L)^{\mathrm{T}}]^{\mathrm{T}}$, and the normalization factor is given by $\log Z(\boldsymbol{\lambda}_0) = \sum_l \log Z(\boldsymbol{\lambda}_{0,l})$.

Many commonly used predictive models can be written in the form of (2.5). For example, a generalized linear model with a sparsity-favoring Laplace prior can be recovered by setting $\mathbf{U}_i = \mathbf{x}_i$ for the likelihood terms,

and $\mathbf{U}_j = \mathbf{e}_j$ and $t_j(z_j, \phi) = \frac{1}{\phi}\exp\left(\frac{1}{\phi}|z_j|\right)$ for the prior factors, where $\mathbf{e}_j$ is the j:th unit vector in $\mathbb{R}^d$ so that $z_j = \theta_j$, . On the other hand, a typical Gaussian process model is obtained by defining only one prior factor with $\mathbf{U}_j = \mathbf{I}$ so that $p(\boldsymbol{\theta}|\mathbf{X}, \boldsymbol{\phi}) = t_j(\boldsymbol{\theta}, \boldsymbol{\phi}) = \mathcal{N}(\mathbf{m}(\mathbf{X}), \mathbf{K}(\mathbf{X}, \mathbf{X}))$, where $j = n + 1$, $\mathbf{m}(\mathbf{X})$ is the prior mean function, and $\mathbf{K}(\mathbf{X}, \mathbf{X})$ the prior covariance function that controls the smoothness properties of the latent function $\mathbf{f}(\mathbf{x})$ through hyperparameters $\boldsymbol{\phi}$. For the likelihood terms the i:th transformation $\mathbf{U}_i$ is chosen so that it collects all the latent function values $\mathbf{f}_i = \mathbf{f}(\mathbf{x}_i)$ associated with the input-output pair $\{\mathbf{x}_i, \mathbf{y}_i\}$ from the vector $\mathbf{f} = \boldsymbol{\theta}$ containing all the latent values.

In the following the dependence on the inputs $\mathbf{X}$ is omitted from the notation, because $\mathbf{X}$ is known and the approximate inference can be summarized using $\{\mathbf{U}_i\}_{i=1}^{n+m}$. Denoting each likelihood term with $t_i(\mathbf{U}_i^{\mathrm{T}}\boldsymbol{\theta}, \boldsymbol{\phi})$, the posterior distribution (2.5) can be written as

$$p(\boldsymbol{\theta}, \boldsymbol{\phi}, |\mathcal{D}) = Z^{-1} \prod_{i=1}^{n+m} t_i(\mathbf{U}_i^{\mathrm{T}}\boldsymbol{\theta}, \boldsymbol{\phi}_{\mathcal{A}_i}) \prod_{l=1}^{L} p(\phi_l), \tag{2.8}$$

where $\boldsymbol{\phi}_{\mathcal{A}_i} = \{\phi_l | l \in \mathcal{A}_i\}$ contains all the hyperparameters associated with the i:th term. For example, $\boldsymbol{\phi}_{\mathcal{A}_i}$ can contain the hyperparameters related to the likelihood terms ($i \leq n$) or the prior terms ($i > n$), respectively. The factors $t_i(\mathbf{U}_i^{\mathrm{T}}\boldsymbol{\theta}, \boldsymbol{\phi}_{\mathcal{A}_i})$ are from now referred to as site functions (or sites).

2.3 Expectation Propagation

This section presents a brief derivation and a summary of an EP algorithm suitable for approximate inference with the model structure defined in Section 2.2. In addition, alternative provably-convergent algorithms are discussed in Section 2.3.7.

2.3.1 Structure of the Posterior Approximation

EP is used to approximate the posterior distribution (2.8) with

$$p(\boldsymbol{\theta}, \boldsymbol{\phi}|\mathcal{D}) \approx Z_{\mathrm{EP}}^{-1} \prod_{i=1}^{n+m} \tilde{Z}_i \tilde{t}_{\boldsymbol{\theta},i}(\boldsymbol{\theta}) \tilde{t}_{\boldsymbol{\phi},i}(\boldsymbol{\phi}) p(\boldsymbol{\phi}) = q(\boldsymbol{\theta}) q(\boldsymbol{\phi}), \tag{2.9}$$

where each analytically intractable site function is approximated with a site approximation that can be factored between $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$:

$$t_i(\mathbf{U}_i^{\mathrm{T}}\boldsymbol{\theta}, \boldsymbol{\phi}_{\mathcal{A}_i}) \approx \tilde{t}_i(\boldsymbol{\theta}, \boldsymbol{\phi}) = \tilde{Z}_i \tilde{t}_{\boldsymbol{\theta},i}(\boldsymbol{\theta}) \tilde{t}_{\boldsymbol{\phi},i}(\boldsymbol{\phi}), \tag{2.10}$$

and $\tilde{Z}_i$ is a scalar scaling parameter, which is needed to form an EP approximation for the marginal likelihood of the observed data $Z = p(\mathbf{Y}|\mathbf{X}) \approx$

Z_{EP}. Independent posterior approximations are chosen for $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, because it results in computationally more efficient computations in Publication IV. This is a common assumption in many EP approaches for hierarchical models [see, e.g., Hernández-Lobato et al., 2008, Hernández-Lobato et al., 2011]. If the model under consideration permits feasible integrations for determining the moments of the tilted distributions (2.20), a fully-coupled approximation could be obtained, for example, by incorporating the hyperparameters into $\boldsymbol{\theta}$ using a suitable parameterization and adjusting $\mathbf{U}_i$ accordingly. Gaussian site approximations are assumed for $\boldsymbol{\theta}$:

$$\tilde{t}_{\boldsymbol{\theta},i}(\boldsymbol{\theta}) = \exp\left(-\frac{1}{2}\boldsymbol{\theta}^{\mathrm{T}}\tilde{\mathbf{Q}}_i\boldsymbol{\theta} + \tilde{\mathbf{h}}_i^{\mathrm{T}}\boldsymbol{\theta}\right), \tag{2.11}$$

where $\tilde{\mathbf{Q}}$ is a $d \times d$ site precision matrix, $\tilde{\mathbf{h}}$ a $d \times 1$ site location vector. For the hyperparameters $\boldsymbol{\phi}$, site term approximations conjugate with the prior (2.7) are chosen:

$$\tilde{t}_{\boldsymbol{\phi},i}(\boldsymbol{\phi}) = \exp\left(\tilde{\boldsymbol{\lambda}}_i^{\mathrm{T}}\mathbf{g}(\boldsymbol{\phi})\right) = \prod_{l=1}^{L}\tilde{t}_{\boldsymbol{\phi}_l,i}(\boldsymbol{\phi}_l) = \prod_{l=1}^{L}\exp\left(\tilde{\boldsymbol{\lambda}}_{i,l}^{\mathrm{T}}\mathbf{g}_l(\boldsymbol{\phi}_l)\right) \tag{2.12}$$

where $\mathbf{g}(\boldsymbol{\phi}) = [\mathbf{g}_1(\boldsymbol{\phi}_1)^{\mathrm{T}}, ..., \mathbf{g}_L(\boldsymbol{\phi}_L)^{\mathrm{T}}]^{\mathrm{T}}$ are the sufficient statistics of the prior (2.7), and $\tilde{\boldsymbol{\lambda}}_i = [\tilde{\boldsymbol{\lambda}}_{i,l}^{\mathrm{T}}, ..., \tilde{\boldsymbol{\lambda}}_{i,L}^{\mathrm{T}}]^{\mathrm{T}}$ is a vector of site parameters analogous to the natural parameters $\boldsymbol{\lambda}_0$ of the prior. If some of the site terms $t_i(\mathbf{U}_i^{\mathrm{T}}\boldsymbol{\theta}, \boldsymbol{\phi})$ are already in the factored form of the term approximations (2.10), no EP approximations are required for those terms and the parameters of the site approximations can be equated with the corresponding natural parameters of the exact sites.

Multiplying the site approximations (2.11) together according to (2.9) gives the following Gaussian approximation for $\boldsymbol{\theta}$:

$$q(\boldsymbol{\theta}) = Z(\mathbf{h}, \mathbf{Q})^{-1}\psi(\boldsymbol{\theta}, \mathbf{h}, \mathbf{Q}) = \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{2.13}$$

where the Gaussian distribution is written by defining the exponential term as

$$\psi(\boldsymbol{\theta}, \mathbf{h}, \mathbf{Q}) = \exp\left(-\frac{1}{2}\boldsymbol{\theta}^{\mathrm{T}}\mathbf{Q}\boldsymbol{\theta} + \mathbf{h}^{\mathrm{T}}\boldsymbol{\theta}\right), \tag{2.14}$$

and the normalization factor (or the partition function) as

$$\begin{aligned}\log Z(\mathbf{h}, \mathbf{Q}) &= \log\int \psi(\boldsymbol{\theta}, \mathbf{h}, \mathbf{Q})d\boldsymbol{\theta} \\ &= \frac{d}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{Q}| + \frac{1}{2}\mathbf{h}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{h}.\end{aligned} \tag{2.15}$$

The approximate mean vector can be written as $\boldsymbol{\mu} = \mathbf{Q}^{-1}\mathbf{h}$ and the covariance matrix as $\boldsymbol{\Sigma} = \mathbf{Q}^{-1}$ using the location vector $\mathbf{h}$ and the precision

matrix $\mathbf{Q}$:

$$\begin{aligned}\mathbf{h} &= \sum_{i=1}^{n+m} \tilde{\mathbf{h}}_i \\ \mathbf{Q} &= \sum_{i=1}^{n+m} \tilde{\mathbf{Q}}_i.\end{aligned} \tag{2.16}$$

Because the prior (2.7) can be factorized similarly to the site term approximations (2.12), a factorized posterior approximation is obtained for the hyperparameters:

$$q(\boldsymbol{\phi}) = \prod_{l=1}^{L} q(\boldsymbol{\phi}_l) = \prod_{l=1}^{L} Z(\boldsymbol{\lambda}_l)^{-1} \exp\left(\boldsymbol{\lambda}_l^{\mathrm{T}} \mathbf{g}_l(\boldsymbol{\phi}_l)\right), \tag{2.17}$$

where $Z(\boldsymbol{\lambda}_l) = \int \exp\left(\boldsymbol{\lambda}_l^{\mathrm{T}} \mathbf{g}_l(\boldsymbol{\phi}_l)\right) d\boldsymbol{\lambda}_l$ and the natural parameters are given by

$$\boldsymbol{\lambda}_l = \boldsymbol{\lambda}_{0,l} + \sum_{i=1}^{n+m} \tilde{\boldsymbol{\lambda}}_{i,l}. \tag{2.18}$$

2.3.2 General EP Algorithm

The standard EP algorithm [Minka, 2001b] updates the parameters of the site approximations and the posterior approximation $q(\boldsymbol{\theta}, \boldsymbol{\phi})$ sequentially. At each iteration, first a proportion η of the i:th site term is removed from the posterior approximation to obtain a cavity distribution:

$$q_{-i}(\boldsymbol{\theta}, \boldsymbol{\phi}) = q_{-i}(\boldsymbol{\theta}) q_{-i}(\boldsymbol{\phi}) \propto q(\boldsymbol{\theta}) q(\boldsymbol{\phi}) \tilde{t}_{\boldsymbol{\theta},i}(\boldsymbol{\theta})^{-\eta} \tilde{t}_{\boldsymbol{\phi},i}(\boldsymbol{\phi})^{-\eta}, \tag{2.19}$$

where $\eta \in (0, 1]$ is a fraction parameter that can be adjusted to implement fractional (or power) EP updates [Minka, 2004, 2005]. When $\eta = 1$ and $i \leq n$, the cavity distribution (2.19) can be thought of as a leave-one-out (LOO) posterior approximation where the contribution of the i:th likelihood term $p(\mathbf{y}_i|\mathbf{z}_i, \boldsymbol{\phi})$ is removed from $q(\boldsymbol{\theta}, \boldsymbol{\phi})$. Then, the i:th site approximation is replaced with the exact site term to form a tilted distribution

$$\hat{p}_i(\boldsymbol{\theta}, \boldsymbol{\phi}) = \hat{Z}_i^{-1} q_{-i}(\boldsymbol{\theta}, \boldsymbol{\phi}) t_i(\mathbf{U}_i^{\mathrm{T}} \boldsymbol{\theta}, \boldsymbol{\phi}_{\mathcal{A}_i})^{\eta}, \tag{2.20}$$

where $\hat{Z}_i = \int q_{-i}(\boldsymbol{\theta}, \boldsymbol{\phi}) t_i(\mathbf{U}_i^{\mathrm{T}} \boldsymbol{\theta}, \boldsymbol{\phi}_{\mathcal{A}_i})^{\eta} d\boldsymbol{\theta} d\boldsymbol{\phi}$ is a normalization factor, which in case $i \leq n$ can also be thought of as an approximation for the LOO predictive density of the excluded data point $\mathbf{y}_i$. The tilted distribution can be regarded as a more refined approximation to the true posterior distribution. Next, the algorithm attempts to match the approximate posterior distribution $q(\boldsymbol{\theta}, \boldsymbol{\phi})$ with $\hat{p}_i(\boldsymbol{\theta}, \boldsymbol{\phi})$ by finding a member of the chosen

approximate family $\hat{q}_i(\boldsymbol{\theta}, \boldsymbol{\phi})$ that satisfies

$$\hat{q}_i(\boldsymbol{\theta}, \boldsymbol{\phi}) = \arg\min_{q_i} \mathrm{KL}\left(\hat{p}_i(\boldsymbol{\theta}, \boldsymbol{\phi}) || q_i(\boldsymbol{\theta}, \boldsymbol{\phi})\right), \tag{2.21}$$

where KL denotes the Kullback-Leibler divergence and q_i can be factored as $q_i(\boldsymbol{\theta}, \boldsymbol{\phi}) = q_i(\boldsymbol{\theta}) q_i(\boldsymbol{\phi})$. Then, the parameters of the i:th site terms are updated so that the moments of $q(\boldsymbol{\theta}, \boldsymbol{\phi})$ match with $\hat{q}_i(\boldsymbol{\theta}, \boldsymbol{\phi})$:

$$\hat{q}_i(\boldsymbol{\theta}, \boldsymbol{\phi}) \equiv q(\boldsymbol{\theta}, \boldsymbol{\phi}) \propto q_{-i}(\boldsymbol{\theta}) q_{-i}(\boldsymbol{\phi}) \tilde{t}_{\boldsymbol{\theta},i}(\boldsymbol{\theta})^{\eta} \tilde{t}_{\boldsymbol{\phi},i}(\boldsymbol{\phi})^{\eta}. \tag{2.22}$$

Finally, the posterior approximation $q(\boldsymbol{\theta}, \boldsymbol{\phi})$ is updated according to the changes in the site parameters. These steps are repeated for all sites in some suitable order until convergence.

From now on, we refer to the previously described EP update scheme, where the posterior approximations $q(\boldsymbol{\theta})$ and $q(\boldsymbol{\phi})$ are refreshed after each of the $n+m$ site updates, as sequential EP. If the posterior approximations are updated only after new site parameter values have been determined for all the n likelihood sites or m prior sites, we refer to parallel EP [see, e.g., van Gerven et al., 2009].

When approximations $q(\boldsymbol{\theta})$ and $q(\boldsymbol{\phi})$ belong to the exponential family, the KL minimization step (2.21) is equal to matching the expected sufficient statistics of $q(\boldsymbol{\theta})$ and $q(\boldsymbol{\phi})$ with the corresponding marginal expectations with respect to $\hat{p}_i(\boldsymbol{\theta}, \boldsymbol{\phi})$ [Minka, 2001b, 2005, Seeger, 2005]. For the chosen approximate family, the KL divergence of (2.21) can be written as

$$\begin{aligned}
\mathrm{KL}\left(\hat{p}_i(\boldsymbol{\theta}, \boldsymbol{\phi}) || q_i(\boldsymbol{\theta}, \boldsymbol{\phi})\right) &= \int \hat{p}_i(\boldsymbol{\theta}, \boldsymbol{\phi}) \log \left(\frac{\hat{p}_i(\boldsymbol{\theta}, \boldsymbol{\phi})}{q_i(\boldsymbol{\theta}) \prod_{l=1}^{L} q_i(\boldsymbol{\phi}_l)} \right) d\boldsymbol{\theta} d\boldsymbol{\phi} \\
&= \frac{1}{2} \mathrm{Tr}\left(\mathrm{E}_{\hat{p}_i}\left(\boldsymbol{\theta}\boldsymbol{\theta}^{\mathrm{T}} \right) \mathbf{Q}_i \right) - \mathbf{h}_i^{\mathrm{T}} \mathrm{E}_{\hat{p}_i}\left(\boldsymbol{\theta}\right) + \log Z(\mathbf{h}_i, \mathbf{Q}_i) \\
&\quad - \sum_{l=1}^{L} \left(\boldsymbol{\lambda}_{i,l}^{\mathrm{T}} \mathrm{E}_{\hat{p}_i}\left(g(\boldsymbol{\phi}_l)\right) - \log Z(\boldsymbol{\lambda}_{i,l}) \right) + C
\end{aligned} \tag{2.23}$$

where $q_i(\boldsymbol{\theta}) = \mathcal{N}(\mathbf{Q}_i^{-1}\mathbf{h}_i, \mathbf{Q}_i^{-1})$, $q_i(\boldsymbol{\phi}_l) = Z(\boldsymbol{\lambda}_{i,l})^{-1} \exp\left(\boldsymbol{\lambda}_{i,l}^{\mathrm{T}} g(\boldsymbol{\phi}) \right)$ for $l = 1, ..., L$, $\mathrm{E}_{\hat{p}_i}$ denotes expectation with respect to $\hat{p}_i(\boldsymbol{\theta}, \boldsymbol{\phi})$, and C does not depend on $\boldsymbol{\lambda}_{i,l}$, $\mathbf{h}_i$, or $\mathbf{Q}_i$. Computing the derivatives of (2.23) with respect to $\mathbf{h}_i$, $\mathbf{Q}_i$, and $\boldsymbol{\lambda}_{i,l}$ using the well known result that with exponential families the expected sufficient statistics can be obtained by differentiating the log partition function (see equations (2.15) and (2.17)) with respect to the natural parameters:

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{h}_i} \log Z(\mathbf{h}_i, \mathbf{Q}_i) &= \mathrm{E}_{q_i}(\boldsymbol{\theta}) = \mathbf{Q}_i^{-1}\mathbf{h}_i \\
\frac{\partial}{\partial \mathbf{Q}_i} \log Z(\mathbf{h}_i, \mathbf{Q}_i) &= -\frac{1}{2} \mathrm{E}_{q_i}(\boldsymbol{\theta}\boldsymbol{\theta}^{\mathrm{T}}) = -\frac{1}{2}\mathbf{Q}_i^{-1} \\
\frac{\partial}{\partial \boldsymbol{\lambda}_{i,l}} \log Z(\boldsymbol{\lambda}_{i,l}) &= \mathrm{E}_{q_i}(\mathbf{g}(\boldsymbol{\phi}_l)),
\end{aligned} \tag{2.24}$$

and setting the derivatives to zero gives the following conditions for the minimizing distribution $\hat{q}_i(\boldsymbol{\theta}, \boldsymbol{\phi}) = \hat{q}_i(\boldsymbol{\theta}) \prod_{l=1}^{L} \hat{q}_i(\boldsymbol{\phi}_l)$:

$$\begin{aligned} \mathrm{E}_{\hat{p}_i}(\boldsymbol{\theta}) &= \mathrm{E}_{\hat{q}_i}(\boldsymbol{\theta}) = \hat{\boldsymbol{\mu}}_i \\ \mathrm{E}_{\hat{p}_i}\left(\boldsymbol{\theta}\boldsymbol{\theta}^{\mathrm{T}}\right) &= \mathrm{E}_{\hat{q}_i}\left(\boldsymbol{\theta}\boldsymbol{\theta}^{\mathrm{T}}\right) = \hat{\boldsymbol{\Sigma}}_i + \hat{\boldsymbol{\mu}}_i \hat{\boldsymbol{\mu}}_i^{\mathrm{T}} \\ \mathrm{E}_{\hat{p}_i}(\mathbf{g}_l(\boldsymbol{\phi}_l)) &= \mathrm{E}_{\hat{q}_i}(\mathbf{g}_l(\boldsymbol{\phi}_l)) = \hat{\mathbf{s}}_i(\hat{\boldsymbol{\lambda}}_{i,l}) \quad l = 1, ..., L. \end{aligned} \tag{2.25}$$

where $\hat{q}_i(\boldsymbol{\theta}) = \mathcal{N}(\hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\Sigma}}_i)$ with $\hat{\boldsymbol{\mu}}_i = \hat{\mathbf{Q}}_i^{-1}\hat{\mathbf{h}}_i$ and $\hat{\boldsymbol{\Sigma}}_i = \hat{\mathbf{Q}}_i^{-1}\hat{\mathbf{h}}_i$, and $\hat{q}_i(\boldsymbol{\phi}_l) = Z(\hat{\boldsymbol{\lambda}}_{i,l})^{-1} \exp\left(\hat{\boldsymbol{\lambda}}_{i,l}^{\mathrm{T}} g(\boldsymbol{\phi})\right)$. Equation (2.25) shows that, because of the factorized approximate family, only the expected sufficient statistics of $q(\boldsymbol{\theta})$ and $q(\boldsymbol{\phi}_1), ..., q(\boldsymbol{\phi}_L)$ have to be made equal with the corresponding expectations with respect to the marginal tilted distributions $\hat{p}_i(\boldsymbol{\theta})$ and $\hat{p}_i(\boldsymbol{\phi}_1), ..., \hat{p}_i(\boldsymbol{\phi}_L)$. For updating the Gaussian approximation $q(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, this requires determining the mean $\hat{\boldsymbol{\mu}}_i$ and covariance $\hat{\boldsymbol{\Sigma}}_i$ of the marginal tilted distribution $\hat{p}_i(\boldsymbol{\theta})$, and updating the site parameters $\tilde{\mathbf{h}}_i$ and $\tilde{\mathbf{Q}}_i$ according to (2.22) so that $\hat{\boldsymbol{\mu}}_i$ and $\hat{\boldsymbol{\Sigma}}_i$ are consistent with $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, respectively. For the hyperparameter approximations $q(\boldsymbol{\phi}_l)$, equation (2.25) requires determining the expected sufficient statistics $\hat{\mathbf{s}}_i = \mathrm{E}_{\hat{p}_i}(\mathbf{g}_l(\boldsymbol{\phi}_l))$ with respect to the marginal tilted distributions $\hat{p}_i(\boldsymbol{\phi}_l)$ and finding a $\hat{\boldsymbol{\lambda}}_{i,l}$ such that $\mathrm{E}_{\hat{q}_i}(\mathbf{g}_l(\boldsymbol{\phi}_l)) = \hat{\mathbf{s}}_i$. Provided that a suitable minimal representation is chosen for the natural parameterization of $q(\boldsymbol{\phi}_l)$ (the same as with $\hat{q}_i(\boldsymbol{\phi}_l)$), there exists a bijective mapping between the natural parameters $\hat{\boldsymbol{\lambda}}_{i,l}$ and the moment parameters $\hat{\mathbf{s}}_i$ from which $\hat{\boldsymbol{\lambda}}_{i,l}$ can be solved [Seeger, 2005].[2] After determining $\hat{\boldsymbol{\lambda}}_{i,l}$ the site parameters $\tilde{\boldsymbol{\lambda}}_{i,l}$ are updated so that $q(\boldsymbol{\phi}_l)$ is consistent with $\hat{q}_i(\boldsymbol{\phi}_l)$ according to (2.22), which is equivalent to $\hat{\boldsymbol{\lambda}}_{i,l} = \boldsymbol{\lambda}_{-i,l} + \eta\tilde{\boldsymbol{\lambda}}_{i,l}$ for an approximation belonging to the exponential family.

Because the i:th site term $t_i(\mathbf{U}_i^{\mathrm{T}}\boldsymbol{\theta}, \boldsymbol{\phi}_{\mathcal{A}_i})$ depends only on a subset $\boldsymbol{\phi}_{\mathcal{A}_i} = \{\boldsymbol{\phi}_l | l \in \mathcal{A}_i\}$ of the hyperparameters and the approximate posterior distribution (2.17) can be factored between $\boldsymbol{\phi}_1, ..., \boldsymbol{\phi}_L$, the tilted distribution (2.20) can also be factored as $\hat{p}_i(\boldsymbol{\theta}, \boldsymbol{\phi}) = \hat{p}_i(\boldsymbol{\theta}, \boldsymbol{\phi}_{\mathcal{A}_i}) \prod_{l \notin \mathcal{A}_i} q_{-i}(\boldsymbol{\phi}_l)$, where

$$\hat{p}_i(\boldsymbol{\theta}, \boldsymbol{\phi}_{\mathcal{A}_i}) = \hat{Z}_i^{-1} t_i(\mathbf{U}_i^{\mathrm{T}}\boldsymbol{\theta}, \boldsymbol{\phi}_{\mathcal{A}_i})^{\eta} q_{-i}(\boldsymbol{\theta}) \prod_{l \in \mathcal{A}_i} q_{-i}(\boldsymbol{\phi}_l). \tag{2.26}$$

Consequently, for $l \notin \mathcal{A}_i$, the marginal expectations of $\mathbf{g}_l(\boldsymbol{\phi}_l)$ with respect to the tilted distribution reduce to the corresponding cavity expectations: $\mathrm{E}_{\hat{p}_i}(\mathbf{g}_l(\boldsymbol{\phi}_l)) = \mathrm{E}_{q_{-i}}(\mathbf{g}_l(\boldsymbol{\phi}_l))$ for $l \notin \mathcal{A}_i$. It follows that for $l \notin \mathcal{A}_i$ no site parameter updates are required for the i:th site term, because $\hat{q}_l(\boldsymbol{\phi}_l) =$

[2] Results from $\nabla_{\boldsymbol{\lambda}_l} \log Z(\boldsymbol{\lambda}_l) = \mathrm{E}(\mathbf{g}_l(\boldsymbol{\phi}_l))$ and the convexity of the log partition function: $\nabla^2_{\boldsymbol{\lambda}_l} \log Z(\boldsymbol{\lambda}_l) = \mathrm{Var}_q(\mathbf{g}_l(\boldsymbol{\phi}_l)) > 0$.

$q_{-i}(\boldsymbol{\phi}_l)$ and the EP update equation (2.22) results in zero site parameters $\tilde{\boldsymbol{\lambda}}_{i,l} = \mathbf{0}$, as expected.

2.3.3 EP with Gaussian Approximations

The practical feasibility of EP depends on the structure of the site terms and the choice of the approximating family done in Section 2.3.1. The main requirements are 1) that the integrations with respect to $\hat{p}_i(\boldsymbol{\theta}, \boldsymbol{\phi}_{\mathcal{A}_i})$ in equation (2.25) can be carried out efficiently and 2) that the chosen approximate family $q(\boldsymbol{\theta}, \boldsymbol{\phi})$ is closed under the marginalizations required to determine the approximations for linear transforms $\mathbf{z}_i = \mathbf{U}_i^{\mathrm{T}}\boldsymbol{\theta}$ and subsets $\boldsymbol{\phi}_{\mathcal{A}_i}$, and that the associated computations are tractable [see also Seeger, 2005]. The first condition requires that integrations with respect to $\hat{p}_i(\boldsymbol{\theta}, \boldsymbol{\phi}_{\mathcal{A}_i})$ can be done analytically over $\boldsymbol{\theta}$ and/or $\boldsymbol{\phi}_{\mathcal{A}_i}$, and that the remaining non-analytical integrations are so low-dimensional that numerical quadrature methods can be utilized efficiently. The second condition is met because the approximation can be factored between $\boldsymbol{\theta}$ and $\boldsymbol{\phi}_1, ..., \boldsymbol{\phi}_L$, and the Gaussian approximation for $\boldsymbol{\theta}$ is closed under linear transformations.

With a Gaussian approximation for $\boldsymbol{\theta}$ additional computational savings and lower-dimensional site parameters are obtained, when the site terms $t_i(\mathbf{U}^{\mathrm{T}}\boldsymbol{\theta}, \boldsymbol{\phi}_{\mathcal{A}_i})$ depend on $\boldsymbol{\theta}$ only through lower-dimensional random variables $\mathbf{z}_i$ resulting from a linear transformation $\mathbf{z}_i = \mathbf{U}_i^{\mathrm{T}}\boldsymbol{\theta}$ [see, e.g., Seeger, 2005, Cseke and Heskes, 2011]. This property has been utilized with linear classifiers and other linear models with general likelihoods [Minka, 2001a, Qi et al., 2004, Seeger et al., 2007, Hernández-Lobato et al., 2008, van Gerven et al., 2009, 2010], and a good summary of the Gaussian EP algorithm is presented by Cseke and Heskes [appendix C, 2011]. With GPs and other latent Gaussian models, where each site depends on a subset of $\boldsymbol{\theta}$, the transformation is defined so that $\mathbf{U}_i$ picks the desired components and the EP algorithm reduces to matching and propagating the associated marginal moments (see. e.g, Rasmussen and Williams [2006], Seeger and Jordan [2004], Seeger et al. [2006], Girolami and Zhong [2007] and Publications II and III).

In the following a short description is given on the EP updates resulting from equations (2.19) – (2.22) with Gaussian approximations $q(\boldsymbol{\theta})$, and the complete EP algorithm with general exponential family approximations for $\boldsymbol{\phi}$ is presented in Section 2.3.4. Equation (2.19) results in a Gaussian cavity distribution for $\boldsymbol{\theta}$, denoted by $q_{-i}(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}_{-i}, \boldsymbol{\Sigma}_{-i})$, because

multiplying two members of the same exponential family results in an unnormalized member of same family (the same applies for $q(\boldsymbol{\phi}_l)$) [Minka, 2005, Seeger, 2005, see, e.g,]. The normalization term of the marginal tilted distribution (2.26) associated with the i:th site can be written as

$$\begin{aligned}\hat{Z}_i &= \int t_i(\mathbf{U}_i^{\mathrm{T}}\boldsymbol{\theta}, \boldsymbol{\phi}_{\mathcal{A}_i})^{\eta}\mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}_{-i}, \boldsymbol{\Sigma}_{-i}) \prod_{l\in\mathcal{A}_i} q_{-i}(\boldsymbol{\phi}_l)d\boldsymbol{\theta}d\boldsymbol{\phi}_{\mathcal{A}_i} \\ &= \int t_i(\mathbf{z}_i, \boldsymbol{\phi}_{\mathcal{A}_i})^{\eta}\mathcal{N}(\mathbf{z}_i|\mathbf{m}_{-i}, \mathbf{V}_{-i}) \prod_{l\in\mathcal{A}_i} q_{-i}(\boldsymbol{\phi}_l)d\mathbf{z}_i d\boldsymbol{\phi}_{\mathcal{A}_i} \qquad (2.27)\end{aligned}$$

where $\mathbf{m}_{-i} = \mathbf{U}_i^{\mathrm{T}}\boldsymbol{\mu}_{-i}$, and $\mathbf{V}_{-i} = \mathbf{U}_i^{\mathrm{T}}\boldsymbol{\Sigma}_{-i}\mathbf{U}_i$ can be interpreted as the mean and covariance of the cavity distribution of $\mathbf{z}_i$ induced by $q_{-i}(\boldsymbol{\theta})$ and it is denoted by $q_{-i}(\mathbf{z}_i) = \mathcal{N}(\mathbf{z}_i|\mathbf{m}_{-i}, \mathbf{V}_{-i})$. Differentiating the both integrals in (2.27) first once and then twice with respect to $\boldsymbol{\mu}_{-i}$, and equating the both results gives[3]

$$\begin{aligned}\hat{\boldsymbol{\Sigma}}_i^{-1} &= \boldsymbol{\Sigma}_{-i}^{-1} + \mathbf{U}_i\tilde{\mathbf{T}}_i\mathbf{U}_i^{\mathrm{T}} \\ \hat{\boldsymbol{\Sigma}}_i^{-1}\hat{\boldsymbol{\mu}}_i &= \boldsymbol{\Sigma}_{-i}^{-1}\boldsymbol{\mu}_{-i} + \mathbf{U}_i\tilde{\boldsymbol{\nu}}_i, \qquad (2.28)\end{aligned}$$

where $\hat{\boldsymbol{\mu}}_i$ and $\hat{\boldsymbol{\Sigma}}_i$ are the mean and covariance matrix of the marginal $\hat{p}_i(\boldsymbol{\theta})$ defined in (2.25), $\hat{\mathbf{m}}_i = \mathbf{U}_i^{\mathrm{T}}\hat{\boldsymbol{\mu}}_i$ and $\hat{\mathbf{V}}_i = \mathbf{U}_i^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_i\mathbf{U}_i$ are the corresponding moments of $\hat{p}_i(\mathbf{z}_i) = \hat{Z}_i^{-1}t_i(\mathbf{z}_i, \boldsymbol{\phi}_{\mathcal{A}_i})^{\eta}\mathcal{N}(\mathbf{z}_i|\mathbf{m}_{-i}, \mathbf{V}_{-i})\prod_{l\in\mathcal{A}_i} q_{-i}(\boldsymbol{\phi}_l)$, and the lower dimensional location and precision contributions are given by $\tilde{\boldsymbol{\nu}}_i = \hat{\mathbf{V}}_i^{-1}\hat{\mathbf{m}}_i - \mathbf{V}_{-i}\mathbf{m}_{-i}$ and $\tilde{\mathbf{T}}_i = \hat{\mathbf{V}}_i^{-1} - \mathbf{V}_{-i}^{-1}$.

Equation (2.28) shows that the i:th site term's contributions to the natural parameters of $\hat{q}(\boldsymbol{\theta})$ defined by (2.22) and (2.25) can be determined by computing only the moments of the lower dimensional cavity and tilted distributions related to $\mathbf{z}_i$. Furthermore, using the result (2.28) with (2.25), the moment consistency condition (2.22), and the original definition of $q(\boldsymbol{\theta})$ in (2.13), the parameters of the site approximation (2.11) resulting from the EP update can be written in a more economical form as

$$\begin{aligned}\tilde{\mathbf{Q}}_i &= \mathbf{U}_i\tilde{\mathbf{T}}_i\mathbf{U}_i^{\mathrm{T}} \\ \tilde{\mathbf{h}}_i &= \mathbf{U}_i\tilde{\boldsymbol{\nu}}_i, \qquad (2.29)\end{aligned}$$

where $\tilde{\mathbf{T}}_i = \eta^{-1}(\hat{\mathbf{V}}_i^{-1} - \mathbf{V}_{-i}^{-1})$ and $\tilde{\boldsymbol{\nu}}_i = \eta^{-1}(\hat{\mathbf{V}}_i^{-1}\hat{\mathbf{m}}_i - \mathbf{V}_{-i}\mathbf{m}_{-i})$, that is, the new site parameters are determined by the tilted and cavity moments of $\mathbf{z}_i$. Furthermore, using the definition $\mathbf{m}_{-i} = \mathbf{U}_i^{\mathrm{T}}\boldsymbol{\mu}_{-i}$ and $\mathbf{V}_{-i} = \mathbf{U}_i^{\mathrm{T}}\boldsymbol{\Sigma}_{-i}\mathbf{U}_i$ with equation (2.19), shows that the cavity moments of $\mathbf{z}_i$ can be computed using only rank-d_i matrix computations, where d_i is the number of

[3] An analogous result can be derived also by differentiating with respect to the natural parameters of $q_{-i}(\boldsymbol{\theta})$.

components in $\mathbf{z}_i$. Put together, a Gaussian approximate family with site terms dependent on lower-dimensional linearly transformed random variables $\mathbf{z}_i$, results in computational more economical site approximations and EP updates. The resulting EP algorithm for updating the approximations $q(\boldsymbol{\theta})$ simultaneously with the hyperparameter approximations $q(\boldsymbol{\phi}_l)$ is summarized in Section 2.3.4.

2.3.4 Algorithm Description

With the chosen model structure and approximate family, the EP algorithm can be implemented as follows. First, initialize the site parameters $\tilde{\boldsymbol{\nu}}_i$, $\tilde{\mathbf{T}}_i$, and $\tilde{\boldsymbol{\lambda}}_i$, together with the approximations $q(\boldsymbol{\theta})$, $q(\boldsymbol{\phi}_l)$, $l = 1, ..., L$. Then iterate the following steps at a chosen order for each $i = 1, ..., n + m$:

1. Determine the parameters of the cavity distribution (2.19). Compute first the mean $\mathbf{m}_i$ and the covariance $\mathbf{V}_i$ of the approximate marginal distribution of the transformed variable $\mathbf{z}_i = \mathbf{U}_i^{\mathrm{T}}\boldsymbol{\mu}$: $q(\mathbf{z}_i) = \mathcal{N}(\mathbf{m}_i, \mathbf{V}_i)$, where $\mathbf{m}_i = \mathbf{U}_i^{\mathrm{T}}\boldsymbol{\mu}$ and $\mathbf{V}_i = \mathbf{U}_i^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{U}_i$. The cavity distribution for the transformed variable $\mathbf{z}_i$ is then given by $q_{-i}(\mathbf{z}_i) = \mathcal{N}(\mathbf{m}_{-i}, \mathbf{V}_{-i})$, where

$$\begin{aligned} \mathbf{m}_{-i} &= \mathbf{V}_{-i}(\mathbf{V}_i^{-1}\mathbf{m}_i - \eta\tilde{\boldsymbol{\nu}}_i) \\ \mathbf{V}_{-i} &= \left(\mathbf{V}_i^{-1} - \eta\tilde{\mathbf{T}}_i\right)^{-1}. \end{aligned} \tag{2.30}$$

   The cavity distributions for the hyperparameters associated with the i:th site, $\boldsymbol{\phi}_{\mathcal{A}_i} = \{\boldsymbol{\phi}_l | l \in \mathcal{A}_i\}$, can be written as

$$q_{-i}(\boldsymbol{\phi}_l) = Z(\boldsymbol{\lambda}_{-i,l})^{-1} \exp\left(\boldsymbol{\lambda}_{-i,l}^{\mathrm{T}}\mathbf{g}_l(\boldsymbol{\phi}_l)\right) \quad l \in \mathcal{A}_i, \tag{2.31}$$

   where the natural parameters can be computed from the parameters of the approximate posterior (2.17) and the site approximations (2.12) as

$$\boldsymbol{\lambda}_{-i,l} = \boldsymbol{\lambda}_l - \eta\tilde{\boldsymbol{\lambda}}_{i,l}. \tag{2.32}$$

2. Compute the sufficient statistics with respect to the i:th tilted distribution (2.21). For the the transformed variable $\mathbf{z}_i$, determine (or approximate) the marginal mean $\hat{\mathbf{m}}_i = \mathrm{E}_{\hat{p}_i}(\mathbf{z}_i)$ and covariance $\hat{\mathbf{V}}_i = \mathrm{E}_{\hat{p}_i}(\mathbf{z}_i\mathbf{z}_i^{\mathrm{T}}) - \hat{\mathbf{m}}_i\hat{\mathbf{m}}_i^{\mathrm{T}}$ with respect to the the marginal tilted distribution:

$$\hat{p}_i(\mathbf{z}_i, \boldsymbol{\phi}_{\mathcal{A}_i}) = \hat{Z}_i^{-1} t_i(\mathbf{z}_i, \boldsymbol{\phi}_{\mathcal{A}_i})^{\eta} q_{-i}(\mathbf{z}_i) \prod_{l \in \mathcal{A}_i} q_{-i}(\boldsymbol{\phi}_l). \tag{2.33}$$

   For the hyperparameters $\boldsymbol{\phi}_l$, $l \in \mathcal{A}_i$, determine the parameters $\hat{\boldsymbol{\lambda}}_{i,l}$ of a

member of the chosen approximate family (2.17), denoted by

$$\hat{q}_i(\boldsymbol{\phi}_l) = Z(\hat{\boldsymbol{\lambda}}_{i,l})^{-1} \exp\left(\hat{\boldsymbol{\lambda}}_{i,l}^{\mathrm{T}} \mathbf{g}_l(\boldsymbol{\phi}_l)\right), \tag{2.34}$$

that satisfies $\mathrm{E}_{\hat{p}_i}(\mathbf{g}_l(\boldsymbol{\phi}_l)) = \mathrm{E}_{\hat{q}_i}(\mathbf{g}_l(\boldsymbol{\phi}_l))$.

If the normalization term $\hat{Z}_i$ can be computed analytically, the sufficient statistics can be determined by differentiating $\log \hat{Z}_i$ with respect to the cavity parameters: e.g., for the hyperparameters $\boldsymbol{\phi}_l$, the expected sufficient statistics can be computed as:

$$\mathrm{E}_{\hat{p}_i}\left(\mathbf{g}_l(\phi_l)\right) = \nabla_{\boldsymbol{\lambda}_{-i,l}} \log \hat{Z}_i + \mathrm{E}_{q_{-i}}(\mathbf{g}_l(\phi_l)). \tag{2.35}$$

Otherwise, the necessary moments have to approximated with a suitable method such as numerical quadrature integration.

3. Update the site parameters according to (2.22) by damping the updates with $\delta \in (0, 1]$. For the site approximations $\tilde{t}_{\boldsymbol{\theta},i}(\boldsymbol{\theta})$ this results in the following updates:

$$\begin{aligned} \tilde{\mathbf{T}}_i^{\text{new}} &= (1-\delta)\tilde{\mathbf{T}}_i + \delta\eta^{-1}(\hat{\mathbf{V}}_i^{-1} - \mathbf{V}_{-i}^{-1}) \\ &= \tilde{\mathbf{T}}_i + \delta\eta^{-1}(\hat{\mathbf{V}}_i^{-1} - \mathbf{V}_i^{-1}) \end{aligned} \tag{2.36}$$

$$\begin{aligned} \tilde{\boldsymbol{\nu}}_i^{\text{new}} &= (1-\delta)\tilde{\boldsymbol{\nu}}_i + \delta\eta^{-1}(\hat{\mathbf{V}}_i^{-1}\hat{\mathbf{m}}_i - \mathbf{V}_{-i}^{-1}\mathbf{m}_{-i}) \\ &= \tilde{\boldsymbol{\nu}}_i + \delta\eta^{-1}(\hat{\mathbf{V}}_i^{-1}\hat{\mathbf{m}}_i - \mathbf{V}_i^{-1}\mathbf{m}_i). \end{aligned} \tag{2.37}$$

For the site approximations $\tilde{t}_{\boldsymbol{\phi}_l,i}(\boldsymbol{\phi}_l)$ this results in the following updates:

$$\tilde{\boldsymbol{\lambda}}_{i,l}^{\text{new}} = (1-\delta)\tilde{\boldsymbol{\lambda}}_{i,l} + \delta\eta^{-1}(\hat{\boldsymbol{\lambda}}_{i,l}^{-1} - \boldsymbol{\lambda}_{-i,l}) = \tilde{\boldsymbol{\lambda}}_{i,l} + \delta\eta^{-1}(\hat{\boldsymbol{\lambda}}_{i,l}^{-1} - \boldsymbol{\lambda}_l). \tag{2.38}$$

4. If sequential EP is used, apply rank-one posterior updates on $q(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and update $q(\boldsymbol{\phi}_l) = Z(\boldsymbol{\lambda}_l)^{-1} \exp(\boldsymbol{\lambda}_l^{\mathrm{T}} g_l(\boldsymbol{\phi}_l))$ for $l \in \mathcal{A}_i$.

2.3.5 Fractional Updates and Damping

The EP update procedure (2.19)– (2.22) with $\eta \neq 1$ can be interpreted as an iterative approach for minimizing a family of divergence measures called α-divergence parameterized with $\alpha = \eta$ [Minka, 2005]. The α-divergence between an exact posterior $p(\boldsymbol{\theta})$ and an approximation $q(\boldsymbol{\theta})$, denoted by $D_\alpha\left(p(\boldsymbol{\theta})|q(\boldsymbol{\theta})\right)$, can be minimized by guessing an initial $q(\boldsymbol{\theta})$

and updating it repeatedly according to

$$\begin{aligned} \hat{q}(\boldsymbol{\theta}) &= \text{proj}\left[p(\boldsymbol{\theta})^{\alpha} q(\boldsymbol{\theta})^{1-\alpha}\right] \\ q(\boldsymbol{\theta})^{\text{new}} &= \hat{q}(\boldsymbol{\theta})^{\delta} q(\boldsymbol{\theta})^{1-\delta} \end{aligned} \tag{2.39}$$

where $\delta \in (0, 1]$ is a damping factor and the proj-operation denotes the minimization of $\mathrm{KL}\left(p(\boldsymbol{\theta})^{\alpha} q(\boldsymbol{\theta})^{1-\alpha} | q(\boldsymbol{\theta})\right)$ similarly to (2.23) – (2.25) [Minka, 2005]. Here the KL-divergence minimization is done so that also the normalization constants of $q(\boldsymbol{\theta})$ and $p(\boldsymbol{\theta})$ are matched. The scheme is not guaranteed to converge but with a sufficient amount of damping it usually converges similarly to fractional EP [see, e.g., Seeger, 2008]. The local KL minimization of the fractional EP update procedure (2.19)– (2.22) can be written in the same form as (2.39):

$$\begin{aligned} \hat{q}_i(\boldsymbol{\theta}, \boldsymbol{\phi}) = \tilde{t}_i(\boldsymbol{\theta}, \boldsymbol{\phi})^{\text{new}} q_{-i}(\boldsymbol{\theta}, \boldsymbol{\phi}) &= \text{proj}\left[\hat{p}_i(\boldsymbol{\theta}, \boldsymbol{\phi})^{\alpha} q_i(\boldsymbol{\theta}, \boldsymbol{\phi})^{1-\alpha}\right] \\ &= \text{proj}\left[t_i(\mathbf{U}_i^{\mathrm{T}} \boldsymbol{\theta}, \boldsymbol{\phi}_{\mathcal{A}_i})^{\alpha} \tilde{t}_i(\boldsymbol{\theta}, \boldsymbol{\phi})^{1-\alpha} q_{-i}(\boldsymbol{\theta}, \boldsymbol{\phi})\right], \end{aligned} \tag{2.40}$$

where $q_{-i}(\boldsymbol{\theta}, \boldsymbol{\phi}) \propto q(\boldsymbol{\theta}, \boldsymbol{\phi}) \tilde{t}_i(\boldsymbol{\theta}, \boldsymbol{\phi})^{-1}$, $\tilde{t}_i(\boldsymbol{\theta}, \boldsymbol{\phi}) = \tilde{Z}_i \tilde{t}_{\boldsymbol{\theta},i}(\boldsymbol{\theta}) \tilde{t}_{\boldsymbol{\phi},i}(\boldsymbol{\phi})$ is the site approximation (2.10), $\hat{p}_i(\boldsymbol{\theta}, \boldsymbol{\phi}) = q_{-i}(\boldsymbol{\theta}, \boldsymbol{\phi}) t_i(\mathbf{U}_i^{\mathrm{T}} \boldsymbol{\theta}, \boldsymbol{\phi}_{\mathcal{A}_i})$ is an unnormalized tilted distribution, and $q_i(\boldsymbol{\theta}, \boldsymbol{\phi}) = q_{-i}(\boldsymbol{\theta}, \boldsymbol{\phi}) \tilde{t}_i(\boldsymbol{\theta}, \boldsymbol{\phi})$ an unnormalized approximation. Equation (2.40) is equivalent to (2.21) with $\eta = \alpha$, because $\tilde{t}_i(\boldsymbol{\theta}, \boldsymbol{\phi})^{1-\alpha} q_{-i}(\boldsymbol{\theta}, \boldsymbol{\phi}) \propto q(\boldsymbol{\theta}, \boldsymbol{\phi}) \tilde{t}_i(\boldsymbol{\theta}, \boldsymbol{\phi})^{-\alpha}$ is equivalent with the definition of the cavity distribution (2.19), and choosing $\delta = \eta$ results in the moment consistency condition (2.22).

The standard EP based on minimizing $\mathrm{KL}(\hat{p}_i(\boldsymbol{\theta}, \boldsymbol{\phi}) | q_i(\boldsymbol{\theta}, \boldsymbol{\phi}))$ is obtained by setting the fraction parameter to $\eta = 1$ in equations (2.19)– (2.22) whereas choosing a smaller value produces a slightly different approximation that puts less emphasis on preserving all the nonzero probability mass of the tilted distributions [Minka, 2005]. Choosing $\eta < 1$ tries to represent the uncertainty resulting from possible multiple modes of $\hat{p}_i(\boldsymbol{\theta}, \boldsymbol{\phi})$ but ignores modes far away from the main probability mass, which results in a tendency to underestimate variances. The limit $\eta \to 0$ can be interpreted as minimization of the reverse KL, $\mathrm{KL}(q_i(\boldsymbol{\theta}, \boldsymbol{\phi}) | \hat{p}_i(\boldsymbol{\theta}, \boldsymbol{\phi}))$, which is the standard assumption in various VB approaches discussed in Section 2.4. Larger values of η put more emphasis on preserving the overall uncertainty in $\hat{p}_i(\boldsymbol{\theta}, \boldsymbol{\phi})$ but in multimodal cases this can lead to very large tilted variances. Depending on the application, this may not be the best choice, if, e.g., accurate predictions are obtained only using one of the modes (see Minka [2005] and Publication II). In practice, decreasing η can alleviate convergence problems resulting from possible multimodalities and also

improve the overall numerical stability of the algorithm, because part of the precision of each site approximation is left in the cavity distribution at each iteration (Minka [2005], Seeger [2008], Publication II).

There is no theoretical convergence guarantee for the standard EP algorithm but damping the site parameter updates can help to achieve convergence in harder problems [Minka and Lafferty, 2002, Heskes and Zoeter, 2002]. In damping, the natural site parameters are updated to a convex combination of the old values and the new values resulting from (2.22) adjusted by the damping factor $\delta \in (0, 1]$ (compare with equations (2.37) – (2.34)). This is also equivalent with (2.39). The convergence problems are usually seen as increasing oscillations over iterations in the site parameter values and they may occur, for example, if there are inaccuracies in the tilted moment evaluations, or if the approximate distribution is not a suitable proxy for the true posterior, for example, due to multimodalities (for visualizations, see Publications II and III). In the experiments of Publications II and IV, parallel update scheme was found to require larger amount of damping, which can be explained by the fact that the local minimization (2.22) are derived from the sequential scheme.

2.3.6 The Marginal Likelihood Approximation

A numerically robust implementation for evaluating the EP marginal likelihood approximation with Gaussian approximations can be done, e.g., by following Cseke and Heskes [appendix C, 2011]. In the following an analogous expression is summarized for the chosen approximate family. An EP approximation for the marginal likelihood is given by

$$Z = \int \prod_{i=1}^{n+m} t_i(\mathbf{U}_i^{\mathrm{T}}\boldsymbol{\theta}, \boldsymbol{\phi}_{\mathcal{A}_i}) p(\boldsymbol{\phi}) d\boldsymbol{\theta} d\boldsymbol{\phi} \approx \int \prod_i \tilde{t}_i(\boldsymbol{\theta}, \boldsymbol{\phi}) p(\boldsymbol{\phi}) d\boldsymbol{\theta} d\boldsymbol{\phi}, \tag{2.41}$$

where the exact terms are simply replaced by their approximations. Taking into account the factorized form of the site approximations, $\tilde{t}_i(\boldsymbol{\theta}, \boldsymbol{\phi}) = \tilde{Z}_i \tilde{t}_{\boldsymbol{\theta},i}(\boldsymbol{\theta}) \tilde{t}_{\boldsymbol{\phi},i}(\boldsymbol{\phi})$, and definitions (2.11) and (2.12), the log marginal likelihood can be written as

$$\begin{aligned} \log Z_{\mathrm{EP}} &= \sum_i \log \tilde{Z}_i + \log \int \prod_i \tilde{t}_{\boldsymbol{\theta},i}(\boldsymbol{\theta}) d\boldsymbol{\theta} + \sum_{l=1}^{L} \log \int \prod_{i|l \in \mathcal{A}_i} \tilde{t}_{\boldsymbol{\phi}_l,i}(\boldsymbol{\phi}) p(\boldsymbol{\phi}_l) d\boldsymbol{\phi}_l \\ &= \sum_i \log \tilde{Z}_i + \log Z(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + \sum_{l=1}^{L} \left(\log Z(\boldsymbol{\lambda}_l) - \log Z(\boldsymbol{\lambda}_{0,l}) \right), \end{aligned} \tag{2.42}$$

where $\log Z(\boldsymbol{\lambda}_l) = \int \exp(\boldsymbol{\lambda}_l^{\mathrm{T}} g_l(\boldsymbol{\phi}_l)) d\boldsymbol{\phi}$ and $\log Z(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is defined analogously to (2.15):

$$\log Z(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{d}{2} \log 2\pi + \frac{1}{2} \log |\boldsymbol{\Sigma}| + \frac{1}{2} \boldsymbol{\mu}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}. \tag{2.43}$$

Numerical robust expressions for the site normalization terms $\tilde{Z}_i$ can be derived by using the local moment matching conditions (see (2.22)) for the normalization constants:

$$\begin{aligned} \hat{Z}_i &= \int t_i(\mathbf{z}_i, \boldsymbol{\phi}_{\mathcal{A}_i})^\eta q_{-i}(\boldsymbol{\theta}, \boldsymbol{\phi}) d\boldsymbol{\theta} d\boldsymbol{\phi} = \tilde{Z}_i^\eta \int \tilde{t}_{\mathbf{z},i}(\boldsymbol{\theta})^\eta \tilde{t}_{\boldsymbol{\phi},i}(\boldsymbol{\phi})^\eta q_{-i}(\boldsymbol{\theta}, \boldsymbol{\phi}) d\boldsymbol{\theta} d\boldsymbol{\phi} \\ &= \tilde{Z}_i^\eta \int \psi(\mathbf{z}_i, \eta\tilde{\boldsymbol{\nu}}_i, \eta\tilde{\mathbf{T}}_i) q_{-i}(\mathbf{z}_i) d\mathbf{z}_i \int \tilde{t}_{\boldsymbol{\phi},i}(\boldsymbol{\phi})^\eta q_{-i}(\boldsymbol{\phi}) d\boldsymbol{\phi}, \end{aligned} \tag{2.44}$$

which results in

$$\begin{aligned} \log \tilde{Z}_i = & \frac{1}{\eta} \Big(\log \hat{Z}_i + \log Z(\mathbf{m}_{-i}, \mathbf{V}_{-i}) - \log Z(\mathbf{m}_i, \mathbf{V}_i) \\ & + \sum_{l=1}^{L} \left(\log Z(\boldsymbol{\lambda}_{-i,l}) - \log Z(\boldsymbol{\lambda}_l) \right) \Big), \end{aligned} \tag{2.45}$$

where $\hat{Z}_i$ are the normalization terms of (2.26). The site normalization terms (2.45) can be evaluated by saving the necessary parameters at each EP update of the algorithm described in Section (2.3.4). Approximation (2.42) can be recomputed without significant additional cost after each sequential or parallel posterior update step, where quantities such as $\log |\boldsymbol{\Sigma}| = -\log |\mathbf{Q}|$ and $\boldsymbol{\mu} = \boldsymbol{\Sigma}\mathbf{h} = \mathbf{Q}^{-1}\mathbf{h}$ can be computed or updated using the same Cholesky decompositions or rank one updates, respectively.

2.3.7 Provably Convergent Double-Loop Algorithms

When standard EP does not converge, it is possible to find approximations satisfying the moment matching conditions (2.22) or equivalently (2.25) using provably convergent double-loop algorithms [Minka, 2001b,c, Heskes and Zoeter, 2002, Opper and Winther, 2005]. For example, Heskes and Zoeter [2002] present simulation results with linear dynamical systems where useful approximations are found using a double-loop algorithm when damped EP fails to converge. With GP models, visual comparisons of the converge properties of a double-loop algorithm with both sequential and parallel EP are presented in Publication II.

To simplify the notation, the double-loop algorithms are considered using only one parameter vector $\boldsymbol{\theta}$. The exact posterior (2.8) is replaced with $p(\boldsymbol{\theta}|\mathcal{D}) = Z^{-1} \prod_{i=1}^{n} t_i(\boldsymbol{\theta})$, where Z is the marginal likelihood and $t_i(\boldsymbol{\theta})$ are

the site terms. The posterior is approximated with

$$q(\boldsymbol{\theta}) = Z_{\text{EP}}^{-1} \prod_{i=1}^{n} \tilde{t}_i(\boldsymbol{\theta}) = Z(\boldsymbol{\lambda})^{-1} \exp(\boldsymbol{\lambda}^{\mathrm{T}} \mathbf{g}(\boldsymbol{\theta})), \tag{2.46}$$

where $\tilde{t}_i(\boldsymbol{\theta}) = \tilde{Z}_i \exp(\tilde{\boldsymbol{\lambda}}_i^{\mathrm{T}} \mathbf{g}(\boldsymbol{\theta}))$, $\mathbf{g}(\boldsymbol{\theta})$ is a vector of sufficient statistics, $\boldsymbol{\lambda} = \sum_i \tilde{\boldsymbol{\lambda}}_i$ is a vector of natural parameters, and the normalization term is given by $Z(\boldsymbol{\lambda}) = \int \exp(\boldsymbol{\lambda}^{\mathrm{T}} \mathbf{g}(\boldsymbol{\theta})) d\boldsymbol{\theta}$. Defining, e.g., $\mathbf{g}(\boldsymbol{\theta}) = [\boldsymbol{\theta}^{\mathrm{T}}, -\frac{1}{2}(\boldsymbol{\theta} \otimes \boldsymbol{\theta})^{\mathrm{T}}]^{\mathrm{T}}$ and $\boldsymbol{\lambda} = [\mathbf{h}^{\mathrm{T}}, \mathrm{vec}(\mathbf{Q})^{\mathrm{T}}]^{\mathrm{T}}$, where $\mathrm{vec}(\mathbf{Q})$ denotes the vertical concatenation of the columns of $\mathbf{Q}$, $\boldsymbol{\lambda}^{\mathrm{T}} \mathbf{g}(\boldsymbol{\theta}) = \mathbf{h}^{\mathrm{T}} \boldsymbol{\theta} - \frac{1}{2} \boldsymbol{\theta}^{\mathrm{T}} \mathbf{Q} \boldsymbol{\theta}$ and consequently the approximation (2.13) can be written as $q(\boldsymbol{\theta}) = Z(\boldsymbol{\lambda})^{-1} \exp(\boldsymbol{\lambda}^{\mathrm{T}} \mathbf{g}(\boldsymbol{\theta}))$. The hyperparameters ϕ could be included in the notation by concatenating their natural parameters in $\boldsymbol{\lambda}$ and sufficient statistics in $\mathbf{g}(\boldsymbol{\theta})$.

The fixed points of the EP algorithm that satisfy the moment consistency conditions (2.25), correspond to the stationary points of the following objective function

$$\min_{\boldsymbol{\lambda}} \max_{\boldsymbol{\lambda}_-} \left(\frac{n}{\eta} - 1 \right) \log Z(\boldsymbol{\lambda}) - \frac{1}{\eta} \sum_{i=1}^{n} \log \hat{Z}_i(\boldsymbol{\lambda}_{-i}), \tag{2.47}$$

where $\boldsymbol{\lambda}_- = \{\boldsymbol{\lambda}_{-i}\}_{i=1}^n$, $\hat{Z}_i(\boldsymbol{\lambda}_{-i}) = \int t_i(\boldsymbol{\theta})^{\eta} \exp\left(\boldsymbol{\lambda}_{-i}^{\mathrm{T}} \mathbf{g}(\boldsymbol{\theta})\right) d\boldsymbol{\theta}$ analogously with the normalization constants of the tilted distributions (2.27), and the min-max problem is solved subject to the constraint $(n - \eta)\boldsymbol{\lambda} = \sum_{i=1}^n \boldsymbol{\lambda}_{-i}$ [Minka, 2001b,c, 2005]. Substituting the constraint into (2.47) and taking the derivatives with respect to $\boldsymbol{\lambda}_{-i}$ using the results $\nabla_{\boldsymbol{\lambda}} \log Z(\boldsymbol{\lambda}) = \mathrm{E}_q(\mathbf{g}(\boldsymbol{\theta}))$ and $\nabla_{\boldsymbol{\lambda}_{-i}} \log \hat{Z}_i(\boldsymbol{\lambda}_{-i}) = \mathrm{E}_{\hat{p}_i}(\mathbf{g}(\boldsymbol{\theta}))$, where the tilted distribution is defined as $\hat{p}_i(\boldsymbol{\theta}) = \hat{Z}_i(\boldsymbol{\lambda}_{-i})^{-1} t_i(\boldsymbol{\theta})^{\eta} \exp\left(\boldsymbol{\lambda}_{-i}^{\mathrm{T}} \mathbf{g}(\boldsymbol{\theta})\right)$, results in analogous moment consistency conditions with (2.25): $\mathrm{E}_{\hat{p}_i}(\mathbf{g}(\boldsymbol{\theta})) = \mathrm{E}_q(\mathbf{g}(\boldsymbol{\theta}))$ for $i = 1, ..., n$. The min-max problem (2.47) could be solved using, e.g., gradient-based methods or trying similar message passing iterations as with regular EP [Minka, 2001a,b]. $\log Z(\boldsymbol{\lambda})$ and $\log \hat{Z}_i(\boldsymbol{\lambda}_{-i})$ are convex functions of the natural parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}_{-i}$, because $\nabla^2_{\boldsymbol{\lambda}} \log Z(\boldsymbol{\lambda}) = \mathrm{Cov}_q(\mathbf{g}(\boldsymbol{\theta}))$ and $\nabla^2_{\boldsymbol{\lambda}_{-i}} \hat{Z}_i(\boldsymbol{\lambda}_{-i}) = \mathrm{Cov}_{\hat{p}_i}(\mathbf{g}(\boldsymbol{\theta}))$. Consequently, assuming that both $\hat{p}_i(\boldsymbol{\theta})$ and $q(\boldsymbol{\theta})$ are proper probability density functions, the objective (2.47) is a sum of a convex and a concave part with respect to the parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}_-$.

An alternative double-loop objective can be obtained from (2.47) by separating the first term into a concave and a convex part as

$$\min_{\boldsymbol{\lambda}} \max_{\boldsymbol{\lambda}_-} -\log Z(\boldsymbol{\lambda}) - \frac{1}{\eta} \sum_{i=1}^{n} \log \hat{Z}_i(\boldsymbol{\lambda}_{-i}) + \frac{n}{\eta} \log Z(\boldsymbol{\lambda}_s), \tag{2.48}$$

where the parameters of the convex part are denoted with $\boldsymbol{\lambda}_s$. Writing the constraint as $\boldsymbol{\lambda} = \eta^{-1} \sum_{i=1}^n (\boldsymbol{\lambda} - \boldsymbol{\lambda}_{-i}) = \eta^{-1} \sum_{i=1}^n (\boldsymbol{\lambda}_s - \boldsymbol{\lambda}_{-i}) = \sum_{i=1}^n \tilde{\boldsymbol{\lambda}}_i$,

where $\boldsymbol{\lambda}_s = \boldsymbol{\lambda}_{-i} + \eta\tilde{\boldsymbol{\lambda}}_i$, analogous definitions are recovered for the site parameters and the cavity parameters with the standard EP (see, e.g., (2.32)). Equation (2.48) can be interpreted as the double-loop objective of the expectation consistent (EC) approach for approximate inference, extended according to the description of Heskes et al. [2005]. The parameters of the convex part, $\boldsymbol{\lambda}_s$, define a surrogate distribution $q_s(\boldsymbol{\theta}) = Z(\boldsymbol{\lambda}_s)^{-1}\exp(\boldsymbol{\lambda}_s^{\mathrm{T}}\mathbf{g}(\boldsymbol{\theta}))$, which at convergence shares the same expected sufficient statistics with the posterior approximation $q(\boldsymbol{\theta})$. When $q(\boldsymbol{\theta})$ is Gaussian and the site terms depend on $\boldsymbol{\theta}$ through linear transformation of the form $\mathbf{z}_i = \mathbf{U}_i^{\mathrm{T}}\boldsymbol{\theta}$, the surrogate distributions can be transformed into equivalent marginal approximations for the latent values $\mathbf{z}_i$ similarly to the standard EP as described in Section 2.3.3. The objective functions (2.47) and (2.48) are equivalent with $-\log Z_{\mathrm{EP}}$ defined by (2.42) and (2.45), and (2.48) is also equivalent to the expectation consistent (EC) free energy approximation presented by Opper and Winther [2005]. A unifying view of the EC and EP approximations as well as connections to the Bethe free energies are presented by Heskes et al. [2005].

Equation (2.48) suggests a double-loop algorithm where the inner loop consist of maximization with respect to $\boldsymbol{\lambda}_-$ with fixed $\boldsymbol{\lambda}_s$ and the outer loop of minimization with respect to $\boldsymbol{\lambda}_s$. The inner maximization affects only the first two terms and ensures that the marginal moments of the current posterior approximation $q(\boldsymbol{\theta})$ are equal to the moments of the tilted distributions $\hat{p}_i(\boldsymbol{\theta})$ for fixed $\boldsymbol{\lambda}_s$. The outer minimization ensures that the moments of $q_s(\boldsymbol{\theta})$ are equal to marginal moments of $q(\boldsymbol{\theta})$. At convergence, $q(\boldsymbol{\theta})$, $\hat{p}_i(\boldsymbol{\theta})$, and $q_s(\boldsymbol{\theta})$ share the same values for the expected sufficient statistics $\mathrm{E}(\mathbf{g}(\boldsymbol{\theta}))$. The inner maximization can be done using generic gradient-based optimization methods or by trying similar message passing iterations as with the regular EP algorithm (see, e.g., Minka [2001c] and Publication II). Once the inner optimum is found, the outer minimization can be done by bounding the concave part from above with a linear function of $\boldsymbol{\lambda}_s$ and minimizing the resulting upper bound: $\boldsymbol{\lambda}_s^{\mathrm{new}} = \arg\min_{\boldsymbol{\lambda}_s}\{-\frac{n}{\eta}\boldsymbol{\lambda}_s^{\mathrm{T}}\hat{\mathbf{s}} + \frac{n}{\eta}\log Z_s(\boldsymbol{\lambda}_s)\}$, where $\hat{\mathbf{s}} = \mathrm{E}_q(\mathbf{g}(\boldsymbol{\theta})) = \mathrm{E}_{\hat{p}_i}(\mathbf{g}(\boldsymbol{\theta}))$. This corresponds to updating $\boldsymbol{\lambda}_s$ so that the expected sufficient statistic of $q_s(\boldsymbol{\theta})$ are consistent with $q(\boldsymbol{\theta})$. If $t_i(\boldsymbol{\theta})$ are bounded, the objective is bounded from below and consequently there exists stationary points satisfying these expectation consistency constraints [Minka, 2001b, Opper and Winther, 2005]. In the case of multiple stationary points the solution with the smallest free energy can be chosen.

Since the first two terms in (2.48) are concave functions of $\boldsymbol{\lambda}_-$ and $\tilde{\boldsymbol{\lambda}} = \{\tilde{\boldsymbol{\lambda}}_i\}_{i=1}^n$, the inner maximization problem is concave with respect to $\boldsymbol{\lambda}_-$ (or equivalently $\tilde{\boldsymbol{\lambda}}$) after substitution of the constraints $\boldsymbol{\lambda} = \sum_{i=1}^n \tilde{\boldsymbol{\lambda}}_i = \eta^{-1} \sum_{i=1}^n (\boldsymbol{\lambda}_s - \boldsymbol{\lambda}_{-i})$ [Opper and Winther, 2005]. However, because the Hessian of the first term is given by $\nabla^2_{\boldsymbol{\lambda}_{-i}} \hat{Z}_i(\boldsymbol{\lambda}_{-i}) = \mathrm{Cov}_{\hat{p}_i}(\mathbf{g}(\boldsymbol{\theta}))$, the inner loop optimization with respect to $\boldsymbol{\lambda}_-$ is well defined and concave only if the tilted distributions $\hat{p}_i(\boldsymbol{\theta})$ are proper probability distributions that result in positive definite covariances for $\mathbf{g}(\boldsymbol{\theta})$. Therefore, to ensure that the products of $q_{-i}(\boldsymbol{\theta})$ and the sites $t_i(\boldsymbol{\theta})$ are proper distributions and that the inner-loop moment matching remains meaningful in case of Gaussian approximations, the cavity precisions $\mathbf{V}_{-i}^{-1} = \mathbf{V}_i^{-1} - \eta \tilde{\mathbf{T}}_i$ (see equation (2.31)) have to be kept positive definite during the iterations. Because decreasing η improves the conditioning of $\mathbf{V}_{-i}$, fractional updates can improve the numerical stability of the algorithm. Furthermore, since the cavity distributions related to the likelihood sites can be regarded as approximations of the LOO predictive distributions of $\boldsymbol{\theta}$, a positive definite cavity variance $\mathbf{V}_{-i}$ for a certain site would correspond to a situation where $q(\mathbf{U}_i^{\mathrm{T}}\boldsymbol{\theta}|\mathbf{y}_{-i}, \mathbf{X})$ would have infinite variance in some direction(s) of the space of $\mathbf{z}_i$, which is not sensible from practical modeling perspective. On the other hand, the site precisions $\tilde{\mathbf{T}}_i$ may become negative with non-log-concave sites, which correspond to a local increase of posterior uncertainty resulting, e.g., from an outlying observation [Publication II].

2.4 Variational Mean-Field (VMF)

This section summarizes the VMF method for approximate inference and compares its properties with EP. The VMF approximation is obtained by minimizing the global reverse KL-divergence

$$\mathrm{KL}(q(\boldsymbol{\theta}, \boldsymbol{\phi})|p(\boldsymbol{\theta}, \boldsymbol{\phi}, |\mathcal{D})) \tag{2.49}$$

with respect to $q(\boldsymbol{\theta}, \boldsymbol{\phi})$. With the chosen factorization assumption for the approximate posterior, $q(\boldsymbol{\theta}, \boldsymbol{\phi}) = q(\boldsymbol{\theta}) \prod_{l=1}^L q(\boldsymbol{\phi}_l)$, the solution can be obtained by iterating the following updates for $q(\boldsymbol{\theta})$ and $q(\boldsymbol{\phi}_l)$, $l = 1, ..., L$,

until convergence [Jordan et al., 1999, Jaakkola, 2000]:

$$
\begin{aligned}
q(\boldsymbol{\theta})^{\text{new}} &\propto \exp\left(\int q(\boldsymbol{\phi}) \log p(\mathbf{Y}, \boldsymbol{\theta}, \boldsymbol{\phi}|\mathbf{X}) d\boldsymbol{\phi}\right) \qquad (2.50)\\
&\propto \exp\left(\sum_{i=1}^{n+m} \int \prod_{l \in \mathcal{A}_i} q(\boldsymbol{\phi}_l) \log t_i(\mathbf{U}_i^{\mathrm{T}}\boldsymbol{\theta}, \boldsymbol{\phi}_{\mathcal{A}_i}) d\boldsymbol{\phi}_{\mathcal{A}_i}\right)\\
q(\boldsymbol{\phi}_l)^{\text{new}} &\propto \exp\left(\int q(\boldsymbol{\theta}) q(\boldsymbol{\phi}_{-l}) \log p(\mathbf{Y}, \boldsymbol{\theta}, \boldsymbol{\phi}|\mathcal{D}) d\boldsymbol{\theta} d\boldsymbol{\phi}_{-l}\right)\\
&\propto \exp\left(\sum_{l \in \mathcal{A}_i} \int q(\boldsymbol{\theta}) \prod_{k \in \mathcal{A}_i \backslash l} q(\boldsymbol{\phi}_k) \log t_i(\mathbf{U}_i^{\mathrm{T}}\boldsymbol{\theta}, \boldsymbol{\phi}_{\mathcal{A}_i}) d\boldsymbol{\theta} d\boldsymbol{\phi}_{\mathcal{A}_i \backslash l}\right) p(\boldsymbol{\phi}_l),
\end{aligned}
$$

where $\boldsymbol{\phi}_{-l} = \{\boldsymbol{\phi}_k\}_{k \neq l}$ and $\mathcal{A}_i \backslash l = \{k \in \mathcal{A}_i | k \neq l\}$. These iterations are guaranteed to converge to a local maximum of the variational lower bound $\mathcal{L}(q) \leq \log Z$ defined by $\mathcal{L}(q) = \mathrm{E}_q(\log p(\mathbf{Y}, \boldsymbol{\theta}, \boldsymbol{\phi}|\mathbf{X})) - \mathrm{E}_q(\log q(\boldsymbol{\theta}, \boldsymbol{\phi}))$.

If the site terms $t_i(\mathbf{U}_i^{\mathrm{T}}\boldsymbol{\theta}, \boldsymbol{\phi}_{\mathcal{A}_i})$ are chosen suitably, the expectations with respect to $q(\boldsymbol{\theta})$ and $q(\boldsymbol{\phi}_l)$ can be computed analytically and the updated posterior approximations will remain in a tractable family of distribution. For example, simple analytically tractable computations are obtained by assuming a linear predictor $z_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\theta}$ and a Gaussian scale-mixture observation model $y_i \sim \mathcal{N}(y_i|z_i, \phi_i)$ with unknown noise variances ϕ_i for each observation (analogous to a Student-t model [Tipping and Lawrence, 2003]), and using a posterior approximation that can be factored as $q(\boldsymbol{\theta}, \boldsymbol{\phi}) = q(\boldsymbol{\theta}) \prod_i q(\phi_i)$. Assigning a multivariate Gaussian prior $t_{n+1}(\boldsymbol{\theta}) = p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ to the coefficients, and taking expectations of $\log t_i = -\frac{1}{2\phi_i}(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\theta})^2 - \frac{1}{2}\log(2\pi\phi_i)$ with respect to $q(\phi_i)$, it is straightforward to show that the approximate posterior distribution of $\boldsymbol{\theta}$ will be of the same form as (2.13): $q(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}^{-1} = \mathbf{Q} = \sum_{i=1}^n \mathbf{x}_i \mathrm{E}(\phi_i^{-1})\mathbf{x}_i^{\mathrm{T}} + \boldsymbol{\Sigma}_0^{-1}$ and $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} = \mathbf{h} = \sum_{i=1}^n \mathbf{x}_i \mathrm{E}(\phi_i^{-1}) y_i + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0$. Assigning independent conjugate inverse-Gamma priors to the residual variances, $p(\phi_i) \propto \phi_i^{-\alpha_0 - 1}\exp(\phi_i/\beta_0)$ for $i = 1, ..., n$, and using (2.50) results in inverse-Gamma posteriors $q(\phi_i) \propto \phi_i^{-\alpha-1}\exp(\phi_i/\beta)$, where $\alpha = \alpha_0 + \frac{1}{2}$ and $\beta = \beta_0 + \frac{1}{2}((y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\mu})^2 + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{x}_i)$. A similar VMF approach is used in the comparisons of different approximate inference methods for robust GP regression with the Student-t model in Publications I and II, where it is found that compared to EP the VMF approach tends to underestimate the posterior uncertainties in cases when the posterior distribution is multimodal (VMF represents only one of the possible modes of $p(\boldsymbol{\theta}|\mathcal{D})$). However, generally the complexity of the approximations (2.50) is not constrained, unless a projection on an exponential family distribution is done in a similar way as with EP in equation (2.21) [Minka, 2005].

An analogous variational approximation with the global VMF solution (2.50) can be formed also by using a variational message passing (VMP) algorithm that updates the approximations for the site terms $t_i(\mathbf{U}_i^{\mathrm{T}}\boldsymbol{\theta}, \boldsymbol{\phi}_{\mathcal{A}_i})$ one at a time [Winn and Bishop, 2005]. Such algorithm can be obtained from the fractional EP algorithm by reversing the KL minimization of equation (2.21):

$$\hat{q}_i(\boldsymbol{\theta}, \boldsymbol{\phi}) = \arg\min_{q_i} \mathrm{KL}\left(q_i(\boldsymbol{\theta}, \boldsymbol{\phi}) | \hat{p}_i(\boldsymbol{\theta}, \boldsymbol{\phi})\right), \tag{2.51}$$

where $\hat{p}_i(\boldsymbol{\theta}, \boldsymbol{\phi})$ is defined similarly to equations (2.19) and (2.20) with $\eta = 1$, that is, $\hat{p}_i(\boldsymbol{\theta}, \boldsymbol{\phi}) \propto t_i(\mathbf{U}_i^{\mathrm{T}}\boldsymbol{\theta}, \boldsymbol{\phi}_{\mathcal{A}_i}) q_{-i}(\boldsymbol{\theta}) \prod_{l=1}^{L} q_{-i}(\boldsymbol{\phi}_l)$, where $q_{-i}(\boldsymbol{\theta}) \propto q(\boldsymbol{\theta}) \tilde{t}_{\boldsymbol{\theta},i}(\boldsymbol{\theta})^{-1}$ and $q_{-i}(\boldsymbol{\phi}_l) \propto q(\boldsymbol{\phi}_l) \tilde{t}_{\boldsymbol{\phi}_l,i}(\boldsymbol{\phi}_l)^{-1}$. The approximation $q_i(\boldsymbol{\theta}, \boldsymbol{\phi})$ is of the same form as the approximating family: $q_i(\boldsymbol{\theta}, \boldsymbol{\phi}) = q_i(\boldsymbol{\theta}) \prod_l q_i(\boldsymbol{\phi}_l)$, where $q_i(\boldsymbol{\theta}) \propto q_{-i}(\boldsymbol{\theta}) \tilde{t}_{\boldsymbol{\theta},i}(\boldsymbol{\theta})$ and $q_i(\boldsymbol{\phi}_l) \propto q_{-i}(\boldsymbol{\phi}_l) \tilde{t}_{\boldsymbol{\phi}_l,i}(\boldsymbol{\phi}_l)$. The reverse-KL minimization of (2.51) can also be interpreted as an iterative approach for minimizing a more general divergence measure called α-divergence at the limit $\alpha \to 0$ with $\alpha = \eta$ [Minka, 2005]. Analogously to the global KL minimizer (2.50), an iterative solution to the local minimization (2.51) can be written as $\hat{q}_i(\boldsymbol{\theta}) \propto q_{-i}(\boldsymbol{\theta}) \tilde{t}_{\boldsymbol{\theta},i}(\boldsymbol{\theta})^{\mathrm{new}}$ and $\hat{q}_i(\boldsymbol{\phi}_l) \propto q_{-i}(\boldsymbol{\phi}_l) \tilde{t}_{\boldsymbol{\phi}_l,i}(\boldsymbol{\phi}_l)^{\mathrm{new}}$ for $l \in \mathcal{A}_i$, where the new site approximations are given by

$$\begin{aligned} \tilde{t}_{\boldsymbol{\theta},i}(\boldsymbol{\theta})^{\mathrm{new}} &\propto \exp\left(\int \prod_{l \in \mathcal{A}_i} q(\boldsymbol{\phi}_l) \log t_i(\mathbf{U}_i^{\mathrm{T}}\boldsymbol{\theta}, \boldsymbol{\phi}_{\mathcal{A}_i}) d\boldsymbol{\phi}_{\mathcal{A}_i} \right) \\ \tilde{t}_{\boldsymbol{\phi}_l,i}(\boldsymbol{\phi}_l)^{\mathrm{new}} &\propto \exp\left(\int q(\boldsymbol{\theta}) \prod_{k \in \mathcal{A}_i \setminus l} q(\boldsymbol{\phi}_k) \log t_i(\mathbf{U}_i^{\mathrm{T}}\boldsymbol{\theta}, \boldsymbol{\phi}_{\mathcal{A}_i}) d\boldsymbol{\theta} d\boldsymbol{\phi}_{\mathcal{A}_i \setminus l} \right). \end{aligned} \tag{2.52}$$

In equation (2.52) the expectations are taken with respect to the previous approximations $q(\boldsymbol{\theta})$ and $q(\boldsymbol{\phi}_l)$. Compared to the fractional EP solution, $\eta \neq 0$, the VMP algorithm ($\eta \to 0$) has the special property that the stationary solutions of the message passing algorithm based on the local KL minimizations are also stationary points of the minimization of the global divergence (2.49) [Minka, 2005, Knowles and Minka, 2011]. The VMP algorithm can also be extended to non-conjugate models by adopting suitable bounds on $\log t_i(\mathbf{U}_i^{\mathrm{T}}\boldsymbol{\theta}, \boldsymbol{\phi}_{\mathcal{A}_i})$ or $\mathrm{E}_q(\log t_i(\mathbf{U}_i^{\mathrm{T}}\boldsymbol{\theta}, \boldsymbol{\phi}_{\mathcal{A}_i}))$ and solving the natural site parameters by minimizing an upper bound on the local KL divergence (2.51), or alternatively using numerical quadratures to solve the local KL minimizations directly [Minka, 2005, Knowles and Minka, 2011].

One key difference between the EP and VMF solutions is the way the integration is done over the remaining parameters when approximate

marginal distributions $q(\boldsymbol{\theta})$ and $q(\boldsymbol{\phi}_l)$ are determined using factorized posterior approximations. In EP, the expected sufficient statistics of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}_l$ are determined by integration over $\boldsymbol{\phi}$ or $\boldsymbol{\theta}$ in the tilted distribution $\hat{p}_i(\boldsymbol{\theta}, \boldsymbol{\phi})$ according to (2.25), and the approximations $q(\boldsymbol{\theta})$ and $q(\boldsymbol{\phi}_l)$ are subsequently set consistent with these sufficient statistics. The EP approximation relies on the assumption that during the iterations the tilted distributions $\hat{p}_i(\boldsymbol{\theta}, \boldsymbol{\phi})$, $i = 1, ..., n + m$, form an increasingly improving sequence of approximations for the true posterior $p(\boldsymbol{\theta}, \boldsymbol{\phi}|\mathcal{D})$, and that equation (2.25) preserves a sufficiently good representation of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}_l$ in the form of expected marginal sufficient statistics. In other words, EP does an iterative approximate integration over the uncertainty on the hyperparameters $\boldsymbol{\phi}$ while forming the approximate marginal distribution $q(\boldsymbol{\theta})$.

In the VMP update (2.52), the approximations for $q(\boldsymbol{\theta})$ and $q(\boldsymbol{\phi}_l)$ are formed by taking expectations of $\log t_i(\mathbf{U}_i^{\mathrm{T}}\boldsymbol{\theta}, \boldsymbol{\phi}_{\mathcal{A}_i})$ with respect to the current approximations $q(\boldsymbol{\phi}_{\mathcal{A}_i})$ or $q(\boldsymbol{\theta})$, respectively. This update can be clarified by making the commonly used assumption that the site terms are in the exponential family conditioned either on $\boldsymbol{\phi}$ or $\{\boldsymbol{\theta}, \boldsymbol{\phi}_{-l}\}$, and conjugate with the chosen approximations for $q(\boldsymbol{\theta})$ and $q(\boldsymbol{\phi}_l)$, respectively [Winn and Bishop, 2005]. For example, with the chosen Gaussian approximation for $\boldsymbol{\theta}$, the site terms could be selected so that conditioned on $\boldsymbol{\phi}$ they could be written as

$$\log t_i(\mathbf{U}_i^{\mathrm{T}}\boldsymbol{\theta}, \boldsymbol{\phi}_{\mathcal{A}_i}) = -\frac{1}{2}\mathbf{z}_i^{\mathrm{T}}\mathbf{Q}_i(\boldsymbol{\phi}_{\mathcal{A}_i})\mathbf{z}_i + \mathbf{h}_i(\boldsymbol{\phi}_{\mathcal{A}_i})^{\mathrm{T}}\mathbf{z}_i + C(\boldsymbol{\phi}_{\mathcal{A}_i}), \tag{2.53}$$

where $\mathbf{z}_i = \mathbf{U}_i^{\mathrm{T}}\boldsymbol{\theta}$. Taking expectations with respect to $q(\boldsymbol{\phi})$ according to (2.52), results in Gaussian site terms $\tilde{t}_{\boldsymbol{\theta},i}$, where $\mathbf{Q}_i(\boldsymbol{\phi}_{\mathcal{A}_i})$ and $\mathbf{h}_i(\boldsymbol{\phi}_{\mathcal{A}_i})$ are replaced with their expectations with respect to the current approximation $q(\boldsymbol{\phi})$. Similarly, in the previous example with a linear model and a Student-t likelihood, the posterior approximation of the coefficients was formed as if observations were made according to $y_i \sim \boldsymbol{\theta}^{\mathrm{T}}\mathbf{x}_i + e_i$, where $e_i \sim \mathcal{N}(0, \mathrm{E}_q(\phi_i^{-1})^{-1})$. This differs clearly from the integration over $\hat{p}_i(\boldsymbol{\phi})$ in the corresponding EP update and generally this can be seen as a tendency to underestimate the joint posterior uncertainty on $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ (for illustrations and comparisons see, e.g., Minka [2005], Bishop [2006], Nickisch and Rasmussen [2008], Publications II and III).

2.5 Local Variational Bounds (LVB)

Because the free form updates (2.50) and (2.52) of the VMF and VMP approximations require a priori factorization assumptions and suitable conjugate-exponential model specifications, and the complexity of the resulting marginal approximations are not generally constrained, various alternative VB approaches have been proposed. Many of them are based on using suitable parametric bounds for the site terms to form a computationally tractable lower bound for the model evidence Z [see, e.g. Jaakkola and Jordan, 1996, Jordan et al., 1999, Gibbs and MacKay, 2000, Seeger and Wipf, 2010, Khan et al., 2012]. Also many of the proposed approaches are based on direct minimization of the global KL divergence (2.49) (or the variational free energy) using, e.g., gradient-based methods [see, e.g., Raiko et al., 2007, Nickisch and Rasmussen, 2008, Opper and Archambeau, 2009, Honkela et al., 2010]. This section summarizes briefly the former LVB approach, which is used for comparisons with the EP approach in GP regression with the Student-t model in Publication II. The implementation is similar to the binary GP classification described by Gibbs and MacKay [2000], and extensive comparisons with other approximate methods for GP classification have been done by Nickisch and Rasmussen [2008].

In the following it is assumed that the hyperparameters $\boldsymbol{\phi}$ are fixed or their type-II MAP estimates (2.4) are determined using the LVB marginal likelihood approximation. LVB is based on forming for each site term a lower bound $b_i(\mathbf{z}_i, \boldsymbol{\Gamma}_i)$ that is conjugate with the approximate family $q(\boldsymbol{\theta})$:

$$t_i(\mathbf{U}_i^{\mathrm{T}}\boldsymbol{\theta}, \boldsymbol{\phi}_{\mathcal{A}_i}) \geq \exp\left(-\frac{1}{2}\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\Gamma}_i^{-1}\mathbf{z}_i + \mathbf{b}_i(\boldsymbol{\Gamma}_i)^{\mathrm{T}}\mathbf{z}_i - \frac{1}{2}h_i(\boldsymbol{\Gamma}_i)\right) = b_i(\mathbf{z}_i, \boldsymbol{\Gamma}_i) \tag{2.54}$$

where $\mathbf{z}_i = \mathbf{U}_i^{\mathrm{T}}\boldsymbol{\theta}$ and the expressions of $\mathbf{b}_i$ and h_i as a function of the free parameters $\boldsymbol{\Gamma}_i$ depend on the chosen model [for examples of commonly used site terms (or potentials) see, Nickisch, 2010]. The local bounds $b_i(\mathbf{z}_i, \boldsymbol{\Gamma}_i)$ can be used to form an analytically tractable lower bound on the marginal likelihood, $Z = p(\mathbf{y}|\mathbf{X}, \boldsymbol{\phi}) = \int p(\mathbf{y}, \boldsymbol{\theta}|\mathbf{X}, \boldsymbol{\phi}) d\boldsymbol{\theta} \geq Z_{\mathrm{LVB}}(\boldsymbol{\Gamma})$, which in turn can be maximized with respect to the parameters $\boldsymbol{\Gamma}_i$ to obtain a posterior approximation $q(\boldsymbol{\theta})$ and a marginal likelihood approximation for hyperparameter inference. After determining the variational parameters, the posterior approximation $q(\boldsymbol{\theta})$ can be formed similarly to (2.13), where $\mathbf{Q} = \sum_{i=1}^{n+m} \mathbf{U}_i \boldsymbol{\Gamma}_i^{-1} \mathbf{U}_i^{\mathrm{T}}$ and $\mathbf{h} = \sum_{i=1}^{n+m} \mathbf{U}_i \boldsymbol{\Gamma}_i^{-1} \mathbf{b}_i$. LVB can also be regarded as a special case of the direct KL minimization approaches [Nickisch and

Rasmussen, 2008, Opper and Archambeau, 2009, Nickisch, 2010], where also $\mathbf{b}_i$ are optimized as free parameters.

In many commonly used models the site terms depend only on scalar random variables $z_i = \mathbf{u}_i^{\mathrm{T}}\boldsymbol{\theta}$ [Nickisch, 2010], which results in only one scalar parameter γ_i for each site term and facilitates the optimization. For example, with the Student-t observation model studied in Publication II, the site terms are defined as $t_i(z_i, \nu, \sigma^2) \propto (1+\nu^{-1}(y_i - z_i)^2\sigma^{-2})^{-(\nu+1)/2}$, where z_i is the latent function value $f_i = f(\mathbf{x}_i)$, y_i the observation, σ^2 the noise magnitude, and ν the degrees of freedom parameter. Only scalar scale parameters γ_i need to be optimized and the location parameters $b_i(\gamma_i)$ are determined by the corresponding observations: $b_i = y_i/\gamma_i$. Also multivariate bounds have been proposed, e.g., in multi-class classification [Chai, 2012], but it should be noted that in the approach of Chai [2012], the bounds are defined directly on the expectations $\mathrm{E}_q(\log t_i)$, which can presumably result in tighter bounds and more stable performance in some cases [Knowles and Minka, 2011].

2.6 Laplace Approximation (LA)

The method was already proposed by Laplace [1774] for approximating integrals of the form $\int_a^b \exp(f(x))dx$, where $f(x)$ is twice-differentiable. Since then it has been used widely for approximating different integrals over intractable posterior distributions to determine approximate predictive distributions and type-II hyperparameter estimates for various models including generalized linear models [see, e.g., Bishop, 2006], neural networks [Mackay, 1995, Bishop, 2006], Gaussian processes and other latent Gaussian models (LGMs) [Williams and Barber, 1998, Rasmussen and Williams, 2006, Rue et al., 2009]. Most of these approaches are based on determining first type-II MAP estimates for the hyperparameters $\boldsymbol{\phi}$ by gradient-based optimization using Laplace's approximation of the marginal likelihood $p(\mathbf{y}|\mathbf{X}, \boldsymbol{\phi})$. The MAP-II estimates $\hat{\boldsymbol{\phi}}$ are subsequently used to approximate the predictive density with $p(\mathbf{y}_*|\mathbf{X}_*, \hat{\boldsymbol{\phi}}) = \int p(\mathbf{y}_*|\mathbf{X}_*, \boldsymbol{\theta}, \hat{\boldsymbol{\phi}})q(\boldsymbol{\theta}|\mathcal{D}, \hat{\boldsymbol{\phi}})d\boldsymbol{\theta}$, where the approximate conditional posterior distribution $q(\boldsymbol{\theta}|\mathcal{D}, \hat{\boldsymbol{\phi}})$ is determined using LA.

The Laplace approximation of $p(\boldsymbol{\theta}|\mathcal{D}, \hat{\boldsymbol{\phi}})$ is constructed by determining the mode $\hat{\boldsymbol{\theta}}$ of the conditional posterior using, e.g., Newton's method or conjugate gradient optimization, and making a second order Taylor ap-

proximation of $\log p(\boldsymbol{\theta}|\mathcal{D}, \boldsymbol{\phi})$ around the mode:

$$\log p(\boldsymbol{\theta}|\mathcal{D}, \boldsymbol{\phi}) = \sum_{i=1}^{n+m} \log t_i(\mathbf{U}_i^{\mathrm{T}}\boldsymbol{\theta}, \boldsymbol{\phi}_{\mathcal{A}_i}) - \log Z(\boldsymbol{\phi}) \tag{2.55}$$
$$\approx \sum_{i=1}^{n+m} \left(\log t_i(\hat{\mathbf{z}}_i, \boldsymbol{\phi}_{\mathcal{A}_i}) - \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^{\mathrm{T}}\mathbf{U}_i\mathbf{W}_i\mathbf{U}_i^{\mathrm{T}}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \right) - \log Z(\boldsymbol{\phi})$$

where $\hat{\mathbf{z}}_i = \mathbf{U}_i^{\mathrm{T}}\hat{\boldsymbol{\theta}}$, $\mathbf{W}_i = -\nabla^2_{\mathbf{z}_i} \log t_i(\mathbf{z}_i, \boldsymbol{\phi}_{\mathcal{A}_i})|_{\hat{\mathbf{z}}_i}$ and $\log Z(\boldsymbol{\phi}) = p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\phi})$ is the log marginal likelihood. The first order terms vanish because the derivative $\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\mathcal{D}, \boldsymbol{\phi}) = \sum_i \nabla_{\boldsymbol{\theta}} \log t_i(\mathbf{U}_i^{\mathrm{T}}\boldsymbol{\theta}, \boldsymbol{\phi}_{\mathcal{A}_i})$ is zero at the mode. Note also that the quadratic approximation of (2.55) is exact for all Gaussian site terms and that for fully-Gaussian posterior distributions LA is exact. Collecting the second order terms from (2.55) gives a Gaussian posterior approximation $q(\boldsymbol{\theta})$ that can be defined similarly to (2.13) by denoting $\tilde{\mathbf{Q}}_i = \mathbf{U}_i\mathbf{W}_i\mathbf{U}_i^{\mathrm{T}}$ and $\tilde{\mathbf{h}}_i = \mathbf{U}_i\mathbf{W}_i\mathbf{U}_i^{\mathrm{T}}\hat{\boldsymbol{\theta}}$. The mean of the approximation will become equal to the model, that is, $\boldsymbol{\mu} = \hat{\boldsymbol{\theta}}$, and the covariance is given by the negative inverse Hessian $\boldsymbol{\Sigma} = (\sum_{i=1}^{n+m} \tilde{\mathbf{Q}}_i)^{-1}$. Note that the covariance $\boldsymbol{\Sigma}$ will always be positive definite if the optimization has terminated at a local mode. Furthermore, with log-concave models the Hessian will remain positive definite during the optimization. With non-log-concave models such as the Student-t model studied in Publications I and II, additional stabilizations or alternative algorithms such as the EM algorithm [see, e.g., Gelman et al., 2004, Bishop, 2006] are needed in the optimization.

Type-II MAP estimates of $\boldsymbol{\phi}$ can be determined by approximating the conditional marginal likelihood $p(\mathbf{y}|\mathbf{X}, \boldsymbol{\phi})$ with Laplace's method and combining it with the the prior 2.7 to obtain an approximation for the marginal posterior $p(\boldsymbol{\phi}|\mathcal{D})$, and optimizing it using gradient-based methods [see, e.g., Rasmussen and Williams, 2006, Nickisch and Rasmussen, 2008]. Taking the exponential of the quadratic approximation (2.55) and integrating both sides over $\boldsymbol{\theta}$ gives

$$\log Z(\boldsymbol{\phi}) = \sum_{i=1}^{n+m} \log t_i(\hat{\mathbf{z}}_i, \boldsymbol{\phi}_{\mathcal{A}_i}) + \frac{1}{2}\log|\boldsymbol{\Sigma}| + \frac{d}{2}\log(2\pi), \tag{2.56}$$

where d is the dimension of $\boldsymbol{\theta}$. For optimization purposes, the gradients of $\log Z(\boldsymbol{\phi})$ can be solved analytically for models, where the logarithms of the site terms are twice differentiable. With non-differentiable site terms such as the Laplace distribution additional modification are required to determine the LA approximation and the derivatives of $\log Z(\boldsymbol{\phi})$ [Williams, 1995]. Typically computing the derivatives of $\log Z(\boldsymbol{\phi})$ are math-

ematically more involved, because also the implicit derivatives with respect to $\hat{\phi}$ and $\mathbf{W}_i$ have to be taken into account in addition to explicit derivatives of the expression (2.56). In contrast, with EP only the explicit derivatives of the approximate marginal likelihood are required, because the implicit derivatives with respect to the natural parameters of the site approximations and the cavity distributions cancel each other at the stationary solutions of the algorithm [Seeger, 2005, Opper and Winther, 2005]. This tractable property of EP enables also straightforward computation of the derivatives with the nested EP approximations proposed in Publication III.

The LA approximation can also be obtained using a same kind of message passing algorithm as EP and VMP. This framework called Laplace propagation was proposed by [Smola et al., 2004] and it is based on propagating the moments resulting from subsequent local Laplace approximations of $\hat{p}_i(\boldsymbol{\theta}) \propto q_{-i}(\boldsymbol{\theta}) t_i(\mathbf{U}_i^{\mathrm{T}}\boldsymbol{\theta}, \boldsymbol{\phi}_{\mathcal{A}_i})$ defined analogously to the EP tilted distributions (2.20), and updating local site approximations $\tilde{t}_i(\mathbf{U}_i^{\mathrm{T}}\boldsymbol{\theta})$ analogous to (2.11) based on these local approximations. Laplace propagation can lead to computational savings with models, where the site terms depend only on a small subset of $\boldsymbol{\theta}$ such as with typical graphical models [Smola et al., 2004].

2.7 Improving the Approximate Marginal Distributions

Recently, Rue et al. [2009] proposed various techniques inspired by Tierney and Kadane [1986] to improve upon the previously described LA framework with Gaussian Markov random field models.[4] They described computationally efficient ways to determine numerical non-Gaussian (and non-symmetric) approximations for the marginal distributions of the latent variables (the components of $\boldsymbol{\theta}$) through subsequent (or nested) use of Laplace's method, and to approximate integration over the hyperparameters $\boldsymbol{\phi}$ using grid-based numerical methods and the LA approximation for $p(\boldsymbol{\theta}|\mathcal{D}, \boldsymbol{\phi})$. They were able to obtain very accurate approximations compared with MCMC methods with many practically relevant models that result, for example, in skew marginal posterior densities.

[4]Note that Rue et al. [2009] use term "Gaussian approximation" for the local quadratic approximation at the mode (our LA approximation) and term "Laplace approximation" for the numerical non-symmetric marginal approximations obtained using the approach of Tierney and Kadane [1986].

EP-based corrections for the marginal distributions of the latent values with LGMs have been proposed by Paquet et al. [2009]. Later Cseke and Heskes [2011] described a unifying interpretation of the LA and EP based marginal corrections for LGMs, and proposed alternative computationally cheaper or more accurate approaches for approximating posterior marginals with both LA and EP. The comparisons of Cseke and Heskes [2011] show that when the site functions lead to very skewed posterior distributions, the approximate marginals obtained with EP are clearly more accurate compared with the mode-based approximations of LA. Furthermore, EP can handle more general site functions that are not differentiable or are defined only at discrete locations of the parameter space without additional modifications. However, in many practical modeling problems with large number of observations and log-concave posterior densities, LA can result in sufficiently accurate inference as illustrated by the comparisons of Rue et al. [2009].

One potential application of the marginal corrections could be improving the predictive density estimates in cases, where each of the future observations depends only on a subset of the latent values denoted here by $\boldsymbol{\theta}_*$: $p(\mathbf{y}_*|\mathbf{X}_*, \hat{\boldsymbol{\phi}}) = \int p(\mathbf{y}_*|\mathbf{X}_*, \boldsymbol{\theta}_*, \hat{\boldsymbol{\phi}}) q(\boldsymbol{\theta}_*|\mathcal{D}, \hat{\boldsymbol{\phi}}) d\boldsymbol{\theta}_*$. Potential skewness or heavier tails of $p(\boldsymbol{\theta}_*|\mathcal{D}, \hat{\boldsymbol{\phi}})$ could be taken into account in the integration by using either LA or EP based corrections to determine a numerical approximation for $q(\boldsymbol{\theta}_*|\mathcal{D}, \hat{\boldsymbol{\phi}})$. However, this may become computationally demanding when the predictions have to be computed quickly or when the subset $\boldsymbol{\theta}_*$ contains many components of $\boldsymbol{\theta}$. In the multi-class GP classification studied in Publication III, $\boldsymbol{\theta}_*$ has one component for each class leading to high dimensional non-analytical integrations, which is why an alternative approach was used to approximate the integration over all components of $\boldsymbol{\theta}$ separately for each test input $\mathbf{x}_*$ and each component of $\mathbf{y}_*$. This approximation is based on the general LA-based approach suitable for approximating expectations of positive functions also proposed originally by [Tierney and Kadane, 1986]. The comparisons of Publication III show that these corrections clearly improve the predictive density estimates over the standard LA approach. However, with EP the marginal corrections were not found necessary, because the Gaussian EP estimates of $q(\boldsymbol{\theta}_*|\mathcal{D}, \hat{\boldsymbol{\phi}})$ resulted in very accurate predictions compared to MCMC. The same applies also for the EP comparisons of Publication II. One probable explanation for this is that for integration over $\boldsymbol{\theta}_*$ it suffices that the marginal approximation captures the relevant probability mass

of the true distribution. The mean and covariance approximation of EP seems to be better suited for this purpose compared to the mode-based LA approximation as also discussed by Paquet et al. [2009] and Cseke and Heskes [2011].

3. Approximate Inference in Case Studies

This section summarizes the approximate inference approaches studied in the three Bayesian modeling applications considered in Publications I–IV, and connects them to the generic description of various approximate inference methods presented in Section 2.

3.1 Gaussian Process Regression with a Student-t Likelihood

Publications I and II consider approximate inference in GP regression with the heavy-tailed Student-t observation model, which enables robust inference on the unobserved latent function values $f_i = f(\mathbf{x}_i)$ in the presence of outlying observations [see, e.g., Liu and Rubin, 1995, West, 1984, Geweke, 1993, Gelman et al., 2004]. An MCMC approach based on hybrid Monte Carlo (HMC) sampling was proposed by Neal [1997], and later Kuss [2006] described a VMF approach that utilizes the scale-mixture representation of the Student-t model analogously to the linear regression approach described by Tipping and Lawrence [2003] (a short description of the linear model approach is given also in Section 2.4). Kuss [2006] also presented various comparisons with other robust observation models including finite mixtures of Gaussians and the Laplace distribution.

The Student-t observation model is given by

$$p(y_i|f_i,\nu,\sigma^2) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\nu\pi}\sigma}\left(1+\frac{(y_i-f_i)^2}{\nu\sigma^2}\right)^{-(\nu+1)/2}, \qquad (3.1)$$

where y_i is a scalar observation associated with the latent function value $f_i = f(\mathbf{x}_i)$, ν is the degrees of freedom parameter and σ^2 the scale parameter [Gelman et al., 2004]. The scale σ^2 controls the overall variance of the distribution and ν the thickness of the tails: as ν decreases, the tails get thicker and a larger proportion of the observations can be classified as outliers. The latent values $f(\mathbf{x})$ are given a zero-mean GP prior,

which by definition, implies a that any finite subset of latent variables, $\mathbf{f} = [f_1, ..., f_n]^{\mathrm{T}}$, has a multivariate Gaussian distribution:

$$p(\mathbf{f}|\mathbf{X}, \boldsymbol{\phi}_K) = \mathcal{N}(\mathbf{f}|\mathbf{0}, \mathbf{K}), \tag{3.2}$$

where $\mathbf{K}$ is a covariance matrix whose elements are defined by a covariance function $k(\mathbf{x}, \mathbf{x}')$ as $\mathbf{K}_{i,j} = k(\mathbf{x}, \mathbf{x}'|\boldsymbol{\phi}_K)$ [Rasmussen and Williams, 2006]. The covariance function encodes the modeler's prior assumptions on the latent function such as the smoothness and the scale of the variation, and it can be chosen freely as long as the covariance matrices it produces are symmetric and positive semi-definite. The properties of the prior are controlled by adjustable hyperparameters $\boldsymbol{\phi}_K$. All the experiments are done using a squared-exponential covariance function, which is infinitely differentiable and stationary meaning that it produces very smooth functions, which tend to the prior mean at the regions of the input space with no observations.

The model can be written in the general form defined in Section 2.2 as follows: The latent values are denoted as $\boldsymbol{\theta} = \mathbf{f}$ and the hyperparameters as $\boldsymbol{\phi} = \{\nu, \sigma^2, \boldsymbol{\phi}_K\}$. The likelihood sites for $i = 1, ..., n$ are defined as $t_i(\mathbf{U}_i^{\mathrm{T}}\boldsymbol{\theta}, \boldsymbol{\phi}_{\mathcal{A}_i}) = p(y_i|f_i, \nu, \sigma^2)$, where $\boldsymbol{\phi}_{\mathcal{A}_i} = \{\nu, \sigma^2\}$, $\mathbf{U}_i = \mathbf{e}_i$, and $\mathbf{e}_i$ is the i:th unit vector of $\boldsymbol{\theta} \in \mathbb{R}^n$. Only one prior site is defined for $i = n + 1$ as $t_{n+1}(\mathbf{U}_i^{\mathrm{T}}\boldsymbol{\theta}, \boldsymbol{\phi}_{\mathcal{A}_i}) = \mathcal{N}(\mathbf{0}, \mathbf{K})$, where $\mathbf{U}_i = \mathbf{I}$ and $\boldsymbol{\phi}_{\mathcal{A}_i} = \{\boldsymbol{\phi}_K\}$. With all approximate inference methods except MCMC, approximate integration is done only over $\boldsymbol{\theta}$ and type-II MAP estimates are used for the hyperparameters $\boldsymbol{\phi}$, which is common framework with GP models [Rasmussen and Williams, 2006, Nickisch and Rasmussen, 2008].

The challenge with the Student-t model is that the conditional posterior distribution of the latent function $p(\mathbf{f}|\mathcal{D}, \boldsymbol{\phi})$ values may contain multiple modes and that the potential outlying observations result in local increases in the approximate posterior uncertainty on the corresponding latent function values with the LA and EP approximations. The latter property can be seen as negative precision contributions in the approximate posterior covariances contrary to the always non-negative contributions with log-concave models such as the logit and probit used in binary GP-classification [Nickisch and Rasmussen, 2008]. More specifically, the scalar precision contributions related to outlying observations become negative, which with LA can be observed as negative values of $\mathbf{W}_i$ defined in (2.55), and with EP as negative site precisions $\tilde{\mathbf{T}}_i = \tilde{\tau}_i$ (2.30). This requires some additional care when implementing the LA and EP approximations following the standard algorithms described by Rasmussen and

Williams [2006] and can also result in clearly different behavior between the approximate methods (for comparisons, see the Examples 1 and 2 in Publication II).

The LA approximation (2.55) requires a robust and efficient method for determining the conditional mode $\hat{\boldsymbol{\theta}} = \hat{\mathbf{f}}$ of the latent function values given the hyperparameters, and a robust way for determining the approximate marginal likelihood (2.56) in case the Hessian $-\nabla^2_{\mathbf{f}} \log p(\mathbf{f}|\mathbf{X}, \boldsymbol{\phi}) = \mathbf{W} + \mathbf{K}$, where $\mathbf{W} = \mathrm{diag}(W_1, ..., W_n)$, is close to singular at the local mode, caused, e.g., by some large negative precision contributions W_i. Publication I describes a robust implementation of LA using the EM algorithm [see, e.g., Gelman et al., 2004] for determining the conditional mode and computational modifications based on controlled rank-one Cholesky updates that enable robust evaluation of the posterior covariance $\boldsymbol{\Sigma} = (\mathbf{K}^{-1} + \mathbf{W})^{-1}$ and the marginal likelihood approximation (2.56) in case $-\nabla^2_{\mathbf{f}} \log p(\mathbf{f}|\mathbf{X}, \boldsymbol{\phi})$ is poorly conditioned at the mode $\hat{\mathbf{f}}$. By experimental comparisons with MCMC and the commonly used VMF approximation [Tipping and Lawrence, 2003, Kuss, 2006] Publication I also shows that LA provides a good alternative for VMF in terms of speed and accuracy.

An interesting connection between the EM algorithm and a stabilized version of Newton's method with the Student-t model was brought forward by Hannes Nickisch (personal communication). The unconstrained Newton update step for the mode estimate $\hat{\mathbf{f}}$ can be written as

$$\hat{\mathbf{f}}^{\mathrm{new}} = (\mathbf{K}^{-1} + \mathbf{W})^{-1}(\mathbf{W}\hat{\mathbf{f}} + \nabla_{\mathbf{f}} \log p(\mathbf{y}|\mathbf{f}, \nu, \sigma^2)|_{\hat{\mathbf{f}}}), \tag{3.3}$$

where $\mathbf{W} = \mathrm{diag}(W_1, ..., W_n)$ contains the second order derivatives of the likelihood terms, $W_i = -\nabla^2_{f_i} \log p(y_i|f_i, \nu, \sigma^2)|_{\hat{f}_i}$, on its diagonal. Newton updates according to (3.3) can lead to overly large unconstrained steps and numerically unstable matrix computations, because W_i can become negative. This can be clarified by writing W_i as a difference between two non-negative terms as

$$W_i = \mathrm{E}(V_i^{-1}|y_i, \hat{f}_i, \nu, \sigma^2) - 2(v+1)r_i^2(r_i^2 + \nu\sigma^2)^{-2}, \tag{3.4}$$

where $r_i = y_i - \hat{f}_i$ and $\mathrm{E}(V_i^{-1}|y_i, \hat{f}_i, \nu, \sigma^2) = (\nu+1)(r_i^2 + \nu\sigma^2)^{-1} > 0$ are the expected inverse residual variances associated with the scale mixture representation of the Student-t model (see equations (10)–(12) in Publication I). If the Newton updates (3.3) are stabilized successively by adding a non-negative ridge $2(v+1)r_i^2(r_i^2 + \nu\sigma^2)^{-2} \geq 0$ to each W_i, the EM update of equation (13) in Publication I is recovered, which can be seen by noting

that $\nabla_{f_i} \log p(y_i|f_i, \sigma^2, \nu)|_{\hat{f}_i} = (\nu+1)r_i(r_i^2 + \nu\sigma^2)^{-1}$. Therefore the EM updates with the Student-t model are equivalent to Newton updates, where the potentially negative W_i are constrained to positive values.

At first, the implementation of the standard EP algorithm summarized in Section 2.3.4 seems straightforward for the Student-t model, because the prior term ($i = n+1$) is already in the Gaussian family conditioned on $\boldsymbol{\phi}$ and each of the Student-t sites depend only on a scalar $\mathbf{z}_i = f_i$. The moments of the tilted distributions can be approximated using one-dimensional quadrature integrations and all the cavity computations in (2.30) and the site parameter updates (2.36)–(2.37) can be done using only scalar operations. Integration over the hyperparameters ν and σ^2 could also be done using the framework presented in Section 2.3.4, but this would require three-dimensional quadrature integrations. Integration over the prior parameters $\boldsymbol{\phi}_K$ would be computationally challenging, because the prior site depends on all the latent values $\mathbf{f}$ and $\boldsymbol{\phi}_K$ nonlinearly, which would require potentially very high-dimensional numerical integrations over $\boldsymbol{\phi}_K$ to approximate the required tilted moments (2.25).

The practical stability problems with the non-log-concave Student-t sites arise from possible negative site precision changes $\Delta\tilde{\tau}_i = \delta\eta^{-1}(\hat{V}_i^{-1} - V_i^{-1})$ resulting from the EP updates (2.36). Large negative precision changes $\Delta\tilde{\tau}_i$ correspond to local increases of the posterior uncertainty ($\hat{V}_i > V_i$) and these are often related to multimodalities in the tilted distributions: For example, if the i:th observation is not clearly a regular observation nor an outlier, the tilted distribution can have two modes, one related to y_i and another related to the cavity $q_{-i}(f_i)$. After either a sequential or a parallel posterior update the negative changes $\Delta\tilde{\tau}_i$ may result in negative cavity variances in other sites, if the associated latent values are a priori correlated with the i:th site.

Publication II uses simple regression examples to illustrate these stability problems together with related convergence problems, which can be seen as nondecreasing oscillations of the site parameters during the iterations. Publication II also explains how damping and fractional updates can alleviate these problems, and describes a robust EP implementation based on parallel EP updates [van Gerven et al., 2009] that relies on provably convergent double-loop iterations to ensure convergence in difficult cases. In addition, several practical modifications are described to improve the robustness and to facilitate the computations with the parallel EP updates and the double-loop iterations. Finally the predictive perfor-

mance of EP is assessed by comparisons with VMF [Kuss, 2006], LVB [Gibbs and MacKay, 2000, Nickisch and Rasmussen, 2008], and MCMC [Neal, 1997, Gelman et al., 2004, Vanhatalo and Vehtari, 2007] using several real-world data sets. It is shown that compared with VMF, LVB, and LA, EP provides more accurate predictions in terms of mean log predictive densities with similar computational cost.

The double-loop formulation of Publication II can be obtained from the more general objective (2.48) by defining the sufficient statistics as $\mathbf{g}_i(\boldsymbol{\theta}) = [f_1, -\frac{1}{2}f_1^2, ..., f_n, -\frac{1}{2}, f_n^2]^{\mathrm{T}}$ and the site parameters as $\tilde{\boldsymbol{\lambda}}_i = \mathbf{e}_i \otimes [\tilde{\lambda}_{i,1}, \tilde{\lambda}_{i,1}]^{\mathrm{T}}$ for the likelihood sites $i = 1, ..., n$. The Gaussian prior site $p(\boldsymbol{\theta}) = t_{n+1}(\boldsymbol{\theta}) = \mathcal{N}(\mathbf{f}|\mathbf{0}, \mathbf{K})$ can be incorporated in the approximation $q(\boldsymbol{\theta})$ exactly by defining $q(\boldsymbol{\theta}) = Z(\boldsymbol{\lambda})^{-1} p(\boldsymbol{\theta}) \exp(\boldsymbol{\lambda}^{\mathrm{T}} \mathbf{g}(\boldsymbol{\theta}))$, where $Z(\boldsymbol{\lambda}) = \int p(\boldsymbol{\theta}) \exp(\boldsymbol{\lambda}^{\mathrm{T}} \mathbf{g}(\boldsymbol{\theta})) d\boldsymbol{\theta}$. The prior sites will be included implicitly in the tilted distributions $\hat{p}_i(\boldsymbol{\theta})$, because the surrogate distribution $q_s(\boldsymbol{\theta}) = \prod_{i=1}^{n} q_s(f_i)$ is kept consistent with the marginal distributions of $q(\boldsymbol{\theta})$ at the outer-loop updates.

3.2 Gaussian Process Classification with the Multinomial Probit

Publication III considers approximate inference in multi-class GP classification, which has been studied extensively using both the softmax [1] and the multinomial probit models. MCMC approaches have been described using both the softmax [Neal, 1998] and the multinomial probit models [Girolami and Rogers, 2006]. With the analytic approximation methods considered in Section 2, the softmax model has been used only with LA [Williams and Barber, 1998] and LVB [Chai, 2012], and the multinomial probit with EP [Seeger and Jordan, 2004, Seeger et al., 2006, Girolami and Zhong, 2007]. Alternative EP approaches based on threshold functions have been described by Kim and Ghahramani [2006] and Hernández-Lobato et al. [2011].

The main focus of Publication III is on accurate and computationally efficient EP implementation for the multinomial probit model for which the likelihood terms are defined as

$$p(y_i|\mathbf{f}_i) = \int \mathcal{N}(u_i|0,1) \prod_{j=1, j\neq y_i}^{c} \Phi(u_i + f_i^{y_i} - f_i^j) du_i, \tag{3.5}$$

where $\Phi(x)$ denotes the cumulative density function of the standard normal distribution, $\mathbf{f}_i = [f_i^1, ..., f_i^c]^{\mathrm{T}}$ contains the latent function values associated with the i:th observation $y_i \in \{1, 2, ...c\}$, which encodes the correct

[1] The softmax model is also known as the multinomial logistic model.

class label for input $\mathbf{x}_i$. The latent values related to all the observations are denoted as $\mathbf{f} = [f_1^1, ..., f_n^1, f_1^2, ..., f_n^2, ..., f_1^c, ..., f_n^c]^{\mathrm{T}}$, and they are given a zero-mean GP prior:

$$p(\mathbf{f}|\mathbf{X}, \boldsymbol{\phi}_K) = \mathcal{N}(\mathbf{f}|\mathbf{0}, \mathbf{K}), \tag{3.6}$$

where $\mathbf{K} = \mathrm{blkdiag}(\mathbf{K}^1, ..., \mathbf{K}^c)$ is a block-diagonal covariance matrix formed from the prior covariances $\mathbf{K}^j$ assigned to the latent values associated with each of the classes $j = 1, ..., c$. A squared-exponential covariance function with a common set of hyperparameters $\boldsymbol{\phi}_K$ is used for all $\mathbf{K}^j$, $j = 1, ..., c$, similarly to the LA approach described by Rasmussen and Williams [2006].

The model can be written in the general form defined in Section 2.2 as follows: The latent values are denoted as $\boldsymbol{\theta} = \mathbf{f}$ and the hyperparameters as $\boldsymbol{\phi} = \boldsymbol{\phi}_K$. The likelihood sites for $i = 1, ..., n$ are defined as $t_i(\mathbf{U}_i^{\mathrm{T}}\boldsymbol{\theta}, \boldsymbol{\phi}_{\mathcal{A}_i}) = p(y_i|\mathbf{f}_i)$, where $\boldsymbol{\phi}_{\mathcal{A}_i} = \emptyset$, and the transformation is given by $\mathbf{U}_i = \mathbf{I}_c \otimes \mathbf{e}_i$, where $\mathbf{e}_i$ is the i:th unit vector with n elements, and $\mathbf{I}_c$ a $c \times c$ identity matrix ($\mathbf{U}_i$ simply collects all the latent values associated with observation i from $\mathbf{f}$). Only one prior site is defined for $i = n + 1$ as $t_{n+1}(\mathbf{U}_i^{\mathrm{T}}\boldsymbol{\theta}, \boldsymbol{\phi}_{\mathcal{A}_i}) = \mathcal{N}(\mathbf{0}, \mathbf{K})$, where $\mathbf{U}_i = \mathbf{I}$ and $\boldsymbol{\phi}_{\mathcal{A}_i} = \{\boldsymbol{\phi}_K\}$. With all approximate inference methods except MCMC, approximate integration is done only over $\boldsymbol{\theta}$ and type-II MAP estimates are used for the hyperparameters $\boldsymbol{\phi}$ according to the commonly used inference framework with GP models [Rasmussen and Williams, 2006, Nickisch and Rasmussen, 2008].

The challenge with the multinomial probit model is that each likelihood term (3.5) and hence also each tilted distribution (2.20), depend on multiple latent values $\mathbf{f}_i$, which is why a straightforward EP implementation following Section 2.3.4 requires c-dimensional numerical integrals to approximate the required tilted moments (2.25). In addition, a straightforward adaptation of the algorithm of Section 2.3.4, would result in c-dimensional site precision structures $\tilde{\mathbf{T}}_i$ (2.36) and a $nc \times nc$ posterior covariance $\boldsymbol{\Sigma}$. This would become computationally prohibitive with large data set and many target classes. One possibility to facilitate the computations is to use a posterior approximation that can factored between the latent values related to the different classes, which requires at least one two-dimensional quadrature and $2c - 1$ one-dimensional quadratures per site update and results in a similar block-diagonal covariance matrix as the prior covariance $\mathbf{K}$ defined in (3.6) [Seeger et al., 2006, Girolami and Zhong, 2007].

Publication III proposes an alternative nested EP approach that does

not require quadratures, result in a similar computationally tractable posterior representation as the LA approach [Williams and Barber, 1998, Rasmussen and Williams, 2006], and represent accurately all between class posterior correlations between the latent values. The approach is based on utilizing the special structure of the multinomial probit likelihood terms (3.5) as follows. Each probit term in (3.5) depends on a linear combination of the latent values $\mathbf{f}_i$ and the auxiliary variable u_i, denoted from now on with $\mathbf{w}_i^{\mathrm{T}}\tilde{\mathbf{b}}_{i,j} = u_i + f_i^{y_i} - f_i^j$, where $\mathbf{w}_i = [\mathbf{f}_i^T, u_i]^T$ and the fixed transformation can be written as $\tilde{\mathbf{b}}_{i,j} = [(\mathbf{e}_{y_i} - \mathbf{e}_j)^T, 1]^T$ using c-dimensional unit vectors $\mathbf{e}_{y_i}$ and $\mathbf{e}_j$. Multiplying the cavity distributions $q_{-i}(\mathbf{f}_i)$ with the site (3.5) and removing the marginalization over u_i results in the following augmented tilted distribution:

$$\hat{p}(\mathbf{w}_i) = \hat{Z}_i^{-1}\mathcal{N}(\mathbf{w}_i|\boldsymbol{\mu}_{\mathbf{w}}, \boldsymbol{\Sigma}_{\mathbf{w}}) \prod_{j=1, j\neq y_i}^{c} \Phi(\mathbf{w}_i^T\tilde{\mathbf{b}}_{i,j}), \tag{3.7}$$

where $\boldsymbol{\mu}_{\mathbf{w}} = [\mathbf{m}_{-i}^T, 0]^T$, $\boldsymbol{\Sigma}_{\mathbf{w}} = \mathrm{blkdiag}(\mathbf{V}_{-i}, 1)$, and $\mathbf{V}_{-i}$ together with $\mathbf{m}_{-i}$ are defined in (2.30). The marginal mean and covariance of $\mathbf{f}_i$ with respect to $\hat{p}_i(\mathbf{w}_i)$ correspond to the required tilted moments (2.25) with respect to $\hat{p}_i(\mathbf{f}_i)$, and they can be approximated efficiently using an inner EP algorithm following Section 2.3.4.

Because the probit terms of (3.7) depend only on $c-1$ scalar random variables $z_{i,j} = \mathbf{w}_i^T\tilde{\mathbf{b}}_{i,j}$, each inner EP approximation can be represented using only $c-1$ scalar site precision parameters and $c-1$ scalar site location parameters according to the result (2.29). Publication III shows that by writing the expression of the outer EP site precision $\tilde{\mathbf{T}}_i$ using the scalar inner EP parameters, a similar posterior precision structure is obtained as with LA using the softmax model: $\tilde{\mathbf{T}}_i = \mathrm{diag}(\boldsymbol{\pi}_i) - \boldsymbol{\pi}_i(\mathbf{1}_c^{\mathrm{T}}\boldsymbol{\pi}_i)^{-1}\boldsymbol{\pi}_i^{\mathrm{T}}$, which enables posterior computations scaling as $\mathcal{O}(cn^3)$ (see, e.g, the implementation described by Rasmussen and Williams [2006]).

Because running the inner EP approximations until convergence for each of the n outer EP site approximations can become computationally demanding with large c, Publication III introduces also an incremental update scheme where damped updates are done on the scalar site parameters of the inner EP approximations only once at each outer EP iteration. By experimental comparisons with the non-incremental update scheme, it is demonstrated that the incremental scheme leads to convergence in comparable number of outer EP iterations. The experiments of Publication III show that nested EP produces more accurate approximations for the marginal distributions of the latent values compared with LA and

VMF, and that with fixed hyperparameters the approximate marginals are also very accurate compared with MCMC. Predictive comparisons between LA, VMF, EP, and MCMC using several real-world data sets show that EP is the most consistent method with respect MCMC in terms of predictive densities, but the differences are small if only classification accuracy is concerned.

3.3 Neural Network Regression with Sparsity-promoting Priors

Publication IV considers approximate EP inference with multi-layer perceptron (MLP) networks with sparsity-promoting hierarchial priors on the input weights. A similar inference problem was studied earlier by Mackay [1995] who described an automatic relevance determination (ARD) approach for Neural Networks (NNs), where individual relevance parameters are assigned to the weights associated with the different input features. Approximate inference on the network weights conditioned on the hyperparameters including the noise level and the feature relevance parameters, was performed using LA, and type-II MAP estimates of the hyperparameters were determined based on the approximate marginal likelihood (2.56). Another ARD approach was proposed by Neal [1996], where approximate MCMC integration is performed over all the model parameters including both the weights and the relevance hyperparameters. Williams [1995] described an alternative sparsity-favoring approach based on Laplace priors and LA approximation for the weights. Similarly to the classical Lasso regularization with linear models [Tibshirani, 1994], the mode-based LA approximation can produce truly sparse weight estimates without separate relevance hyperparameters.

As already discussed in Section 1, the main motivation for Publication IV is to study whether computationally efficient nonlinear predictors with flexible input priors could be constructed by adapting the existing EP methodology presented for sparse linear models [Seeger, 2008, Hernández-Lobato et al., 2008, van Gerven et al., 2009] to finite-parametric NNs with a linear input-layer. In contrast with the GP models studied in Sections 3.1 and 3.2, for which the posterior computations scale as $\mathcal{O}(n^3)$ and approximate integration over the hyperparameters requires either MCMC approximations or multidimensional grid-based methods [Rue et al., 2009, Cseke and Heskes, 2011], the NN approach could enable posterior computations scaling linearly in n and efficient EP integration

over the weights with sparsity-promoting priors.

Publication IV considers two-layer NNs where the unknown function value $f_i = f(\mathbf{x}_i)$ related to a d-dimensional input vector $\mathbf{x}_i$ is modeled as

$$f(\mathbf{x}_i) = \sum_{k=1}^{K} v_k g(\mathbf{w}_k^{\mathrm{T}} \mathbf{x}_i) + v_0 = \mathbf{v}^{\mathrm{T}} \mathbf{g}(\mathbf{h}_i), \tag{3.8}$$

where $g(x)$ is a nonlinear activation function, K the number of hidden units, and v_0 the output bias. Vector $\mathbf{w}_k = [w_{k,1}, w_{k,2}, ..., w_{k,d}]^{\mathrm{T}}$ contains the input layer weights related to hidden unit k and v_k is the corresponding output layer weight. The right-hand side of (3.8) is obtained by denoting the input-layer activations as $\mathbf{h}_i = \mathbf{h}(\mathbf{x}_i) = \tilde{\mathbf{x}}_i^{\mathrm{T}} \mathbf{w}$, where $\mathbf{w} = [\mathbf{w}_1^{\mathrm{T}}, ..., \mathbf{w}_K^{\mathrm{T}}]^{\mathrm{T}}$ collects all the input layer weights, $\tilde{\mathbf{x}}_i = \mathbf{I}_K \otimes \mathbf{x}_i$, and $\mathbf{g}(\mathbf{h}_i)$ applies the nonlinear transformation $g(x)$ on each component of $\mathbf{h}_i$ according to $\mathbf{g}(\mathbf{h}_i) = [g(h_{i,1}), g(h_{i,2}), ..., g(h_{i,K}), 1]^{\mathrm{T}}$ (the last element corresponds to the output bias v_0). Publication IV focuses on regression problems with scalar observations y_i distributed according to a Gaussian observation model $p(y_i|f_i, \sigma^2) = \mathcal{N}(y_i|f_i, \sigma^2)$, where σ^2 is the noise variance. A Gaussian hyperprior is assigned to $\theta = \log(\sigma^2)$: $p(\theta) = \mathcal{N}(\mu_{\phi,0}, \sigma_{\phi,0}^2)$, which corresponds to a log-normal prior for σ^2.

To construct flexible sparsity-promoting prior framework, hierarchical priors are assigned to the input layer weights, $p(w_j|\phi_{l_j})$, where w_j is the j:th element of $\mathbf{w}$, and ϕ_{l_j} is a joint hyperparameter controlling the prior variance of all input weights belonging to group $l_j \in \{1, ..., L\}$ (Index variable l_j defines the group in which the weight $\mathbf{w}_j$ belongs to). Gaussian hyperpriors are chosen for the hierarchical scale parameters: $p(\phi_l) = \mathcal{N}(\mu_{\phi,0}, \sigma_{\phi,0}^2)$, which requires that the prior variances of w_j are modeled using a suitable transform such as $\mathrm{Var}(w_j|\phi_{l_j}) = \exp(\phi_{l,j})$. The approach enables flexible definition of weight priors with different sparseness properties such as independent Laplace priors with a common scale parameter or Gaussian automatic relevance determination (ARD) priors with different relevance parameters for all inputs as described in Publication IV. To prevent potential unidentifiability problems resulting from symmetric activations functions $g(x) = -g(-x)$, the output weights are constrained to positive values by assigning left-truncated heavy-tailed priors to them: $p(v_k|\sigma_{v,0}^2) = 2t_\nu(v_k|0, \sigma_{v,0}^2)$, where $v_k \geq 0$ for $k = 1, ..., K$, and $t_\nu(v_k|0, \sigma_{v,0}^2)$ denotes a Student-t distribution with degrees of freedom ν, mean zero, and scale parameter $\sigma_{v,0}^2$. A zero-mean Gaussian prior with fixed variance is assigned to the output bias v_0.

The model can be written in the general form defined in Section 2.2 as

follows: The network weights are denoted with $\boldsymbol{\theta} = [\mathbf{w}^{\mathrm{T}}, \mathbf{v}^{\mathrm{T}}]$ and the hyperparameters with $\boldsymbol{\phi} = \{\theta, \phi_1, ..., \phi_L\}$. The likelihood sites for $i = 1, ..., n$ are defined as $t_i(\mathbf{U}_i^{\mathrm{T}}\boldsymbol{\theta}, \boldsymbol{\phi}_{\mathcal{A}_i}) = p(y_i|f_i, \sigma^2)$, where $\boldsymbol{\phi}_{\mathcal{A}_i} = \sigma^2$, and the transformation is given by $\mathbf{U}_i = \mathrm{blkdiag}(\tilde{\mathbf{x}}_i, \mathbf{I}_{K+1})$, where $\tilde{\mathbf{x}}_i = \mathbf{I}_K \otimes \mathbf{x}_i$, and $\mathbf{I}_{K+1}$ is a $K + 1 \times K + 1$ identity matrix. The transformed random variables $\mathbf{z}_i$ associated with the i:th likelihood site can now be written as $\mathbf{z}_i = \mathbf{U}_i^{\mathrm{T}}\boldsymbol{\theta} = [\mathbf{h}_i^{\mathrm{T}}, \mathbf{v}^{\mathrm{T}}]^{\mathrm{T}}$. The prior sites related to the input weights $\mathbf{w}$ are defined for $j = 1, ..., Kd$ and $i = n + j$ as $t_i(\mathbf{U}_i^{\mathrm{T}}\boldsymbol{\theta}, \boldsymbol{\phi}_{\mathcal{A}_i}) = p(w_j|\phi_{l_j})$, where $\mathbf{U}_i$ picks the j:th component of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}_{\mathcal{A}_i} = \{\phi_{l_j}\}$. Similarly, the prior sites related to the output weights are defined for $k = 0, ..., K$ and $i = n+Kd+k+1$ as $t_i(\mathbf{U}_i^{\mathrm{T}}\boldsymbol{\theta}, \boldsymbol{\phi}_{\mathcal{A}_i}) = p(v_k)$, where $\mathbf{U}_i$ picks the $(Kd + k + 1)$:th component of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}_{\mathcal{A}_i} = \emptyset$. Compared with the GP models considered in Publications I–III, a key technical difference in the NN approach of Publication IV is that EP approximations are formed, in addition to the non-Gaussian likelihood terms, also for the prior terms of the network weights. In addition, approximate EP integration is done simultaneously over all the hyperparameters $\boldsymbol{\phi} = \{\theta, \phi_1, ..., \phi_L\}$ using a factorized Gaussian approximation for each component of $\boldsymbol{\phi}$.

The challenge in the EP implementation is to construct a sufficiently accurate and computationally efficient Gaussian approximations for the likelihood terms that depend in non-linear manner from the $(2K + 1)$-dimensional transformed random variables $\mathbf{z}_i$. Previously, such Gaussian approximations for NN models have been formed using the extended Kalman filter (EKF) [de Freitas, 1999] and the unscented Kalman filter (UKF) [Wan and van der Merwe, 2000]. Alternative mean field approaches possessing similar characteristic with EP have been proposed by Opper and Winther [1996] and Winther [2001]. Similarly to Publication III, an EP algorithm for approximating the likelihood sites requires determining the marginal means and covariances of $\mathbf{z}_i$ with respect to the multivariate tilted distributions given by $\hat{p}_i(\mathbf{z}_i, \theta) \propto p(y_i|f_i, \theta)^{\eta} q_{-i}(\mathbf{z}_i) q_{-i}(\theta)$. Because determining these moments requires $(2K + 2)$-dimensional integrations, which may quickly become infeasible as K increases, Publication IV proposes a computationally more convenient posterior approximation that can be factored between the output weights $\mathbf{v}$ and the input weight $\mathbf{w}_k$ according to $q(\boldsymbol{\theta}) = q(\mathbf{v}) \prod_{k=1}^{K} q(\mathbf{w}_k)$. This approximation enables efficient computations of $\mathrm{E}(f_i)$, $\mathrm{Cov}(f_i)$, and $\mathrm{Cov}(\mathbf{v}, f_i)$ with respect to $q_{-i}(\boldsymbol{\theta}) = q_{-i}(\mathbf{v}) \prod_{k=1}^{K} q_{-i}(\mathbf{w}_k)$, which are subsequently used to approximate the tilted moments of $\mathbf{v}$ in a similar way as is done in the approxi-

mate linear filtering paradigm of the UKF filter [Wan and van der Merwe, 2000]. The tilted moments of the hidden unit activations $h_{i,k}$ are estimated by assuming $f_i = \sum_{k=1}^{K} v_k g(h_{i,k}) + v_0$ approximately normally distributed with respect to the cavity distributions $q_{-i}(\mathbf{v}, \mathbf{w})$ conditioned on one of the components of $\mathbf{h}_i$ at a time. A similar assumption was used by Ribeiro and Opper [2011] to form factorizing EP approximation for linear perceptrons.

Once reliable likelihood term approximations are obtained, forming EP approximations for the prior sites is rather straightforward and it can be done in a similarly way as with the existing EP approaches for sparse linear models. One practical requirement for efficient computations is that the conditional means $\mathrm{E}(w_j|\phi_{l_j})$ and variances $\mathrm{Var}(w_j|\phi_{l_j})$ with respect to the tilted distributions associated with the hierarchical prior sites can be computed analytically (otherwise two-dimensional quadratures are required).

Because of the previously described approximations, the resulting EP approach of Publication IV requires only one-dimensional numerical quadratures for determining the moments of the tilted distributions and results in a computationally efficient algorithm, whose complexity scales linearly with respect to both n and K. The complexity of the algorithm scales similarly to an ensemble of independent sparse linear models and also the resulting approximate predictive model can be interpreted as a nonlinear combination of independent sparse linear models associated with each hidden unit. Experiments with simulated regression problems demonstrate that the proposed approach enables robust integration over the posterior uncertainty of the input weights and the hierarchical scale parameters, and that the method can avoid potential overfitting problems related to point-estimate based ARD frameworks. Furthermore, the approach can learn strongly nonlinear input effects in multivariate regression problems and approximate the associated feature relevances correctly. However, in predictive comparisons using real-world data sets, the EP approach performs slightly worse compared with two alternative models with ARD priors including a NN inferred using MCMC [Neal, 1996], and an infinite GP network based on type-II MAP estimates of the relevance parameters [Rasmussen and Williams, 2006]. This behavior may be partly explained by the simple zero-initializations of the input weights used in the experiments, because good initializations of the weight values are known to be important in training NN models [see, e.g., Erhan et al., 2010].

4. Discussion

The introductory part of this thesis has reviewed various analytical methods for approximate Bayesian inference assuming a general and flexible predictive modeling framework. Both the theoretical properties of the methods and the practical accuracy of the resulting approximations have been discussed using the existing theoretical literature and the experimental results of Publications I–IV together with references to the experiments done by other authors. The main focus has been on describing the properties of EP and connecting the existing work on the method with the novel EP implementations proposed for approximate inference in the case studies considered in Publication II–IV.

One of the main arguments against the practical feasibility of the standard EP is the lack of formal convergence proof. The experiments of Publication II with the parallel-EP implementation that relies on convergent double-loop iterations in difficult cases together with the existing work on convergent double-loop algorithms [Minka, 2001c, Heskes and Zoeter, 2002, Opper and Winther, 2005, Seeger and Nickisch, 2011, Hernández-Lobato and Hernández-Lobato, 2011] show that by careful implementation accurate predictions can be obtained with EP also in multimodal inference problems with non-log-concave site functions. As further confirmation of the practical accuracy and efficiency of EP, it has recently been adopted in several machine learning toolboxes [Rasmussen and Nickisch, 2010, Nickisch, 2012, Minka et al., 2012, Vanhatalo et al., 2013]. Compared with LA, VMF, and LVB, EP achieves better predictive performance in several real-world data sets in the robust regression application studied in Publications I and II. Furthermore, EP is also able to quantify better the increased uncertainty on the latent function $f(\mathbf{x})$ in multimodal cases as demonstrated by simulated regression examples. However, in such cases the EP approximation can result in significant false uncertainty on

$f(\mathbf{x})$ in the input-space regions between and opposite sides of the modes in contrast to the possible false certainty provided by the LA or VB approximations that summarize only one of the modes depending on the initializations of the algorithms.

One of the main challenges in practical EP implementations is that determining the moments of the tilted distributions (2.20) may not be computationally feasible when the site terms depend on high-dimensional transformed variables $\mathbf{z}_i$ and/or on a large number of hyperparameters $\boldsymbol{\phi}_{\mathcal{A}_i}$. The commonly encountered multi-class classification is a good example of such inference problem. The nested EP approach proposed in Publication III is an appealing alternative for the existing approaches that rely on either multi-dimensional numerical quadratures or factored approximations to facilitate the inference [Seeger and Jordan, 2004, Girolami and Zhong, 2007]. The nested EP relies on an augmented integral representation of the multinomial probit model in a similar spirit as many VMF approximations [see, e.g., Tipping and Lawrence, 2003, Girolami and Rogers, 2006], but it can represent accurately all the posterior correlations between the latent variables in contrast with the factored VB approximations. Therefore, the same concept could be expanded also for other models where the site terms can be reformulated as integral representations of simple terms that depend only on one-dimensional random variables resulting from linear transformations of the model parameters. Furthermore, the multi-class EP approach of Publication III could be readily extended for linear models and the coefficients associated with the different classes could be coupled using the hierarchical prior framework described in Publication IV.

An appealing property of the GP models studied in Publications I–III is that they enable convenient integration over the latent function space with priors that correspond to infinitely complex models. For example, the squared exponential and the neural network covariance functions can be derived from a radial basis network and a multi-layer perceptron network, respectively, at the limit of infinitely many hidden units with Gaussian priors on the weights [Rasmussen and Williams, 2006]. Unfortunately, this flexibility does not come without a price. As discussed earlier, the inherent complexity of the posterior computations scale cubically with respect to the number of observations n. In addition, because of the complex functional dependencies between the latent values and the hyperparameters through the GP prior, approximate integration over the hyperparam-

eters can become challenging in problems with a large number of input features and feature-specific relevance parameters. On the other hand, with finite-parametric MLP networks studied in Publication IV, approximate inference on the likelihood sites using, e.g., EP or LA scales linearly with respect to n, and once a Gaussian likelihood approximation is determined, inference on the coefficient priors is straightforward with the existing methods proposed for linear models [see, e.g., Seeger, 2008, Hernández-Lobato et al., 2008, van Gerven et al., 2010]. Although the GP models achieved slightly better predictive accuracy compared with the EP-based finite networks in the experiments with real-world data, the EP approach performed relatively well in a very challenging estimation problem, if one keeps in mind the good results in the simulated experiments and the fact that not much emphasis was put on more elaborate initialization schemes, which are known to be very beneficial [see, e.g., Erhan et al., 2010]. Therefore the NN approach described in Publication IV presents an interesting alternative for the recently popular GP models, because the NN approach can be interpreted as an approximate framework that forms flexible nonlinear predictors from multiple sparse linear models and the approach can be extended to general activation functions or fixed interaction terms between the linear input layer models. The NN framework could be particularly well suited for problems where interpretable models with relative simple nonlinear latent functional dependencies and flexible hierarchical input priors are favored.

As an additional minor adaption of the methods described in Publication IV, the approximate EP integration over the hyperparameters associated with the likelihood sites could be readily extended for GP models, where the likelihood sites depend only on one or two hyperparameters ϕ and the tilted moment integrations over the latent values f_i conditioned ϕ can be done analytically.[1] Another possible extension of the NN framework are multi-layered (deep) models, which could be approximated in the same fashion by approximating the hidden unit activations in each layer independent of each other. In such cases, more elaborate initialization schemes such as the ones reviewed by Erhan et al. [2010] become probably essential.

[1] Similar ideas have already been proposed for linear models by Hernández-Lobato et al. [2008].

Bibliography

Hagai Attias. A variational Bayesian framework for graphical models. In *Advances in Neural Information Processing Systems 12*, pages 209–215. MIT Press, 2000.

Thomas Bayes. An essay towards solving a problem in the doctrine of chances. *Philosphical Transactions*, 53:370–418, 1763.

Matthew J. Beal. *Variational algorithms for approximate Bayesian inference*. PhD thesis, The Gatsby Computational Neuroscience Unit, University College London, 2003.

José M. Bernardo and Adrian F. M. Smith. *Bayesian Theory*. John Wiley & Sons, Ltd., 2000.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer Science +Business Media, LLC, 2006.

Kian Ming A. Chai. Variational multinomial logit Gaussian process. *Journal of Machine Learning Research*, 13:1745–1808, 2012.

Botond Cseke and Tom Heskes. Approximate marginals in latent Gaussian models. *Journal of Machine Learning Research*, 12:417–454, 2011.

Joao F. G. de Freitas. *Bayesian Methods for Neural Networks*. PhD thesis, University of Cambridge, 1999.

Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio. Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*, 11:625–660, 2010.

Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, second edition, 2004.

John Geweke. Bayesian treatment of the independent Student-t linear model. *Journal of Applied Econometrics*, 8:519–540, 1993.

Mark N. Gibbs and David J. C. MacKay. Variational Gaussian process classifiers. In *IEEE Transactions on Neural Networks*, pages 1458–1464, 2000.

Mark Girolami and Simon Rogers. Variational Bayesian multinomial probit regression with Gaussian process priors. *Neural Computation*, 18:1790–1817, 2006.

Mark Girolami and Mingjun Zhong. Data integration for classification problems employing Gaussian process priors. In *Advances in Neural Information Processing Systems 19*, pages 465–472. The MIT Press, 2007.

Daniel Hernández-Lobato, José M. Hernández-Lobato, and A. Suárez. Expectation propagation for microarray data classification. *Pattern Recognition Letters*, 31(12):1618–1626, 2010.

Daniel Hernández-Lobato, José M. Hernández-Lobato, and Pierre Dupont. Robust multi-class Gaussian process classification. In *Advances in Neural Information Processing Systems 24*, pages 280–288, 2011.

José M. Hernández-Lobato and Daniel Hernández-Lobato. Convergent expectation propagation in linear models with spike-and-slab priors. Technical report, arXiv:1112.2289 [stat.ML], 2011.

José M. Hernández-Lobato, Tjeerd Dijkstra, and Tom Heskes. Regulator discovery from gene expression time series of malaria parasites: a hierarchical approach. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 649–656, Cambridge, MA, 2008. MIT Press.

Tom Heskes and Onno Zoeter. Expectation propagation for approximate inference in dynamic Bayesian networks. In A. Darwiche and N. Friedman, editors, *Uncertainty in Artificial Intelligence: Proceedings of the Eighteenth Conference (UAI-2002)*, pages 216–233, San Francisco, CA, 2002. Morgan Kaufmann Publishers.

Tom Heskes, Manfred Opper, Wim Wiegerinck, Ole Winther, and Onno Zoeter. Approximate inference techniques with expectation constraints. *Journal of Statistical Mechanics: Theory and Experiment*, 2005:P11015, 2005.

Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. arXiv:1207.0580 [cs.NE], 2012.

Antti Honkela, Tapani Raiko, Mikael Kuusela, Matti Tornio, and Juha Karhunen. Approximate riemannian conjugate gradient learning for fixed-form variational Bayes. *Journal of Machine Learning Research*, 11:3235–3268, 2010.

Tommi S. Jaakkola. Tutorial on variational approximation methods. In *Advanced Mean Field Methods: Theory and Practice*, pages 129–159. MIT Press, 2000.

Tommi S. Jaakkola and Michael I. Jordan. A variational approach to Bayesian logistic regression models and their extensions. In *Proceedings of the Sixth International Workshop on Artificial Intelligence and Statistics*, 1996.

Tommi S. Jaakkola and Michael I. Jordan. Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10:25–37, 2000.

Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. In *Machine Learning*, pages 183–233. MIT Press, 1999.

Jari Kaipio and Erkki Somersalo. *Statistical and computational inverse problems*. Springer Science+Business Media, third edition, 2005.

Emtiyaz Khan, Shakir Mohamed, and Kevin Murphy. Fast Bayesian inference for non-conjugate Gaussian process regression. In P. Bartlett, F.C.N. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 3149–3157. 2012.

Hyun-Chul Kim and Zoubin Ghahramani. Bayesian Gaussian process classification with the EM-EP algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):1948–1959, 2006.

David A. Knowles and Tom Minka. Non-conjugate variational message passing for multinomial and binary regression. In J. Shawe-Taylor, R.S. Zemel, P. Bartlett, F.C.N. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 1701–1709, 2011.

Malte Kuss. *Gaussian Process Models for Robust Regression, Classification, and Reinforcement Learning*. PhD thesis, Technische Universität Darmstadt, 2006.

Jouko Lampinen and Aki Vehtari. Bayesian approach for neural networks - review and case studies. *Neural Networks*, 14(3):7–24, 2001.

Pierre Simon Laplace. Mémoire sur la probabilité des causes par les évènemens. *Mémoires de Mathématique et de Physique, Tome Sixiéme*, pages 621–656, 1774. English translation by S. M. Stiegler 1986. Memoir on the probability of causes of events, *Statistical Science*, 1(19):364-378.

Chuanhai Liu and Donald B. Rubin. ML estimation of the t distribution using EM and its extensions, ECM and ECME. *Statistica Sinica*, 5:19–39, 1995.

David J. C. Mackay. Probable networks and plausible predictions – a review of practical Bayesian methods for supervised neural networks. *Network: Computation in Neural Systems*, 6(3):469–505, 1995.

Thomas Minka. Bayesian linear regression. Technical report, Massachusetts Institute of Technology, 2000.

Thomas Minka. *A Family of Algorithms for Approximate Bayesian Inference*. PhD thesis, Massachusetts Institute of Technology, 2001a.

Thomas Minka. Expectation propagation for approximate Bayesian inference. In *UAI '01: Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, pages 362–369, San Francisco, CA, USA, 2001b. Morgan Kaufmann Publishers Inc.

Thomas Minka. The EP energy function and minimization schemes. Technical report, 2001c.

Thomas Minka. Power EP. Technical report, Microsoft Research, Cambridge, 2004.

Thomas Minka. Divergence measures and message passing. Technical report, Microsoft Research, Cambridge, 2005.

Thomas Minka and John Lafferty. Expectation propagation for the generative aspect model. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*, pages 352–359. Morgan Kaufmann, 2002.

Thomas Minka, John Winn, John Guiver, and David Knowles. Infer.NET 2.5, 2012. Microsoft Research Cambridge. http://research.microsoft.com/infernet.

Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

Radford M. Neal. *Bayesian Learning for Neural Networks*. Springer, 1996.

Radford M. Neal. Monte Carlo Implementation of Gaussian Process Models for Bayesian Regression and Classification. Technical Report 9702, Dept. of statistics and Dept. of Computer Science, University of Toronto, 1997.

Radford M. Neal. Regression and classification using Gaussian process priors (with discussion). In *Bayesian Statistics 6*, pages 475–501. Oxford University Press, 1998.

Hannes Nickisch. *Bayesian Inference and Experimental Design for Large Generalised Linear Models*. PhD thesis, Technische Universität Berlin, Berlin, 2010.

Hannes Nickisch. glm-ie: Generalised linear models inference & estimation toolbox. *Journal of Machine Learning Research*, 13:1699–1703, 2012.

Hannes Nickisch and Carl E. Rasmussen. Approximations for binary Gaussian process classification. *Journal of Machine Learning Research*, 9:2035–2078, October 2008.

Anthony O'Hagan and Jonathan Forster. *Kendals Advanced Theory of Statistics, Volume 2B: Bayesian Inference*. Arnold, second edition, 2004.

Manfred Opper and Cédric Archambeau. The variational Gaussian approximation revisited. *Neural Computation*, 21(3):786–792, 2009.

Manfred Opper and Ole Winther. Mean field approach to Bayes learning in feedforward neural networks. *Physical Review Letters*, 76:1964–1967, Mar 1996.

Manfred Opper and Ole Winther. Expectation Consistent Approximate Inference. *Journal of Machine Learning Research*, 6:2177–2204, 2005.

Ulrich Paquet, Ole Winther, and Manfred Opper. Perturbation corrections in approximate inference: Mixture modelling applications. *Journal of Machine Learning Research*, 10:1263–1304, 2009.

Yuan (Alan) Qi, Thomas P. Minka, Rosalind W. Picard, and Zoubin Ghahramani. Predictive automatic relevance determination by expectation propagation. In *Proceedings of Twenty-first International Conference on Machine Learning*, pages 671–678, 2004.

Joaquin Quiñonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6(3):1939–1959, December 2005.

Tapani Raiko, Harri Valpola, Markus Harva, and Juha Karhunen. Building blocks for variational Bayesian learning of latent variable models. *Journal of Machine Learning Research*, 8:155–201, 2007.

Carl E. Rasmussen and Hannes Nickisch. Gaussian processes for machine learning (GPML) toolbox. *Journal of Machine Learning Research*, 11:3011–3015, 2010.

Carl E. Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

Fabiano Ribeiro and Manfred Opper. Expectation propagation with factorizing distributions: A Gaussian approximation and performance results for simple models. *Neural computation*, 23(4):1047–1069, 2011.

Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer, second edition, 2004.

Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal statistical Society (Series B)*, 71(2):1–35, 2009.

Simo Särkkä. *Recursive Bayesian Inference on Stochastic Differential Equations*. PhD thesis, Helsinki University of Technology, Report B54, 2006.

Matthias Seeger. Expectation propagation for exponential families. Technical report, Max Planck Institute for Biological Cybernetics, Tübingen, Germany, 2005.

Matthias Seeger. Bayesian inference and optimal design for the sparse linear model. *Journal of Machine Learning Research*, 9:759–813, 2008.

Matthias Seeger and Michael Jordan. Sparse Gaussian process classification with multiple classes. Technical report, University of California, Berkeley, CA, 2004.

Matthias Seeger and Hannes Nickisch. Fast convergent algorithms for expectation propagation approximate Bayesian inference. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 652–660. JMLR W&CP, vol. 15, 2011.

Matthias Seeger and David Wipf. Variational Bayesian inference techniques. *IEEE Signal Processing Magazine*, 27(6):81–91, 2010.

Matthias Seeger, Neil Lawrence, and Ralf Herbrich. Efficient nonparametric Bayesian modelling with sparse Gaussian process approximations. Technical report, Max Planck Institute for Biological Cybernetics, Tübingen, Germany, 2006.

Matthias Seeger, Sebastian Gerwinn, and Matthias Bethge. Bayesian inference for sparse generalized linear models. In *European Conference on Machine Learning,ECML*, pages 298–309, 2007.

Alexander Smola, Vishy Vishwanathan, and Eleazar Eskin. Laplace propagation. In Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

Luke Tierney and Joseph B. Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86, 1986.

Michael E. Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, dec 2001.

Michael E. Tipping and Neil D. Lawrence. A variational approach to robust Bayesian interpolation. In *Proceedings of the IEEE International Workshop on Neural Networks for Signal Processing.*, pages 229–238. IEEE, 2003.

Marcel van Gerven, Botond Cseke, Robert Oostenveld, and Tom Heskes. Bayesian source localization with the multivariate Laplace prior. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1901–1909, 2009.

Marcel van Gerven, Botond Cseke, Floris de Lange, and Tom Heskes. Efficient Bayesian multivariate fMRI analysis using a sparsifying spatio-temporal prior. *NeuroImage*, 50:150–161, 2010.

Jarno Vanhatalo and Aki Vehtari. Sparse log Gaussian processes via MCMC for spatial epidemiology. *JMLR Workshop and Conference Proceedings*, 1:73–89, 2007.

Jarno Vanhatalo, Ville Pietiläinen, and Aki Vehtari. Approximate inference for disease mapping with sparse Gaussian processes. *JMLR Workshop and Conference Proceedings*, 29(15):1580–1607, 2010.

Jarno Vanhatalo, Jaakko Riihimäki, Jouni Hartikainen, Pasi Jylänki, Ville Tolvanen, and Aki Vehtari. GPstuff: Bayesian modeling with gaussian processes. *Journal of Machine Learning Research*, 14:1175–1179, 2013.

Aki Vehtari and Janne Ojanen. A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142–228, 2012.

Eric A. Wan and Rudolph van der Merwe. The unscented Kalman filter for nonlinear estimation. In *Proceedings of IEEE Symposium on Adaptive Systems for Signal Processing, Communications, and Control (AS-SPCC)*, pages 153–158, 2000.

Mike West. Outlier models and prior distributions in Bayesian linear regression. *Journal of the Royal Statistical Society (Series B)*, 46(3):431–439, 1984.

Christopher K. I. Williams. Computation with infinite neural networks. *Neural Computation*, 10(5):1203–1216, 1998.

Christopher K. I. Williams and David Barber. Bayesian classification with Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1342–1351, 1998.

Peter M. Williams. Bayesian regularisation and pruning using a Laplace prior. *Neural Computation*, 7(1):117–143, 1995.

John Winn and Christopher M. Bishop. Variational message passing. *Journal of Machine Learning Research*, 6:661–694, 2005.

Ole Winther. Computing with finite and infinite networks. In *Advances in Neural Information Processing Systems 13 (NIPS'2000)*, pages 336–342. MIT press, 2001.

David Wipf and Srikantan Nagarajan. A new view of automatic relevance determination. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1625–1632. MIT Press, Cambridge, MA, 2008.

David Wipf, Bhaskar D. Rao, and Srikantan Nagarajan. Latent variable Bayesian models for promoting sparsity. *Information Theory, IEEE Transactions on*, 57(9):6236 –6255, sept. 2011.

Aalto-DD 152/2013

ISBN 978-952-60-5354-7
ISBN 978-952-60-5355-4 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934
ISSN 1799-4942 (pdf)

Aalto University
School of Science
Department of Biomedical Engineering and Computational Science

BUSINESS + ECONOMY

ART + DESIGN + ARCHITECTURE

SCIENCE + TECHNOLOGY

CROSSOVER

DOCTORAL DISSERTATIONS