Fast Convergent Algorithms for Expectation Propagation Approximate Bayesian Inference

Matthias Seeger
Ecole Polytechnique Fédérale de Lausanne
INR 112, Station 14
1015 Lausanne, Switzerland

Hannes Nickisch
Max Planck Institute for Intelligent Systems
Spemannstraße 38
72076 Tübingen, Germany

Abstract

We propose a novel algorithm to solve the expectation propagation relaxation of Bayesian inference for continuous-variable graphical models. In contrast to most previous algorithms, our method is provably convergent. By marrying convergent EP ideas from [15] with covariance decoupling techniques [23, 13], it runs at least an order of magnitude faster than the most common EP solver.

1 Introduction

A growing number of challenging machine learning applications require decision-making from incomplete data (e.g., stochastic optimization, active sampling, robotics), which relies on quantitative representations of uncertainty (e.g., Bayesian posterior, belief state) and is out of reach of the commonly used paradigm of learning as point estimation on hand-selected data. While Bayesian inference is harder than point estimation in general, it can be relaxed to *variational* optimization problems which can be computationally competitive, if only they are treated with the algorithmic state-of-the-art established for the latter.

In this paper, we propose a novel algorithm for the expectation propagation (EP; or adaptive TAP, or expectation consistent (EC)) relaxation [14, 11, 15], which is both much faster than the commonly used sequential EP algorithm, and is provably convergent (the sequential algorithm lacks such a guarantee). Our method builds on the convergent double loop algorithm of [15], but runs orders of magnitude faster. We gain a deeper understanding of EP (or EC) as optimization problem, unifying it with covariance decoupling ideas [23, 13], and allowing for "point estimation" algorithmic progress to be brought to bear on this powerful approximate inference formulation.

Suppose that observations $\boldsymbol{y} \in \mathbb{R}^m$ are modelled as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{u} + \boldsymbol{\varepsilon}$, where $\boldsymbol{u} \in \mathbb{R}^n$ are latent variables of interest, $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ is Gaussian noise, and $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ is the design matrix. For example, $\boldsymbol{u}$ can be an image to be reconstructed from $\boldsymbol{y}$ (e.g., Fourier coefficients in magnetic resonance imaging [22]), further examples are found in [18]. The prior distribution has the form $P(\boldsymbol{u}) \propto \prod_{i=1}^q t_i(s_i)$ with *non-Gaussian* potentials $t_i(\cdot)$, and $\boldsymbol{s} := \boldsymbol{B}\boldsymbol{u}$ for a matrix $\boldsymbol{B}$. A well-known example are *Laplace sparsity priors* defined by $t_i(s_i) = e^{-\tau_i|s_i|}$ [18], where $\boldsymbol{B}$ collects simple filters (e.g., derivatives, wavelet coefficients). This formal setup also encompasses binary classification ($\boldsymbol{u}$ classifier weights, $\prod_{i=1}^q t_i(s_i)$ the classification likelihood [13]) or spiking neuron models [5]. The posterior distribution is

$$P(\boldsymbol{u}|\boldsymbol{y}) = Z^{-1} N(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{u}, \sigma^2\boldsymbol{I}) \prod_{i=1}^q t_i(s_i), \quad (1)$$

$Z := \int N(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{u}, \sigma^2\boldsymbol{I}) \prod_{i=1}^q t_i(s_i)\, d\boldsymbol{u}$ the *partition function* for $P(\boldsymbol{u}|\boldsymbol{y})$, and $\boldsymbol{s} = \boldsymbol{B}\boldsymbol{u}$. *Bayesian inference* amounts to computing moments of $P(\boldsymbol{u}|\boldsymbol{y})$ and/or $\log Z$. Hyperparameters $\boldsymbol{f}$ can be learned by maximizing $\log Z(\boldsymbol{f})$ [10] (e.g., motion deblurring by blind deconvolution [9]). In Bayesian experimental design (or active learning) [13], $\boldsymbol{X}$ is built up sequentially by greedily maximizing expected information scores. These applications require posterior covariance information beyond any single point estimate.

The expectation propagation relaxation along with known algorithms is described in Section 2, scalable inference techniques are reviewed in Section 3. We develop our novel algorithm in Section 4, provide a range of real-world experiments (image deblurring and reconstruction) in Section 5, and close with a discussion (Section 6). Upon publication, code for our algorithm will be released into the public domain.

Appearing in Proceedings of the 14^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2011, Ft. Lauderdale, FL, USA.

2 Expectation Propagation

Expectation propagation (EP) [11, 15] stands out among variational inference approximations. First, it is more generally applicable than most others. Second, a range of empirical studies indicate that EP can be a far more accurate approximation to Bayesian inference than today's competitors of comparable running time [7, 12]. Consequently, EP has been applied to a wide range of problems: binary and multi-way classification [11, 20], neuronal spiking models [5], ordinal regression [2], semi-supervised learning [8], sparse linear models and ICA [6, 18], and inference in Ising models [15] (the algorithm developed in this paper can be applied in all these cases).[1] On the other hand, EP is more difficult to handle than most other methods, for a number of reasons. It is not an optimization problem based on a bound on $\log Z$ (1), but constitutes a search for a saddle point [15]. Moreover, its stationary equations are more complicated in structure than commonly used bounds. Finally, running EP can be numerically challenging [18, 1].

In the sequel, we describe the variational optimization problem behind (fractional) EP, details can be found in [11, 15, 18]. The goal is to fit the posterior distribution $P(\boldsymbol{u}|\boldsymbol{y})$ from (1) by a *Gaussian* of the form

$$Q(\boldsymbol{u}|\boldsymbol{y}) := Z_Q^{-1} N(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{u}, \sigma^2\boldsymbol{I}) e^{\boldsymbol{b}^T\boldsymbol{s} - \frac{1}{2}\boldsymbol{s}^T(\mathrm{diag}\,\boldsymbol{\pi})\boldsymbol{s}},$$
$$\mathrm{Cov}_Q[\boldsymbol{u}|\boldsymbol{y}]^{-1} = \boldsymbol{A} := \sigma^{-2}\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{B}^T(\mathrm{diag}\,\boldsymbol{\pi})\boldsymbol{B}, \quad (2)$$

where $Z_Q := \int N(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{u}, \sigma^2\boldsymbol{I}) e^{\boldsymbol{b}^T\boldsymbol{s} - \frac{1}{2}\boldsymbol{s}^T(\mathrm{diag}\,\boldsymbol{\pi})\boldsymbol{s}}\, d\boldsymbol{u}$, $\boldsymbol{s} = \boldsymbol{B}\boldsymbol{u}$. $Q(\boldsymbol{u}|\boldsymbol{y})$ depends on the variational parameters $\boldsymbol{b}$ and $\boldsymbol{\pi} \succeq \boldsymbol{0}$, collected as $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{b})$ below. Let marginal distributions $N(\mu_i, \rho_i)$ be indexed by moment parameters $\boldsymbol{\mu}$, $\boldsymbol{\rho}$, $\eta \in (0, 1]$ a fractional parameter (while standard EP uses $\eta = 1$, $\eta < 1$ can improve numerical stability of Gaussian EP [18]). For $i \in \{1, \ldots, q\}$, denote $\kappa_i = \kappa_i(s_i) := b_i s_i - \frac{1}{2}\pi_i s_i^2$. The *cavity marginal* is $Q_{-i}(s_i) \propto N(s_i|\mu_i, \rho_i) e^{-\eta\kappa_i}$, the *tilted marginal* $\hat{P}_i(s_i) \propto Q_{-i}(s_i) t_i(s_i)^\eta$. While $\hat{P}_i(s_i)$ is not a Gaussian, its moments (mean and variance) can be computed tractably. An EP fixed point $(\boldsymbol{\pi}, \boldsymbol{b})$ satisfies *expectation consistency* [15]: if $N(\mu_i, \rho_i) = Q(s_i|\boldsymbol{y})$, then $\hat{P}_i(s_i)$ and $Q(s_i|\boldsymbol{y})$ have the same mean and variance for all $i = 1, \ldots, q$. The corresponding (negative free) energy function is

$$\phi(\boldsymbol{\pi}, \boldsymbol{b}, \boldsymbol{\mu}, \boldsymbol{\rho}) := -2\log Z_Q$$
$$- \tfrac{2}{\eta}\sum\nolimits_{i=1}^{q} \left(\log \mathrm{E}_{Q_{-i}}[t_i(s_i)^\eta] - \log \mathrm{E}_{Q_{-i}}[e^{\eta\kappa_i}]\right),$$

where Z_Q is the partition function of $Q(\boldsymbol{u}|\boldsymbol{y})$ (see Eq. 2). If we define $\boldsymbol{\mu}$, $\boldsymbol{\rho}$ in terms of $\boldsymbol{\pi}$, $\boldsymbol{b}$ (by requiring that $N(\mu_i, \rho_i) = Q(s_i|\boldsymbol{y})$), it is easy to see that $\nabla_{\boldsymbol{\pi}}\phi = \nabla_{\boldsymbol{b}}\phi = \boldsymbol{0}$ implies expectation consistency. However, this dependency tends to be broken intermediately in most EP algorithms. A schematic overview of the expectation consistency conditions is as follows (notations $\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_-, \boldsymbol{s}_*, \boldsymbol{z}$ are introduced in subsequent sections; $\overset{\mathrm{MM}}{\longleftrightarrow}$ denotes Gaussian moment matching):

$$\begin{array}{ccc} \overbrace{N(\mu_i, \rho_i)}^{\tilde{\boldsymbol{\theta}} \leftrightarrow (\boldsymbol{\mu}, \boldsymbol{\rho})} & \rightarrow & \overbrace{Q_{-i}(s_i) \propto N(s_i|\mu_i, \rho_i) e^{-\eta\kappa_i}}^{\boldsymbol{\theta}_- (= \tilde{\boldsymbol{\theta}} - \eta\boldsymbol{\theta})} \\ & & \downarrow \\ \underbrace{Q(s_i|\boldsymbol{y})}_{= N(s_{*i}, z_i)} & \overset{\mathrm{MM}}{\longleftrightarrow} & \hat{P}_i(s_i) \propto Q_{-i}(s_i) t_i(s_i)^\eta \end{array} \quad (3)$$

The total criterion $\phi(\boldsymbol{\pi}, \boldsymbol{b}, \boldsymbol{\mu}(\boldsymbol{\pi}, \boldsymbol{b}), \boldsymbol{\rho}(\boldsymbol{\pi}, \boldsymbol{b}))$ is neither convex nor concave [15].

The most commonly used *sequential* EP algorithm visits each potential $i \in \{1, \ldots, q\}$ in turn, first updating μ_i, ρ_i, then π_i, b_i based on one iteration[2] of $\partial_{\pi_i}\phi = \partial_{b_i}\phi = 0$ [11, 15]. For models of moderate size n, a numerically robust implementation maintains the inverse covariance matrix $\boldsymbol{A}$ (2) as representation of $Q(\boldsymbol{u}|\boldsymbol{y})$. A sweep over all potentials costs $O(q\,n^2)$. If memory costs of $O(n^2)$ are prohibitive, we can determine μ_i, ρ_i on demand by solving a linear system with $\boldsymbol{A}$, in which case a sweep requires q such systems. The sequential EP algorithm is too slow to be useful for many applications. Notably, all publications for EP we are aware of (with the exception of two references discussed in the sequel) employ this method, generally known as "the EP algorithm".

In [4], a *parallel* variant of EP is applied to rather large models of a particular structure. They alternate between updates of all $\boldsymbol{\mu}$, $\boldsymbol{\rho}$ and all $\boldsymbol{\pi}$, $\boldsymbol{b}$, the latter by one iteration of $\partial_{\boldsymbol{\pi}}\phi = \partial_{\boldsymbol{b}}\phi = \boldsymbol{0}$ (these equations decouple w.r.t. $i = 1, \ldots, q$). The most expensive step per iteration by far is the computation of marginal variances $\boldsymbol{\rho}$, which is feasible only for the very sparse matrices $\boldsymbol{A}$ specific to their application. Neither sequential nor parallel algorithm come with a convergence proof.

A provably convergent double loop algorithm for EP is given by Opper&Winther in [15]. For its derivation, we need to consider a *natural parameterization* of the problem. The underlying reason for this is that log partition functions like $\log Z_Q$ (2) are simple convex functions in natural parameters, and derivatives w.r.t. the latter result in posterior expectations. Collect $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{b})$ and recall that $\kappa_i = b_i s_i - \frac{1}{2}\pi_i s_i^2$. Let $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\pi}}, \tilde{\boldsymbol{b}})$ be natural parameters corresponding to $\boldsymbol{\mu}, \boldsymbol{\rho}$ ($\tilde{\pi}_i = 1/\rho_i$, $\tilde{b}_i = \mu_i/\rho_i$), and $\tilde{\kappa}_i = \tilde{b}_i s_i - \frac{1}{2}\tilde{\pi}_i s_i^2$, so that $N(s_i|\mu_i, \rho_i) = Z_i^{-1} e^{\tilde{\kappa}_i}$, where $Z_i = \int e^{\tilde{\kappa}_i}\, ds_i$ is the

[1] A comprehensive bibliography can be found at `http://research.microsoft.com/en-us/um/people/minka/papers/ep/roadmap.html`.

[2] "One iteration" means solving for π_i, b_i, assuming that the cavity distribution $Q_{-i}(s_i)$ is fixed (ignoring its dependence on π_i, b_i).

normalization constant. With $\boldsymbol{\theta}_- = (\boldsymbol{\pi}_-, \boldsymbol{b}_-) = \tilde{\boldsymbol{\theta}} - \eta\boldsymbol{\theta}$ and $\kappa_{-i} = b_{-i}s_i - \frac{1}{2}\pi_{-i}s_i^2 = \tilde{\kappa}_i - \eta\kappa_i$, we have that $Q_{-i}(s_i) \propto e^{\kappa_{-i}}$ and $\hat{P}_i(s_i) = \hat{Z}_i^{-1}e^{\kappa_{-i}}t_i(s_i)^\eta$ with $\hat{Z}_i = \int e^{\kappa_{-i}}t_i(s_i)^\eta\, ds_i$. If $\phi_\cap(\boldsymbol{\theta}_-, \tilde{\boldsymbol{\theta}}) := -\frac{2}{\eta}\sum_i \log\hat{Z}_i - 2\log Z_Q$ and $\phi_\cup(\tilde{\boldsymbol{\theta}}) := \frac{2}{\eta}\sum_i \log Z_i$, we have that $\phi(\boldsymbol{\theta}_-, \tilde{\boldsymbol{\theta}}) = \phi_\cap(\boldsymbol{\theta}_-, \tilde{\boldsymbol{\theta}}) + \phi_\cup(\tilde{\boldsymbol{\theta}})$, where $\phi_\cap(\boldsymbol{\theta}_-, \tilde{\boldsymbol{\theta}})$ is jointly concave[3], while $\phi_\cup(\tilde{\boldsymbol{\theta}})$ is convex. Define $\phi(\tilde{\boldsymbol{\theta}}) := \max_{\boldsymbol{\theta}_-}\phi(\boldsymbol{\theta}_-, \tilde{\boldsymbol{\theta}})$. The Opper&Winther algorithm (locally) minimizes $\phi(\tilde{\boldsymbol{\theta}})$ via two nested loops. The inner loop (IL) is the concave maximization $\boldsymbol{\theta}_- \leftarrow \operatorname{argmax}_{\boldsymbol{\theta}_-}\phi_\cap(\boldsymbol{\theta}_-, \tilde{\boldsymbol{\theta}})$ for fixed $\tilde{\boldsymbol{\theta}}$. An outer loop (OL) iteration consists of an IL followed by an update of $\tilde{\boldsymbol{\theta}}$: $\boldsymbol{\mu} \leftarrow \mathrm{E}_Q[\boldsymbol{s}|\boldsymbol{y}]$, $\boldsymbol{\rho} \leftarrow \mathrm{Var}_Q[\boldsymbol{s}|\boldsymbol{y}]$. Within the schema (3), the IL ensures expectation consistency $\overset{\mathrm{MM}}{\longleftrightarrow}$ in the lower row, while the OL update equates marginals in the left column. While this algorithm provably converges to a stationary point of $\phi(\tilde{\boldsymbol{\theta}})$ whenever the criterion is lower bounded [15], it is expensive to run, as variance computations $\mathrm{Var}_Q[\boldsymbol{s}|\boldsymbol{y}]$ are required frequently during the IL optimization (convergence and properties are discussed in the Appendix). Finally, since $\boldsymbol{\theta} = \eta^{-1}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_-)$, concave maximization w.r.t. $\boldsymbol{\theta}_-$ for fixed $\tilde{\boldsymbol{\theta}}$ can equivalently be seen as concave maximization w.r.t. $\boldsymbol{\theta}$. We will do the latter for notational convenience in the sequel.

3 Scalable Variational Inference

Scalable algorithms for a variational inference relaxation[4] different from EP have been proposed in [13, 21] (this relaxation is called VB in the sequel, for "Variational Bounding"). They can be used whenever all potentials are super-Gaussian, meaning that $t_i(s_i) = \max_{\pi_i>0} e^{b_i s_i - \frac{1}{2}\pi_i s_i^2 - h_i(\pi_i)/2}$ for some $h_i(\pi_i)$, which implies the bound $-2\log Z \le \phi^{\mathrm{VB}}(\boldsymbol{\pi}) := -2\log Z_Q + h(\boldsymbol{\pi})$ on the log partition function of $P(\boldsymbol{u}|\boldsymbol{y})$ (up to an additive constant), where $h(\boldsymbol{\pi}) := \sum_i h_i(\pi_i)$. Note that in this relaxation, $\boldsymbol{b}$ is fixed up front ($\boldsymbol{b} = \boldsymbol{0}$ if all potentials $t_i(s_i)$ are even), and $\boldsymbol{\pi}$ are the sole variational parameters. They proceed in two steps. First, $-2\log Z_Q = \log|\boldsymbol{A}| + \min_{\boldsymbol{u}_*} R(\boldsymbol{\pi}, \boldsymbol{b}, \boldsymbol{u}_*)$ (up to an additive constant), where $R(\boldsymbol{\pi}, \boldsymbol{b}, \boldsymbol{u}_*) := \sigma^{-2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}_*\|^2 + \boldsymbol{s}_*^T(\operatorname{diag}\boldsymbol{\pi})\boldsymbol{s}_* - 2\boldsymbol{b}^T\boldsymbol{s}_*$, $\boldsymbol{s}_* = \boldsymbol{B}\boldsymbol{u}_*$. Second, since $\boldsymbol{\pi} \mapsto \log|\boldsymbol{A}|$ is a concave function, Fenchel duality [17, ch. 12] implies that $\log|\boldsymbol{A}| = \min_{\boldsymbol{z}} \boldsymbol{z}^T\boldsymbol{\pi} - g^*(\boldsymbol{z})$ for some $g^*(\boldsymbol{z})$. The variational problem becomes

$$\begin{aligned} & \min_{\boldsymbol{\pi}\succ\boldsymbol{0}} \phi^{\mathrm{VB}}(\boldsymbol{\pi}) \qquad (4) \\ = \; & \min_{\boldsymbol{z}\succ\boldsymbol{0}} \min_{\boldsymbol{\pi}\succ\boldsymbol{0}, \boldsymbol{u}_*} \boldsymbol{z}^T\boldsymbol{\pi} - g^*(\boldsymbol{z}) + R(\boldsymbol{\pi}, \boldsymbol{b}, \boldsymbol{u}_*) + h(\boldsymbol{\pi}). \end{aligned}$$

[3] Log partition functions ($\log\hat{Z}_i$, $\log Z_Q$) are convex in their natural parameters, and $\boldsymbol{\theta} = \eta^{-1}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_-)$ is linear.

[4] In contrast to EP, this relaxation is convex iff all $t_i(s_i)$ are log-concave [13, 21].

It is solved by a double loop algorithm, alternating between inner loop (IL) minimizations w.r.t. $\boldsymbol{\pi}, \boldsymbol{u}_*$ for fixed $\boldsymbol{z}$ and outer loop (OL) updates of $\boldsymbol{z}$ and $g^*(\boldsymbol{z})$.

The important difference to both the double loop algorithm of [15] and the parallel algorithm of [4] lies in the *decoupling* transformation $\log|\boldsymbol{A}| = \min_{\boldsymbol{z}} \boldsymbol{z}^T\boldsymbol{\pi} - g^*(\boldsymbol{z})$. $\phi^{\mathrm{VB}}(\boldsymbol{\pi})$ is hard to minimize due to the coupling term $\log|\boldsymbol{A}|$. For example, $\nabla_{\boldsymbol{\pi}}\log|\boldsymbol{A}| = \operatorname{diag}(\boldsymbol{B}\boldsymbol{A}^{-1}\boldsymbol{B}^T) = \mathrm{Var}_Q[\boldsymbol{s}|\boldsymbol{y}]$ requires Gaussian variance computations, which are very expensive in practice [21]. But $\log|\boldsymbol{A}|$ is replaced by a fixed linear function in each IL problem, where we can eliminate $\boldsymbol{\pi}$ analytically and are left with a *penalized least squares* problem of the form $\min_{\boldsymbol{u}_*} \sigma^{-2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}_*\|^2 - \sum_i \psi_i(s_{*i})$, easy to solve with standard algorithms that do not need Gaussian variances at all. To understand the decoupling transformation more generally, consider minimizing (4) w.r.t. each variable in turn, keeping the others fixed. The solutions are $\boldsymbol{u}_* = \mathrm{E}_Q[\boldsymbol{u}|\boldsymbol{y}]$ (means) and $\boldsymbol{z} = \nabla_{\boldsymbol{\pi}}\log|\boldsymbol{A}| = \mathrm{Var}_Q[\boldsymbol{s}|\boldsymbol{y}]$ (variances). The role of decoupling is to *split between computations of means and variances* [21]: the latter, much more expensive to obtain in general, are required at OL update points only, much less frequently than the former (means) which are obtained by solving a single linear system.

Most applications cited at the beginning of Section 2 come with potentials which are not super-Gaussian, but can easily be handled with EP. Moreover, in super-Gaussian siatuations, EP seems to be substantially more accurate than VB as approximation to Bayesian inference [7, 12]. To construct an efficient EP solver, we have to make use of decoupling in a similar fashion, so to *minimize the number of Gaussian variances computations*, while retaining provable convergence.

4 Speeding up Expectation Propagation

A fast and convergent EP algorithm is obtained by marrying the double loop algorithm of [15] with the decoupling trick of [13]. During its course, $\tilde{\boldsymbol{\theta}}$ (or $\boldsymbol{\mu}, \boldsymbol{\rho}$) will mainly be fixed, and we will drop it from notation accordingly (but recall that the $\hat{Z}_i$ depend on it). Moreover, we will typically work with $\boldsymbol{\theta} = (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_-)/\eta$ rather than $\boldsymbol{\theta}_-$. Then,

$$\begin{aligned} & \phi_\cap(\boldsymbol{\theta}) \qquad (5) \\ = \; & \min_{\boldsymbol{z},\boldsymbol{u}_*} \underbrace{\boldsymbol{z}^T\boldsymbol{\pi} - g^*(\boldsymbol{z}) + R(\boldsymbol{\pi}, \boldsymbol{b}, \boldsymbol{u}_*) - 2\eta^{-1}\sum_i \log\hat{Z}_i}_{=:\phi_\cap(\boldsymbol{v},\boldsymbol{\theta}),\ \boldsymbol{v}=(\boldsymbol{z},\boldsymbol{u}_*)} \\ = \; & \min_{\boldsymbol{z},\boldsymbol{u}_*} \sigma^{-2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}_*\|^2 - \sum_i \psi_i(s_{*i}, \pi_i, b_i) - g^*(\boldsymbol{z}), \\ & \psi_i := -(z_i + s_{*i}^2)\pi_i + 2b_i s_{*i} + 2\eta^{-1}\log\hat{Z}_i. \end{aligned}$$

With $\boldsymbol{v} = (\boldsymbol{z}, \boldsymbol{u}_*)$ and $\phi_\cap(\boldsymbol{\theta}) = \min_{\boldsymbol{v}} \phi_\cap(\boldsymbol{\theta}, \boldsymbol{v})$, the IL problem of [15] is $\max_{\boldsymbol{\theta}} \min_{\boldsymbol{v}} \phi_\cap$. As shown in the Appendix, $\phi_\cap(\boldsymbol{\theta}, \boldsymbol{v})$ is a closed proper *concave-convex function* (convex in $\boldsymbol{v}$ for each $\boldsymbol{\theta}$, concave in $\boldsymbol{\theta}$ for each $\boldsymbol{v}$) [17]. Strong duality holds: $\max_{\boldsymbol{\theta}} \min_{\boldsymbol{v}} \phi_\cap = \min_{\boldsymbol{v}} \max_{\boldsymbol{\theta}} \phi_\cap$, so the IL problem is equivalent to

$$\min_{\boldsymbol{z}} \left(\min_{\boldsymbol{u}_*} \sigma^{-2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}_*\|^2 - \sum\nolimits_i \psi_i(s_{*i}) \right) - g^*(\boldsymbol{z}),$$
$$\psi_i(s_i) := \min_{\pi_i, b_i} \psi_i(s_i, \pi_i, b_i). \tag{6}$$

This problem is jointly convex in $\boldsymbol{z}, \boldsymbol{u}_*$ (note that $\psi_i(s_{*i})$ is concave as minimum of concave functions, and the minimization over π_i, b_i is a jointly convex problem). Solving the inner problem of (6) for fixed $\boldsymbol{z}$ is a simple and very efficient penalized least squares building block, denoted by $(\boldsymbol{u}_*, \boldsymbol{\theta}) \leftarrow \mathtt{PLS}(\boldsymbol{z}, \tilde{\boldsymbol{\theta}})$ in the sequel. Note that at its solution, $\boldsymbol{u}_* = \mathrm{E}_Q[\boldsymbol{u}|\boldsymbol{y}]$, where $Q(\boldsymbol{u}|\boldsymbol{y})$ is indexed by $\boldsymbol{\theta}$.

This means that the problem addressed in [15] can be written in the form $\min_{\boldsymbol{z}, \tilde{\theta}} \phi(\boldsymbol{z}, \tilde{\boldsymbol{\theta}})$. The significance is the same as in Section 3: both $\phi(\boldsymbol{z}, \tilde{\boldsymbol{\theta}})$ and $\min_{\tilde{\boldsymbol{\theta}}} \phi(\boldsymbol{z}, \tilde{\boldsymbol{\theta}})$ (local minimum) for fixed $\boldsymbol{z}$ can be determined very efficiently. The dominating cost of computing Gaussian variances is concentrated in the update of $\boldsymbol{z}$. Two main ideas lead to the algorithm we propose here. First, we descend on $\phi(\boldsymbol{z}, \tilde{\boldsymbol{\theta}})$ rather than $\phi(\tilde{\boldsymbol{\theta}}) = \min_{\boldsymbol{z}} \phi(\boldsymbol{z}, \tilde{\boldsymbol{\theta}})$ [15], saving on variance computations. One iteration of our method determines $\boldsymbol{z} \leftarrow \mathrm{Var}_Q[\boldsymbol{s}|\boldsymbol{y}]$, then a local minimum $\min_{\tilde{\boldsymbol{\theta}}} \phi(\boldsymbol{z}, \tilde{\boldsymbol{\theta}})$ in a convergent way. Empirically, such "optimistic" iterations seem to always descend on $\phi(\boldsymbol{z}, \tilde{\boldsymbol{\theta}})$ until convergence to a stationary point of $\phi(\tilde{\boldsymbol{\theta}})$, but just as for the sequential or parallel algorithm, we cannot establish this rigorously. At this point, the second idea is to rely on the inner loop optimization of [15] in order to enforce descent eventually. We obtain a provably convergent algorithm by combining optimistic steps $\min_{\tilde{\boldsymbol{\theta}}} \phi(\boldsymbol{z}, \tilde{\boldsymbol{\theta}})$ for fixed $\boldsymbol{z}$ with the rigorous but slow mechanism of [15]. As most, if not all optimistic steps produce sufficient descent in practice, provable convergence comes almost for free (in contrast to [15], where it carries a large price tag).

To flesh out this notion, denote[5] $\phi(\boldsymbol{\theta}, \boldsymbol{z}, \tilde{\boldsymbol{\theta}}) = \boldsymbol{z}^T \boldsymbol{\pi} - g^*(\boldsymbol{z}) + (\min_{\boldsymbol{u}_*} R(\boldsymbol{\pi}, \boldsymbol{b}, \boldsymbol{u}_*)) - 2\eta^{-1} \sum_i \log \hat{Z}_i$, and $\phi(\boldsymbol{z}, \tilde{\boldsymbol{\theta}}) = \max_{\boldsymbol{\theta}} \phi(\boldsymbol{\theta}, \boldsymbol{z}, \tilde{\boldsymbol{\theta}})$. Note that $\phi(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) = \min_{\boldsymbol{z}} \phi(\boldsymbol{\theta}, \boldsymbol{z}, \tilde{\boldsymbol{\theta}})$, moreover $\max_{\boldsymbol{\theta}} \phi(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) = \min_{\boldsymbol{z}} \max_{\boldsymbol{\theta}} \phi(\boldsymbol{\theta}, \boldsymbol{z}, \tilde{\boldsymbol{\theta}})$ by strong duality. First, $\phi(\tilde{\boldsymbol{\theta}}) \le \phi(\boldsymbol{z}, \tilde{\boldsymbol{\theta}})$, so that $\phi(\boldsymbol{z}, \tilde{\boldsymbol{\theta}})$ is lower bounded if $\phi(\tilde{\boldsymbol{\theta}})$ is (which, like [15], we assume). Next, as shown in the Appendix, we can very efficiently minimize $\phi(\boldsymbol{z}, \tilde{\boldsymbol{\theta}})$ locally w.r.t. $\tilde{\boldsymbol{\theta}}$ by setting $\boldsymbol{\rho} \leftarrow \boldsymbol{z}$, then iterating between $(\boldsymbol{u}_*, \boldsymbol{\theta}) \leftarrow \mathtt{PLS}(\boldsymbol{z}, \tilde{\boldsymbol{\theta}})$ and $\boldsymbol{\mu} \leftarrow \boldsymbol{s}_* = \boldsymbol{B}\boldsymbol{u}_* = \mathrm{E}_Q[\boldsymbol{s}|\boldsymbol{y}]$. In the sequel, we denote this subalgorithm by $\tilde{\boldsymbol{\theta}}' \leftarrow \mathtt{updateTTil}(\boldsymbol{z}, \tilde{\boldsymbol{\theta}})$. While `updateTTil` may call `PLS` multiple times, it does not require expensive Gaussian variance computations. An "optimistic" step of our algorithm updates $\boldsymbol{z}' \leftarrow \mathrm{Var}_Q[\boldsymbol{s}|\boldsymbol{y}]$, then $\tilde{\boldsymbol{\theta}}' \leftarrow \mathtt{updateTTil}(\boldsymbol{z}', \tilde{\boldsymbol{\theta}})$, at the cost of one variance computation. Within the schema (3), we update $\boldsymbol{z}$, set $\boldsymbol{\rho} \leftarrow \boldsymbol{z}$, then attain expectation consistency and $\boldsymbol{\mu} \stackrel{!}{=} \mathrm{E}_Q[\boldsymbol{s}|\boldsymbol{y}] = \boldsymbol{s}_* = \boldsymbol{B}\boldsymbol{u}_*$ for fixed variances $\boldsymbol{z}$, $\boldsymbol{\rho}$.

Suppose we are at a point $\boldsymbol{z}, \tilde{\boldsymbol{\theta}}$ (and $\boldsymbol{\theta}$), so that $\tilde{\boldsymbol{\theta}}$ is a local minimum point of $\phi(\boldsymbol{z}, \tilde{\boldsymbol{\theta}})$. How can we descend: $\phi(\boldsymbol{z}', \tilde{\boldsymbol{\theta}}') < \phi(\boldsymbol{z}, \tilde{\boldsymbol{\theta}})$ unless $\tilde{\boldsymbol{\theta}}$ is a stationary point of $\phi(\tilde{\boldsymbol{\theta}})$? Let $\boldsymbol{\theta}^{(1)} = \boldsymbol{\theta}$. The optimistic step would be $\boldsymbol{z}^{(1)} = \mathrm{Var}_Q[\boldsymbol{s}|\boldsymbol{y}]$, then $\tilde{\boldsymbol{\theta}}' \leftarrow \mathtt{updateTTil}(\boldsymbol{z}^{(1)}, \tilde{\boldsymbol{\theta}})$. If $\phi(\boldsymbol{z}^{(1)}, \tilde{\boldsymbol{\theta}}')$ is sufficiently smaller than $\phi(\boldsymbol{z}, \tilde{\boldsymbol{\theta}})$, we are done with our descent step: $\boldsymbol{z}' = \boldsymbol{z}^{(1)}$. Otherwise, we run one iteration $\boldsymbol{\theta}^{(1)} \to \boldsymbol{\theta}^{(2)}$ of the inner optimization $\max_{\boldsymbol{\theta}} \phi(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})$ of [15]. This requires variance computations, while $\boldsymbol{z}^{(1)}$ can be reused (and $\boldsymbol{z}^{(2)}$ may already be computed). We set $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}^{(2)}$ and attempt another optimistic step: $\boldsymbol{z}^{(2)}$, $\mathtt{updateTTil}(\boldsymbol{z}^{(2)}, \tilde{\boldsymbol{\theta}})$. Without intervening descent, we would eventually obtain $\boldsymbol{\theta}^{(k)} = \max_{\boldsymbol{\theta}} \phi(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})$, thus $\boldsymbol{z}^{(k)} = \mathrm{argmin}_{\boldsymbol{z}'} \phi(\boldsymbol{z}', \tilde{\boldsymbol{\theta}})$. If no descent happens from there, $\tilde{\boldsymbol{\theta}}$ must be a stationary point of $\phi(\tilde{\boldsymbol{\theta}})$ (see [15] and Appendix).

Note that in most cases in practice, our algorithm does not run into the inner optimization of [15] even once. Yet the possibility of doing so is what makes our convergence proof work. Algorithm 1 provides a schema.

A word of warning about the inner optimization $\max_{\boldsymbol{\theta}} \phi(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})$. From (6), it is tempting to iterate between $\boldsymbol{z} \leftarrow \mathrm{Var}_Q[\boldsymbol{s}|\boldsymbol{y}]$ and $(\boldsymbol{u}_*, \boldsymbol{\theta}) \leftarrow \mathtt{PLS}(\boldsymbol{z}, \tilde{\boldsymbol{\theta}})$. However, this does not lead to descent and typically fails in practice. As seen in Section 3, the update of $\boldsymbol{z}$ serves to refit an *upper* bound, suitable for *minimizing*, but not *maximizing* over $\boldsymbol{\theta}$. In our algorithm, this problem is compensated by the minimization over $\tilde{\boldsymbol{\theta}}$: optimistic steps seem to always descend.

4.1 Computational Details

In this section, we provide details for computational primitives required in Algorithm 1. First, we show how to efficiently compute `PLS`, i.e. solve the inner problem in (6) for fixed $\boldsymbol{z} \succ \boldsymbol{0}$. As all $\psi_i(s_{*i})$ are concave, this is a convex penalized least squares problem, for which many very efficient solvers are available. A slight technical challenge comes from the implicit definition of the regularizer: evaluating ψ_i and its derivatives entails a bivariate convex minimization.

In our experiments, we employ a standard gradient-

[5] In the sequel, we will eliminate $\boldsymbol{u}_*$ by minimization in our notation. Since strong duality holds, we can move $\min_{\boldsymbol{u}_*}$ outside when solving `PLS` (6) at any time (for fixed $\boldsymbol{z}, \tilde{\boldsymbol{\theta}}$).

Algorithm 1 Double loop EP algorithm.
The part shaded in grey was never accessed in our experiments (see text for comments).

$\Delta(a,b) := (b-a)/\max\{|a|,|b|,10^{-9}\}$.
Iterate over $\boldsymbol{z}, \tilde{\boldsymbol{\theta}} \leftrightarrow (\boldsymbol{\mu}, \boldsymbol{\rho})$.
repeat
  $\boldsymbol{\pi}^{(1)} = \boldsymbol{\pi}$.
  for $k = 1, 2, \ldots$ **do**
    $\boldsymbol{z}^{(k)} = \mathrm{Var}_Q[\boldsymbol{s}|\boldsymbol{y}]$.
    $(\tilde{\boldsymbol{\theta}}', \boldsymbol{\theta}') \leftarrow$ `updateTTil`$(\boldsymbol{z}^{(k)}, \tilde{\boldsymbol{\theta}})$.
    if $\Delta(\phi(\boldsymbol{z}^{(k)}, \tilde{\boldsymbol{\theta}}'), \phi(\boldsymbol{z}, \tilde{\boldsymbol{\theta}})) > \varepsilon$ **then**
      Sufficient descent: $\boldsymbol{z} \leftarrow \boldsymbol{z}^{(k)}$, $\tilde{\boldsymbol{\theta}} \leftarrow \tilde{\boldsymbol{\theta}}'$, $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}'$. Leave loop over k.
    else
      Run iteration of $\max_{\boldsymbol{\theta}} \phi(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})$: $\boldsymbol{\theta}^{(k)} \to \boldsymbol{\theta}^{(k+1)}$. Set $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}^{(k+1)}$.
      if $|\Delta(\phi(\boldsymbol{\theta}^{(k+1)}, \tilde{\boldsymbol{\theta}}), \phi(\boldsymbol{\theta}^{(k)}, \tilde{\boldsymbol{\theta}}))| < \varepsilon$ **then**
        Converged to stationary point $\tilde{\boldsymbol{\theta}}$: Terminate algorithm.
      end if
    end if
  end for
until Maximum number of iterations done

based Quasi-Newton optimizer. Suppose we are at $\boldsymbol{u}_*$ and have determined the maximizer $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{b})$. If $f(\boldsymbol{u}_*) = \sigma^{-2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}_*\|^2 - \sum_i \psi_i(s_{*i})$, then $\psi_i'(s_{*i}) = \partial_{s_{*i}}\psi_i(s_{*i}, \pi_i, b_i) = 2(b_i - \pi_i s_{*i})$, so that $\nabla_{\boldsymbol{u}_*} f(\boldsymbol{u}_*) = 2\sigma^{-2}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{u}_* - \boldsymbol{y}) + 2\boldsymbol{B}^T(\boldsymbol{\pi} \circ \boldsymbol{s}_* - \boldsymbol{b})$, at the cost of one matrix-vector multiplication (MVM) with $\boldsymbol{X}^T\boldsymbol{X}$, $\boldsymbol{B}^T$, $\boldsymbol{B}$ respectively (here, "$\circ$" denotes the componentwise product). For the bivariate minimizations, the derivatives are $\partial_{b_i}\psi_i = 2(s_{*i} - \mathrm{E}_{\hat{P}_i}[s_i])$, $\partial_{\pi_i}\psi_i = -(z_i + s_{*i}^2) + \mathrm{E}_{\hat{P}_i}[s_i^2]$: we have to adjust b_i, π_i so that mean and variance of $\hat{P}_i$ coincides with s_{*i} and z_i. Details for the computation of $\hat{P}_i$ are given in [18]. In our implementation, we initialize the minimization by two standard EP updates, then run Newton's algorithm (details are given in a longer paper). Even for large q, these bivariate minimizations can often be done more rapidly than MVMs with $\boldsymbol{X}^T\boldsymbol{X}$. Moreover, they can be solved in parallel on graphics hardware.

The inner optimization $\max_{\boldsymbol{\theta}} \phi(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})$ of [15] can be addressed by any convex solver. We employ Quasi-Newton once more. The gradients are $\partial_{\boldsymbol{b}}\phi(\boldsymbol{\pi}, \boldsymbol{b}, \tilde{\boldsymbol{\theta}}) = 2((\mathrm{E}_{\hat{P}_i}[s_i]) - (\mathrm{E}_Q[s_i]))$, $\partial_{\boldsymbol{\pi}}\phi(\boldsymbol{\pi}, \boldsymbol{b}, \tilde{\boldsymbol{\theta}}) = (\mathrm{E}_Q[s_i^2]) - (\mathrm{E}_{\hat{P}_i}[s_i^2])$. This computation entails $\boldsymbol{z} = \mathrm{Var}_Q[\boldsymbol{s}|\boldsymbol{y}]$. Note that with a standard solver, a sufficient increase in $\phi(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})$ (for fixed $\tilde{\boldsymbol{\theta}}$) may require a number of $\mathrm{Var}_Q[\boldsymbol{s}|\boldsymbol{y}]$ computations. We are not aware of an effective way to decouple this problem as in Section 3.

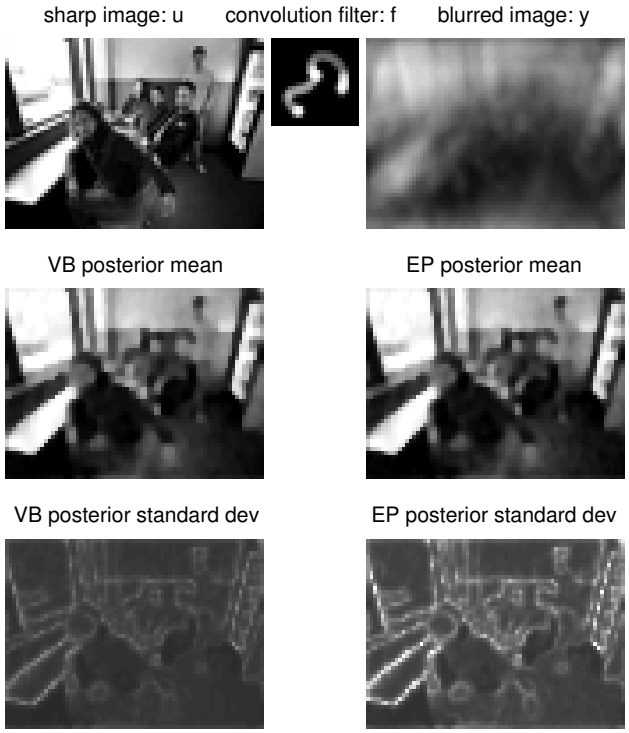

Figure 1: Deconvolution setting and resulting marginals (variances on $\boldsymbol{u}$, not on $\boldsymbol{s}$). $\boldsymbol{u}$ is 48×73, pixels, the kernel $\boldsymbol{f}$ is 22×25 ($n = 3504$, $q = 10512$, $\tau_a = \tau_r = 15$, $\sigma^2 = 10^{-5}$).

Gaussian Variances

Finally, how do we compute Gaussian variances $\boldsymbol{z} = \mathrm{Var}_Q[\boldsymbol{s}|\boldsymbol{y}] = \mathrm{diag}(\boldsymbol{B}\boldsymbol{A}^{-1}\boldsymbol{B}^T)$? This is by far the most expensive computation in all EP algorithms discussed here: our main contribution is a novel convergent algorithm which requires few of these calls. In our experiments, n is a few thousand, $q \approx 3n$, and we can maintain an $n \times n$ matrix in memory. We use the identity

$$\boldsymbol{z} = \mathrm{diag}\left(\boldsymbol{B}\boldsymbol{A}^{-1}\sum\nolimits_i \boldsymbol{\delta}_i\boldsymbol{\delta}_i^T\boldsymbol{B}\right) = \sum\nolimits_i (\boldsymbol{B}\boldsymbol{A}^{-1}\boldsymbol{\delta}_i) \circ (\boldsymbol{B}\boldsymbol{\delta}_i),$$

where $\boldsymbol{\delta}_i = (\mathrm{I}_{\{j=i\}})_j$. We compute the Cholesky decomposition $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^T$, then $\boldsymbol{A}^{-1}$ from $\boldsymbol{L}$, using LAPACK code, then accumulate $\boldsymbol{z}$ by $2n$ MVMs with $\boldsymbol{B}$.

If n is larger than 10^4 or so, this approach is not workable anymore. If $\boldsymbol{A}$ is very sparse, it may possess a sparse Cholesky decomposition which can be determined efficiently, in which case $\boldsymbol{z}$ is determined easily [4]. However, for typical image reconstruction models, $\boldsymbol{X}$ is dense. For the VB relaxation of Section 3, variances have been approximated by the Lanczos algorithm [22, 13]. It is noted in [19] that variances are strongly (but selectively) underestimated in this way, and consequences for the VB double loop algorithm

are established there: in a nutshell, while outcomes are qualitatively different, the algorithm behaviour remains reasonable. In contrast, if any of the EP algorithms discussed in this paper are run with Lanczos variance approximations, they exhibit highly erratic behaviour. Parallel EP [4] rapidly diverges, our variant ends in numerical breakdown. While we are lacking a complete explanation for these failures at present, it seems evident that the expectation consistency conditions, whose structure is more complicated than the simple VB bound, do not tolerate strong variance errors. Our observation underlines the thesis of [19, 21]. Robustness to variance errors of the kind produced by Lanczos becomes an important asset of variational inference relaxations, at least if large scale inference is to be addressed. The EP relaxation, as it stands, does not seem to be robust in this sense. Explaining this fact, and possibly finding a *robust modification* of the expectation consistency conditions, remain important topics for future research.

5 Experiments

5.1 Expectation Propagation vs. VB

In the following experiment, we compare approximate inference outcomes of EP (Section 2) and VB (Section 3), complementing previous studies [7, 12]. We address the (non-blind) deconvolution problem for image deblurring (details ommitted here are found in [9]): $\boldsymbol{u} \in \mathbb{R}^n$ represent the desired sharp image, $\boldsymbol{X} = (\mathrm{diag}\, \tilde{\boldsymbol{f}})\boldsymbol{F}_n$, where $\boldsymbol{F}_n$ is the $n \times n$ discrete Fourier transform (DFT)[6], $\tilde{\boldsymbol{f}} = \boldsymbol{F}_n \boldsymbol{f}$ the spectrum of the blur kernel $\boldsymbol{f}$, and $\boldsymbol{y} = \boldsymbol{F}_n \tilde{\boldsymbol{y}}$, $\tilde{\boldsymbol{y}}$ the blurry image. Our model setup is similar to what was previously used in [19]: $P(\boldsymbol{u})$ is a Laplace sparsity prior (see Section 1), the transform $\boldsymbol{B}$ consists of an orthonormal wavelet transform $\boldsymbol{B}_a$ and horizontal/vertical differences $\boldsymbol{B}_r$ ("total variation"), corresponding prior parameters are τ_a, τ_r. Recall that $\boldsymbol{b}$ is fixed[7] depending on the $t_i(\cdot)$ in VB: since they are even, $\boldsymbol{b} = \boldsymbol{0}$. In contrast, they are free variational parameters in EP. Posterior marginals, as approximated by EP and VB, are shown in Figure 1, while we compare parameters $\boldsymbol{b}$, $\boldsymbol{\pi}$ in Figure 2.

The EP and VB approximations are substantially different. While the means are visually similar, EP's posterior variances are larger and show a more pronounced structure. An explanation is offered by the striking differences in final parameters $\boldsymbol{b}$, $\boldsymbol{\pi}$. Roughly, π_i scales the degree of penalization of s_i [18]. While both EP and VB strongly penalize certain coefficients, VB (in contrast to EP) seems to universally penalize all s_i (all $\pi_{\mathrm{VB},i} > 10$), thus may produce small variances simply by overpenalization. EP clearly makes use of $\boldsymbol{b}$, which allow to control the posterior mean independent of the covariance: a mechanism not available for VB. It is important to note that our findings are in line with those in [12], who found that VB strongly underapproximated marginal variances (they obtained the ground truth by expensive Monte Carlo simulations). As noted in Section 1, it is often the posterior uncertainty estimates (covariances) which give Bayesian decision-making an edge over point estimation approaches.

[6] Strictly speaking, we encode $\mathbb{C}$ by $\mathbb{R}^2$, and $\boldsymbol{F}_n$ is the "real-to-complex" DFT (closely related to the discrete cosine transform). Both $\tilde{\boldsymbol{f}}$ and $\tilde{\boldsymbol{y}}$ are Hermitian and can be stored as $\mathbb{R}^n$ vectors.

[7] This is an inherent feature of the variational bound, which would cease to be valid if $\boldsymbol{b}$ were optimized over.

5.2 EP Timing Comparison

In this section, we provide timing comparisons between EP algorithms discussed in this paper. Our setup is much the same as in Section 5.1, but both the choice of $\boldsymbol{X}$ and data is taken from [19]. The problem is inference over images $\boldsymbol{u} \in \mathbb{R}^n$ from "Cartesian MRI" measurements (discrete Fourier coefficients) $\boldsymbol{y} \in \mathbb{C}^m$, so that $\boldsymbol{X} = \boldsymbol{I}_{J,\cdot}\boldsymbol{F}_n$, where J is an index selecting acquired coefficients (in fact, complete columns in DF space ("phase encodes") are sampled, according to a design optimized for natural images). The prior is the same as used above.

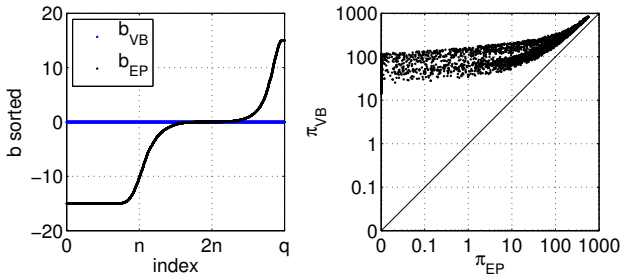

Figure 2: Final parameters for deconvolution. Left: $\boldsymbol{b}$ sorted ($\boldsymbol{b}_{\mathrm{VB}} = \boldsymbol{0}$ by construction). Right: $\boldsymbol{\pi}$.

In our first experiment, we use 64×64 images ($n = 4096$, $q = 12160$) and a design $\boldsymbol{X}$ sampling 16 columns ($m = 1024$, 4 times undersampled). We compare the sequential and parallel EP algorithms with our novel fast (convergent) EP method. We chose not to include results for the double loop algorithm of [15], since it runs even slower than the sequential method (see comments in Section 4.1). Our results are averaged over 20 different images (the $\boldsymbol{y}$ vectors are noisy acquisitions, $\sigma^2 = 10^{-3}$, but the same across methods). Moreover, $\tau_a = 0.04/\sigma$, $\tau_r = 0.08/\sigma$ (same values as in [19]). Timing runs were done on an otherwise unloaded standard desktop machine. For each run, we

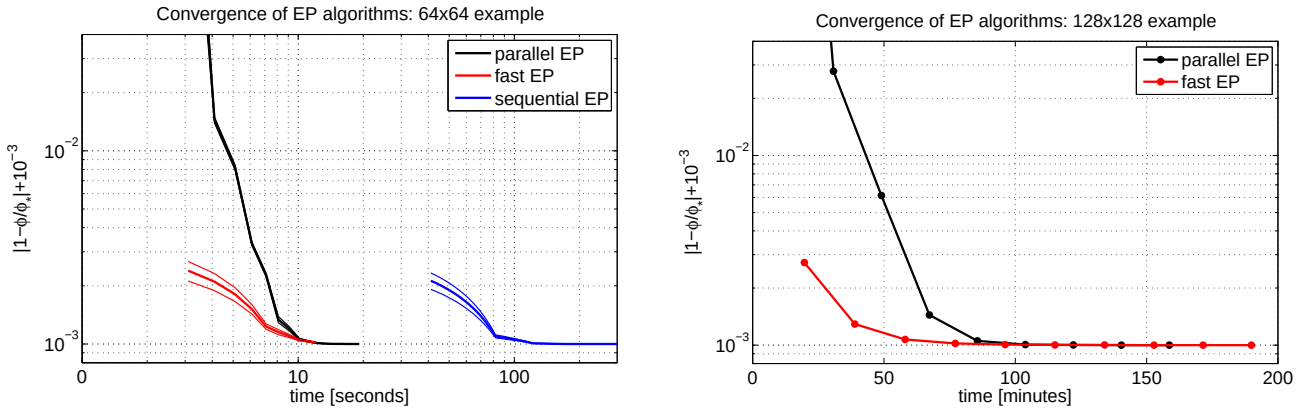

Figure 3: Timing comparison of EP algorithms for inference over greyscale images. Left: 64×64 images. Right: 128×128 image. Shown is relative distance to EP energy stationary point $|(\phi - \phi_*)/\phi_*|$ as function of running time (left: mean, two std. over 20 different images).
Algorithms: sequential EP (Section 2; left only), parallel EP (Section 2), and fast EP (our method).

stored tupels (T_j, ϕ_j) at the end of each outer iteration (for sequential EP, this is a sweep over all potentials), T_j elapsed time (in secs), ϕ_j the EP energy value attained. On a fixed image, all methods eventually attained the same energy value[8] (say, ϕ_*), and we show $(T_j, |(\phi_j - \phi_*)/\phi_*|)$. Results are presented in Figure 3, left. First, the sequential algorithm is not competitive with the others. At a time when the others converged, it is roughly 1/4 through its first sweep (while requiring about four sweeps to converge). Second, the parallel and our fast EP algorithm converge in about the same time. However, ours does so much more smoothly and attains a near optimal solution more quickly.

In a second experiment, we use a single 128×128 image ($n = 16384$, $q = 48896$) and a design $\boldsymbol{X}$ sampling 36 columns (≈ 3.5 times undersampled). We compare the parallel with our fast EP algorithm, since the sequential method is clearly infeasible at this scale. Here, $\sigma^2 = 2 \cdot 10^{-4}$, $\tau_a = 0.04/\sigma$, $\tau_r = 0.08/\sigma$. Results are presented in Figure 3, right. On this larger problem, our algorithm converges significantly faster.

Our method (fast EP in Figure 3) is provably convergent, while parallel EP (and sequential EP) lacks such a guarantee. Beyond, the main difference between fast and parallel EP lies in how thoroughly variance computations are exploited. Fast EP spends more effort between them, solving $\min_{\tilde{\boldsymbol{\theta}}} \phi(\boldsymbol{z}, \tilde{\boldsymbol{\theta}}) = \min_{\tilde{\boldsymbol{\theta}}} \max_{\boldsymbol{\theta}} \phi(\boldsymbol{\theta}, \boldsymbol{z}, \tilde{\boldsymbol{\theta}})$, while parallel EP simply does a single EP update. Our method therefore incurs an overhead, which motivates the results for 64×64 images. However, this overhead is modest (each step of PLS costs $O(q + n \log n)$), while the cost for variances, at $O(n(n^2 + q))$, grows very fast. The overhead for fast EP pays off in the 128×128 image example, due to the fact that it requires about two variance computations less than parallel EP to attain convergence. Notably, the overhead cost can still be greatly reduced by running different algorithms (see Section 6) or parallelizing the computations of the $\psi_i(s_{*i})$, which is not done in our implementation.

[8] While this is not guaranteed by present EP convergence theory, it happened in all our cases.

6 Discussion

We proposed a novel, provably convergent algorithm to solve the expectation propagation relaxation of Bayesian inference. Based on the insight that the most expensive computations by far in any variational method concern Gaussian variances, we exploit a decoupling trick previously used in [23, 13] in order to minimize the number of such computations. Our method is at least an order of magnitude faster than the commonly used sequential EP algorithm, and improves on parallel EP [4], the previously fastest solver we are aware of, both in running time and guaranteed convergence. Moreover, it is in large parts similar to recent algorithms for other relaxations [13], which allows for transfer of efficient code. While the sequential EP algorithm is most widely used today, our results indicate that this is wasteful even for small and medium size problems and should be avoided in the future.

There are numerous avenues for future work. First, for problems of the general form discussed in Section 5, the central penalized least squares primitive PLS could be solved more efficiently by employing modern augmented Lagrangian techniques, such as the ADMM algorithm reviewed in [3] (today's most efficient sparse

deconvolution algorithms are based on this technique), and by parallelizing the innermost bivariate optimization problems leading to $\psi_i(s_{*i})$ and its derivatives. Such measures would bring down the (already modest) overhead of our technique, compared to parallel EP. Moreover, we aim to resolve whether the "optimistic steps" our algorithm is mainly based on, provably lead to descent by themselves (this would render the fallback on [15], shaded in Algorithm 1, obsolete, thus simplify the code).

Known EP algorithms (including ours presented here) break down in the presence of substantial Gaussian variance approximation errors, in contrast to algorithms for simpler relaxations which behave robustly. If Bayesian imaging applications, such as those in Section 5, are to be run at realistic sizes, variance errors cannot be avoided. The most important future direction is therefore to understand the reason for this non-robustness of EP algorithms (or even the expectation-consistency conditions as such) and to seek for alternatives which combine the accuracy of this relaxation with good behaviour in the presence of typical Gaussian variances approximation errors [19, 21].

Appendix

We start by reviewing the convergence proof for the EP double loop algorithm of Section 2 [15]. The problem is $\min_{\tilde{\boldsymbol{\theta}}} \max_{\boldsymbol{\theta}_-} \phi_\cap(\boldsymbol{\theta}_-, \tilde{\boldsymbol{\theta}}) + \phi_\cup(\tilde{\boldsymbol{\theta}})$. Now, $\phi_\cap(\tilde{\boldsymbol{\theta}}) = \max_{\boldsymbol{\theta}_-} \phi_\cap(\boldsymbol{\theta}_-, \tilde{\boldsymbol{\theta}})$ is concave. If $\boldsymbol{\theta}_- = \operatorname{argmin} \phi(\boldsymbol{\theta}_-, \tilde{\boldsymbol{\theta}})$, then $\phi(\tilde{\boldsymbol{\theta}'}) \le R(\tilde{\boldsymbol{\theta}'}) := \phi_\cap(\boldsymbol{\theta}_-, \tilde{\boldsymbol{\theta}}) - \boldsymbol{g}^T(\tilde{\boldsymbol{\theta}'} - \tilde{\boldsymbol{\theta}}) + \phi_\cup(\tilde{\boldsymbol{\theta}'})$, where $\boldsymbol{g} = -\nabla_{\tilde{\boldsymbol{\theta}}} \phi_\cap(\tilde{\boldsymbol{\theta}}) = -\partial_{\tilde{\boldsymbol{\theta}}} \phi_\cap(\boldsymbol{\theta}_-, \tilde{\boldsymbol{\theta}})$ [17, ch. 12]. If $\boldsymbol{\theta} = \eta^{-1}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_-)$, then $\boldsymbol{g} = \partial_{\tilde{\boldsymbol{\theta}}} 2 \log Z_Q = \eta^{-1}(\mathrm{E}_Q[\boldsymbol{s}|\boldsymbol{y}], -\frac{1}{2}\mathrm{E}_Q[\boldsymbol{s}^2|\boldsymbol{y}])$. Now, $\phi_\cup(\tilde{\boldsymbol{\theta}}) = R(\tilde{\boldsymbol{\theta}})$, and $R(\tilde{\boldsymbol{\theta}'})$ is convex, its minimum defined by $\nabla_{\tilde{\boldsymbol{\theta}}'} \phi_\cup(\tilde{\boldsymbol{\theta}'}) = \boldsymbol{g}$. Therefore, minimizing $R(\tilde{\boldsymbol{\theta}'})$ leads to $\phi(\tilde{\boldsymbol{\theta}'}) < \phi(\tilde{\boldsymbol{\theta}})$, unless $\boldsymbol{g} = \nabla_{\tilde{\boldsymbol{\theta}}} \phi_\cup(\tilde{\boldsymbol{\theta}})$, thus $\nabla_{\tilde{\boldsymbol{\theta}}} \phi(\tilde{\boldsymbol{\theta}}) = \boldsymbol{0}$. Since the sequence $\phi(\tilde{\boldsymbol{\theta}})$ is nonincreasing and lower bounded, it must converge to a stationary point. To determine $\boldsymbol{g}$, note that if $\boldsymbol{u}_*$ is the minimizer in (6), then $\mathrm{E}_Q[\boldsymbol{s}|\boldsymbol{y}] = \boldsymbol{s}_* = \boldsymbol{B}\boldsymbol{u}_*$ and $\mathrm{E}_Q[\boldsymbol{s}^2|\boldsymbol{y}] = \boldsymbol{s}_*^2 + \mathrm{Var}_Q[\boldsymbol{s}|\boldsymbol{y}]$. Moreover, since $\phi(\tilde{\boldsymbol{\theta}'})$ is the sum of log partition functions of $N(\mu_i, \rho_i)$, the equation $\nabla_{\tilde{\boldsymbol{\theta}}'} \phi_\cup(\tilde{\boldsymbol{\theta}'}) = \boldsymbol{g}$ is solved by $\boldsymbol{\mu}' = \boldsymbol{s}_*$, $\boldsymbol{\rho}' = \mathrm{Var}_Q[\boldsymbol{s}|\boldsymbol{y}]$.

Importantly, exactly the same argument establishes the convergence (to a stationary point) of $\min_{\tilde{\boldsymbol{\theta}}'} \phi(\boldsymbol{z}, \tilde{\boldsymbol{\theta}'})$ for any fixed $\boldsymbol{z} \succ \boldsymbol{0}$, thus the computation of `updateTTil` in Section 4. We only have to replace $\log|\boldsymbol{A}(\boldsymbol{\pi})|$ by $\boldsymbol{z}^T\boldsymbol{\pi} - g^*(\boldsymbol{z})$ (both are concave in $\boldsymbol{\theta}$, therefore concave in $(\boldsymbol{\theta}_-, \tilde{\boldsymbol{\theta}})$), noting that the gradient w.r.t. $\boldsymbol{\pi}$ changes from $\nabla_{\boldsymbol{\pi}} \log|\boldsymbol{A}| = \mathrm{Var}_Q[\boldsymbol{s}|\boldsymbol{y}]$ to $\nabla_{\boldsymbol{\pi}}(\boldsymbol{z}^T\boldsymbol{\pi} - g^*(\boldsymbol{z})) = \boldsymbol{z}$. The only difference to the algorithm of [15] just discussed is that $\boldsymbol{\rho}$ is updated to $\boldsymbol{z}$, not to $\mathrm{Var}_Q[\boldsymbol{s}|\boldsymbol{y}]$, so that variances do not have to be computed.

Next, we establish the properties of the inner loop problem $\max_{\boldsymbol{\theta}} \phi_\cap(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})$ (Eqs. 5, 6). In particular, we prove that strong duality holds. Recall that $\boldsymbol{v} = (\boldsymbol{z}, \boldsymbol{u}_*)$ and $\phi_\cap(\boldsymbol{v}, \boldsymbol{\theta})$ from (5). We begin by extending $\phi_\cap(\boldsymbol{v}, \boldsymbol{\theta})$ for all values of $\boldsymbol{z}$ and $\boldsymbol{\pi}$ [17]. First, $g^*(\boldsymbol{z}) = \inf_{\boldsymbol{\pi}} \boldsymbol{z}^T\boldsymbol{\pi} - \log|\boldsymbol{A}(\boldsymbol{\pi})|$ is the *concave* dual function of $\log|\boldsymbol{A}(\boldsymbol{\pi})|$. Since $\log|\boldsymbol{A}| \to \infty$ whenever any $\pi_i \to \infty$ [21], then $g^*(\boldsymbol{z}) \to -\infty$ as any $z_i \searrow 0$, and $\phi_\cap := +\infty$ if any $z_i \le 0$. Moreover, $\phi_\cap := -\infty$ if $\boldsymbol{z} \succ \boldsymbol{0}$ and any $\pi_i < 0$, and $\phi_\cap(\boldsymbol{v}, \boldsymbol{\pi}, \boldsymbol{b}) := \lim_{\tilde{\boldsymbol{\pi}} \searrow \boldsymbol{\pi}} \phi_\cap(\boldsymbol{v}, \tilde{\boldsymbol{\pi}}, \boldsymbol{b})$ for any $\boldsymbol{\pi} \succeq \boldsymbol{0}$. With these extensions, it is easy to see that $\phi_\cap(\boldsymbol{v}, \boldsymbol{\theta})$ is a closed proper concave-convex function [17, ch. 33]: convex in $\boldsymbol{v}$ for each $\boldsymbol{\theta}$, concave in $\boldsymbol{\theta}$ for each $\boldsymbol{v}$. Note that we always have that $\max_{\boldsymbol{\theta}} \min_{\boldsymbol{v}} \phi_\cap \le \min_{\boldsymbol{v}} \max_{\boldsymbol{\theta}} \phi_\cap$ (weak duality). In order to establish equality (strong duality), we show that $\phi_\cap(\cdot, \boldsymbol{\theta})$ do not have a common nonzero direction of recession. Given that, strong duality follows from [17, Theorem 37.3].

Theorem 1 *Let $\phi(\boldsymbol{v}, \boldsymbol{\theta})$ be defined as in (5), and extended to a closed proper concave-convex function. If $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{b})$ is such that $\boldsymbol{\pi} \succ \boldsymbol{0}$ and $\boldsymbol{A}(\boldsymbol{\pi})$ is positive definite, then $\phi(\cdot, \boldsymbol{\theta})$ has no nonzero direction of recession. For any $\boldsymbol{d} \ne \boldsymbol{0}$ and any $\boldsymbol{v}$ so that $\phi(\boldsymbol{v}, \boldsymbol{\theta}) < \infty$:*

$$\lim_{t \to \infty} \frac{\phi(\boldsymbol{v} + t\boldsymbol{d}, \boldsymbol{\theta}) - \phi(\boldsymbol{v}, \boldsymbol{\theta})}{t} > 0.$$

Proof Write $F(\boldsymbol{v}) = \phi_\cap(\boldsymbol{v}, \boldsymbol{\theta})$ for brevity, and pick any $\boldsymbol{d} \ne \boldsymbol{0}$. $\boldsymbol{d}$ is a direction of recession iff $\lim_{t\to\infty}(F(\boldsymbol{v} + t\boldsymbol{d}) - F(\boldsymbol{v}))/t \le 0$ for some $\boldsymbol{v}$ [17, Theorem 8.5]. Pick any $\boldsymbol{v} = (\boldsymbol{z}, \boldsymbol{u}_*)$, $\boldsymbol{z} \succ \boldsymbol{0}$, and let $\boldsymbol{d} = (\boldsymbol{d}_z, \boldsymbol{d}_u)$. If $\boldsymbol{d}_u \ne \boldsymbol{0}$, then $F(\boldsymbol{v} + t\boldsymbol{d}) = \Omega(t^2)$ by the positive definite quadratic part. If $(\boldsymbol{d}_z)_i < 0$ for any i, then there is some $t_0 > 0$ so that $(\boldsymbol{z} + t\boldsymbol{d}_z)_i$ is negative and $F(\boldsymbol{v} + t\boldsymbol{d}) = \infty$ for all $t \ge t_0$. This leaves us with $\boldsymbol{d}_u = \boldsymbol{0}$, $\boldsymbol{d}_z \succeq \boldsymbol{0}$, so that $(\boldsymbol{d}_z)_i > 0$ for some i. Let $\tilde{\boldsymbol{\pi}} = \boldsymbol{\pi} - (\pi_i/2)\boldsymbol{\delta}_i$. By definition, $g^*(\boldsymbol{z} + t\boldsymbol{d}_z) \le (\boldsymbol{z} + t\boldsymbol{d}_z)^T\tilde{\boldsymbol{\pi}} - \log|\boldsymbol{A}(\tilde{\boldsymbol{\pi}})|$, therefore

$$\begin{aligned}
&\frac{F(\boldsymbol{v} + t\boldsymbol{d}) - F(\boldsymbol{v})}{t} = \boldsymbol{d}_z^T\boldsymbol{\pi} + \frac{g^*(\boldsymbol{z}) - g^*(\boldsymbol{z} + t\boldsymbol{d}_z)}{t} \\
&\ge \boldsymbol{d}_z^T(\boldsymbol{\pi} - \tilde{\boldsymbol{\pi}}) + \frac{g^*(\boldsymbol{z}) + \log|\boldsymbol{A}(\tilde{\boldsymbol{\pi}})| - \boldsymbol{z}^T\tilde{\boldsymbol{\pi}}}{t} \\
&= \pi_i(\boldsymbol{d}_z)_i/2 + \frac{g^*(\boldsymbol{z}) + \log|\boldsymbol{A}(\tilde{\boldsymbol{\pi}})| - \boldsymbol{z}^T\tilde{\boldsymbol{\pi}}}{t},
\end{aligned}$$

which is positive as $t \to \infty$. ∎

References

[1] D. Barber. Expectation correction for smoothing in switching linear Gaussian state space models. *Journal of Machine Learning Research*, 7:2515–2540, 2006.

[2] W. Chu and Z. Ghahramani. Gaussian processes for ordinal regression. *Journal of Machine Learning Research*, 6:1019–1041, 2005.

[3] P. Combettes and J. Pesquet. Proximal splitting methods in signal processing. In H. Bauschke, R. Burachik, P. Combettes, V. Elser, D. Luke, and H. Wolkowicz, editors, *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*. Springer, 2010.

[4] M. van Gerven, B. Cseke, F. de Lange, and T. Heskes. Efficient Bayesian multivariate fMRI analysis using a sparsifying spatio-temporal prior. *Neuroimage*, 50:150–161, 2010.

[5] S. Gerwinn, J. Macke, M. Seeger, and M. Bethge. Bayesian inference for spiking neuron models with a sparsity prior. In Platt et al. [16].

[6] P. Hojen-Sorensen, O. Winther, and L. Hansen. Mean field approaches to independent component analysis. *Neural Computation*, 14:889–918, 2002.

[7] M. Kuss and C. Rasmussen. Assessing approximate inference for binary Gaussian process classification. *Journal of Machine Learning Research*, 6:1679–1704, 2005.

[8] N. Lawrence and M. Jordan. Semi-supervised learning via Gaussian processes. In L. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*. MIT Press, 2005.

[9] A. Levin, Y. Weiss, F. Durand, and W. Freeman. Understanding and evaluating blind deconvolution algorithms. In *Computer Vision and Pattern Recognition*, 2009.

[10] D. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[11] T. Minka. Expectation propagation for approximate Bayesian inference. In J. Breese and D. Koller, editors, *Uncertainty in Artificial Intelligence 17*. Morgan Kaufmann, 2001.

[12] H. Nickisch and C. Rasmussen. Approximations for binary Gaussian process classification. *Journal of Machine Learning Research*, 9:2035–2078, 2008.

[13] H. Nickisch and M. Seeger. Convex variational Bayesian inference for large scale generalized linear models. In A. Danyluk, L. Bottou, and M. Littman, editors, *International Conference on Machine Learning 26*, volume 382, pages 761–768. ACM, 2009.

[14] M. Opper and O. Winther. Adaptive and self-averaging Thouless-Anderson-Palmer mean field theory for probabilistic modeling. *Physical Review E*, 64(056131), 2001.

[15] M. Opper and O. Winther. Expectation consistent approximate inference. *Journal of Machine Learning Research*, 6:2177–2204, 2005.

[16] J. Platt, D. Koller, Y. Singer, and S. Roweis, editors. *Advances in Neural Information Processing Systems 20*. Curran Associates, 2008.

[17] R. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

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

[19] M. Seeger. Gaussian covariance and scalable variational inference. In J. Fürnkranz and T. Joachims, editors, *International Conference on Machine Learning 27*. Omni Press, 2010.

[20] M. Seeger and M. I. Jordan. Sparse Gaussian process classification with multiple classes. Technical Report 661, Department of Statistics, University of California at Berkeley, 2004.

[21] M. Seeger and H. Nickisch. Large scale Bayesian inference and experimental design for sparse linear models. *SIAM Journal of Imaging Sciences*, 4(1):166–199, 2011.

[22] M. Seeger, H. Nickisch, R. Pohmann, and B. Schölkopf. Bayesian experimental design of magnetic resonance imaging sequences. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1441–1448. Curran Associates, 2009.

[23] D. Wipf and S. Nagarajan. A new view of automatic relevance determination. In Platt et al. [16], pages 1625–1632.