Journal of Machine Learning Research 14 (2013) 2239-2286 Submitted 9/12; Revised 4/13; Published 8/13

Gaussian Kullback-Leibler Approximate Inference

Edward Challis E.CHALLIS@CS.UCL.AC.UK
David Barber D.BARBER@CS.UCL.AC.UK
Department of Computer Science
University College London
London, WC1E 6BT, UK

Editor: Manfred Opper

Abstract

We investigate Gaussian Kullback-Leibler (G-KL) variational approximate inference techniques for Bayesian generalised linear models and various extensions. In particular we make the following novel contributions: sufficient conditions for which the G-KL objective is differentiable and convex are described; constrained parameterisations of Gaussian covariance that make G-KL methods fast and scalable are provided; the lower bound to the normalisation constant provided by G-KL methods is proven to dominate those provided by local lower bounding methods; complexity and model applicability issues of G-KL versus other Gaussian approximate inference methods are discussed. Numerical results comparing G-KL and other deterministic Gaussian approximate inference methods are presented for: robust Gaussian process regression models with either Student-*t* or Laplace likelihoods, large scale Bayesian binary logistic regression models, and Bayesian sparse linear models for sequential experimental design.

Keywords: generalised linear models, latent linear models, variational approximate inference, large scale inference, sparse learning, experimental design, active learning, Gaussian processes

1. Introduction

For a vector of parameters $\mathbf{w} \in \mathbb{R}^D$, in a large class of probabilistic models we require the inferential quantities

$$p(\mathbf{w}) = \frac{1}{Z}\mathcal{N}(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \prod_{n=1}^{N} \phi_n(\mathbf{w}), \tag{1}$$

$$Z = \int \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \prod_{n=1}^{N} \phi_n(\mathbf{w}) d\mathbf{w}, \tag{2}$$

where $p(\mathbf{w})$ is a multivariate real valued probability density function, $\mathcal{N}(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a multivariate Gaussian density with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and $\{\phi_n\}_{n=1}^N$ are positive, real valued, non-Gaussian potential functions.

The range of models that require us to compute these quantities is broad. In the Bayesian setting an important class is Bayesian generalised linear models (GLMs) for which examples include: sparse Bayesian linear models, where the Gaussian term is the likelihood and $\{\phi_n\}_{n=1}^N$ are factors of the sparse prior (Park and Casella, 2008); Gaussian process models, where the Gaussian term is a prior over latent function values and $\{\phi_n\}_{n=1}^N$ are factors of the non-Gaussian likelihood (Vanhatalo et al., 2009); and binary logistic regression, where the Gaussian term is a prior on the parameter

vector and $\{\phi_n\}_{n=1}^N$ are logistic sigmoid likelihood functions (Jaakkola and Jordan, 1997). In the context of unsupervised learning examples include: independent components analysis, where the Gaussian term is the conditional density of the signals and $\{\phi_n\}_{n=1}^N$ are factors of the sparse density on the latent sources $\mathbf{w}$ (Girolami, 2001); and binary or categorical factor analysis models where the Gaussian term is the density on the latent variables and $\{\phi_n\}_{n=1}^N$ are factors of the binary or multinomial conditional distribution (Tipping, 1999; Marlin et al., 2011).

In Bayesian supervised learning, Z is the marginal likelihood, otherwise termed the evidence, and the target density $p(\mathbf{w})$ is the posterior of the parameters conditioned on the data. Evaluating Z is essential for the purposes of model comparison, hyperparameter estimation, active learning and experimental design. Indeed, any marginal function of the posterior such as a moment, or a predictive density estimate also implicitly requires Z.

In unsupervised learning, Z is the model likelihood obtained by marginalising out the hidden variables $\mathbf{w}$ and $p(\mathbf{w})$ is the density of the hidden variables conditioned on the visible variables. $p(\mathbf{w})$ is required to optimise model parameters using either expectation maximisation or gradient ascent methods.

Computing Z, in either the Bayesian or unsupervised learning setting, is typically intractable due to the size of most problems of practical interest, which is usually much greater than one both in the dimension D and the number of potential functions N. Methods that can efficiently approximate these quantities are thus required.

Due to the importance of this model class, a great deal of effort has been dedicated to finding accurate approximations to $p(\mathbf{w})$ and Z. Whilst there are many different possible approximation routes, including sampling, consistency methods such as expectation propagation and perturbation techniques such as the Laplace method, our focus here is on a technique that lower-bounds Z and makes a Gaussian approximation to the target density $p(\mathbf{w})$.

We obtain a Gaussian approximation to $p(\mathbf{w})$ and a lower-bound on $\log Z$ by minimising the Kullback-Leibler divergence between the approximating Gaussian density and $p(\mathbf{w})$. Gaussian Kullback-Leibler approximate inference, which is how we refer to this procedure, is not new (Saul et al., 1996; Barber and Bishop, 1998; Seeger, 1999b; Kuss and Rasmussen, 2005; Opper and Archambeau, 2009). However, as we outline in the following subsection, we provide a number of theoretical and practical developments regarding its application.

1.1 Overview

In Section 2 we provide an introduction and overview of Gaussian Kullback-Leibler (G-KL) approximate inference methods for problems of the form of Equation (2) and describe a large class of models for which G-KL inference is feasible.

In Section 3 we address G-KL bound optimisation. We provide conditions on the potential functions $\{\phi_n\}_{n=1}^N$ for which the G-KL bound is smooth and concave. Thus we provide conditions for which optimisation using Newton's method will exhibit quadratic convergence rates and using quasi-Newton methods superlinear convergence rates.

In Section 4 we discuss the complexity of G-KL bound and gradient computations required to perform approximate inference. To make G-KL approximate inference scalable we present constrained parameterisations of covariance.

In Section 5 we compare G-KL approximate inference to other Gaussian approximate inference methods. We prove that the G-KL lower-bound is tighter than the bound offered by local

lower-bounding methods. We also discuss and compare computational scaling properties and model applicability issues.

In Section 6 we apply the G-KL procedure to three popular machine learning models. First, we consider the problem of Gaussian process regression with noise robust non-conjugate likelihoods. Second, we apply G-KL approximate inference to large Bayesian binary classification tasks. Third, we consider sequential experimental design in Bayesian sparse linear models. In these experiments we aim to assess the performance of the G-KL procedure in terms of speed, accuracy of inference and predictive performance. Results are compared to other deterministic Gaussian approximate inference procedures.

2. Gaussian KL Approximate Inference

The primary assumption of this work is that a target density of the form of Equation (2) with unbounded support in $\mathbb{R}^D$ is reasonably approximated by a Gaussian. Many approximate inference methods make this assumption, for example the Laplace approximation (see Barber, 2012 for a recent introduction), expectation propagation with an assumed Gaussian approximating density (Minka, 2001) and local variational bounding methods (Jaakkola and Jordan, 1997). This paper considers the method of fitting a Gaussian to $p(\mathbf{w})$ by minimising the Kullback-Leibler divergence between the two densities.

The Kullback-Leibler (KL) divergence for two probability density functions $q(\mathbf{w})$ and $p(\mathbf{w})$ is defined as

$$\mathrm{KL}(q(\mathbf{w})|\,p(\mathbf{w})) := \int_{\mathcal{W}} q(\mathbf{w}) \log \frac{q(\mathbf{w})}{p(\mathbf{w})} d\mathbf{w}, \tag{3}$$

where $\mathcal{W}$ is the support of $q(\mathbf{w})$. The KL divergence has the properties: $\mathrm{KL}(q(\mathbf{w})|\,p(\mathbf{w})) \geq 0$ for all $p(\mathbf{w})$ and $q(\mathbf{w})$, $\mathrm{KL}(q(\mathbf{w})|\,p(\mathbf{w})) = 0$ iff $q(\mathbf{w}) = p(\mathbf{w})$ almost everywhere, and $\mathrm{KL}(q(\mathbf{w})|\,p(\mathbf{w})) \neq \mathrm{KL}(p(\mathbf{w})|\,q(\mathbf{w}))$ for $q(\mathbf{w}) \neq p(\mathbf{w})$. The KL divergence, whilst not being a true metric, is thus a measure of the discrepancy between two probability distributions.

G-KL approximate inference proceeds by fitting the 'variational' Gaussian, $q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m},\mathbf{S})$, to the target, $p(\mathbf{w})$, by minimising $\mathrm{KL}(q(\mathbf{w})|\,p(\mathbf{w}))$ with respect to the moments $\mathbf{m}$ and $\mathbf{S}$. Substituting Equation (1) into Equation (3), using the fact that the KL divergence is non-negative, we obtain the bound $\log Z \geq \mathcal{B}_{KL}(\mathbf{m},\mathbf{S})$ where

$$\mathcal{B}_{KL}(\mathbf{m},\mathbf{S}) := \underbrace{-\left\langle \log q(\mathbf{w}) \right\rangle_{q(\mathbf{w})}}_{\text{entropy}} + \underbrace{\left\langle \log \mathcal{N}(\mathbf{w}|\boldsymbol{\mu},\boldsymbol{\Sigma}) \right\rangle_{q(\mathbf{w})}}_{\text{Gaussian potential}} + \underbrace{\sum_{n=1}^{N} \left\langle \log \phi_n(\mathbf{w}) \right\rangle_{q(\mathbf{w})}}_{\text{site potentials}}, \tag{4}$$

and $\left\langle f(x) \right\rangle_{p(x)}$ denotes taking the expectation of $f(x)$ with respect to the density $p(x)$. Unless otherwise stated all expectations should be assumed to be taken with respect to $q(\mathbf{w})$ and so we omit this subscript in the following notation.

The first and second terms in Equation (4) are integrals that admit simple analytic forms, the last term in general does not. The entropy of the variational Gaussian distribution is equal to $\frac{1}{2}\log\det(2\pi e\mathbf{S})$. The second term is the Gaussian expectation of a negative quadratic in $\mathbf{w}$ which is itself a negative quadratic in $\mathbf{m},\mathbf{S}$. Whilst the site potentials are not, in full generality, easy to evaluate in the next section we describe a large class of models for which they can be computed efficiently and so for which the G-KL bound is tractable.

2.1 Tractable G-KL Approximations

To evaluate the G-KL bound, Equation (4), we are required to compute $\sum_{n=1}^{N} \langle \log \phi_n(\mathbf{w}) \rangle$. For generic potential functions $\{\phi_n\}_{n=1}^{N}$ computing the required integrals is not always a numerically accessible task. However, in many practical problems of interest each potential function ϕ_n takes the form

$$\phi_n(\mathbf{w}) = \phi_n(\mathbf{w}^\mathsf{T}\mathbf{h}_n), \tag{5}$$

for fixed vectors $\mathbf{h}_n$. We refer to such potentials as site projections. We note that the linear projection of a Gaussian random vector is also Gaussian distributed. That is to say if $y = \mathbf{w}^\mathsf{T}\mathbf{h}$ where $\mathbf{w} \sim \mathcal{N}(\mathbf{w}|\mathbf{m},\mathbf{S})$ and $\mathbf{h}$ is fixed then $y \sim \mathcal{N}\left(y|\mathbf{h}^\mathsf{T}\mathbf{m}, \mathbf{h}^\mathsf{T}\mathbf{S}\mathbf{h}\right)$. We can use this result to express $\left\langle \log \phi_n(\mathbf{w}^\mathsf{T}\mathbf{h}_n) \right\rangle$ as a one-dimensional integral

$$\left\langle \log \phi_n(\mathbf{w}^\mathsf{T}\mathbf{h}_n) \right\rangle = \langle \log \phi_n(x) \rangle_{\mathcal{N}(x|m_n,s_n^2)} = \langle \log \phi_n(m_n + z s_n) \rangle_{\mathcal{N}(z|0,1)} \tag{6}$$

with $m_n := \mathbf{m}^\mathsf{T}\mathbf{h}_n$ and $s_n^2 := \mathbf{h}_n^\mathsf{T}\mathbf{S}\mathbf{h}_n$ (this result is presented in Barber and Bishop (1998) and Kuss and Rasmussen (2005) and Appendix A of this paper). The required integral can then be readily computed either analytically (for example $\phi(x) \propto e^{-|x|}$) or more generally using any one-dimensional numerical integration routine.

For models of the form of Equation (2), with each potential $\phi_n(\mathbf{w})$ a site projection, the G-KL bound can thus be expressed as

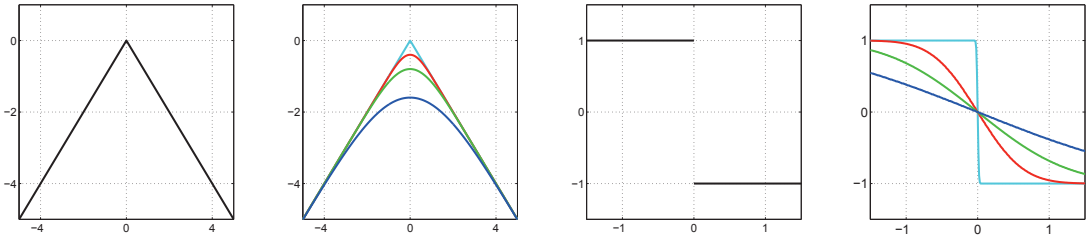
$$\mathcal{B}_{KL}(\mathbf{m},\mathbf{S}) = \underbrace{\frac{1}{2}\log\det(2\pi e\mathbf{S})}_{\text{entropy}} + \underbrace{\sum_{n=1}^{N} \langle \log \phi_n(m_n + z s_n) \rangle_{\mathcal{N}(z|0,1)}}_{\text{site projection potentials}} - \underbrace{\frac{1}{2}\left[\log\det(2\pi\boldsymbol{\Sigma}) + (\mathbf{m}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{m}-\boldsymbol{\mu}) + \operatorname{trace}\left(\boldsymbol{\Sigma}^{-1}\mathbf{S}\right)\right]}_{\text{Gaussian potential}}. \tag{7}$$

2.1.1 NON-GAUSSIAN MODELS

G-KL approximate inference is not limited to models where the target density has a Gaussian potential $\mathcal{N}(\mathbf{w}|\boldsymbol{\mu},\boldsymbol{\Sigma})$. The bound can be evaluated in the more general case $p(\mathbf{w}) \propto \prod_{n=1}^{N} \phi_n(\mathbf{w}^\mathsf{T}\mathbf{h}_n)$ with each ϕ_n non-Gaussian. One concrete example of this scenario is binary logistic regression with a Laplace prior on the parameter vector $\mathbf{w}$. In this context each ϕ_n potential function corresponds to either the logistic sigmoid factors specifying the likelihood or the $e^{-|w_d|}$ Laplace factors specifying the prior. When $p(\mathbf{w}) \propto \prod_{n=1}^{N} \phi_n(\mathbf{w}^\mathsf{T}\mathbf{h}_n)$ the G-KL bound consists only of the first two terms in Equation (7).

3. G-KL Bound Optimisation

G-KL approximate inference proceeds to obtain the tightest lower-bound to $\log Z$ and the 'closest' Gaussian approximation to $p(\mathbf{w})$ by maximising $\mathcal{B}_{KL}(\mathbf{m},\mathbf{S})$ with respect to the moments $\mathbf{m}$ and $\mathbf{S}$ of the variational Gaussian density. Therefore, to realise the benefits of G-KL approximate inference we require stable and scalable algorithms to optimise the bound. To this end we now show that for a broad class of models the G-KL objective is both differentiable and concave.



(a) $-|x|$ (b) $-\langle|x|\rangle$ (c) $-\text{sgn}(x)$ (d) $-\langle\text{sgn}(x)\rangle$

Figure 1: Non-differentiable functions and their Gaussian expectations. Figures (a) and (c) plot the non-differentiable function $\psi(x) = -|x|$ and the non-continuous function $\psi(x) = -\text{sgn}(x)$. Figures (b) and (c) plot the expectations of those functions for Gaussian distributed x as a function of the Gaussian mean m: $\langle\psi(x)\rangle_{\mathcal{N}(x|m,\sigma^2)}$. The expectations are smooth w.r.t. the Gaussian mean. As the variance of the Gaussian tends to zero the expectation converges to the underlying function value. Gaussian expectations taken w.r.t. $\mathcal{N}\left(x|m,\sigma^2\right)$ where $\sigma = 0.0125, 0.5, 1, 2$.

3.1 G-KL Bound Differentiability

Whilst the target density of our model may not be differentiable in $\mathbf{w}$ the G-KL bound with respect to the variational moments $\mathbf{m}, \mathbf{S}$ frequently is. See Figure 1 for a depiction of this phenomenon for two, simple, non-differentiable functions. The G-KL bound is in fact smooth for potential functions that are neither differentiable nor continuous (for example they have jump discontinuities). In Appendix C we show that the G-KL bound is smooth for potential functions that are piecewise smooth with a finite number of discontinuities, and where the logarithm of each piecewise segment is a quadratic. This class of functions includes the widely used Laplace density amongst others.

3.2 G-KL Bound Concavity

If each site potential $\{\phi_n\}_{n=1}^N$ is log-concave then the G-KL bound $\mathcal{B}_{KL}(\mathbf{m}, \mathbf{S})$ is jointly concave with respect to the variational Gaussian mean $\mathbf{m}$ and $\mathbf{C}$ the upper triangular Cholesky decomposition of covariance such that $\mathbf{S} = \mathbf{C}^\mathsf{T}\mathbf{C}$. We say that $f(x)$ is log-concave if $\log f(x)$ is concave in x.

Since the bound depends on the logarithm of $\prod_{n=1}^N \phi_n$ without loss of generality we may take $N = 1$. Ignoring constants with respect to $\mathbf{m}$ and $\mathbf{C}$, we can write the G-KL bound as

$$\mathcal{B}_{KL}(\mathbf{m}, \mathbf{C}) \overset{\text{c.}}{=} \sum_{d=1}^{D} \log C_{dd} - \frac{1}{2}\mathbf{m}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\mathbf{m} + \boldsymbol{\mu}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\mathbf{m} - \frac{1}{2}\text{trace}\left(\boldsymbol{\Sigma}^{-1}\mathbf{C}\mathbf{C}^\mathsf{T}\right) + \left\langle \log\phi(\mathbf{w}^\mathsf{T}\mathbf{h})\right\rangle. \quad (8)$$

Excluding $\left\langle \log\phi(\mathbf{w}^\mathsf{T}\mathbf{h})\right\rangle$ from the expression above all terms are concave functions exclusively in either $\mathbf{m}$ or $\mathbf{C}$. Since the sum of concave functions on distinct variables is jointly concave, the first four terms of Equation (8) represent a jointly concave contribution to the bound.

To complete the proof[1] we need to show that $\left\langle \log \phi(\mathbf{w}^\mathsf{T}\mathbf{h}) \right\rangle$ is jointly concave in $\mathbf{m}$ and $\mathbf{C}$. Log-concavity of $\phi(x)$ is equivalent to the statement that for any $x_1, x_2 \in \mathbb{R}$ and any $\theta \in [0,1]$

$$\log \phi(\theta x_1 + (1-\theta)x_2) \geq \theta \log \phi(x_1) + (1-\theta)\log \phi(x_2). \tag{9}$$

Therefore, to show that $\mathcal{E}(\mathbf{m}, \mathbf{C}) := \left\langle \log \phi(\mathbf{w}^\mathsf{T}\mathbf{h}) \right\rangle_{\mathcal{N}(\mathbf{w}|\mathbf{m}, \mathbf{C}^\mathsf{T}\mathbf{C})}$ is concave it suffices to show for any $\theta \in [0,1]$ that

$$\mathcal{E}(\theta\mathbf{m}_1 + (1-\theta)\mathbf{m}_2, \theta\mathbf{C}_1 + (1-\theta)\mathbf{C}_2) \geq \theta\mathcal{E}(\mathbf{m}_1, \mathbf{C}_1) + (1-\theta)\mathcal{E}(\mathbf{m}_2, \mathbf{C}_2).$$

This can be done by making the substitution $\mathbf{w} = \theta\mathbf{m}_1 + (1-\theta)\mathbf{m}_2 + (\theta\mathbf{C}_1 + (1-\theta)\mathbf{C}_2)^\mathsf{T}\mathbf{z}$, giving

$$\begin{aligned}\mathcal{E}(\theta\mathbf{m}_1 + (1-\theta)\mathbf{m}_2, \theta\mathbf{C}_1 + (1-\theta)\mathbf{C}_2) = \int \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I}) \times \\ \log \phi\left(\theta\mathbf{h}^\mathsf{T}\left(\mathbf{m}_1 + \mathbf{C}_1^\mathsf{T}\mathbf{z}\right) + (1-\theta)\mathbf{h}^\mathsf{T}\left(\mathbf{m}_2 + \mathbf{C}_2^\mathsf{T}\mathbf{z}\right)\right) d\mathbf{z}.\end{aligned}$$

Using concavity of $\log \phi(x)$ with respect to x and Equation (9) with $\mathbf{w}_1 = \mathbf{m}_1 + \mathbf{C}_1^\mathsf{T}\mathbf{z}$ and $\mathbf{w}_2 = \mathbf{m}_2 + \mathbf{C}_2^\mathsf{T}\mathbf{z}$ we have that

$$\begin{aligned}\mathcal{E}(\theta\mathbf{m}_1 + (1-\theta)\mathbf{m}_2, \theta\mathbf{C}_1 + (1-\theta)\mathbf{C}_2) &\geq \theta \int \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I}) \log \phi\left(\mathbf{h}^\mathsf{T}\left(\mathbf{m}_1 + \mathbf{C}_1^\mathsf{T}\mathbf{z}\right)\right) d\mathbf{z} \\ &\quad + (1-\theta) \int \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I}) \log \phi\left(\mathbf{h}^\mathsf{T}\left(\mathbf{m}_2 + \mathbf{C}_2^\mathsf{T}\mathbf{z}\right)\right) d\mathbf{z} \\ &= \theta\mathcal{E}(\mathbf{m}_1, \mathbf{C}_1) + (1-\theta)\mathcal{E}(\mathbf{m}_2, \mathbf{C}_2).\end{aligned}$$

Thus the G-KL bound is jointly concave in $\mathbf{m}, \mathbf{C}$ provided all site potentials $\{\phi_n\}_{n=1}^N$ are log-concave.

With consequence to the theoretical convergence rates of gradient based optimisation procedures, the bound is also strongly-concave. A function $f(\mathbf{x})$ is strongly-concave if there exists some $c < 0$ such that for all $\mathbf{x}$, $\nabla^2 f(\mathbf{x}) \preceq c\mathbf{I}$ (Boyd and Vandenberghe, 2004, Section 9.1.2).[2] For the G-KL bound the constant c can be assessed by inspecting the covariance of the Gaussian potential, $\mathbf{\Sigma}$. If we arrange the set of all G-KL variational parameters as a vector formed by concatenating $\mathbf{m}$ and the non-zero elements of the column's of $\mathbf{C}$ then the Hessian of $\left\langle \log \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}, \mathbf{\Sigma}) \right\rangle$ is a block diagonal matrix. Each block of this Hessian is either $-\mathbf{\Sigma}^{-1}$ or its submatrix $\left[-\mathbf{\Sigma}^{-1}\right]_{i:D,i:D}$, where $i = 2, \ldots, D$. The set of eigenvalues of a block diagonal matrix is the union of the eigenvalues of each of the block matrices' eigenvalues. Furthermore, the eigenvalues of each submatrix are bounded by the upper and lower eigenvalues of $-\mathbf{\Sigma}^{-1}$. Therefore $\nabla^2 \mathcal{B}_{KL}(\mathbf{m}, \mathbf{S}) \succeq c\mathbf{I}$ where c is -1 times the smallest eigenvalue of $\mathbf{\Sigma}^{-1}$. The sum of a strongly-concave function and a concave function is strongly-concave and thus the G-KL bound as a whole is strongly-concave.

3.3 Summary

In this section, and in Appendix C, we have provided conditions for which the G-KL bound is strongly concave, smooth, has closed sublevel sets and Lipschitz continuous Hessians. Under these

1. This proof was provided by Michalis K. Titsias and simplifies the original presentation made in (Challis and Barber, 2011).
2. We say for square matrices $\mathbf{A}$ and $\mathbf{B}$ that $\mathbf{A} \preceq \mathbf{B}$ iff $\mathbf{B} - \mathbf{A}$ is positive semidefinite.

conditions optimisation of the G-KL bound will have quadratic convergence rates using Newton's method and super-linear convergence rates using quasi-Newton methods (Nocedal and Wright, 2006; Boyd and Vandenberghe, 2004). For larger problems, where cubic scaling properties arising from the approximate Hessian calculations required by quasi-Newton methods are infeasible, we will use limited memory quasi-Newton methods, nonlinear conjugate gradients or Hessian free Newton methods to optimise the G-KL bound.

Concavity with respect to the G-KL mean is clear and intuitive—for any fixed G-KL covariance the G-KL bound as a function of the mean can be interpreted as a Gaussian blurring of $\log p(\mathbf{w})$—see Figure 1. As $\mathbf{S} = \nu^2\mathbf{I} \to \mathbf{0}$ then $\mathbf{m}^* \to \mathbf{w}^{MAP}$ where $\mathbf{m}^*$ is the optimal G-KL mean and $\mathbf{w}^{MAP}$ is the maximum a posteriori (MAP) parameter setting.

Another deterministic Gaussian approximate inference procedure for models of the form of Equation (2) are local variational bounding methods (discussed at further length in Section 5.1.1). For log-concave potentials local variational bounding methods, which optimise a different criterion with a different parameterisation to the G-KL bound, have also been shown to result in a convex optimisation problem (Seeger and Nickisch, 2011b). To the best of our knowledge, local variational bounding and G-KL approximate inference methods are the only known concave variational inference procedures for models of the form of Equation (2).

Whilst G-KL bound optimisation and MAP estimation share conditions under which they are concave problems, the G-KL objective is often differentiable when the MAP objective is not. Non-differentiable potentials are used throughout machine learning and statistics. Indeed, the practical utility of such non-differentiable potentials in statistical modelling has driven a lot of research into speeding up algorithms to find the mode of these densities—for example see Schmidt et al. (2007). Despite recent progress these algorithms tend to have slower convergence rates than quasi-Newton methods on smooth, strongly-convex objectives with Lipschitz continuous gradients and Hessians.

One of the significant practical advantages of G-KL approximate inference over MAP estimation and the Laplace approximation is that the target density is not required to be differentiable. With regards to the complexity of G-KL bound optimisation, whilst an additional cost is incurred over MAP estimation from specifying and optimising the variance of the approximation, a saving is made in the number of times the objective and its gradients need to be computed. Quantifying the net saving (or indeed cost) of G-KL optimisation over MAP estimation is an interesting question reserved for later work.

4. Complexity : G-KL Bound and Gradient Computations

In the previous section we provided conditions for which the G-KL bound is strongly concave and differentiable and so provided conditions for which G-KL bound optimisation using quasi-Newton methods will exhibit super-linear convergence rates. Whilst such convergence rates are highly desirable they do not in themselves guarantee that optimisation is scalable. An important practical consideration is the numerical complexity of the bound and gradient computations required by any gradient ascent optimisation procedure.

Discussing the complexity of G-KL bound and gradient evaluations in full generality is complex we therefore restrict ourselves to considering one particularly common case. We consider models where the covariance of the Gaussian potential in Equation (2) is spherical, $\mathbf{\Sigma} = \nu^2\mathbf{I}$, and each potential function is a site projection, $\phi_n(\mathbf{w}) = \phi_n(\mathbf{w}^\mathsf{T}\mathbf{h}_n)$. For models that do not satisfy this assumption, in Appendix D we present a full breakdown of the complexity of bound and gradient

computations for each G-KL covariance parameterisation presented in Section 4.1.3 and a range of parameterisations for the Gaussian potential $\mathcal{N}(\mathbf{w}|\mathbf{m},\mathbf{\Sigma})$.

Note that problems where $\mathbf{\Sigma}$ is not a scaling of the identity can be reparameterised to an equivalent problem for which it is. For some problems this reparameterisation can provide significant reductions in complexity. The procedure, the domains for which it is suitable, and the possible computational savings it provides are discussed at further length in Appendix E.

For Cholesky factorisations of covariance, $\mathbf{S} = \mathbf{C}^\mathsf{T}\mathbf{C}$, of dimension D the bound and gradient contributions from the $\log\det(\mathbf{S})$ and $\mathrm{trace}(\mathbf{S})$ terms in Equation (7) scale $O(D)$ and $O\left(D^2\right)$ respectively. Terms in Equation (7) that are a function exclusively of the G-KL mean, $\mathbf{m}$, scale at most $O(D)$ and are the cheapest to evaluate. The computational bottleneck arises from the projected variational variances $s_n^2 = \|\mathbf{C}^\mathsf{T}\mathbf{h}_n\|^2$ required to compute each $\left\langle \log\phi_n(\mathbf{w}^\mathsf{T}\mathbf{h}_n)\right\rangle$ term. Computing all such projected variances scales $O\left(ND^2\right)$.[3]

A further computational expense is incurred from computing the N one dimensional integrals required to evaluate $\sum_{n=1}^N \left\langle \log\phi_n(\mathbf{w}^\mathsf{T}\mathbf{h}_n)\right\rangle$. These integrals are computed either numerically or analytically depending on the functional form of ϕ_n. Regardless, this computation scales $O(N)$, possibly though with a significant prefactor. When numerical integration is required, we note that since $\left\langle \log\phi_n(\mathbf{w}^\mathsf{T}\mathbf{h}_n)\right\rangle$ can be expressed as $\langle \log\phi_n(m_n + z s_n)\rangle_{\mathcal{N}(z|0,1)}$ we can usually assert that the integrand's significant mass lies for $z \in [-5,5]$ and so that quadrature will yield sufficiently accurate results at modest computational expense. For all the experiments considered here we used fixed width rectangular quadrature and performing these integrals was not the principal bottleneck. For modelling scenarios where this is not the case we note that a two dimensional lookup table can be constructed, at a one off cost, to approximate $\langle \log\phi(m+zs)\rangle_{\mathcal{N}(z|0,1)}$ and its derivatives as a function of m and s.

Thus for a broad class of models the G-KL bound and gradient computations scale $O\left(ND^2\right)$ for general parameterisations of the covariance $\mathbf{S} = \mathbf{C}^\mathsf{T}\mathbf{C}$. In many problems of interest the fixed vectors $\mathbf{h}_n$ are sparse. Letting L denote the number of non-zero elements in each vector $\mathbf{h}_n$, computing $\left\{s_n^2\right\}_{n=1}^N$ scales now $O(NDL)$ where frequently $L \ll D$. Nevertheless, such scaling for the G-KL method can be prohibitive for large problems and so constrained parameterisations are required.

4.1 Constrained Parameterisations of G-KL Covariance

Unconstrained G-KL approximate inference requires storing and optimising $\frac{1}{2}D(D+1)$ parameters to specify the G-KL covariance's Cholesky factor $\mathbf{C}$. In many settings this can be prohibitive. To this end we now consider constrained parameterisations of covariance that reduce both the time and space complexity of G-KL procedures.

Gaussian densities can be parameterised with respect to the covariance or its inverse the precision matrix. A natural question to ask is which of these is best suited for G-KL bound optimisation. Unfortunately, the G-KL bound is neither concave nor convex with respect to the precision matrix. What is more, the complexity of computing the ϕ_n site potential contributions to the bound increases for the precision parameterised G-KL bound. Thus the G-KL bound seems more naturally parameterised in terms of covariance than precision.

3. We note that since a Gaussian potential, $\mathcal{N}(\mathbf{w}|\boldsymbol{\mu},\mathbf{\Sigma})$, can be written as a product over D site projection potentials computing $\left\langle \log\mathcal{N}(\mathbf{w}|\boldsymbol{\mu},\mathbf{\Sigma})\right\rangle$ will in general scale $O\left(D^3\right)$—see Appendix B.3.2.

4.1.1 OPTIMAL G-KL COVARIANCE STRUCTURE

As originally noted by Seeger (1999a), the optimal structure for the G-KL covariance can be assessed by calculating the derivative of $\mathcal{B}_{KL}(\mathbf{m},\mathbf{S})$ with respect to $\mathbf{S}$ and equating it to zero. Doing so, $\mathbf{S}$ satisfies

$$\mathbf{S}^{-1} = \boldsymbol{\Sigma}^{-1} + \mathbf{H}\boldsymbol{\Gamma}\mathbf{H}^{\mathsf{T}}, \tag{10}$$

where $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_n]$ and $\boldsymbol{\Gamma}$ is diagonal such that

$$\Gamma_{nn} = \left\langle \left(z^2 - 1\right) \frac{\log \phi_n(m_n + z s_n)}{2 s_n^2} \right\rangle_{\mathcal{N}(z|0,1)}. \tag{11}$$

$\boldsymbol{\Gamma}$ depends on $\mathbf{S}$ through the projected variance terms $s_n^2 = \mathbf{h}_n^{\mathsf{T}}\mathbf{S}\mathbf{h}_n$ and Equation (10) does not provide a closed form expression to solve for $\mathbf{S}$. Furthermore, iterating Equation (10) is not guaranteed to converge to a fixed point or uniformly increase the bound. Indeed this iterative procedure frequently diverges. We are free, however, to directly optimise the bound by treating the diagonal entries of $\boldsymbol{\Gamma}$ as variational parameters and thus change the number of parameters required to specify $\mathbf{S}$ from $\frac{1}{2}D(D+1)$ to N. This procedure, whilst possibly reducing the number of free parameters, requires us to compute $\log\det(\mathbf{S})$ where $\mathbf{S}$ has no convenient structure and so in general scales $O\left(D^3\right)$—infeasible when $D \gg 1$.

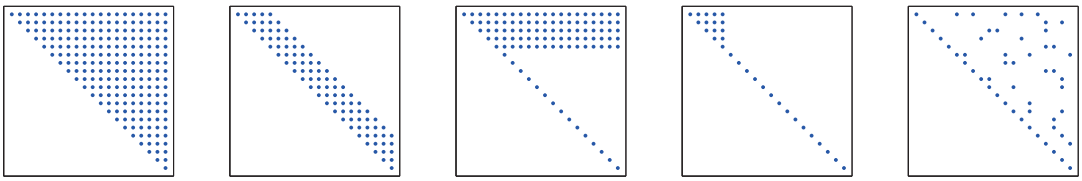
A further consequence of using this parameterisation of covariance is that the bound is non-concave. We know from Seeger and Nickisch (2011b) that parameterising $\mathbf{S}$ according to Equation (10) renders $\log\det(\mathbf{S})$ concave with respect to $(\Gamma_{nn})^{-1}$. However the site projection potentials are not concave with respect to $(\Gamma_{nn})^{-1}$ thus the bound is neither concave nor convex for this parameterisation resulting in convergence to a possibly local optimum. Non-convexity and $O\left(D^3\right)$ scaling motivates the search for better parameterisations of covariance. In Appendix B we provide equations for each term of the G-KL bound and its gradient for each of the covariance parameterisations considered below.

4.1.2 FACTOR ANALYSIS

Parameterisations of the form $\mathbf{S} = \boldsymbol{\Theta}\boldsymbol{\Theta}^{\mathsf{T}} + \operatorname{diag}\left(\mathbf{d}^2\right)$ can capture the K leading directions of variance for a $D \times K$ dimensional loading matrix $\boldsymbol{\Theta}$. Unfortunately this parameterisation renders the G-KL bound non-concave. Non-concavity is due to the entropic contribution $\log\det(\mathbf{S})$ which is not even unimodal. All other terms in the bound remain concave under this factorisation. Provided one is happy to accept convergence to possibly local optima, this is still a useful parameterisation. Computing the projected variances with $\mathbf{S}$ in this form scales $O(NDK)$ and evaluating $\log\det(\mathbf{S})$ and its derivative scales $O\left(K^2(K+D)\right)$.

4.1.3 CONSTRAINED CONCAVE PARAMETERISATIONS

Below we present constrained parameterisations of covariance which reduce both the space and time complexity of G-KL bound optimisation whilst preserving concavity. To reiterate, the computational scaling figures for the bound and gradient computations listed below correspond to evaluating the projected G-KL variance terms, the bottleneck for models with an isotropic Gaussian potential $\boldsymbol{\Sigma} = \nu^2\mathbf{I}$. The scaling properties for other models are presented in Appendix D. The constrained parameterisations below have different qualities regarding the expressiveness of the variational Gaussian approximation. We note that a zero at the $(i,j)^{th}$ element of covariance specifies



(a) Full (b) Banded (c) Chevron (d) Subspace (e) Sparse

Figure 2: Sparsity structure for constrained concave Cholesky decompositions of covariance.

a marginal independence relation between parameters w_i and w_j. Conversely, a zero at the $(i,j)^{th}$ element of precision corresponds to a independence relation between parameters w_i and w_j when conditioned on the other remaining parameters.

Banded Cholesky. The simplest option is to constrain the Cholesky matrix to be banded, that is $C_{ij} = 0$ for $j > i + B$ where B is the bandwidth. Doing so reduces the cost of a single bound or gradient computation to $O(NDB)$. Such a parameterisation describes a sparse covariance matrix and assumes zero covariance between variables that are indexed out of bandwidth. The precision matrix for banded Cholesky factorisations of covariance will in general be non-sparse.

Chevron Cholesky. We constrain $\mathbf{C}$ such that $C_{ij} = \Theta_{ij}$ when $j \geq i$ and $i \leq K$, $C_{ii} = d_i$ for $i > K$ and 0 otherwise. We refer to this parameterisation as the chevron Cholesky since the sparsity structure has a broad inverted 'V' shape—see Figure 2. Generally, this constrained parameterisation results in a non-sparse covariance but sparse precision. This parameterisation is not invariant to index permutations and so not all covariates have the same representational power. For a Cholesky matrix of this form bound and gradient computations scale $O(NDK)$.

Sparse Cholesky. In general the bound and gradient can be evaluated more efficiently if we impose any fixed sparsity structure on the Cholesky matrix $\mathbf{C}$. In certain modelling scenarios we know a priori which variables are marginally dependent and independent and so may be able construct a sparse Cholesky matrix to reflect that domain knowledge. This is of use in cases where a low band width index ordering cannot be found. For a sparse Cholesky matrix with DK non-zero elements bound and gradient computations scale $O(NDK)$.

Subspace Cholesky. Another reduced parameterisation of covariance can be obtained by considering arbitrary rotations in parameter space, $\mathbf{S} = \mathbf{E}^\mathsf{T}\mathbf{C}^\mathsf{T}\mathbf{C}\mathbf{E}$ where $\mathbf{E}$ is a rotation matrix which forms an orthonormal basis over $\mathbb{R}^D$. Substituting this form for the covariance into Equation (8) and for $\mathbf{\Sigma} = \nu^2\mathbf{I}$ we obtain, up to a constant,

$$\mathcal{B}_{KL}(\mathbf{m},\mathbf{C}) \stackrel{c.}{=} \sum_i \log C_{ii} - \frac{1}{2\nu^2}\left[\|\mathbf{C}\|^2 + \|\mathbf{m}\|^2\right] + \frac{1}{\nu^2}\boldsymbol{\mu}^\mathsf{T}\mathbf{m} + \sum_n \langle \log \phi(m_n + z s_n)\rangle_z$$

where $s_n = \|\mathbf{C}^\mathsf{T}\mathbf{E}^\mathsf{T}\mathbf{h}_n\|$. One may reduce the computational burden by decomposing $\mathbf{E}$ into two submatrices such that $\mathbf{E} = [\mathbf{E}_1, \mathbf{E}_2]$ where $\mathbf{E}_1$ is $D \times K$ and $\mathbf{E}_2$ is $D \times L$ for $L = (D-K)$. Constraining $\mathbf{C}$ such that $\mathbf{C} = \text{blkdiag}\,(\mathbf{C}_1, c\mathbf{I}_{L\times L})$, with $\mathbf{C}_1$ a $K \times K$ Cholesky matrix we have that

$$s_n^2 = \|\mathbf{C}_1^\mathsf{T}\mathbf{E}_1^\mathsf{T}\mathbf{h}_n\|^2 + c^2(\|\mathbf{h}_n\|^2 - \|\mathbf{E}_1^\mathsf{T}\mathbf{h}_n\|^2),$$

meaning that only the K subspace vectors in $\mathbf{E}_1$ are needed to compute $\left\{s_n^2\right\}_{n=1}^N$. Since terms such as $\|\mathbf{h}_n\|$ need only be computed once the complexity of bound and gradient computations reduces to

scaling in K not D. Further savings can be made if we use banded subspace Cholesky matrices: for $\mathbf{C}_1$ having bandwidth B each bound evaluation and associated gradient computation scales $O(NBK)$.

The success of the subspace Cholesky factorisation depends on how well $\mathbf{E}_1$ captures the leading directions of variance. One simple approach to select $\mathbf{E}_1$ is to use the leading principal components of the 'data set' $\mathbf{H}$. Another option is iterate between optimising the bound with respect to $\{\mathbf{m}, \mathbf{C}_1, c\}$ and $\mathbf{E}_1$. We consider two approaches for optimisation with respect to $\mathbf{E}_1$. The first uses the form for the optimal G-KL covariance, Equation (11). By substituting in the projected mean and variance terms m_n and s_n^2 into Equation (11) we can set $\mathbf{E}_1$ to be a rank K approximation to this $\mathbf{S}$. The best rank K approximation is given by evaluating the smallest K eigenvectors of $\boldsymbol{\Sigma}^{-1} + \mathbf{H}\boldsymbol{\Gamma}\mathbf{H}^{\mathsf{T}}$. For very large sparse problems $D \gg 1$ we approximate this using the iterative Lanczos methods described by Seeger and Nickisch (2010). For smaller non-sparse problems more accurate approximations are available. The second approach is to optimise the G-KL bound directly with respect to $\mathbf{E}_1$ under the constraint that the columns of $\mathbf{E}_1$ are orthonormal. One route to achieving this is to use a projected gradient ascent method. Each of these methods and the associated subspace G-KL gradients are presented in greater detail in Appendix B.4.

5. Comparing Gaussian Approximate Inference Procedures

Due to their favourable computational and analytical properties multivariate Gaussian densities are used by many deterministic approximate inference routines. For models of the form of Equation (2) three popular, deterministic, Gaussian, approximate inference techniques are local variational bounding, Laplace approximations, and expectation propagation with an assumed Gaussian density. In this section we briefly review and compare these methods to the G-KL procedure.

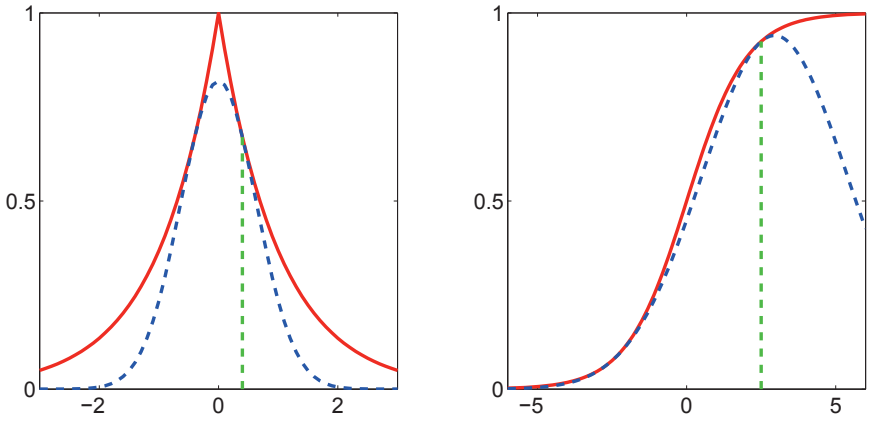
Of the three Gaussian approximate inference methods listed above only one, local variational bounding, provides a lower-bound to the normalisation constant Z. In Section 5.1 we give a brief overview of local bounding procedures and show that the G-KL lower-bound dominates the local lower-bound on $\log Z$.

In Section 5.2 we discuss the applicability of each Gaussian approximate inference method. Specifically we describe the computational scaling properties of each of the algorithms and the potential functions to which they can successfully be applied

5.1 Gaussian Lower-Bounds

An attractive property of G-KL approximate inference is that it provides a strict lower-bound on $\log Z$. Lower-bounding procedures are particularly useful for a number of theoretical and practical reasons. The primary theoretical advantage is that it provides concrete exact knowledge about Z and thus also the target density $p(\mathbf{w})$. Lower-bounds may also be used in conjunction with upper bounds to form bounds on marginal quantities of interest (Gibbs and MacKay, 2000). Thus the tighter the lower-bound on $\log Z$ the more informative it is. Practically, optimising a lower-bound is often a more numerically stable task than the criteria provided by other deterministic approximate inference methods.

Another well studied route to obtaining a lower-bound for problems of the form of Equation (2) are so called local variational bounding procedures, see for example: Jaakkola and Jordan (1997), Gibbs and MacKay (2000), Girolami (2001), Palmer et al. (2006), and Nickisch and Seeger (2009). Whilst both G-KL and local procedures have been discussed in the literature for some time, little work has been done to elucidate the relation between them. In Section 5.1.1 we give an overview



(a) (b)

Figure 3: Exponentiated quadratic lower-bounds for two super-Gaussian potential functions: (a) Laplace potential and lower-bound with operating point at 0.5; (b) Logistic sigmoid potential and lower-bound with operating point at 2.5.

of local variational bounding procedures. In Section 5.1.2 we prove that G-KL provides a tighter lower-bound on Z than local lower-bounding methods.

5.1.1 LOCAL VARIATIONAL BOUNDS

Local variational procedures lower-bound Z by replacing each potential ϕ_n in Equation (2) with a function that lower-bounds it and that renders the integral as a whole analytically tractable. Tractability is obtained by using exponentiated quadratic lower-bounds for each non-Gaussian site potential $\{\phi_n\}_{n=1}^N$. Local variational bounding procedures that use exponentiated quadratic site bounds return a Gaussian approximation to the target density $p(\mathbf{w})$.

Site potentials ϕ_n are known to have tight exponentiated quadratic lower-bounds provided they are super-Gaussian (Palmer et al., 2006). A function $f(x)$ is said to be super-Gaussian if $\exists b \in \mathbb{R}$ s.t. for $g(x) := \log f(x) - bx$ is even, convex and decreasing as a function of $y = x^2$. A number of potential functions of significant practical utility are super-Gaussian, examples include: the logistic sigmoid $\phi(x) = (1+\exp(-x))^{-1}$, the Laplace density $\phi(x) \propto \exp(-|x|)$ and the Student's t density—see Figure 3 for plots of these potential functions and their respective lower-bounds.

Each site projection potential function is lower-bounded by an exponentiated quadratic parameterised in $\mathbf{w}$ and a variational parameter ξ_n. Since exponentiated quadratics are closed under multiplication one may bound the product of site potentials by an exponentiated quadratic also

$$\prod_n \phi_n(\mathbf{w}^\mathsf{T}\mathbf{h}_n) \geq c(\boldsymbol{\xi}) e^{-\frac{1}{2}\mathbf{w}^\mathsf{T}\mathbf{F}(\xi)\mathbf{w}+\mathbf{w}^\mathsf{T}\mathbf{f}(\boldsymbol{\xi})}, \tag{12}$$

where the matrix $\mathbf{F}(\boldsymbol{\xi})$, vector $\mathbf{f}(\boldsymbol{\xi})$ and scalar $c(\boldsymbol{\xi})$ depend on the specific functions $\{\phi_n\}_{n=1}^N$ and the vectors $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_n]$; and $\boldsymbol{\xi}$ is a vector of length N containing the variational parameters ξ_n. For

any setting of $\mathbf{w}$ there exists a setting of $\boldsymbol{\xi}$ for which the bound is tight. Thus we can obtain a bound on Z by substituting Equation (12) into Equation (2):

$$\begin{aligned} Z &= \int \mathcal{N}(\mathbf{w}|\boldsymbol{\mu},\boldsymbol{\Sigma}) \prod_{n=1}^{N} \phi_n(\mathbf{w}^\mathsf{T}\mathbf{h}_n) d\mathbf{w} \\ &\geq \int \mathcal{N}(\mathbf{w}|\boldsymbol{\mu},\boldsymbol{\Sigma})\, c(\boldsymbol{\xi}) e^{-\frac{1}{2}\mathbf{w}^\mathsf{T}\mathbf{F}(\boldsymbol{\xi})\mathbf{w}+\mathbf{w}^\mathsf{T}\mathbf{f}(\boldsymbol{\xi})} d\mathbf{w} \\ &= c(\boldsymbol{\xi}) \frac{e^{-\frac{1}{2}\boldsymbol{\mu}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}}{\sqrt{\det(2\pi\boldsymbol{\Sigma})}} \int e^{-\frac{1}{2}\mathbf{w}^\mathsf{T}\mathbf{A}\mathbf{w}+\mathbf{w}^\mathsf{T}\mathbf{b}} d\mathbf{w}, \end{aligned} \tag{13}$$

where

$$\mathbf{A} := \boldsymbol{\Sigma}^{-1} + \mathbf{F}(\boldsymbol{\xi}) \quad \text{and} \quad \mathbf{b} := \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \mathbf{f}(\boldsymbol{\xi}). \tag{14}$$

Whilst both $\mathbf{A}$ and $\mathbf{b}$ are functions of $\boldsymbol{\xi}$, we drop this dependency for a more compact notation. One can interpret Equation (13) as a Gaussian approximation to the target density where $p(\mathbf{w}) \approx \mathcal{N}\left(\mathbf{w}|\mathbf{A}^{-1}\mathbf{b},\mathbf{A}^{-1}\right)$. Completing the square in Equation (13) and integrating, we have $\log Z \geq B(\boldsymbol{\xi})$, where

$$\mathcal{B}(\boldsymbol{\xi}) = \log c(\boldsymbol{\xi}) - \frac{1}{2}\boldsymbol{\mu}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \frac{1}{2}\mathbf{b}^\mathsf{T}\mathbf{A}^{-1}\mathbf{b} - \frac{1}{2}\log\det(2\pi\boldsymbol{\Sigma}) - \frac{1}{2}\log\det(2\pi\mathbf{A}).$$

To obtain the tightest bound on $\log Z$ one then maximises $\mathcal{B}(\boldsymbol{\xi})$ with respect to $\boldsymbol{\xi}$.

5.1.2 COMPARING G-KL AND LOCAL BOUNDS

An important question is which method, local or G-KL, gives a tighter lower-bound on $\log Z$. Each bound derives from a fundamentally different criterion and it is not immediately clear which if either is superior. The G-KL bound has been noted before, empirically in the case of binary classification (Nickisch and Rasmussen, 2008) and analytically for the special case of symmetric potentials (Seeger, 2009), to be tighter than the local bound. It is tempting to conclude that such observed superiority of the G-KL method is to be expected since the G-KL bound has potentially unrestricted covariance $\mathbf{S}$ and so a richer parameterisation. However, many problems have more site potentials ϕ_n than Gaussian moment parameters, that is $N > \frac{1}{2}D(D+3)$, and the local bound in such cases has a richer parameterisation than the G-KL.

We derive a relation between the local and G-KL bounds for $\{\phi_n\}_{n=1}^N$ generic super-Gaussian site potentials. We first substitute the local bound on $\prod_{n=1}^N \phi_n(\mathbf{w}^\mathsf{T}\mathbf{h}_n)$, Equation (12), into Equation (4) to obtain a new bound

$$\mathcal{B}_{KL}(\mathbf{m},\mathbf{S}) \geq \tilde{\mathcal{B}}_{KL}(\mathbf{m},\mathbf{S},\boldsymbol{\xi}),$$

where

$$\begin{aligned} 2\tilde{\mathcal{B}}_{KL} = -2\langle \log q(\mathbf{w})\rangle - \log\det(2\pi\boldsymbol{\Sigma}) + 2\log c(\boldsymbol{\xi}) - \left\langle (\mathbf{w}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{w}-\boldsymbol{\mu})\right\rangle \\ - \left\langle \mathbf{w}^\mathsf{T}\mathbf{F}(\boldsymbol{\xi})\mathbf{w}\right\rangle + 2\left\langle \mathbf{w}^\mathsf{T}\mathbf{f}(\boldsymbol{\xi})\right\rangle. \end{aligned}$$

Using Equation (14) this can be written as

$$\tilde{\mathcal{B}}_{KL} = -\langle \log q(\mathbf{w})\rangle - \frac{1}{2}\log\det(2\pi\boldsymbol{\Sigma}) + \log c(\boldsymbol{\xi}) - \frac{1}{2}\boldsymbol{\mu}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \frac{1}{2}\left\langle \mathbf{w}^\mathsf{T}\mathbf{A}\mathbf{w}\right\rangle + \left\langle \mathbf{w}^\mathsf{T}\mathbf{b}\right\rangle.$$

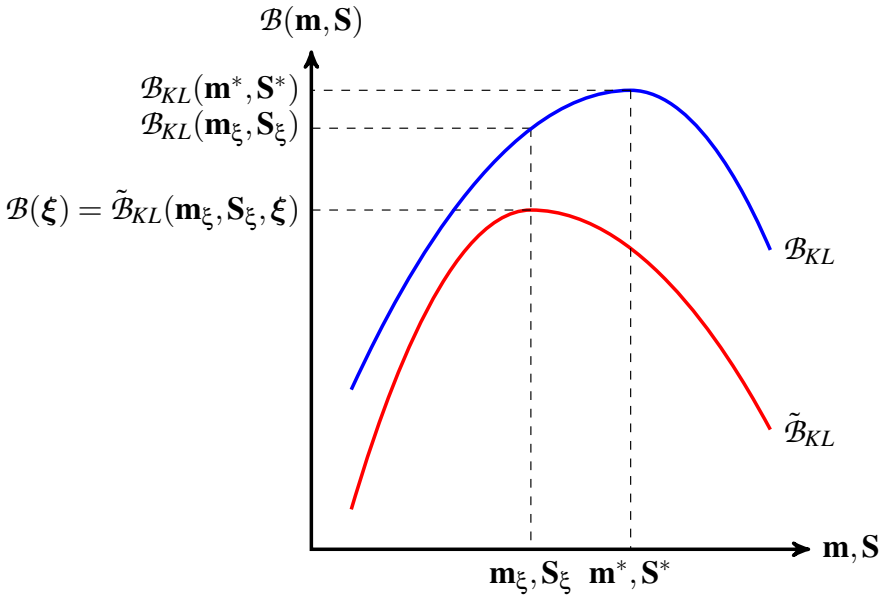


Figure 4: Schematic of the relation between the G-KL bound, $\mathcal{B}_{KL}$ (blue), and the weakened KL bound, $\tilde{\mathcal{B}}_{KL}$ (red), plotted as a function of the Gaussian moments $\mathbf{m}$ and $\mathbf{S}$ with $\boldsymbol{\xi}$ fixed. For any setting of the local site bound parameters $\boldsymbol{\xi}$ we have that $\mathcal{B}_{KL}(\mathbf{m},\mathbf{S}) \geq \tilde{\mathcal{B}}_{KL}(\mathbf{m},\mathbf{S},\boldsymbol{\xi})$. We show in the text that the local bound, $\mathcal{B}(\boldsymbol{\xi})$, is the maximum of the weakened KL bound, that is that $\mathcal{B}(\boldsymbol{\xi}) = \max_{\mathbf{m},\mathbf{S}} \tilde{\mathcal{B}}(\mathbf{m},\mathbf{S},\boldsymbol{\xi})$ with $\mathbf{m}_\xi,\mathbf{S}_\xi = \text{argmax}_{\mathbf{m},\mathbf{S}} \tilde{\mathcal{B}}(\mathbf{m},\mathbf{S},\boldsymbol{\xi})$ in the figure. The G-KL bound can be optimised beyond $\mathcal{B}_{KL}\left(\mathbf{m}_\xi,\mathbf{S}_\xi\right)$ to obtain different, optimal G-KL moments $\mathbf{m}^*$ and $\mathbf{S}^*$ that achieve a tighter lower-bound on $\log Z$.

By defining $\tilde{q}(\mathbf{w}) = \mathcal{N}\left(\mathbf{w}|\mathbf{A}^{-1}\mathbf{b},\mathbf{A}^{-1}\right)$ we obtain

$$\tilde{\mathcal{B}}_{KL} = -\text{KL}(\,q(\mathbf{w})|\,\tilde{q}(\mathbf{w})) - \frac{1}{2}\log\det(2\pi\boldsymbol{\Sigma}) + \log c(\boldsymbol{\xi}) - \frac{1}{2}\boldsymbol{\mu}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \frac{1}{2}\mathbf{b}^\mathsf{T}\mathbf{A}^{-1}\mathbf{b} - \frac{1}{2}\log\det(2\pi\mathbf{A})\,.$$

Since $\mathbf{m},\mathbf{S}$ only appear via $q(\mathbf{w})$ in the KL term, the tightest bound is given when $\mathbf{m},\mathbf{S}$ are set such that $q(\mathbf{w}) = \tilde{q}(\mathbf{w})$. At this setting the KL term in $\tilde{\mathcal{B}}_{KL}$ is zero and $\mathbf{m}$ and $\mathbf{S}$ are given by

$$\mathbf{S}_\xi = \left(\boldsymbol{\Sigma}^{-1} + \mathbf{F}(\boldsymbol{\xi})\right)^{-1}, \qquad \mathbf{m}_\xi = \mathbf{S}_\xi\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \mathbf{f}(\boldsymbol{\xi})\right),$$

and $\tilde{\mathcal{B}}_{KL}\left(\mathbf{m}_\xi,\mathbf{S}_\xi,\boldsymbol{\xi}\right) = \mathcal{B}\left(\boldsymbol{\xi}\right)$. To reiterate, $\mathbf{m}_\xi$ and $\mathbf{S}_\xi$ maximise $\tilde{\mathcal{B}}_{KL}(\mathbf{m},\mathbf{S},\boldsymbol{\xi})$ for any fixed setting of $\boldsymbol{\xi}$. Since $\mathcal{B}_{KL}(\mathbf{m},\mathbf{S}) \geq \tilde{\mathcal{B}}_{KL}(\mathbf{m},\mathbf{S},\boldsymbol{\xi})$ we have that,

$$\mathcal{B}_{KL}(\mathbf{m}_\xi,\mathbf{S}_\xi) \geq \tilde{\mathcal{B}}_{KL}(\mathbf{m}_\xi,\mathbf{S}_\xi,\boldsymbol{\xi}) = \mathcal{B}(\boldsymbol{\xi}).$$

The G-KL bound can be optimised beyond this setting and can achieve an even tighter lower-bound on $\log Z$,

$$\mathcal{B}_{KL}(\mathbf{m}^*, \mathbf{S}^*) = \max_{\mathbf{m},\mathbf{S}} \mathcal{B}_{KL}(\mathbf{m}, \mathbf{S}) \geq \mathcal{B}_{KL}(\mathbf{m}_\xi, \mathbf{S}_\xi).$$

Thus optimal G-KL bounds are provably tighter than both the local variational bound and the G-KL bound calculated using the optimal local bound moments $\mathbf{m}_\xi$ and $\mathbf{S}_\xi$. A graphical depiction of this result is presented in Figure 4.

The experimental results presented in Section 6 show that the improvement in bound values can be significant. Furthermore, constrained parameterisations of covariance, introduced in Section 4, which are required when $D \gg 1$, are also frequently observed to outperform local variational solutions despite the fact that they are not provably guaranteed to do so.

5.2 Complexity and Model Suitability Comparison

We briefly review the core computational bottlenecks and the conditions placed on the potential functions by the local variational bounding, the Laplace approximation and the Gaussian expectation propagation approximate inference methods. A more thorough comparison of these techniques in the context of binary Gaussian Process classification can be found in Nickisch and Rasmussen (2008). Subsequently, we go onto summarise and compare these properties versus the G-KL procedure.

5.2.1 LAPLACE APPROXIMATIONS

Laplace methods, see Barber (2012) for an introduction, approximate the target density with a Gaussian whose mean is centered at the mode of $p(\mathbf{w})$ and whose covariance is the inverse Hessian at the mode of $\log p(\mathbf{w})$. The computational complexity of finding the mode is that of a continuous optimisation problem over D real valued parameters on the joint likelihood objective $\mathcal{N}(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \prod_n \phi_n(\mathbf{w})$. Evaluating the Laplace estimate to $\log Z$ requires computing the determinant of the Hessian, and so scales $O\left(D^3\right)$ which, importantly, only needs to be computed once. To apply the Laplace approximation we require that the target density be twice continuously differentiable, that is we require that each potential function $\{\phi_n\}_{n=1}^N$ be twice continuously differentiable. Provided the Laplace approximation is valid it is generally the fastest of the methods listed here.

5.2.2 LOCAL VARIATIONAL BOUNDING

Local variational bounding methods, as detailed in Section 5.1.1, have N free variational parameters—one for each site potential ϕ_n. Optimising the bound, using either generalised expectation maximisation or gradient based methods, requires solving N linear symmetric $D \times D$ systems. Efficient exact implementations of this method maintain the covariance using its Cholesky factorisation and perform efficient rank one Cholesky updates (Seeger, 2007). Doing so each round of updates scales $O\left(ND^2\right)$. As detailed in Section 5.1.1, local variational bounding procedures are applicable provided tight exponentiated quadratic lower-bounds to the site projection potentials $\{\phi_n\}_{n=1}^N$ exist—that is each site potential is required to be super-Gaussian (Palmer et al., 2006).

Recently scalable approximate solvers for local variational bounding procedures have been developed—see Seeger and Nickisch (2011b) for a review. These methods make use of a number of algorithmic relaxations to reduce the computational burden of local bound optimisation. First, double loop algorithms are employed that reduce the number of times that $\log\det(\mathbf{A})$, see

Section 5.1.1, and its derivative needs to be computed. Second, these algorithms use approximate methods to evaluate the marginal variances that are required to drive local variational bound optimisation. Marginal variances are approximated either by constructing low rank factorisations of $\mathbf{A}$ using iterative Lanczos methods or by perturb and MAP sampling methods (Papandreou and Yuille, 2010; Seeger, 2010; Ko and Seeger, 2012). Both of these approximations can greatly increase the speed of inference and the size of problems to which local procedures can be applied. Unfortunately, these relaxations are not without consequence regarding the quality of approximate inference. For example, the $\log\det(\mathbf{A})$ term is no longer exactly computed and a lower-bound on $\log Z$ is no longer maintained—only an estimate of $\log Z$ is provided. Lanczos approximated marginal variances are often found to be strongly underestimated and bound values strongly overestimated. Whilst the scaling properties are in general problem and user dependent, roughly speaking, these relaxations reduce the computational complexity to scaling $O\left(KD^2\right)$ where K is the rank of the approximate covariance factorisation.

5.2.3 GAUSSIAN EXPECTATION PROPAGATION

Gaussian expectation propagation methods seek to approximate the target density by sequentially matching moments between marginals of the variational Gaussian distribution and a density constructed from the Gaussian approximation and individual site potentials (Minka, 2001). Gaussian expectation propagation (G-EP), for problems of the form of Equation (2), is parameterised using $2N$ free variational parameters, updating each of which requires N rank one $D \times D$ Cholesky updates and the solution of N symmetric D-dimensional linear systems—thus scaling $O\left(ND^2\right)$ assuming $N > D$. Importantly, G-EP optimises neither a convex nor concave objective and is not guaranteed to converge. Whilst G-EP does not require the site projection potentials to be either smooth or super-Gaussian, convergence issues can occur if they are multimodal or not log-concave.

Provably convergent double loop extensions to G-EP have been developed—see Opper and Winther (2005) and references therein for details. Typically these methods are slower than vanilla G-EP implementations. However, recent algorithmic developments have yielded significant speed ups over vanilla G-EP whilst maintaining the convergence guarantees (Seeger and Nickisch, 2011a). Importantly, however, these procedures require the exact solution of rank D symmetric linear systems and thus scale $O\left(D^3\right)$.

5.2.4 G-KL

G-KL approximate inference methods require that each site projection potential has unbounded support on $\mathbb{R}$. Unlike Laplace procedures G-KL is applicable for models with non-differentiable site potentials. Unlike local variational bounding procedures G-KL does not require the site potentials to be super-Gaussian. In contrast to G-EP, which is known to suffer from convergence issues for non log-concave sites, G-KL procedures optimise a strict lower-bound and convergence is guaranteed for gradient ascent optimisation.

When $\{\phi_n\}_{n=1}^N$ are log-concave G-KL bound optimisation is a concave problem and we are guaranteed to converge to the global optimum of the G-KL bound. Local bounding methods have also been shown to be concave problems in this setting (Nickisch and Seeger, 2009). However, as we have shown in Section 5.1, the optimal G-KL bound to $\log Z$ is provably tighter than the local variational bound.

Exact implementations of G-KL approximate inference require storing and optimising over $\frac{1}{2}D(D+3)$ parameters to specify the Gaussian mean and covariance. Often the number of G-KL parameters is greater than that for Laplace, G-EP or local variational bounding methods. However, the computations required by G-KL methods scale similarly to these other Gaussian approximation methods. Empirically, as we show in Section 6, G-KL approximate inference is seen to have comparable convergence speeds to local bounding methods and G-EP.

Importantly, G-KL procedures can be made scalable by using constrained parameterisations of covariance that do not require making a priori factorisation assumptions for the approximate posterior density. Scalable covariance decompositions for G-KL inference maintain a strict lower-bound on $\log Z$ whereas approximate local bound optimisers do not. G-EP, being a fixed point procedure, has been shown to be unstable when using low-rank covariance approximations and appears constrained to scale $O\left(ND^2\right)$ (Seeger and Nickisch, 2011a).

6. Applications

In this section we present results obtained from applying Gaussian KL approximate inference methods to three popular machine learning models. In Section 6.1 we compare deterministic Gaussian approximate inference methods in robust Gaussian process regression models. In Section 6.2 we compare the performance of the constrained parameterisations of G-KL covariance that we presented in Section 4.1.3 in large scale Bayesian logistic regression models. In Section 6.3 we compare Gaussian approximate inference methods to drive sequential experimental design procedures in Bayesian sparse linear models.

6.1 Robust Gaussian Process Regression

Gaussian Processes (GP) are a popular non-parametric approach to supervised learning problems, see Rasmussen and Williams (2006) for a thorough introduction, for which inference falls into the general form of Equation (2). Excluding limited special cases, computing Z and evaluating the posterior density, necessary to make predictions and set hyperparameters, is analytically intractable.

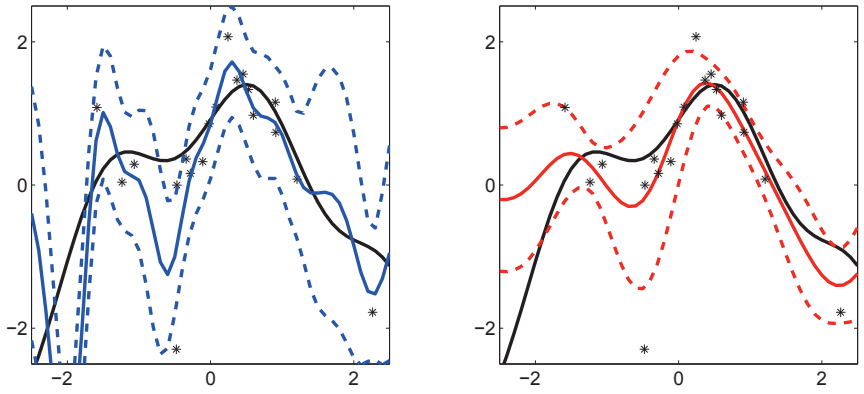
The supervised learning model for fully observed covariates $\mathbf{X} \in \mathbb{R}^{N\times D}$ and corresponding dependent variables $\mathbf{y} \in \mathbb{R}^N$ is specified by the GP prior on the latent function values $\mathbf{w} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and the likelihood $p(\mathbf{y}|\mathbf{w})$. The GP prior moments are constructed by the GP covariance and mean functions which take the covariates $\mathbf{X}$ and a vector of hyperparameters $\boldsymbol{\theta}$ as arguments. The posterior on the latent function values, $\mathbf{w}$, is given by

$$p(\mathbf{w}|\mathbf{y},\mathbf{X},\boldsymbol{\theta}) = \frac{1}{Z} p(\mathbf{y}|\mathbf{w}) \mathcal{N}(\mathbf{w}|\boldsymbol{\mu},\boldsymbol{\Sigma}).$$

The likelihood factorises over data instances, $p(\mathbf{y}|\mathbf{w}) = \prod_{n=1}^{N} \phi(w_n)$, thus the GP posterior is of the form of Equation (1) with site projection potentials of the form of Equation (5).

6.1.1 GP REGRESSION

For GP regression models the likelihood is most commonly Gaussian distributed, equivalent to assuming zero mean additive Gaussian noise. This assumption leads to analytically tractable, indeed Gaussian, forms for the posterior. However, Gaussian additive noise is a strong assumption to make, and is often not corroborated by real world data. Gaussian distributions have thin tales—the density



(a) Gaussian likelihood (b) Student's t likelihood

Figure 5: Gaussian process regression with a squared exponential covariance function and (a) a Gaussian or (b) a Student's t likelihood. Covariance hyperparameters are optimised for a training data set with outliers. Latent function posterior mean (solid) and ± 1 standard deviation (dashed) values are plotted in blue (a) and red (b). The data generating function is plotted in black. The Student's t model makes more conservative interpolated predictions whilst the Gaussian model appears to over-fit the data.

function rapidly tends to zero for values far from the mean—see Figure 6. Outliers in the training set then do not have to be too extreme to negatively affect test set predictive accuracy. This effect can be especially severe for GP models that have the flexibility to incorporate training set outliers to areas of high likelihood—essentially over-fitting the data.

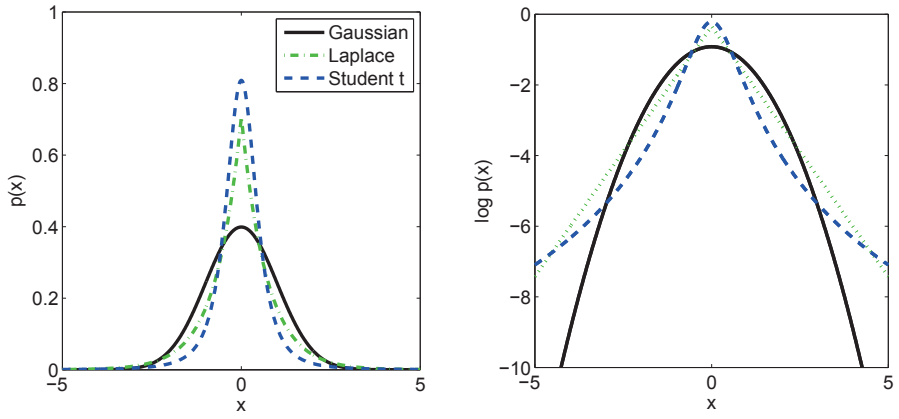
An example of GP regression applied to a data set with outliers is presented in figure 5(a). In this figure a GP prior with squared exponential covariance function coupled with a Gaussian likelihood over-fits the training data and the resulting predicted values differ significantly from the underlying data generating function.

One approach to prevent over-fitting is to use a likelihood that is robust to outliers. Heavy tailed likelihood densities are robust to outliers in that they do not penalise too heavily observations far from the latent function mean. Two distributions are often used in this context: the Laplace otherwise termed the double exponential, and the Student's t. The Laplace probability density function can be expressed as

$$p(y|\mu,\tau) = \frac{1}{2\tau} e^{-|y-\mu|/\tau},$$

where τ controls the variance of the random variable y with $\mathrm{var}(y) = 2\tau^2$. The Student's t probability density function can be written as

$$p(y|\mu,\nu,\sigma^2) = \frac{\Gamma\left(\frac{1}{2}(\nu+1)\right)}{\Gamma\left(\frac{1}{2}\nu\right)\sqrt{\pi\nu\sigma^2}} \left(1 + \frac{(y-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}$$



(a) (b)

Figure 6: Gaussian, Laplace and Student's t densities with unit variance: (a) probability density functions and (b) log probability density functions. Laplace and Student's t densities have stronger peaks and heavier tails than the Gaussian. Student's t with d.o.f. $\nu = 2.5$ and scale $\sigma^2 = 0.2$, Laplace with $\tau = 1/\sqrt{2}$.

where $\nu \in \mathbb{R}^+$ is the degrees of freedom parameter, $\sigma \in \mathbb{R}^+$ the scale parameter, and $\text{var}(y) = \sigma^2\nu/(\nu-2)$ for $\nu > 2$. As the degrees of freedom parameter becomes increasingly large the Student's t distribution converges to the Gaussian distribution. See Figure 6 for a comparison of the Student's t, Laplace and Gaussian density functions.

GP models with outlier robust likelihoods such as the Laplace or the Student's t can yield significant improvements in test set accuracy versus Gaussian likelihood models (Vanhatalo et al., 2009; Jylanki et al., 2011; Opper and Archambeau, 2009). In figure 5(b) we model the same training data as in figure 5(a) but with a heavy tailed Student's t likelihood, the resulting predictive values are more conservative and lie closer to the true data generating function than for the Gaussian likelihood model.

6.1.2 APPROXIMATE INFERENCE

Whilst Laplace and Student's t likelihoods can successfully 'robustify' GP regression models to outliers they also render inference analytically intractable and approximate methods are required. In this section we compare G-KL approximate inference to other deterministic Gaussian approximate inference methods, namely: the Laplace approximation (Lap), local variational bounding (VB) and Gaussian expectation propagation (G-EP).

Each approximate inference method cannot be applied to each likelihood model. Since the Laplace likelihood is not differentiable everywhere Laplace approximate inference is not applicable. Since the Student's t likelihood is not log-concave, indeed the posterior can be multi-modal, vanilla G-EP implementations are numerically unstable (Seeger et al., 2007). Recent work (Jylanki et al.,

		Gauss	Student's t			Laplace		
		Exact	G-KL	VB	Lap	G-KL	VB	G-EP
C. ST	LML	-15 ± 2	-75 ± 2	-240 ± 21	-7 ± 1	8 ± 5	2 ± 2	$--\pm--$
	MSE	1.15 ± 0.2	1.6 ± 0.2	23.8 ± 4	2.2 ± 0.4	1.3 ± 1.1	1.2 ± 1.0	$--\pm--$
	TLP	0.79 ± 0.10	0.73 ± 0.05	-0.65 ± 0.06	0.41 ± 0.03	0.97 ± 0.06	0.91 ± 0.05	$--\pm--$
Friedman	LML	70 ± 6	-159 ± 7	-578 ± 34	-97 ± 4	-69 ± 6	-73 ± 8	$--\pm--$
	MSE	10 ± 3	5 ± 1	17 ± 2	13 ± 1	5 ± 1	3 ± 1	$--\pm--$
	TLP	-0.26 ± 0.09	0.12 ± 0.09	-0.54 ± 0.06	-0.65 ± 0.06	0.07 ± 0.09	0.25 ± 0.11	$--\pm--$
Neal	LML	39 ± 10	-171 ± 14	-962 ± 1	-21 ± 15	-26 ± 9	-27 ± 8	-14 ± 7
	MSE	1.7 ± 0.6	2.9 ± 1.1	4.4 ± 1.3	0.9 ± 0.5	0.9 ± 0.4	0.9 ± 0.4	0.9 ± 0.5
	TLP	0.22 ± 0.12	0.88 ± 0.03	0.36 ± 0.02	0.67 ± 0.08	0.86 ± 0.04	1.13 ± 0.02	0.91 ± 0.04
Boston	LML	51 ± 3	-133 ± 13	-551 ± 37	-53 ± 3	-60 ± 3	-61 ± 3	-53 ± 4
	MSE	26 ± 1	25 ± 2	26 ± 1	23 ± 2	25 ± 2	26 ± 1	22 ± 1
	TLP	-0.74 ± 0.07	-0.44 ± 0.03	-0.58 ± 0.03	-0.44 ± 0.03	-0.52 ± 0.06	-0.51 ± 0.02	-0.46 ± 0.03

Table 1: Gaussian process regression results for different (approximate) inference procedures, likelihood models and data sets. First column section: Gaussian likelihood results with exact inference. Second column section: Student's t likelihood results with G-KL, local variational bounding (VB) and Laplace (Lap) approximate inference. Third column section: Laplace likelihood results with G-KL, VB and Gaussian expectation propagation (G-EP) approximate inference. Each row presents the (approximate or lower-bound) log marginal likelihood (LML), test set mean squared error (MSE), or approximate test set log probability (TLP) values obtained by data set. Table values are the mean and standard error of the values obtained over the 10 random partitions of the data.

2011) has alleviated some of G-EP's convergence issues for Student's t GP regression, however, these extensions are beyond the scope of this work.

Local variational bounding and G-KL procedures are applied to both likelihood models. For local variational bounding, both the Laplace and Student's t densities are super-Gaussian and thus tight exponentiated quadratic lower-bounds exist—see Seeger and Nickisch (2010) for the precise forms that are employed in these experiments. Laplace, local variational bounding and G-EP results are obtained using the `GPML` toolbox (Rasmussen and Nickisch, 2010).[4] G-KL approximate inference is straightforward, for the G-KL approximate posterior $q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m},\mathbf{S})$ the likelihood's contribution to the bound is

$$\langle \log p(\mathbf{y}|\mathbf{w}) \rangle_{q(\mathbf{w})} = \sum_n \left\langle \log \phi_n(m_n + z\sqrt{S_{nn}}) \right\rangle_{\mathcal{N}(z|0,1)}.$$

The equation above is equivalent to Equation (6) with $\mathbf{h}_n = \mathbf{e}_n$ the unit norm basis vector and ϕ_n the likelihood of the n^{th} data point. The expectations for the Laplace likelihood site potentials have simple analytic forms—see Appendix B.2.1. The expectations for the Student's t site potentials are evaluated numerically. All other terms in the G-KL bound have simple analytic forms and computations that scale $\leq O\left(D^3\right)$. G-KL results are obtained, as for all other results in this paper, using the `vgai` Matlab package—see Section 8. For the Laplace likelihood model, which is log-concave, Hessian free Newton methods were used to optimise the G-KL bound. For the Student's t likelihood, which is not log-concave, LBFGS was used to optimise the G-KL bound.

4. The `GPML` toolbox can be downloaded from `www.gaussianprocess.org`.

6.1.3 EXPERIMENTAL SETUP

We consider GP regression with training data $\mathcal{D} = \{(y_n, \mathbf{x}_n)\}_{n=1}^N$ for covariates $\mathbf{x}_n \in \mathbb{R}^D$ and dependent variables $y_n \in \mathbb{R}$. We assume a zero mean Gaussian process prior on the latent function values, $\mathbf{w} = [w_1, ..., w_N]^\mathsf{T} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. The covariance, $\boldsymbol{\Sigma}$, is constructed as the sum of the squared exponential kernel and the independent white noise kernel,

$$\Sigma_{mn} = k(\mathbf{x}_m, \mathbf{x}_n, \boldsymbol{\theta}) = \sigma_{se}^2 e^{-\sum_d (x_{nd} - x_{md})^2 / l_d^2} + \gamma^2 \delta(n, m),$$

where x_{nd} refers to the d^{th} element of the n^{th} covariate, σ_{se}^2 is the 'signal variance' hyperparameter, l_d the squared exponential 'length scale' hyperparameter, and γ the independent white noise hyperparameter (above $\delta(x, y)$ is the Kronecker delta such that $\delta(n, m) = 1$ if $n = m$ and 0 otherwise). Covariance hyperparameters are collected in the vector $\boldsymbol{\theta}$.

We follow the evidence maximisation or maximum likelihood two (ML-II) procedure to estimate the covariance hyperparameters, that is we set covariance hyperparameters to maximise $p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})$. Since $p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})$ cannot be evaluated exactly we use the approximated values offered by each of the approximate inference methods. Covariance hyperparameters are optimised numerically using nonlinear conjugate gradients. The marginal likelihood, $p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})$, is not unimodal and we are liable to converge to a local optimum regardless of which inference method is used. All methods were initialised with the same hyperparameter setting. Hyperparameter derivatives for the G-KL bound are presented in Appendix F.2.

Likelihood hyperparameters were selected to maximise the log predicted probability scores on a held out validation data set. Simultaneous likelihood and covariance ML-II hyperparameter optimisation for the Student's t and Laplace likelihoods yielded poor test set performance regardless of the approximate inference method used (as has been previously reported for Student's t likelihoods in other experiments (Vanhatalo et al., 2009; Jylanki et al., 2011). For the Student's t likelihood model the d.o.f. parameter was fixed with $\nu = 3$.

Results were obtained for the four approximate inference procedures on the four data sets using both the Laplace and the Student's t likelihoods. Two UCI data sets were used:[5] Boston housing and Concrete Slump Test. And two synthetic data sets: Friedman[6] and Neal.[7] Each experiment was repeated over 10 randomly assigned training, validation and test set partitions. The size of each data set is as follows: Concrete Slump Test $D = 9$, $N_{trn} = 50$, $N_{val} = 25$, $N_{tst} = 28$; Boston $D = 13$, $N_{trn} = 100$, $N_{val} = 100$, $N_{tst} = 306$; Friedman $D = 10$, $N_{trn} = 100$, $N_{val} = 100$, $N_{tst} = 100$; Neal $D = 1$, $N_{trn} = 100$, $N_{val} = 100$, $N_{tst} = 100$. Each partition of the data was normalised using the mean and standard deviation statistics of the training data.

To asses the validity of the Student's t and Laplace likelihoods we also implemented GP regression with a Gaussian likelihood and exact inference.

6.1.4 RESULTS

Results are presented in Table 1. Approximate log marginal likelihood (LML), test set mean squared error (MSE) and approximate test set log probability (TLP) mean and standard error values obtained over the 10 partitions of the data are provided. It is important to stress that the TLP values are approximate values for all methods, obtained by summing the approximate log probability of each

5. UCI data sets can be downloaded from `archive.ics.uci.edu/ml/datasets/`.
6. The Friedman data set is constructed as described in Kuss (2006) §5.6.1. and Friedman (1991).
7. The Neal data set is constructed as described in Neal (1997) §7.

test point using the surrogate score presented in Appendix F.1. For G-KL and VB procedures the TLP values are not lower-bounds.

The results confirm the utility of heavy tailed likelihoods for GP regression models. Test set predictive accuracy scores are higher with robust likelihoods and approximate inference methods than with a Gaussian likelihood and exact inference. This is displayed in the lower MSE error and higher TLP scores of the best performing robust likelihood results than for the Gaussian likelihood. Exact inference for the Gaussian likelihood model achieves the greatest LML in all problems except the Concrete Slump Test data. That exact inference with a Gaussian likelihood achieves the strongest LML and weak test set scores implies the ML-II procedure is over-fitting the training data with this likelihood model.

For the Student's t likelihood the performance of each approximate inference method varied significantly. VB results were uniformly the weakest. We conjecture this is an artifact of the squared exponential local site bounds employed by the `gpml` toolbox poorly capturing the non log-concave potential functions mass. For Student's t potentials improved VB performance has been reported by employing bounds that are composed of two terms on disjoint partitions of the domain (Seeger and Nickisch, 2011b), validating their efficacy in the context of Student's t GP regression models is reserved for future work. For the test set metrics G-KL approximate inference achieves the strongest performance.

Broadly, the Laplace likelihood achieved the best results on all data sets. G-EP frequently did not converge for both the Friedman and Concrete Slump Test problems and so results are not presented. Unlike the Student's t likelihood model, results are more consistent across approximate inference methods. G-KL achieves a narrow but consistently superior LML value to VB. Approximate test set predictive values are roughly the same for all inference methods with VB achieving a small advantage.

We reiterate that standard G-EP approximate inference, as implemented in the `GPML` toolbox, was used to obtain these results. The authors did not anticipate convergence issues for G-EP in the GP models considered—the Laplace likelihood model's log posterior is concave and the system has full rank. Power G-EP, as proposed in Minka (2004), has previously been shown to have robust convergence for under determined linear models with Laplace potentials (Seeger, 2008). Similarly, we expect that power G-EP would also exhibit robust convergence in GP models with Laplace likelihoods. Verifying this experimentally and assessing the performance of power G-EP approximate inference in noise robust GP regression models is left for future work.

The G-KL LML uniformly dominates the VB values. This is theoretically guaranteed for a model with fixed hyperparameters and log-concave site potentials, see Section 5.1.2 and Section 3.2. However, the G-KL bound is seen to dominate the local bound even when these conditions are not satisfied. The results show that both G-KL bound optimisation and G-KL hyperparameter optimisation is numerically stable. G-KL approximate inference appears more robust than G-EP and VB—G-KL hyperparameter optimisation always converged, often to a better local optima.

6.1.5 SUMMARY

The results confirm that the G-KL procedure as a sensible route for approximate inference in GP models with non-conjugate likelihoods. The G-KL procedure is generally applicable in this setting and easy to implement for new likelihood models. Indeed, all that is required to implement G-KL approximate inference for a GP regression model is the pointwise evaluation of the univariate

likelihood function $p(y_n|w_n)$. Furthermore, we have seen that G-KL optimisation is numerically robust, in all the experiments G-KL converged and achieved strong performance.

6.2 Bayesian Logistic Regression

In this section we examine the relative performance, in terms of speed and accuracy of inference, of each of the constrained G-KL covariance decompositions presented in Section 4.1.3. As a bench mark, we also compare G-KL approximate inference results to scalable approximate VB methods with marginal variances approximated using iterative Lanczos methods (Seeger and Nickisch, 2011b). Our aim is not make a comparison of deterministic approximate inference methods for Bayesian logistic regression models, see Nickisch and Rasmussen (2008) to that end, but to investigate the time accuracy trade-offs of each of the constrained G-KL covariance parameterisations.

Given a data set, $\mathcal{D} = \{(y_n, \mathbf{x}_n), n = 1, \ldots, N\}$ with class labels $y_n \in \{-1, 1\}$ and covariates $\mathbf{x}_n \in \mathbb{R}^D$, Bayesian logistic regression models the class conditional distribution using $p(y = 1|\mathbf{w}, \mathbf{x}) = \sigma\left(\mathbf{w}^\mathsf{T}\mathbf{x}\right)$, with $\sigma(x) := 1/(1+e^{-x})$ the logistic sigmoid function and $\mathbf{w} \in \mathbb{R}^D$ a vector of parameters. Under a Gaussian prior, $\mathcal{N}(\mathbf{w}|\mathbf{0}, \boldsymbol{\Sigma})$, the posterior is given by

$$p(\mathbf{w}|\mathcal{D}) = \frac{1}{Z}\mathcal{N}(\mathbf{w}|\mathbf{0}, \boldsymbol{\Sigma}) \prod_{n=1}^{N} \sigma\left(y_n \mathbf{w}^\mathsf{T}\mathbf{x}_n\right). \tag{15}$$

Where we have used the symmetry property of the logistic sigmoid such that $p(y = -1|\mathbf{w}, \mathbf{x}) = 1 - p(y = 1|\mathbf{w}, \mathbf{x}) = \sigma\left(-\mathbf{w}^\mathsf{T}\mathbf{x}\right)$. The expression above is of the form of Equation (2) with log-concave site projection potentials $\phi_n(x) = \sigma(x)$ and $\mathbf{h}_n = y_n\mathbf{x}_n$.

6.2.1 EXPERIMENTAL SETUP

We synthetically generate the data sets. The data generating parameter vector $\mathbf{w}^{tr} \in \mathbb{R}^D$ is sampled from a factorising standard normal $\mathbf{w}^{tr} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The covariates, $\{\mathbf{x}_n\}_{n=1}^N$, are generated by first sampling an independent standard normal, then linearly transforming these vectors to impose correlation between some of the dimensions, and finally the data is renormalised so that each dimension has unit variance. The linear transformation matrix we use to impose correlation between covariates is a sparse square matrix generated as the sum of the identity matrix and a sparse matrix with one element from each row sampled from a standard normal. Class labels $y_n \in \{1, -1\}$ are sampled from the likelihood $p(y_n = 1|\mathbf{w}, \mathbf{x}_n) = \sigma(\mathbf{w}^\mathsf{T}\mathbf{x}_n)$. The inferential model's prior and likelihood distributions are set to match the data generating process.

Results are obtained for a range of data set dimensions: $D = 250, 500, 1000$ and $N = \frac{1}{2}D, D, 5D$. We also vary the size of the constrained covariance parameterisations, which is reported as K in the result tables. For chevron Cholesky K refers to the number of non-diagonal rows of $\mathbf{C}$. For subspace Cholesky K is the dimensionality of the subspace. For banded Cholesky K refers to the band width of the parameterisation. For the factor analysis (FA) parameterisation K refers to the number of factor loading vectors. For local variational bounding (VB) approximate inference K refers to the number of Lanczos vectors used to update the variational parameters. The parameter K is varied as a function of the parameter vector dimensionality with $K = 0.05 \times D$ and $K = 0.1 \times D$.

Since the G-KL bound is strongly concave we performed G-KL bound optimisation using Hessian free Newton methods for all the Cholesky parameterised covariance experiments. G-KL bound optimisation was terminated when the largest absolute value of the gradient vector was less than 10^{-3}. For subspace Cholesky we iterated between optimising the subspace parameters $\{\mathbf{m}, \mathbf{C}, c\}$

			$N_{trn}=250$		$N_{trn}=500$		$N_{trn}=2500$	
			$K=25$	$K=50$	$K=25$	$K=50$	$K=25$	$K=50$
Time (s)	G-KL	Chev	0.49±0.02	0.69±0.08	1.25±0.04	1.36±0.04	16.50±0.89	17.31±0.82
		Band	0.96±0.02	1.37±0.02	2.25±0.10	4.06±0.29	24.31±0.96	29.60±1.18
		Sub	0.73±0.01	0.93±0.03	1.41±0.03	1.93±0.04	11.89±0.54	15.26±1.02
		FA	2.05±0.26	2.29±0.21	2.92±0.17	3.47±0.17	20.06±1.51	22.69±2.70
	VB		0.37±0.00	0.47±0.01	0.46±0.02	0.52±0.00	1.56±0.03	1.85±0.01
$\tilde{\mathcal{B}}$	G-KL	Chev	−1.19±0.01	−1.15±0.01	−0.93±0.01	−0.91±0.01	−0.42±0.00	−0.41±0.00
		Band	−1.15±0.01	−1.09±0.01	−0.92±0.01	−0.88±0.01	−0.42±0.00	−0.41±0.00
		Sub	−3.08±0.02	−2.20±0.01	−1.90±0.01	−1.46±0.01	−0.62±0.00	−0.54±0.00
		FA	−1.19±0.01	−1.17±0.01	−0.93±0.01	−0.91±0.01	−0.41±0.00	−0.40±0.00
	VB		−±−	−±−	−±−	−±−	−±−	−±−
$\|\mathbf{w}-\mathbf{w}_{tr}\|_2/D$	G-KL	Chev	0.88±0.00	0.87±0.00	0.84±0.00	0.84±0.00	0.64±0.00	0.64±0.00
		Band	0.87±0.00	0.87±0.00	0.84±0.00	0.84±0.00	0.64±0.00	0.64±0.00
		Sub	0.88±0.00	0.87±0.01	0.87±0.00	0.86±0.00	0.71±0.00	0.70±0.00
		FA	0.88±0.00	0.87±0.01	0.84±0.00	0.84±0.00	0.64±0.00	0.64±0.00
	VB		0.90±0.00	0.89±0.00	0.89±0.00	0.88±0.00	0.72±0.00	0.72±0.00
$\log p(\mathbf{y}^*\|\mathbf{X}^*)/N_{tst}$	G-KL	Chev	−0.58±0.01	−0.58±0.01	−0.50±0.01	−0.49±0.01	−0.18±0.00	−0.18±0.00
		Band	−0.58±0.01	−0.57±0.01	−0.50±0.01	−0.49±0.01	−0.18±0.00	−0.18±0.00
		Sub	−0.72±0.02	−0.65±0.02	−0.63±0.01	−0.59±0.01	−0.20±0.00	−0.20±0.00
		FA	−0.58±0.01	−0.58±0.01	−0.51±0.01	−0.50±0.01	−0.18±0.00	−0.18±0.00
	VB		−0.75±0.02	−0.77±0.02	−0.63±0.01	−0.64±0.01	−0.20±0.00	−0.20±0.00

Table 2: Synthetic Bayesian logistic regression results for a model with unit variance Gaussian prior $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ with $\dim(\mathbf{w}) = 500$, likelihood $p(\mathbf{y}|\mathbf{w}, \mathbf{X}) = \prod_{n=1}^{N_{trn}} \sigma(y_n \mathbf{w}^\mathsf{T} \mathbf{x}_n)$, class labels $y_n \in \{+1, -1\}$ and $N_{tst} = 5000$ test points. G-KL results obtained using chevron Cholesky (Chev), banded Cholesky (Band), subspace Cholesky (Sub) and factor analysis (FA) constrained parameterisations of covariance. Approximate local variational bounding (VB) results are obtained using low-rank factorisations of covariance computed using iterative Lanczos methods. Parameter K denotes the size of the constrained covariance parameterisation.

and updating the subspace basis vectors $\mathbf{E}$ each five times. The subspace vectors were updated using the fixed point iteration with the Lanczos approximation (see Appendix B.4.3 for details). For the FA parameterisation the G-KL bound is not concave so we use LBFGS to perform gradient ascent. All other `minFunc` options were set to default values.

VB approximate inference is achieved using the `glm-ie 1.4` package (Nickisch, 2012). VB inner loop optimisation used nonlinear conjugate gradients with at most 50 iterations. The maximum number of VB outer loop iterations was set to 10. All other VB `glm-ie` optimisation settings were set to default values. All results for these experiments were obtained using Matlab 2011a on a Intel E5450 3Ghz machine with 8 cores and 64GB of RAM.

6.2.2 RESULTS

Results for $D = 500$ are presented in Table 2. For reasons of space, results for $D = 250$ and $D = 1000$ are presented in Table 3 and Table 4 in Appendix G. The tables present average and standard error scores obtained from 10 synthetically generated data sets.

The average convergence time and standard errors of each of the methods is presented in the first row section of the result tables. In the smaller problems considered, the best G-KL times were

achieved by the chevron Cholesky covariance followed by the banded, the subspace and the FA parameterisations in that order.

The recorded banded Cholesky convergence times are seen to scale super-linearly with K. These results are a consequence of the implementation. Whilst chevron and banded parameterisations both scale $O(NDK)$ they access and compute different elements of the data and Cholesky matrices. The chevron gradients can be computed using standard matrix multiplications for which Matlab is highly optimised. The banded parameterisation needs to access matrix elements in a manner not standard to Matlab and so is much slower. This implementational artifact, despite a Matlab mex C implementation, could not be entirely eliminated.

VB and chevron G-KL achieved broadly similar convergence times for the $N \leq D$ and $D \leq 500$ experiments with VB faster in the larger D experiments. VB is significantly faster than G-KL methods for the $N = 5 \times D$ experiments, this is a consequence of the double-loop structure of the VB implementation. Whilst the subspace G-KL method is significantly slower in the smaller problems when $D = 1000$ it is the fastest G-KL method, beating VB in problems where $N \leq D$.

In the result tables, the bound values are normalised by the size of the training set, with $\tilde{\mathcal{B}} = \mathcal{B}/N_{trn}$, to make comparisons across models easier. As the training set size increases the normalised bound value increases, presumably reflecting the fact that the posterior tends to a Gaussian in the limit of large data. Furthermore, the difference in bound values between the parameterisations become smaller as the size of the training set increases.

The G-KL banded covariance parameterisation achieves the strongest bound value with the chevron and factor analysis parameterisations a close second place. The subspace bound values are comparatively poor. This is not unexpected since the subspace parameterisation has a single parameter (denoted c in Section 4.1.3) that specifies the variance in all directions orthogonal to the subspace vectors $\mathbf{E}$. It is known that the density q that minimises $\mathrm{KL}(q|\,p)$ tends to seek out the modes of p and avoid those regions of parameter space where p is close to zero. Therefore the parameter c will tend to the smallest value of the variance of p in the directions orthogonal to the subspace vectors, the resulting G-KL bound value will therefore be greatly underestimated. The approximate VB method does not provide a lower bound when marginal variances are approximated using low-rank methods and therefore values are not reported in the result tables.

Since these results are obtained from data sets sampled from densities with known parameters we can directly asses the accuracy of the posterior parameter estimate against the ground truth. The posterior mean minimises the ℓ^2 loss $\|\mathbf{w}^{tr} - \mathbf{w}\|_2$. Thus, in the third row section of the results table, we report the average error $\|\mathbf{w}^{tr} - \mathbf{m}\|_2$ where $\mathbf{m}$ is the mean of the Gaussian posterior approximation $q(\mathbf{w}) = \mathcal{N}(\mathbf{m}, \mathbf{S})$. To make comparisons easier, the ℓ^2 errors are normalised by the dimensionality of the respective models D. The results show that the the G-KL mean is broadly invariant to the G-KL covariance parameterisations used. VB results are noticeably poorer than the G-KL methods.

Approximate test set log predictive probabilities are presented in the fourth row section of the result tables. This metric is arguably the best suited to measure the global accuracy of the posterior approximations since it is an expectation over the entire support of the approximate posterior (MacKay and Oldfield, 1995). The values reported in the table are approximated using $\log p(\mathbf{y}^*|\mathbf{X}^*) \approx \sum_n \log \left\langle p(y_n^* \mathbf{w}^{\mathsf{T}} \mathbf{x}_n^*) \right\rangle_{q(\mathbf{w})}$. The values presented are normalised by the size of the test set where $N_{tst} = 10 \times D$ in all experiments. The results show that chevron, banded and FA parameterisations achieve the best, and broadly similar, performance. Test set predictive accuracy increases for all methods as a function of the training set size. Subspace G-KL and approximate VB achieve broadly similar and noticeably weaker performance than the other methods.

6.2.3 SUMMARY

The results support the use of the constrained Cholesky covariance parameterisations to drive scalable G-KL approximate inference procedures. Whilst neither the banded nor the chevron Cholesky parameterisations are invariant to permutations of the index set they both achieved the strongest bound values and test set performance. Unfortunately, due to implementational issues, the banded Cholesky parameterisation gradients are slow to compute resulting in slow recorded convergence times. The non-concavity of the factor analysis parameterised covariance resulted in slower recorded convergence times than the concave models. Whilst the subspace G-KL parameterisation had poorer performance in the smaller problems it broadly matched or outperformed the approximate VB method in the largest problems.

6.3 Bayesian Sparse Linear Models

Many problems in machine learning and statistics can be addressed by linear models with sparsity inducing prior distributions. Examples include, feature selection in regression problems (Wipf, 2004), source separation (Girolami, 2001), denoising or deblurring problems (Fergus et al.), and signal reconstruction from a set of under-determined observations (Seeger and Nickisch, 2008). In all of these cases, the prior results in a posteriori parameter estimates that are biased towards sparse solutions. For feature selection problems this assumption can be useful if we believe that only a small subset of the features are necessary to model the data. Using an informative prior is essential in the case of under-determined linear models where there are more sources than signals, in which case hyper-planes in parameter space have equiprobable likelihoods and priors are needed to constrain the space of possible solutions.

Figure 7 depicts the posteriors resulting from an under-determined linear model for a selection of different priors. Since the Laplace prior is log-concave the posterior is unimodal and log-concave. For non log-concave priors the resulting posterior can be multimodal—for instance when $p(\mathbf{w})$ is the Student's t distribution or the sparsity promoting distribution composed from a mixture of Gaussians.

In the case of signal reconstruction, deblurring and source separation sparse priors are used to encode some of the prior knowledge we have about the source signal we wish to recover. Natural images for instance are known to have sparse statistics over a range of linear filters (an example filter being the difference in intensities of neighbouring pixels) (Olshausen and Field, 1996). Sparse priors that encode this knowledge about the statistics of natural images then bias estimates towards settings that share this statistical similarity.

In this section we consider Bayesian sequential experimental design (SED) for the sparse linear model. At each stage of the SED process we approximate the posterior density of the model parameters and then use the approximate posterior to greedily select new, maximally informative measurements. The probabilistic model and experimental design procedure are described in Section 6.3.1 and Section 6.3.2. In Section 6.3.3 we compare approximate inference methods on a small scale artificial SED problem. In Section 6.3.4 we compare G-KL and approximate local variational bounding methods for SED on a $64 \times 64 = 4,096$ pixel natural image problem. Our approach follows that laid out in Seeger and Nickisch (2008), Seeger (2009) and Seeger and Nickisch (2011b).

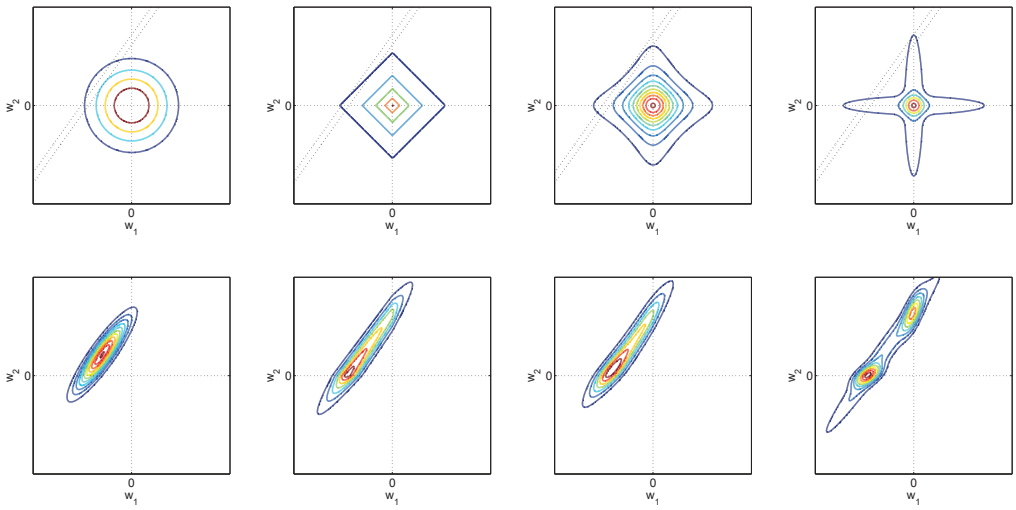


Figure 7: Isocontours of a linear model's prior, likelihood and resulting posterior densities. The top row plots contours of the two dimensional prior (solid) and the Gaussian likelihood (dashed). The second row displays the contours of the posterior induced by the prior and likelihood above it. Column one - a Gaussian prior, column two - a Laplace prior, column three - a Student's t prior and column four a spike and slab prior constructed as a product over dimensions of a two component Gaussian mixture.

6.3.1 PROBABILISTIC MODEL

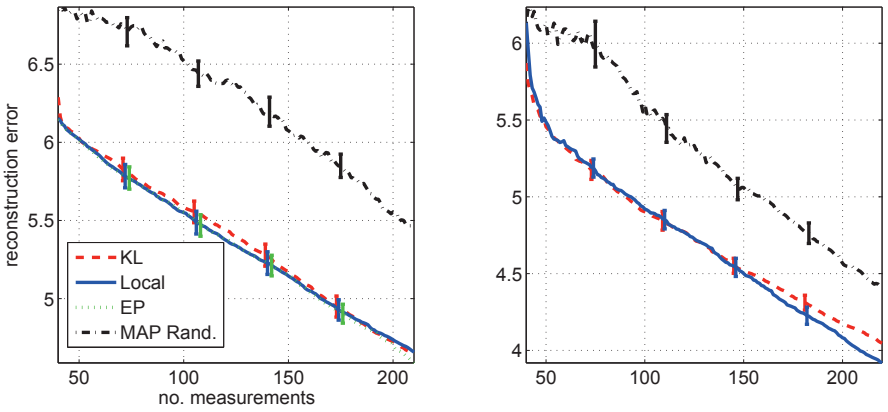
We observe noisy linear measurements $\mathbf{y} \in \mathbb{R}^N$ assumed to be drawn according to $\mathbf{y} = \mathbf{M}\mathbf{w} + \boldsymbol{\nu}$ where $\mathbf{M} \in \mathbb{R}^{N \times D}$ is the linear measurement matrix with $N \ll D$, $\boldsymbol{\nu} \sim \mathcal{N}\left(\mathbf{0}, \nu^2 \mathbf{I}\right)$ is additive Gaussian noise, and $\mathbf{w} \in \mathbb{R}^D$ is the signal that we wish to recover. A sparse prior, here we use either the Laplace or the Student's t, is placed on $\mathbf{s}$ the linear statistics of $\mathbf{w}$ such that $\mathbf{s} = \mathbf{B}\mathbf{w}$. The matrix $\mathbf{B} \in \mathbb{R}^{M \times D}$ is a collection of M linear filters. By placing the prior directly on the statistics, $\mathbf{s}$, the posterior is proportional to the product of the Gaussian likelihood and the sparse prior potentials,

$$p(\mathbf{w}|\mathbf{M}, \mathbf{y}, \boldsymbol{\tau}, \nu^2) \propto \mathcal{N}\left(\mathbf{y}|\mathbf{M}\mathbf{w}, \nu^2 \mathbf{I}\right) p(\mathbf{s}), \quad \text{where} \quad \mathbf{s} = \mathbf{B}\mathbf{w}.$$

Since the priors are placed directly on the statistics $\mathbf{s}$ and not $\mathbf{w}$ they are not normalised densities with respect to $\mathbf{w}$, as a consequence $\mathcal{B}_{KL}(\mathbf{m}, \mathbf{S})$ is no longer a lower-bound to $\log Z$. However, since the normalisation constant of $p(\mathbf{s})$ is constant with respect to $\mathbf{w}$ ignoring this constant does not affect the G-KL approximation to the posterior density.

6.3.2 SEQUENTIAL EXPERIMENTAL DESIGN

SED for the sparse linear model described above is the problem of iteratively choosing which new measurement vectors, $\mathbf{M}^*$, to append to $\mathbf{M}$ so as to maximise subsequent estimation accuracy. Bayesian SED iterates between estimating the posterior density on $\mathbf{w}$, conditioned on current observations, and then using this density to select which new measurements to make. Following



(a) Laplace (b) Student's t

Figure 8: Sequential experimental design for the Bayesian sparse linear model with synthetic signals. Sparse signals, $\mathbf{w}$, are sampled from (a) a Laplace with $\tau = 0.2$ and (b) a Student's t with $\nu = 3$, $\sigma^2 = 0.0267$.

Seeger and Nickisch (2011b) we use the information gain metric to decide which measurement vectors will be maximally informative. Information gain is defined as the difference in differential Shannon information of the posterior density before and after the inclusion of new measurements and their corresponding observations. For the linear model we consider, it is given by

$$I_{gain}(\mathbf{M}^*) = H\left[p(\mathbf{w}|\mathbf{M},\mathbf{y})\right] - H\left[p(\mathbf{w}|\mathbf{M},\mathbf{y},\mathbf{M}^*,\mathbf{y}^*)\right], \tag{16}$$

where $H\left[p(x)\right] := -\left\langle \log p(x) \right\rangle_{p(x)}$ is the Shannon differential entropy.

Since inference is not analytically tractable we cannot access either of the densities required by Equation (16). We can, however, obtain an approximation to the information gain by substituting in a Gaussian approximation to the posterior. Doing so with $p(\mathbf{w}|\mathbf{M},\mathbf{y}) \approx \mathcal{N}(\mathbf{w}|\mathbf{m},\mathbf{S})$ we have $\left\langle \log p(\mathbf{w}|\mathbf{M},\mathbf{y}) \right\rangle \approx \frac{1}{2}\log\det(\mathbf{S}) + c$ with c an additive constant. The second entropy is estimated by Gaussian conditioning on the joint approximate Gaussian density defined as $p(\mathbf{w},\mathbf{y}^*|\mathbf{y}) \propto \mathcal{N}(\mathbf{w}|\mathbf{m},\mathbf{S})\,\mathcal{N}\left(\mathbf{y}^*|\mathbf{M}^*\mathbf{w},\nu^2\mathbf{I}\right)$.

The approximation to the information gain can then be written as

$$I_{gain}(\mathbf{M}^*) \approx \frac{1}{2}\log\det\left(\mathbf{M}^*\mathbf{S}\mathbf{M}^{*\mathsf{T}} + \nu^2\mathbf{I}\right) + c.$$

If we constrain the measurements to have unit norm $I_{gain}(\mathbf{M}^*)$ above will be maximised when the rows of $\mathbf{M}^*$ lie along the leading principal eigenvectors of the approximate posterior covariance $\mathbf{S}$. These eigenvectors are approximated in our experiments using iterative Lanczos methods.

6.3.3 SYNTHETIC SIGNALS

Initially we consider applying sequential experimental design to a sparse signal reconstruction problem using small scale synthetic signals. In this artificial set up we wish to recover some signal

$\mathbf{w}_{tr} \in \mathbb{R}^{512}$ from a set of noisy linear measurement $\mathbf{y} \in \mathbb{R}^m$ where $m \ll 512$. We initialised the experiments with $m_0 = 40$ random unit norm linear measurement vectors $\mathbf{M} \in \mathbb{R}^{m_0 \times 512}$.

In this setup we placed the sparse prior directly on $\mathbf{w}$ with $\mathbf{B} = \mathbf{I}$. Sparse signals, $\mathbf{w}_{tr}$, were sampled independently over dimensions from either the Laplace ($\mu = 0, \tau = 0.2$) or the Student's t ($\nu = 3, \sigma^2 = 0.027$) densities. Noisy linear measurements were sampled from the source signals with $\mathbf{y} \sim \mathcal{N}\left(\mathbf{M}\mathbf{w}_{tr}, \nu^2\mathbf{I}\right)$ and $\nu^2 = 0.005$ throughout. Model priors and likelihoods were fixed to match the data generating densities.

For the Laplace generated signals we applied G-KL, local variational bounding (VB) and power G-EP ($\eta = 0.9$) approximate inference methods. G-EP and VB results were obtained using the publicly available `glm-ie` Matlab toolbox. Since the model is of sufficiently small dimensionality approximate covariance decompositions were not required. For the Student's t generated signals only G-KL and VB approximate inference methods were applied since G-EP is unstable in this setting.

For the Laplace signals, when $D = 512$ and $N = 110$, inference takes 0.3 seconds for VB, 0.6 seconds for G-EP, and 1.6 seconds for G-KL.[8] For the Student's t signals, again with $D = 512$ and $N = 110$, inference takes 0.3 seconds for VB and 6 seconds for G-KL. For Laplace signals, for which the G-KL bound is concave, gradient ascent was performed using a Hessian free Newton method with finite differences approximation for Hessian vector products—see (Nocedal and Wright, 2006, Chapter 7). For the Student's t signals, for which the G-KL bound is not guaranteed to be concave or even unimodal, gradient ascent was performed using nonlinear scaled conjugate gradients. G-KL optimisation was terminated in both settings once the largest absolute value of the bound's gradient was less than 0.01. VB and G-EP were optimised for seven outer loop iterations after which no systematic improvement in the approximate $\log Z$ value was observed.

ℓ^2 norm reconstruction error mean and standard error scores obtained over the 25 experiments conducted are presented in Figure 8. For the Laplace generated signals VB, G-EP and G-KL approximate inference procedures provide broadly the same reconstruction error performance. All sequentially designed procedures outperform MAP estimates with standard normal random measurements. The improved performance comes mainly in the first few iterations of the SED process with all methods achieving broadly similar iterative improvements in reconstruction error after that. For the Student's t prior again VB and G-KL procedures obtain broadly the same performance with G-KL appearing to become slightly less effective towards the end of the experiment.

6.3.4 NATURAL IMAGES

We consider sequential experimental design for the problem of recovering natural images from a set of under-determined noisy linear measurements. This problem is modelled by placing priors on the statistics of natural images that are known to exhibit sparsity. These statistics can be captured by suitable linear projections of the image vector (formed by concatenating the pixel value columns of the image). For the results presented we employ two types of image filter known to exhibit sparse statistics in natural images: finite differences, the difference in intensity values of horizontally or vertically neighbouring pixels; and multi-scale orthonormal wavelet transforms, constructed using the Daubechies four wavelet—see Seeger and Nickisch (2011b) for further details. Both filters can be expressed as extremely sparse vectors, the set of which is collected in the matrix $\mathbf{B}$, giving $\mathbf{B} \in \mathbb{R}^{M \times D}$ where $M = 3 \times D$. Image filters were implemented using the `glm-ie` package. Laplace

8. Experiments were timed using Matlab R2009a on a 32 bit Intel Core 2 Quad 2.5 GHz processor.

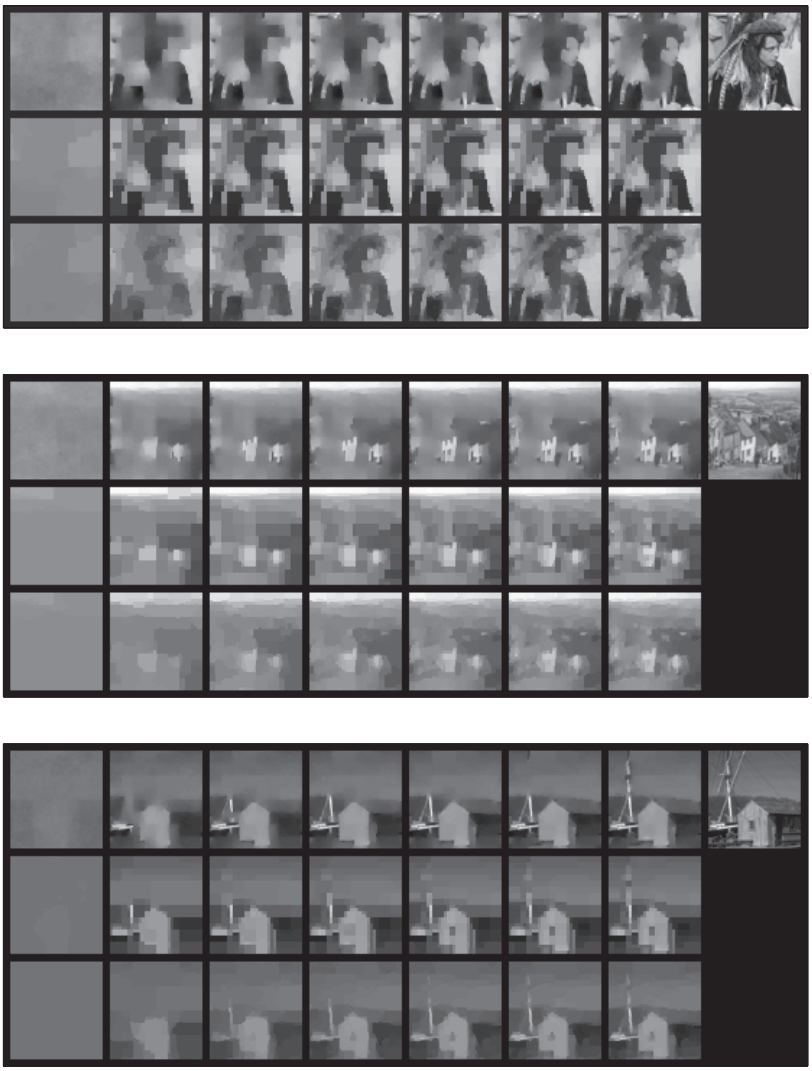


Figure 9: Reconstructed images from the Bayesian sequential experimental design (SED) experiments. We plot the estimated images obtained by each approximate inference procedure at different stages of the SED process. Each pane corresponds to a different underlying image. The true image is shown in the last image of the first row of each pane. Otherwise, the first row of each pane plots the G-KL mean, the second row the VB mean and the third row the MAP reconstruction with randomly selected measurement vectors. The k^{th} column of each pane plots the estimated image using $100 + 300 \times (k-1)$ measurements.

priors placed on each of the linear filter responses had $\tau = 0.1$ for the finite difference filters and

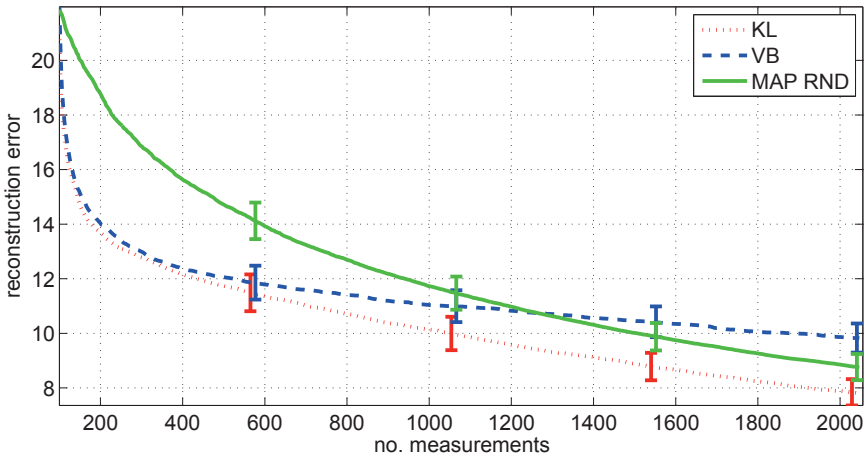


Figure 10: ℓ^2 reconstruction errors for the natural image sequential experimental design task. Mean and standard error scores are presented averaged over 16 different 64×64 pixel images.

$\tau = 0.14$ for the wavelet filters. This experimental approach follows that laid out in Chapter 5 of Nickisch (2010).

We apply the SED procedure detailed above by iteratively approximating the posterior density $p(\mathbf{w}|\mathbf{y}, \mathbf{M}, \mathbf{B}, \boldsymbol{\tau}, \nu^2)$ where: $\mathbf{w} \in \mathbb{R}^D$ corresponds to the unknown image vector; $\mathbf{y} \in \mathbb{R}^N$ the noisy measurements where $N \ll D$; and $\mathbf{M} \in \mathbb{R}^{N \times D}$ is the linear measurement matrix constrained to have rows with unit norm. The measurement matrix is initialised with 100 standard normal randomly sampled vectors normalised to have unit norm. The sequential experimental design process approximates the posterior based on current measurements and the prior, these are then used to select new unit norm linear measurement vectors $\mathbf{M}^* \in \mathbb{R}^{3 \times D}$ to append to $\mathbf{M}$. New observations are then synthetically generated by drawing samples from the Gaussian $\mathbf{y}^* \sim \mathcal{N}\left(\mathbf{M}\mathbf{w}_{tr}, \nu^2\mathbf{I}\right)$. In the experiments conducted we use $64 \times 64 = 4096 = D$ pixel grey scale images. The images were down sampled from a collection frequently used by the vision community,[9] gray scale pixel intensities were linearly transformed to lie in $[-1, 1]$. The likelihood model was fixed with $\nu^2 = 0.005$.

In this larger setting we apply G-KL and VB approximate inference methods only and make use of approximate covariance decompositions. For G-KL approximate inference we use the chevron Cholesky decomposition with 80 non-diagonal rows. The chevron Cholesky parameterisation was chosen due to its strong performance in previous experiments with respect to both convergence time and accuracy of inference—see Section 6.2. VB inference is applied with low rank decompositions of covariance using 80 Lanczos vectors. For the first iteration of the SED procedure, $N_0 = 100$, G-KL converged in 30 seconds and VB in 5 seconds. At each iteration of the SED process each inference procedure was initialised with the posterior from the previous SED iteration. When $N = 2048$ updating the Gaussian approximate posterior took 60 seconds for G-KL and 25 seconds for VB. Convergence of VB inference is difficult to asses since the double loop algorithm with Lanczos approximated covariance is not guaranteed at each iteration to increase the approximated

9. Images were downloaded from `decsai.ugr.es/cvg/dbimagenes/index.php`.

marginal likelihood. We iterated the VB procedure for seven outer loop iterations at which point no systematic increases of approximate marginal likelihood values were observed. Fluctuations in VB approximate marginal likelihood value in subsequent iterations were roughly ± 10. G-KL inference was terminated when the greatest absolute value of the bounds gradient was less than 0.1, at which point G-KL bound values increased by less than 0.5 per iteration. These results highlight a general distinction between the two methods, VB optimisation is an approximate EM algorithm whilst G-KL optimisation in this setting is implemented using an approximate second order gradient ascent procedure. EM is often reported to exhibit rapid convergence to low accuracy solutions but can be very slow at achieving high accuracy solutions (Salakhutdinov et al., 2003).

Reconstruction error results are plotted in Figure 10. We can see that SED offers greater reconstruction accuracy over random designs for a fixed budget of measurements. Up to roughly 400 designed measurement vectors both G-KL and VB procedures achieve similar reconstruction errors, after which the rate of VB iterative performance slows down eventually being overtaken by MAP reconstruction without design (MAP Rand). The reasons for this phenomenon are unclear. As more measurements are added the posterior density will become more spherical, for approximately spherical posteriors the benefit of design over simply adding random measurements is negligible. This could possibly explain the observation that G-KL and the MAP Rand procedures have similar gradients in Figure 10 towards the end of the experiment. Why the performance of VB approximate inference in particular degrades as more observations are added is not clear. One possible explanation is due to the Lanczos covariance approximation, as the posterior becomes increasingly spherical its spectrum will get flatter and the low-rank approximate factorisation may cause degraded Gaussian mean estimation.

Figure 9 displays the estimated deconvolved images at different stages of the SED process. Specifically we plot the G-KL and VB Gaussian mean estimates and the randomly designed MAP estimate. Interestingly, each method displays different visual traits with regards to the quality of the reconstructed image. G-KL estimates have patches with high fidelity and patches with low fidelity and a soft cloudy texture. VB and MAP Rand estimates appear more pixelated than the G-KL estimates with image accuracy more uniform across the image pane.

7. Discussion

We have presented several novel theoretical and practical developments concerning Gaussian Kullback-Leibler approximate inference procedures for models of the form of Equation (2). G-KL approximate inference is seeing a resurgence of interest by the research community—see, for example: Opper and Archambeau (2009), Ormerod and Wand (2012), Honkela et al. (2010) and Graves (2011). The work presented here provides further justification for its application as a Gaussian approximate inference procedure.

G-KL approximate inference's primary strength over other deterministic Gaussian approximate inference methods is the ease with which it can be applied to new models. All that is required to apply G-KL to a model of the form of Equation (2) is the pointwise evaluation of the univariate site projection potentials and that each of these potentials has unbounded support on $\mathbb{R}$. Unlike other deterministic Gaussian approximate inference methods G-KL does not require the site potentials to be differentiable, super-Gaussian or log-concave. Since the G-KL method optimises a strict lower-bound G-KL approximate inference is found to be numerically stable.

A long perceived disadvantage of G-KL approximate inference is the difficulty of optimising the bound with respect to the $O\left(D^2\right)$ parameters needed to specify the G-KL covariance matrix. We have shown, however, that whilst $O\left(D^2\right)$ parameters are required in full generality, the computations needed for bound optimisation compare favourably with other deterministic Gaussian approximate inference procedures. Importantly, we have shown that optimising the G-KL bound is a concave problem for models with log-concave potential functions $\{\phi_n\}_{n=1}^N$.

For larger problems we provided concave constrained parameterisations of covariance that allow G-KL methods to be applied to larger problems without imposing a priori factorisation assumptions on the approximate posterior density. The results presented in Section 6 show that such constrained covariance parameterisations are at least as good as other widely used deterministic methods at capturing posterior covariance. G-KL approximate inference using constrained concave covariance parameterisations have optimisation convergence times comparable to fast approximate variational local bound methods whilst maintaining a strict lower-bound on $\log Z$.

8. Publicly Available Code

A Matlab implementation of the G-KL approximate inference methods described in this paper is publicly available via the mloss.org website at `mloss.org/software/view/308/`. The `vgai` package implements G-KL approximate inference for models of the form of Equation (2) where each potential function is a site projection $\phi_n(\mathbf{w}) = \phi(\mathbf{w}^\mathsf{T}\mathbf{h}_n)$. The toolbox includes implementations of Gaussian, Laplace, Cauchy, Student's t, logistic sigmoid and logistic probit potential functions amongst others. Generic site projection potentials are supported if an implementation of $\psi := \log\phi : \mathbb{R} \to \mathbb{R}$ is provided. The package implements the unconstrained Cholesky, constrained Cholesky and factor analysis parameterisations of covariance discussed in Section 4.1. G-KL bound optimisation is achieved in the `vgai` package using Mark Schmidt's `minFunc` optimisation package.[10]

Acknowledgments

The authors would like to thank the reviewers for their valuable comments which greatly helped improve this manuscript. The authors would also like to thank Michalis K. Titsias for providing a cleaner presentation of the G-KL bound concavity result and Peter Sollich for many helpful comments and suggestions.

Appendix A. Univariate Expectation

For clarity of exposition we present a reworking of the result, as originally presented by Barber and Bishop (1998), that $\int \mathcal{N}(\mathbf{w}|\mathbf{m},\mathbf{S})\,\psi(\mathbf{w}^\mathsf{T}\mathbf{h})d\mathbf{w} = \int \mathcal{N}\left(y|\mathbf{w}^\mathsf{T}\mathbf{h},\mathbf{h}^\mathsf{T}\mathbf{S}\mathbf{h}\right)\psi(y)dy$, where: $\psi : \mathbb{R} \to \mathbb{R}$ is some nonlinear function, $\mathcal{N}(\mathbf{w}|\mathbf{m},\mathbf{S})$ is a multivariate Gaussian density with mean $\mathbf{m} \in \mathbb{R}^D$ and covariance $\mathbf{S} \in \mathbb{R}^{D\times D}$, and $\mathcal{N}\left(y|\mathbf{m}^\mathsf{T}\mathbf{h},\mathbf{h}^\mathsf{T}\mathbf{S}\mathbf{h}\right)$ is a univariate Gaussian with mean $\mathbf{m}^\mathsf{T}\mathbf{h}$ and variance $\mathbf{h}^\mathsf{T}\mathbf{S}\mathbf{h}$.

10. The `minFunc` package can be downloaded from `www.di.ens.fr/~mschmidt/Software/minFunc.html`.

We start by showing that the D-dimensional expectation $\left\langle \psi(\mathbf{w}^\mathsf{T}\mathbf{h}) \right\rangle_{\mathcal{N}(\mathbf{w}|\mathbf{m},\mathbf{S})}$ can be expressed as a univariate integral by making the substitution $\psi(\mathbf{w}^\mathsf{T}\mathbf{h}) = \int \delta(y - \mathbf{w}^\mathsf{T}\mathbf{h})\psi(y)dy$

$$\begin{aligned}
\left\langle \psi(\mathbf{w}^\mathsf{T}\mathbf{h}) \right\rangle_{\mathcal{N}(\mathbf{w}|\mathbf{m},\mathbf{S})} &= \int \mathcal{N}(\mathbf{w}|\mathbf{m},\mathbf{S})\, \psi(\mathbf{w}^\mathsf{T}\mathbf{h}) d\mathbf{w} \\
&= \int \mathcal{N}(\mathbf{w}|\mathbf{m},\mathbf{S}) \int \delta(y - \mathbf{w}^\mathsf{T}\mathbf{h})\psi(y) dy d\mathbf{w} \\
&= \int \underbrace{\int \mathcal{N}(\mathbf{w}|\mathbf{m},\mathbf{S})\, \delta(y - \mathbf{w}^\mathsf{T}\mathbf{h}) d\mathbf{w}}_{:= p(y)} \psi(y) dy.
\end{aligned}$$

We now seek to show that $p(y) \equiv \mathcal{N}\left(y|\mathbf{m}^\mathsf{T}\mathbf{h}, \mathbf{h}^\mathsf{T}\mathbf{S}\mathbf{h}\right)$. First we make the substitution $\mathbf{w} = \mathbf{C}^\mathsf{T}\mathbf{v} + \mathbf{m}$, where $\mathbf{C}$ is the Cholesky decomposition of $\mathbf{S}$ such that $\mathbf{S} = \mathbf{C}^\mathsf{T}\mathbf{C}$, to get

$$p(y) := \int \mathcal{N}(\mathbf{w}|\mathbf{m},\mathbf{S})\, \delta(y - \mathbf{w}^\mathsf{T}\mathbf{h}) d\mathbf{w} = \int \mathcal{N}(\mathbf{v}|\mathbf{0},\mathbf{I})\, \delta(y - \mathbf{v}^\mathsf{T}\mathbf{C}\mathbf{h} - \mathbf{m}^\mathsf{T}\mathbf{h}) d\mathbf{v}.$$

If we now define a basis in the vector space $\mathbf{v}$ with unit normal basis vectors $\{\mathbf{e}_d\}_{d=1}^D$ such that $\mathbf{e}_1$ is parallel to $\mathbf{C}\mathbf{h}$ so that $\mathbf{e}_1^\mathsf{T}\mathbf{C}\mathbf{h} = \|\mathbf{C}\mathbf{h}\|_2$ and so $\mathbf{e}_d^\mathsf{T}\mathbf{C}\mathbf{h} = 0$ when $d \neq 1$. Since $\mathcal{N}(\mathbf{v}|\mathbf{0},\mathbf{I})$ is isotropic the density is invariant to orthonormal transformations $\mathcal{N}(\mathbf{v}|\mathbf{0},\mathbf{I}) = \prod_{d=1}^D \mathcal{N}\left(\mathbf{e}_d^\mathsf{T}\mathbf{v}|0,1\right)$ and so

$$\begin{aligned}
p(y) &= \int \prod_{d=1}^D \mathcal{N}(v_d|0,1)\, \delta(y - \sum_{d=1}^D v_d \mathbf{e}_d^\mathsf{T}\mathbf{C}\mathbf{h} - \mathbf{m}^\mathsf{T}\mathbf{h}) d\mathbf{v} \\
&= \int \mathcal{N}(v_1|0,1)\, \delta(y - v_1 \mathbf{e}_1^\mathsf{T}\mathbf{C}\mathbf{h} - \mathbf{m}^\mathsf{T}\mathbf{h}) dv_1 \\
&= \mathcal{N}\left(y|\mathbf{m}^\mathsf{T}\mathbf{h}, \|\mathbf{C}\mathbf{h}\|_2^2\right) = \mathcal{N}\left(y|\mathbf{m}^\mathsf{T}\mathbf{h}, \mathbf{h}^\mathsf{T}\mathbf{S}\mathbf{h}\right).
\end{aligned}$$

Appendix B. G-KL Bound Gradients

We present the G-KL bound and its gradient for Gaussian and generic site projection potentials with full Cholesky and factor analysis parameterisations of G-KL covariance. Gradients for the chevron, banded and sparse Cholesky covariance parameterisations are implemented simply by placing that Cholesky parameterisation's sparsity mask on the full Cholesky gradient matrix. Subspace Cholesky G-KL gradients and associated optimisation procedures are discussed in Section B.4.

B.1 Entropy

For the Cholesky decomposition of covariance, $\mathbf{S} = \mathbf{C}^\mathsf{T}\mathbf{C}$, the entropy term of the G-KL bound and its gradient with respect to $\mathbf{C}$ are given by

$$\begin{aligned}
-\left\langle \log q(\mathbf{w}) \right\rangle_{q(\mathbf{w})} &= \frac{D}{2}\log(2\pi) + \frac{D}{2} + \sum_{d=1}^D \log(C_{dd}), \\
\frac{\partial}{\partial C_{ij}} -\left\langle \log q(\mathbf{w}) \right\rangle_{q(\mathbf{w})} &= \delta_{ij}\frac{1}{C_{ij}},
\end{aligned}$$

where δ_{ij} is the Kronecker delta.

For the factor analysis (FA) parameterisation of G-KL covariance, $\mathbf{S} = \text{diag}\left(\mathbf{d}^2\right) + \boldsymbol{\Theta}\boldsymbol{\Theta}^\mathsf{T}$ where $\mathbf{d} \in \mathbb{R}^D$ and $\boldsymbol{\Theta} \in \mathbb{R}^{D\times K}$, the entropy is given by,

$$-\left\langle \log q(\mathbf{w}) \right\rangle = \frac{D}{2}\log(2\pi) + \frac{D}{2} + \sum_d \log(d_d) + \frac{1}{2}\log\det\left(\mathbf{I}_{K\times K} + \boldsymbol{\Theta}^\mathsf{T}\text{diag}\left(\frac{1}{\mathbf{d}^2}\right)\boldsymbol{\Theta}\right),$$

admitting the gradients:

$$\frac{\partial}{\partial \mathbf{d}}\left\langle \log q(\mathbf{w}) \right\rangle_{q(\mathbf{w})} = 2\mathbf{d} \odot \text{diag}\left(\mathbf{S}^{-1}\right), \quad \text{and} \quad \frac{\partial}{\partial \boldsymbol{\Theta}}\left\langle \log q(\mathbf{w}) \right\rangle_{q(\mathbf{w})} = 2\mathbf{S}^{-1}\boldsymbol{\Theta}.$$

Where $\odot$ refers to taking the element wise product and $\text{diag}\left(\right)$ refers to either constructing a square diagonal matrix from a column vector or forming a column vector from the diagonal elements of a square matrix. Evaluating $\mathbf{S}^{-1}$ scales $O\left(K^2 D\right)$ using the Woodbury matrix inversion identity:

$$\mathbf{S}^{-1} = \text{diag}\left(\frac{1}{\mathbf{d}^2}\right) - \text{diag}\left(\frac{1}{\mathbf{d}^2}\right)\boldsymbol{\Theta}\left(\mathbf{I}_{K\times K} + \boldsymbol{\Theta}^\mathsf{T}\text{diag}\left(\frac{1}{\mathbf{d}^2}\right)\boldsymbol{\Theta}\right)^{-1}\boldsymbol{\Theta}^\mathsf{T}\text{diag}\left(\frac{1}{\mathbf{d}^2}\right).$$

B.2 Site Projection Potentials

Each site projection potential's contribution to the G-KL bound can be expressed as

$$I_n = \left\langle \log \phi_n(\mathbf{w}^\mathsf{T}\mathbf{h}_n) \right\rangle = \left\langle \log \phi(y) \right\rangle_{\mathcal{N}(y|m_n, s_n^2)} = \left\langle \log \phi(m_n + z s_n) \right\rangle_{\mathcal{N}(z|0,1)},$$

where $m_n = \mathbf{h}_n^\mathsf{T}\mathbf{m}$ and $s_n^2 = \mathbf{h}_n^\mathsf{T}\mathbf{S}\mathbf{h}_n$. In order that general potentials of this form can be easily implemented for different functions ϕ_n we present the gradients according to their chain rule decomposition,

$$\frac{\partial I_n}{\partial \mathbf{m}} = \frac{\partial I_n}{\partial m_n}\frac{\partial m_n}{\partial \mathbf{m}} \quad \text{and} \quad \frac{\partial I_n}{\partial \mathbf{C}} = \frac{\partial I_n}{\partial s_n^2}\frac{\partial s_n^2}{\partial \mathbf{C}}. \tag{17}$$

Expressing I_n and its derivatives as expectations with respect to the standard normal density renders the implementation of numerical integration routines simpler whilst avoiding expressions involving the derivative of the potential function itself. The expectations and their derivatives are given by:

$$\begin{aligned}
I_n &= \int \mathcal{N}(z|0,1)\log\phi_n(m_n + z s_n)dz, \\
\frac{\partial I_n}{\partial m_n} &= \int z\mathcal{N}(z|0,1)\frac{\log\phi_n(m_n + z s_n)}{s_n}dz, \\
\frac{\partial I_n}{\partial s_n^2} &= \int \left(z^2 - 1\right)\mathcal{N}(z|0,1)\frac{\log\phi_n(m_n + z s_n)}{2s_n^2}dz.
\end{aligned}$$

The gradients of $m_n = \mathbf{h}_n^\mathsf{T}\mathbf{m}$ and $s_n^2 = \mathbf{h}_n^\mathsf{T}\mathbf{S}\mathbf{h}_n$ are

$$\frac{\partial m_n}{\partial \mathbf{m}} = \mathbf{h}_n, \quad \text{and} \quad \frac{\partial s_n^2}{\partial \mathbf{C}} = 2\text{triu}\left(\mathbf{C}\mathbf{h}_n\mathbf{h}_n^\mathsf{T}\right),$$

where $\text{triu}\left(\cdot\right)$ is a sparsity mask such that elements below the diagonal are fixed to zero. For FA parameterisations we have

$$\frac{\partial s_n^2}{\partial \mathbf{d}} = 2\mathbf{h}_n^2 \odot \mathbf{d}, \quad \text{and} \quad \frac{\partial s_n^2}{\partial \boldsymbol{\Theta}} = 2\mathbf{h}_n\mathbf{h}_n^\mathsf{T}\boldsymbol{\Theta}.$$

B.2.1 LAPLACE POTENTIALS

The Gaussian expectation of the logarithm of a Laplace potential has a simple analytic expression. Laplace potentials, as considered here, take the product of site projections form. Accordingly, we need only present the derivatives with respect to m_n and s_n^2 so that they can be used in conjunction with Equation (17). We consider the case of a zero mean Laplace density, $p(\mathbf{w}^\mathsf{T}\mathbf{h}_n|\tau) = e^{-|\mathbf{w}^\mathsf{T}\mathbf{h}_n|/\tau}/2\tau$, giving

$$\langle \log p(m_n + z s_n) \rangle_z = -\log(2\tau) - \frac{1}{\tau} \langle |m_n + z s_n| \rangle_z . \tag{18}$$

Laplace potentials with non zero mean, $p(x) = e^{-|x-\eta|/\tau}/2\tau$, can be calculated by making the simple transformation $m_n' = m_n - \eta$. Evaluating the last term of Equation (18) above involves computing the expectation of a rectified univariate Gaussian random variable,

$$\langle |m_n + z s_n| \rangle_z = \left(\frac{2}{\pi}\right)^{\frac{1}{2}} s_n e^{-\frac{1}{2}a_n^2} + m_n \left[1 - 2\Phi(-a_n)\right]$$

where $\Phi(x) := \int_{-\infty}^{x} \mathcal{N}(t|0,1)\, dt$ and $a_n := m_n/s_n$. The corresponding derivatives of which are:

$$\frac{\partial \langle |m_n + z s_n| \rangle}{\partial m_n} = 1 - 2\Phi(-a_n),$$
$$\frac{\partial \langle |m_n + z s_n| \rangle}{\partial s_n^2} = \frac{a_n^2 + 1}{\sqrt{2\pi s_n^2}} e^{-\frac{1}{2}a_n^2} - \frac{a_n^2}{s_n} \mathcal{N}(a_n|0,1).$$

B.3 Gaussian Potentials

For a Gaussian potential $\mathcal{N}(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ the log expectation is given by

$$\left\langle \log \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \right\rangle_{q(\mathbf{w})} = -\frac{1}{2}\left[\log\det(2\pi\boldsymbol{\Sigma}) + (\mathbf{m}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{m}-\boldsymbol{\mu}) + \operatorname{trace}\left(\boldsymbol{\Sigma}^{-1}\mathbf{S}\right)\right].$$

Derivatives with respect to the mean and covariance are:

$$\frac{\partial}{\partial \mathbf{m}} \left\langle \log \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \right\rangle = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \mathbf{m}), \quad \text{and} \quad \frac{\partial}{\partial \mathbf{C}} \left\langle \log \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \right\rangle = -\operatorname{triu}\left(\mathbf{C}\boldsymbol{\Sigma}^{-1}\right).$$

For the FA covariance structure we have,

$$\frac{\partial}{\partial \mathbf{d}} \left\langle \log \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \right\rangle = -\operatorname{diag}\left(\boldsymbol{\Sigma}^{-1}\right) \odot \mathbf{d}, \quad \text{and} \quad \frac{\partial}{\partial \boldsymbol{\Theta}} \left\langle \log \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \right\rangle = -\boldsymbol{\Sigma}^{-1}\boldsymbol{\Theta}.$$

B.3.1 GAUSSIAN LIKELIHOODS

Linear models with additive Gaussian noise have a likelihood potential that can be expressed as $\mathcal{N}\left(\mathbf{y}|\mathbf{H}^\mathsf{T}\mathbf{w}, \boldsymbol{\Sigma}\right)$ where $\mathbf{H} \in \mathbb{R}^{D\times N}$ and $\mathbf{y} \in \mathbb{R}^N$. In this setting typically we assume isotropic noise $\boldsymbol{\Sigma} = \nu^2\mathbf{I}$ and so present gradients for this case only. The expectation of the log of this term has the following algebraic from

$$\left\langle \log \mathcal{N}\left(\mathbf{y}|\mathbf{H}^\mathsf{T}\mathbf{w}, \nu^2\mathbf{I}\right) \right\rangle = -\frac{1}{2}\left[N\log(2\pi\nu^2) + \frac{1}{\nu^2}\left\langle \left(\mathbf{y} - \mathbf{H}^\mathsf{T}\mathbf{w}\right)^\mathsf{T}\left(\mathbf{y} - \mathbf{H}^\mathsf{T}\mathbf{w}\right)\right\rangle\right], \tag{19}$$

where the expectation of the quadratic can be expressed as

$$\left\langle \left(\mathbf{y}-\mathbf{H}^{\mathsf{T}}\mathbf{w}\right)^{\mathsf{T}}\left(\mathbf{y}-\mathbf{H}^{\mathsf{T}}\mathbf{w}\right)\right\rangle = \mathbf{y}^{\mathsf{T}}\mathbf{y} - 2\mathbf{y}^{\mathsf{T}}\mathbf{H}^{\mathsf{T}}\mathbf{m} + \sum_{ij}\left[\mathbf{C}\mathbf{H}\right]_{ij}^{2} + \sum_{i}\left[\mathbf{H}^{\mathsf{T}}\mathbf{m}\right]_{i}^{2}.$$

Equation (19) admits the gradients:

$$\frac{\partial}{\partial \mathbf{m}}\left\langle \log \mathcal{N}\left(\mathbf{y}|\mathbf{H}^{\mathsf{T}}\mathbf{w},\nu^{2}\mathbf{I}\right)\right\rangle = \frac{1}{\nu^{2}}\left(\mathbf{y}^{\mathsf{T}}\mathbf{H}^{\mathsf{T}} - \mathbf{H}\mathbf{H}^{\mathsf{T}}\mathbf{m}\right),$$
$$\frac{\partial}{\partial \mathbf{C}}\left\langle \log \mathcal{N}\left(\mathbf{y}|\mathbf{H}^{\mathsf{T}}\mathbf{w},\nu^{2}\mathbf{I}\right)\right\rangle = -\frac{1}{\nu^{2}}\mathrm{triu}\left(\mathbf{C}\mathbf{H}\mathbf{H}^{\mathsf{T}}\right).$$

For the FA parameterised covariance we have

$$\left\langle \left(\mathbf{y}-\mathbf{H}^{\mathsf{T}}\mathbf{w}\right)^{\mathsf{T}}\left(\mathbf{y}-\mathbf{H}^{\mathsf{T}}\mathbf{w}\right)\right\rangle = \mathbf{y}^{\mathsf{T}}\mathbf{y} - 2\mathbf{y}^{\mathsf{T}}\mathbf{H}^{\mathsf{T}}\mathbf{m} + \sum_{i}\left[\mathbf{H}^{\mathsf{T}}\mathbf{m}\right]_{i}^{2} + \sum_{ij}\left[\mathbf{\Theta}^{\mathsf{T}}\mathbf{H}^{\mathsf{T}}\right]_{ij}^{2} + \sum_{j}\left(\sum_{i}H_{ji}^{2}\right)d_{j}^{2}$$

with corresponding gradients:

$$\frac{\partial}{\partial d_{j}}\left\langle \log \mathcal{N}\left(\mathbf{y}|\mathbf{H}^{\mathsf{T}}\mathbf{w},\nu^{2}\mathbf{I}\right)\right\rangle = -\frac{1}{\nu^{2}}\left(\sum_{i}H_{ji}^{2}\right)d_{j},$$
$$\frac{\partial}{\partial \mathbf{\Theta}}\left\langle \log \mathcal{N}\left(\mathbf{y}|\mathbf{M}\mathbf{w},\nu^{2}\mathbf{I}\right)\right\rangle = -\frac{1}{\nu^{2}}\mathbf{H}\mathbf{H}^{\mathsf{T}}\mathbf{\Theta}.$$

B.3.2 GAUSSIAN POTENTIALS AS SITE PROJECTIONS

The Gaussian potential $\mathcal{N}(\mathbf{w}|\boldsymbol{\mu},\mathbf{\Sigma})$ can be equivalently expressed as a product of D site projection potentials. To see this we use the Cholesky factorisation of the precision matrix $\mathbf{\Sigma}^{-1} = \mathbf{P}^{\mathsf{T}}\mathbf{P}$. Making this substitution, we see that

$$\mathcal{N}(\mathbf{w}|\boldsymbol{\mu},\mathbf{\Sigma}) \propto e^{-\frac{1}{2}(\mathbf{w}-\boldsymbol{\mu})^{\mathsf{T}}\mathbf{P}^{\mathsf{T}}\mathbf{P}(\mathbf{w}-\boldsymbol{\mu})} = e^{-\frac{1}{2}\|\mathbf{P}(\mathbf{w}-\boldsymbol{\mu})\|_{2}^{2}} = \prod_{d=1}^{D} e^{-\frac{1}{2}\left(\mathbf{p}_{d}^{\mathsf{T}}(\mathbf{w}-\boldsymbol{\mu})\right)^{2}}, \tag{20}$$

where the vector $\mathbf{p}_d$ is the d^{th} row vector of $\mathbf{P}$, that is $\mathbf{p}_d^{\mathsf{T}} := \mathbf{P}_{d,:}$. Thus Equation 20 is a product of D site projections with potential function $\phi_d(x) \propto e^{-\frac{1}{2}x^2}$.

B.4 Subspace Covariance Decomposition

We consider optimising the G-KL bound with respect to a covariance matrix parameterised on a subspace of the parameters $\mathbf{w} \in \mathbb{R}^D$. Letting $\mathbf{E} = [\mathbf{E}_1, \mathbf{E}_2]$ be a matrix of orthonormal vectors that span $\mathbb{R}^D$ then we may parameterise the covariance as

$$\mathbf{S}' = \mathbf{E}^{\mathsf{T}}\mathbf{S}\mathbf{E} = [\mathbf{E}_1, \mathbf{E}_2]^{\mathsf{T}}\mathbf{S}\,[\mathbf{E}_1, \mathbf{E}_2],$$

which is equivalent to making an orthonormal transformation in the space of parameters $\mathbf{w}$ using $\mathbf{E}$. If we restrict $\mathbf{S}$ to be block diagonal, $\mathbf{S} = \mathrm{diag}\,(\mathbf{S}_1, \mathbf{S}_2)$, we can write $\mathbf{S}'$ as the sum

$$\mathbf{S}' = \mathbf{E}_1^{\mathsf{T}}\mathbf{S}_1\mathbf{E}_1 + \mathbf{E}_2^{\mathsf{T}}\mathbf{S}_2\mathbf{E}_2.$$

Since $\mathbf{E}$ is orthonormal it does not effect the value or gradient of the entropy's contribution to the bound since $\log\det(\mathbf{S}) = \log\det(\mathbf{S}')$. Provided the Gaussian potential has spherical covariance, $\boldsymbol{\Sigma} = \nu^2\mathbf{I}$, then $\mathbf{E}$ does not effect its contribution the G-KL bound since

$$\text{trace}\left(\boldsymbol{\Sigma}^{-1}\mathbf{S}'\right) = \frac{1}{\nu^2}\text{trace}\left(\mathbf{E}^\mathsf{T}\mathbf{S}\mathbf{E}\right) = \frac{1}{\nu^2}\text{trace}\left(\mathbf{S}\right).$$

Thus we are left to evaluate the projected variance terms $\left\{s_n^2\right\}_{n=1}^N$ required to evaluate the product of site potentials contribution. For $\mathbf{S}$ block diagonal with the second block component spherical, $\mathbf{S}_2 = c^2\mathbf{I}$, the orthonormal basis vectors $\mathbf{E}_2$ do not need to be computed or maintained since

$$s_n^2 = \mathbf{h}_n^\mathsf{T}\mathbf{S}'\mathbf{h}_n = \mathbf{h}_n^\mathsf{T}\mathbf{E}_1^\mathsf{T}\mathbf{S}_1\mathbf{E}_1\mathbf{h}_n + c^2\mathbf{h}_n^\mathsf{T}\mathbf{E}_2^\mathsf{T}\mathbf{E}_2\mathbf{h}_n = \mathbf{h}_n^\mathsf{T}\mathbf{E}_1^\mathsf{T}\mathbf{S}_1\mathbf{E}_1\mathbf{h}_n + c^2\left(\|\mathbf{h}_n\|_2^2 - \|\mathbf{E}_1\mathbf{h}_n\|^2\right).$$

We seek to optimise the G-KL bound w.r.t. to the subspace parameterised variational Gaussian by iterating between optimising the bound with respect to the parameters $\{\mathbf{m}, \mathbf{C}_1, c\}$ and updating the subspace basis vectors $\mathbf{E}_1$. In Section B.4.1 we present the gradients required to optimise the G-KL bound with respect to $\{\mathbf{m}, \mathbf{C}_1, c\}$. In Sections B.4.2 and B.4.3 we consider different routes to optimising the subspace basis $\mathbf{E}_1$.

B.4.1 SUBSPACE CHOLESKY G-KL BOUND GRADIENTS

In this subsection we present the subspace Cholesky G-KL bound gradients. The subspace covariance matrix is given by $\mathbf{S} = \mathbf{E}_1^\mathsf{T}\mathbf{C}_1^\mathsf{T}\mathbf{C}_1\mathbf{E}_1 + c^2\mathbf{E}_2^\mathsf{T}\mathbf{E}_2$, where $\mathbf{C}_1 \in \mathbb{R}^{K\times K}$ is a Cholesky matrix, $c \in \mathbb{R}^+$ and $D = K + L$. Since $\mathbf{E}_2$ does not occur in the expressions presented below, in what follows we omit subscripts and denote $\mathbf{E}_1$ and $\mathbf{C}_1$ as $\mathbf{E}$ and $\mathbf{C}$. We reiterate that the Gaussian potential has spherical covariance $\boldsymbol{\Sigma} = \nu^2\mathbf{I}$. The G-KL bound for the subspace Cholesky covariance parameterisation is given by

$$\begin{aligned}\mathcal{B}_{KL}(\mathbf{m}, \mathbf{C}, c, \mathbf{E}) = \frac{D}{2}\log(2\pi) + \frac{D}{2} + \sum_{k=1}^K \log(C_{kk}) + L\log(c) \\ - \frac{D}{2}\log\left(2\pi\nu^2\right) - \frac{1}{\nu^2}\left[\|\mathbf{m} - \boldsymbol{\mu}\|_2^2 + \text{trace}\left(\mathbf{C}^\mathsf{T}\mathbf{C}\right) + Lc^2\right] \\ + \sum_{n=1}^N \left\langle \log\phi_n(m_n + z s_n)\right\rangle_{\mathcal{N}(z|0,1)}.\end{aligned}$$

The gradient of the G-KL entropy's contribution to the bound is

$$\frac{\partial}{\partial C_{ij}} - \langle\log q(\mathbf{w})\rangle = \delta_{ij}\frac{1}{C_{ij}}, \quad \text{and} \quad \frac{\partial}{\partial c} - \langle\log q(\mathbf{w})\rangle = \frac{L}{c}.$$

The Gaussian potential's contribution to the G-KL bound admits the gradients:

$$\frac{\partial}{\partial \mathbf{C}}\left\langle\log\mathcal{N}\left(\mathbf{w}|\boldsymbol{\mu}, \nu^2\mathbf{I}\right)\right\rangle = -\frac{1}{\nu^2}\mathbf{C}, \quad \text{and} \quad \frac{\partial}{\partial c}\left\langle\log\mathcal{N}\left(\mathbf{w}|\boldsymbol{\mu}, \nu^2\mathbf{I}\right)\right\rangle = -\frac{Lc}{\nu^2}.$$

The site projection potential's contribution to the G-KL bound is computed as in Section B.2 but with the partial derivatives of s_n^2 with respect to $\mathbf{C}$ and c:

$$\frac{\partial s_n^2}{\partial \mathbf{C}} = 2\text{triu}\left(\mathbf{C}\tilde{\mathbf{h}}_n\tilde{\mathbf{h}}_n^\mathsf{T}\right), \quad \frac{\partial s_n^2}{\partial c} = 2c\left(\|\mathbf{h}_n\|_2^2 - \|\tilde{\mathbf{h}}_n\|_2^2\right),$$

where $\tilde{\mathbf{h}}_n := \mathbf{E}\mathbf{h}_n$.

B.4.2 SUBSPACE OPTIMISATION : PROJECTED GRADIENT ASCENT

One route to finding good subspace vectors $\mathbf{E}_1$ is to directly optimise the bound with respect to them. Again we omit subscripts since $\mathbf{E}_2$ makes no contribution to the expressions below. Optimisation is complicated by the fact that we require $\mathbf{E}$ to be orthonormal, that is we require that $\mathbf{E}^\mathsf{T}\mathbf{E} = \mathbf{I}_{K\times K}$. The set of all such orthonormal vectors forms a smooth manifold in $\mathbb{R}^{D\times K}$. A crude but simple approach to optimising the bound with respect to $\mathbf{E}$ is projected gradient ascent—after each gradient step we orthonormalise the updated basis:

$$\mathbf{E}^{new} := \text{orth}\left[\mathbf{E} + \alpha \frac{\partial}{\partial \mathbf{E}} \mathcal{B}_{KL}(\mathbf{m}, \mathbf{C}, \mathbf{E}, c)\right]$$

where $\text{orth}[\cdot]$ denotes an orthonormalisation operator, implemented for instance using a Gram-Schmidt procedure or the singular value decomposition, and α is a parameter controlling the gradient step size.

As described above, when $\mathbf{\Sigma} = \nu^2\mathbf{I}$, the only term in the G-KL bound that depends on $\mathbf{E}$ are the site projection potential functions $\left\langle \log \phi_n(\mathbf{w}^\mathsf{T}\mathbf{h}_n)\right\rangle$. The derivative of the bound then with respect to $\mathbf{E}$ is given by

$$\frac{\partial}{\partial \mathbf{E}} \mathcal{B}_{KL}(\mathbf{m}, \mathbf{C}, \mathbf{E}, c) = \sum_n \frac{\partial}{\partial s_n^2} \left\langle \log \phi(m_n + z s_n)\right\rangle \frac{\partial s_n^2}{\partial \mathbf{E}},$$

where the partial derivative with respect to s_n^2 is given in Section B.2 and

$$\frac{\partial s_n^2}{\partial \mathbf{E}} = \frac{\partial}{\partial \mathbf{E}} \mathbf{h}_n^\mathsf{T}\mathbf{E}^\mathsf{T}\mathbf{C}^\mathsf{T}\mathbf{C}\mathbf{E}\mathbf{h}_n = 2\mathbf{C}^\mathsf{T}\mathbf{C}\mathbf{E}\mathbf{h}_n\mathbf{h}_n^\mathsf{T}.$$

B.4.3 SUBSPACE OPTIMISATION: FIXED POINT ITERATION

Another route to optimising the subspace vectors $\mathbf{E}$ is to use the form for the optimal G-KL covariance matrix presented in Equation (10). Using this method, once we have optimised the bound w.r.t. $\{\mathbf{m}, \mathbf{C}_1, c\}$ we update the subspace vectors $\mathbf{E}$ to be the leading K eigenvectors of $\mathbf{S}$ as defined in Equation (21). Whilst this procedure is not guaranteed to increase the bound in experiments it has yielded strong performance—see for example Section 6.2 and Challis and Barber (2011).

For problems where the Gaussian potential has isotropic variance, $\mathbf{\Sigma} = \nu^2\mathbf{I}$, the form for the optimal G-KL inverse covariance, Equation (10), simplifies to

$$\mathbf{S}^{-1} = \frac{1}{\nu^2}\mathbf{I} + \mathbf{H}\mathbf{\Gamma}\mathbf{H}^\mathsf{T}, \tag{21}$$

where $\mathbf{\Gamma}$ is defined in Equation (11) of Section 11. We now consider two routes to updating the subspace vectors $\mathbf{E}$. First, we consider an approximate eigen decomposition method suitable for smaller non-sparse problems. Second, we consider an iterative Lanczos method better suited to larger sparse problems.

One route to possibly recovering the K leading eigenvectors of $\mathbf{S}$ is to evaluate the K smallest eigenvectors of $\frac{1}{\nu^2}\mathbf{I} + \mathbf{H}\mathbf{\Gamma}\mathbf{H}^\mathsf{T}$. We note that $\mathbf{H}\mathbf{\Gamma}\mathbf{H}^\mathsf{T} \approx \mathbf{H}\mathbf{\Gamma}'\mathbf{H}^\mathsf{T}$ where $\Gamma'_{nn} = \Gamma_{nn}$ if $\Gamma_{nn} > \delta$ and zero otherwise - we set δ small enough such that there are K non zero diagonal elements $\mathbf{\Gamma}'$. If we now calculate the eigen decomposition to $\mathbf{H}\mathbf{\Gamma}'\mathbf{H}^\mathsf{T} = \mathbf{E}\mathbf{\Lambda}\mathbf{E}^\mathsf{T}$ we see that

$$\left[\frac{1}{\nu^2}\mathbf{I} + \mathbf{H}\mathbf{\Gamma}'\mathbf{H}^\mathsf{T}\right]^{-1} = \mathbf{E}\text{diag}\left(\frac{\nu^2}{1 + \lambda'_{nn}\nu^2}\right)\mathbf{E}^\mathsf{T}.$$

For $L \ll D$ we can evaluate the L eigenvectors of $\mathbf{H}\boldsymbol{\Gamma}'\mathbf{H}^\mathsf{T}$ cheaply since the eigenvalues of $\mathbf{X}\mathbf{X}^\mathsf{T}$ coincide with the eigenvalues of $\mathbf{X}^\mathsf{T}\mathbf{X}$.[11] Therefore approximating the K dimensional subspace eigen decomposition reduces to the complexity of decomposing a $K \times K$ matrix. If δ is small enough this method can often outperform approximate iterative decompositions provided the data is non-sparse and of moderate dimensionality.

Iterative Lanczos methods can approximately recover the eigenvectors corresponding to the largest and smallest eigenvalues of a matrix. General details about Lanczos methods can be found in Golub and Van Loan (1996), for the special case of covariance matrices of the form Equation (21) details are provided in Seeger (2010). Iterative Lanczos methods are fast provided the number of eigenvectors we wish to recover is not too large and matrix vector products can be computed efficiently—for example when the matrix has some special structure or is sparse.

Appendix C. Newton Convergence Rate Conditions

Sufficient conditions under which optimising $\mathcal{B}_{KL}(\mathbf{m}, \mathbf{C})$ using Newton's method will exhibit quadratic convergence rates are that $\mathcal{B}_{KL}(\mathbf{m}, \mathbf{C})$ is twice continuously differentiable, strongly concave, has closed sublevel sets and has Lipschitz continuous Hessians on the sublevel sets (Boyd and Vandenberghe, 2004, section 9.5.3). In Section 3.2 we showed that if all ϕ_n are log-concave then the bound is strongly concave in $\mathbf{m}, \mathbf{C}$. In this section we provide conditions for which the other requirements hold.

We consider G-KL inference problems of the form of Equation (2) where $\{\phi_n\}_{n=1}^N$ are site projection potentials that are piecewise exponentiated quadratics, log-concave and have unbounded support on $\mathbb{R}$. Specifically, we show that the required properties hold for potential functions that can be written

$$\phi(x) := \sum_{i=0}^{I} \mathbb{I}\left[x \in (l_i, l_{i+1})\right] \exp(a_i x^2 + b_i x + c_i)$$

where $-\infty = l_0 < l_1, ..., l_{I+1} = \infty$ and $\mathbb{I}[\cdot]$ is an indicator function equal to one when its argument is true and zero otherwise. Note that $\phi(x)$ need not be continuous and can have jump discontinuities at the partition points l_k. For such functions we have that $\log \phi(x) = \sum_{i=0}^{I} \mathbb{I}\left[x \in (l_i, l_{i+1})\right] a_i x^2 + b_i x + c_i$.

C.1 Continuously Differentiable

The expectation of such potentials can then be expressed as a sum of integrals each over a disjoint domain

$$\left\langle \log \phi(\mathbf{w}^\mathsf{T}\mathbf{h}) \right\rangle = \sum_{i=0}^{I} \int_{l_i}^{l_{i+1}} \mathcal{N}\left(z|m, s^2\right) a_i z^2 + b_i z + c_i dz, \tag{22}$$

where $m = \mathbf{m}^\mathsf{T}\mathbf{h}$ and $s^2 = \|\mathbf{C}\mathbf{h}\|_2^2$. Each integral on the right hand side of Equation (22) has a known analytic form which depends on terms of up to order 2 in m, s, standard normal density functions and standard normal cumulative distribution functions—see Marlin et al. (2011) and Herbrich (2005) for their explicit forms and derivatives w.r.t. m, s. As an example, and to make this more concrete, we give the truncated expectation of just the quadratic term $a_i z^2$ below

$$\int_{l_i}^{l_{i+1}} a_i z^2 \mathcal{N}\left(z|m, s^2\right) dz = a_i \left[s^2 \left(\tilde{l}_i \mathcal{N}(\tilde{l}_i) - \tilde{l}_{i+1} \mathcal{N}(\tilde{l}_{i+1}) \right) + \left(s^2 + m^2\right) \left(\Phi(\tilde{l}_{i+1}) - \Phi(\tilde{l}_i) \right) \right],$$

11. To see this consider the eigen equation for $\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{E} = \mathbf{E}\boldsymbol{\Lambda}$ thus $\mathbf{X}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{E} = \mathbf{X}\mathbf{E}\boldsymbol{\Lambda}$.

where $\tilde{l}_i := (l_i - m)/s$, $\mathcal{N}(x)$ is the standard normal density function and $\Phi(x)$ the standard normal cumulative distribution function. The truncated Gaussian expectation of the linear, $b_i z$, and the constant, c_i, terms have similar simpler analytic expressions.

We note that the standard normal density function and the standard normal cumulative density function are both smooth. Thus the expectation in Equation (22) is the sum of smooth functions w.r.t. the parameters m, s. Therefore Equation (22) as a function of $\mathbf{m}, \mathbf{C}$ is the composition of a function that is smooth in m, s and the functions $m = \mathbf{m}^\mathsf{T}\mathbf{h}$ and $s^2 = \|\mathbf{C}\mathbf{h}\|_2^2$ that are smooth in $\mathbf{m}, \mathbf{C}$. By the chain rule, we see that $\left\langle \log \phi(\mathbf{w}^\mathsf{T}\mathbf{h}) \right\rangle$ is smooth with respect to $\mathbf{m}, \mathbf{C}$.

By Lebesgue's dominated convergence theorem, we expect the differentiability of $\left\langle \log \phi(\mathbf{w}^\mathsf{T}\mathbf{h}) \right\rangle$ to hold for a much broader class of potentials ϕ than the piecewise exponentiated quadratic class of functions considered here.

C.2 G-KL Sublevel Sets are Closed

The G-KL sublevel sets, $\mathcal{S}$, are defined

$$\mathcal{S} := \left\{ \mathbf{m} \in \mathbb{R}^D, \mathbf{C} \in \mathbb{R}^{D \times D}_{chol} | \mathcal{B}(\mathbf{m}, \mathbf{C}) \geq \mathcal{B}(\mathbf{m}_0, \mathbf{C}_0) \right\},$$

where $\mathbf{m}_0, \mathbf{C}_0$ are the moments that the G-KL bound optimisation procedure is initialised with and $\mathbb{R}^{D \times D}_{Chol}$ is the set of $D \times D$ upper triangular Cholesky matrices with strictly positive diagonals. Importantly $\mathcal{S}$ is closed since the G-KL bound is a closed function—which is a sufficient condition (Boyd and Vandenberghe, 2004, p.471). A function $f : \mathbb{R}^D \to \mathbb{R}$ with $\mathrm{dom}(f)$ open is closed iff f converges to $-\infty$ along every sequence converging to a boundary point of $\mathrm{dom}(f)$ (Boyd and Vandenberghe, 2004, p.640). The G-KL bound is closed since it is the sum of the entropic term (which up to a constant is equal to $\sum_d \log C_{dd}$), a negative quadratic in $\mathbf{m}, \mathbf{C}$, and $\left\langle \log \phi(\mathbf{w}^\mathsf{T}\mathbf{h}) \right\rangle$ (proven to be jointly concave in $\mathbf{m}, \mathbf{C}$). Thus for any sequence of moments $\{\mathbf{m}_k, \mathbf{C}_k\}$ that converges to the boundary of the G-KL domain we have $\mathcal{B}_{KL}(\mathbf{m}_k, \mathbf{C}_k)$ converging to $-\infty$.

C.3 G-KL Lipschitz Continuous Hessians

We say the Hessian of f is Lipschitz continuous on $\mathcal{S}$ if there exists a constant $L \geq 0$ such that $\forall \mathbf{x}, \mathbf{y} \in \mathcal{S}$

$$\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\|_2 \leq L \|\mathbf{x} - \mathbf{y}\|_2.$$

An equivalent condition is that the Hessian has bounded and continuous derivatives on $\mathcal{S}$. Since the bound is continuously differentiable, since the sublevel sets are closed and since the entropy's contribution to the bound ensures that s^2 is bounded below by a positive constant this property holds.

Appendix D. Complexity of Bound and Gradient Computations

To perform G-KL approximate inference we optimise the G-KL bound, Equation (7), by gradient ascent. In this section we consider the computational scaling properties of single evaluations of the bound and its gradient. We consider each term that depends on the variational parameters $\mathbf{m}$ and $\mathbf{S}$ separately, namely: $\log\det(\mathbf{S})$ from the entropy's contribution, $\mathrm{trace}\left(\boldsymbol{\Sigma}^{-1}\mathbf{S}\right)$ and $\mathbf{m}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\mathbf{m}$ from the Gaussian potential's contribution, and $\left\{m_n, s_n^2\right\}_{n=1}^N$ from the product of site projection potential's contribution.

The G-KL covariance parameterisations we consider are: full Cholesky, diagonal Cholesky, banded Cholesky with bandwidth B, chevron Cholesky with K non-diagonal rows, subspace Cholesky

with K dimensional subspace, sparse Cholesky with DK non-zeros, and factor analysis (FA) with K factor loading vectors. We report only the leading scaling terms and assume, for the sake of clarity, that $N \geq D \geq K, B$ where N is the number of site factors and D is the dimensionality of the parameter vector $\mathbf{w}$. In the last column we report the complexity figures required to compute the projected Gaussian moments $\{m_n, s_n^2\}_{n=1}^N$ where $m_n = \mathbf{m}^\mathsf{T}\mathbf{h}_n$, $s_n^2 = \|\mathbf{C}\mathbf{h}_n\|_2^2$, and $\mathtt{nnz} : \mathbb{R}^D \to \mathbb{N}$ is a function that counts the number of non-zero elements in a vector.

	$\log\det(\mathbf{S})$	trace $(\boldsymbol{\Sigma}^{-1}\mathbf{S})$			$\mathbf{m}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\mathbf{m}$			$\{m_n, s_n^2\}_{n=1}^N$	
		$\boldsymbol{\Sigma}$ - iso	$\boldsymbol{\Sigma}$ - diag	$\boldsymbol{\Sigma}$ - full	$\boldsymbol{\Sigma}$ - iso	$\boldsymbol{\Sigma}$ - diag	$\boldsymbol{\Sigma}$ - full	$\mathtt{nnz}(\mathbf{h}) = D$	$\mathtt{nnz}(\mathbf{h}) = L$
$\mathbf{C}_{full}$	$O(D)$	$O(D^2)$	$O(D^2)$	$O(D^3)$	$O(D)$	$O(D)$	$O(D^2)$	$O(ND^2)$	$O(NDL)$
$\mathbf{C}_{diag}$	$O(D)$	$O(D)$	$O(D)$	$O(D)$	$O(D)$	$O(D)$	$O(D^2)$	$O(ND)$	$O(NL)$
$\mathbf{C}_{band}$	$O(D)$	$O(DB)$	$O(DB)$	$O(D^2B)$	$O(D)$	$O(D)$	$O(D^2)$	$O(NDB)$	$O(NLB)$
$\mathbf{C}_{chev}$	$O(D)$	$O(DK)$	$O(DK)$	$O(D^2K)$	$O(D)$	$O(D)$	$O(D^2)$	$O(NDK)$	$O(NLK)$
$\mathbf{C}_{sub}$	$O(K)$	$O(DK)$	$O(DK)$	$O(K^3)$	$O(D)$	$O(D)$	$O(D^2)$	$O(NK^2)$	$O(NK^2)$
$\mathbf{C}_{spar}$	$O(D)$	$O(DK)$	$O(DK)$	$O(D^2K)$	$O(D)$	$O(D)$	$O(D^2)$	$O(NDK)$	$O(NLK)$
$\mathbf{S}_{FA}$	$O(D^2K)$	$O(DK)$	$O(DK)$	$O(KD^2)$	$O(D)$	$O(D)$	$O(D^2)$	$O(NDK)$	$O(NLK)$

Appendix E. Transformation of Basis

When the model's Gaussian potential, $\mathcal{N}(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, has full covariance optimising the G-KL bound can sometimes be made less expensive by linearly transforming the basis of the parameter vectors $\mathbf{m}$ and $\mathbf{C}$. To do this, essentially we hard code the information contributed to the posterior from the Gaussian potential into our G-KL parameters. That is we parameterise $\mathbf{m}$ and $\mathbf{C}$ as

$$\mathbf{C} = \tilde{\mathbf{C}}\mathbf{P} \quad \text{and} \quad \mathbf{m} = \mathbf{P}^\mathsf{T}\tilde{\mathbf{m}} + \boldsymbol{\mu} \tag{23}$$

where $\mathbf{P}$ is the Cholesky decomposition of the prior covariance such that $\mathbf{P}^\mathsf{T}\mathbf{P} = \boldsymbol{\Sigma}$. For the G-KL moments parameterised this way we each term of the G-KL bound can be evaluated using:

$$\begin{aligned}
-\langle \log q(\mathbf{w}) \rangle &= \log\det\left(\tilde{\mathbf{C}}\right) + \log\det(\mathbf{P}) + \frac{D}{2}\log(2\pi) + \frac{D}{2}, \\
2\left\langle \log \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \right\rangle &= -D\log(2\pi) - D - 2\log\det(\mathbf{P}) - \tilde{\mathbf{m}}^\mathsf{T}\tilde{\mathbf{m}} - \mathrm{trace}\left(\tilde{\mathbf{C}}^\mathsf{T}\tilde{\mathbf{C}}\right), \\
\left\langle \psi(\mathbf{w}^\mathsf{T}\mathbf{h}) \right\rangle &= \int \mathcal{N}(z|0,1)\,\psi(m + zs)dz,
\end{aligned}$$

where $m := \tilde{\mathbf{m}}^\mathsf{T}\tilde{\mathbf{h}} + \boldsymbol{\mu}^\mathsf{T}\mathbf{h}$, $s := \|\tilde{\mathbf{C}}\tilde{\mathbf{h}}\|_2^2$ and $\tilde{\mathbf{h}} := \mathbf{P}\mathbf{h}$. Combining these terms the G-KL bound can be written

$$\mathcal{B}(\mathbf{m}, \mathbf{C}) = \tilde{\mathcal{B}}(\tilde{\mathbf{m}}, \tilde{\mathbf{C}}) = \sum_d \log\left(\tilde{C}_{dd}\right) - \frac{1}{2}\tilde{\mathbf{m}}^\mathsf{T}\tilde{\mathbf{m}} - \frac{1}{2}\sum_{ij}\tilde{C}_{ij}^2 + \sum_n \left\langle \log\phi_n(m_n + zs_n) \right\rangle_{\mathcal{N}(z,0)1}.$$

We are free then to optimise the G-KL bound just with respect to $\tilde{\mathbf{m}}, \tilde{\mathbf{C}}$ at a reduced cost. For a model with a full covariance Gaussian potential and non-sparse $\mathbf{H} = [\mathbf{h}_1, ..., \mathbf{h}_N]$ computing the bound and gradient of $\tilde{\mathcal{B}}(\tilde{\mathbf{m}}, \tilde{\mathbf{C}})$ scales $O\left(D^2 + ND^2\right)$ whereas computing the bound and gradient of the untransformed bound scales $O\left(D^3 + ND^2\right)$—see the table in Appendix D.

This procedure requires some pre-processing—namely the Cholesky decomposition of $\boldsymbol{\Sigma}$ and the 'whitening' of the data set $\tilde{\mathbf{H}} = \mathbf{P}\mathbf{H}$ which scale $O\left(D^3\right)$ and $O\left(ND^2\right)$ respectively. And some post-processing—the final G-KL moments $\mathbf{m}$ and $\mathbf{C}$ are obtained using equations Equation (23) which require a matrix-vector and a matrix-matrix product which scale $O\left(D^2\right)$ and $O\left(D^3\right)$ respectively.

Since during optimisation the bound and its gradient are usually computed many more times than twice, the basis transformation procedure detailed above will result in a significant computational saving. Note that this procedure can speed up G-KL bound optimisation only in settings where $\mathbf{h}_n$ are not sparse. For example Gaussian process regression models, where $\mathbf{h}_n$ are standard normal basis vectors, will not benefit from this reparameterisation since $\tilde{\mathbf{h}}_n = \mathbf{P}\mathbf{h}_n$ are not sparse.

Appendix F. Gaussian Process Regression

In this section we present equations necessary to implement Gaussian process regression models using G-KL approximate inference methods.

F.1 Predictive Density

A Gaussian approximation to the posterior density on the latent function values of the training data may be used to obtain an approximation to the predictive density of the latent function value for a new test point. The GP predictive density to the target variable y_* for a new input $\mathbf{x}_*$ is defined by the integral

$$p(y_*|\mathbf{x}_*,\mathbf{X},\mathbf{y}) = \int p(y_*|w_*)p(w_*|\mathbf{X},\mathbf{y},\mathbf{x}_*)dw_*.$$

The distribution on the test point latent function value, $p(w_*|\mathbf{X},\mathbf{y},\mathbf{x}_*)$, is approximated by marginalising out the training set latent variables using our Gaussian approximate posterior, $\mathcal{N}(\mathbf{w}|\mathbf{m},\mathbf{S}) \approx p(\mathbf{w}|\mathbf{X},\mathbf{y},\boldsymbol{\theta})$, giving

$$\begin{aligned}
p(w_*|\mathbf{X},\mathbf{y},\mathbf{x}_*) &= \int p(w_*|\mathbf{w},\mathbf{X},\mathbf{x}_*)p(\mathbf{w}|\mathbf{y},\mathbf{X})d\mathbf{w} \\
&= \int \mathcal{N}\left(w_*|\boldsymbol{\sigma}_*^\mathsf{T}\boldsymbol{\Sigma}^{-1}\mathbf{w}, \sigma_{**} - \boldsymbol{\sigma}_*^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma}_*\right) p(\mathbf{w}|\mathbf{y},\mathbf{X})d\mathbf{w} \\
&\approx \int \mathcal{N}\left(w_*|\boldsymbol{\sigma}_*^\mathsf{T}\boldsymbol{\Sigma}^{-1}\mathbf{w}, \sigma_{**} - \boldsymbol{\sigma}_*^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma}_*\right) \mathcal{N}(\mathbf{w}|\mathbf{m},\mathbf{S})\, d\mathbf{w} \\
&= \mathcal{N}\left(w_*|\boldsymbol{\sigma}_*^\mathsf{T}\boldsymbol{\Sigma}^{-1}\mathbf{m}, \sigma_{**} - \boldsymbol{\sigma}_*^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma}_* + \boldsymbol{\sigma}_*^\mathsf{T}\boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma}_*\right),
\end{aligned}$$

where $\boldsymbol{\sigma}_*$ and σ_{**} are the prior covariance and variance terms of the test data point $\mathbf{x}_*$. The elements of $\boldsymbol{\sigma}_*$ are calculated by evaluating the covariance function, $k(\mathbf{x},\mathbf{x}')$, between the each of the training covariates and the test point covariate such that $[\boldsymbol{\sigma}_*]_m = k(\mathbf{x}_m,\mathbf{x}_*)$ and $\sigma_{**} = k(\mathbf{x}_*,\mathbf{x}_*)$.

F.2 Hyperparameter Optimisation

For a general likelihood $p(\mathbf{y}|\mathbf{w}) = \prod_{n=1}^N \phi_n(w_n)$ and GP prior $\mathcal{N}(\mathbf{w}|\mathbf{0},\boldsymbol{\Sigma})$ with covariance function $\Sigma_{mn} = k(\mathbf{x}_m,\mathbf{x}_n)$ we get the G-KL bound

$$\begin{aligned}
\mathcal{B}_{KL}(\mathbf{m},\mathbf{C}) = \frac{D}{2} + \sum_n \log C_{nn} - \frac{1}{2}\log\det(\boldsymbol{\Sigma}) - \frac{1}{2}\mathbf{m}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\mathbf{m} - \frac{1}{2}\mathrm{trace}\left(\boldsymbol{\Sigma}^{-1}\mathbf{S}\right) \\
+ \sum_n \left\langle \log\phi(m_n + z\sqrt{S_{nn}}) \right\rangle_{\mathcal{N}(z|0,1)}.
\end{aligned}$$

Taking the derivative of the above expression with respect to the covariance hyperparameters $\boldsymbol{\theta}$ we get

$$\frac{\partial\mathcal{B}_{KL}}{\partial\boldsymbol{\theta}} = -\frac{1}{2}\mathrm{trace}\left(\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\boldsymbol{\theta}}\right) + \frac{1}{2}\mathbf{m}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\boldsymbol{\theta}}\boldsymbol{\Sigma}^{-1}\mathbf{m} + \frac{1}{2}\mathrm{trace}\left(\mathbf{C}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\boldsymbol{\theta}}\boldsymbol{\Sigma}^{-1}\mathbf{C}\right). \tag{24}$$

Note that $\mathbf{m}$ and $\mathbf{C}$ implicitly depend on the covariance hyperparameters $\boldsymbol{\theta}$. However, cross terms such as

$$\frac{\partial \mathcal{B}_{KL}}{\partial \mathbf{m}} \frac{\partial \mathbf{m}}{\partial \boldsymbol{\theta}} \quad \text{or} \quad \frac{\partial \mathcal{B}_{KL}}{\partial \mathbf{C}} \frac{\partial \mathbf{C}}{\partial \boldsymbol{\theta}}$$

do not contribute to Equation (24) at the optimum of the G-KL bound since the gradients of $\mathcal{B}_{KL}$ with respect to $\mathbf{m}$ or $\mathbf{C}$ are zero at this point. Therefore, to evaluate the gradient of $\mathcal{B}_{KL}$ with respect to the covariance hyperparameters first the G-KL bound is optimised with respect to $\mathbf{m}, \mathbf{C}$ with $\boldsymbol{\theta}$ fixed, then at that optimum we use Equation (24) to calculate the derivative with respect to $\boldsymbol{\theta}$.

Appendix G. Bayesian Logistic Regression Results

			$N_{trn} = 125$		$N_{trn} = 250$		$N_{trn} = 1250$	
			$K = 13$	$K = 25$	$K = 13$	$K = 25$	$K = 13$	$K = 25$
Time (s)	G-KL	Chev	0.14±0.01	0.16±0.00	0.32±0.02	0.34±0.01	3.31±0.09	3.38±0.14
		Band	0.21±0.01	0.28±0.01	0.41±0.01	0.53±0.01	4.05±0.09	4.64±0.09
		Sub	0.42±0.05	0.46±0.02	0.69±0.03	0.81±0.04	4.24±0.15	5.17±0.28
		FA	0.75±0.05	0.74±0.05	0.94±0.08	1.12±0.08	6.18±0.61	5.49±0.40
	VB		0.27±0.01	0.28±0.00	0.29±0.00	0.31±0.01	0.46±0.01	0.45±0.00
$\tilde{\mathcal{B}}$	G-KL	Chev	−1.08±0.02	−1.05±0.02	−0.89±0.01	−0.87±0.01	−0.41±0.00	−0.40±0.00
		Band	−1.05±0.02	−1.00±0.01	−0.88±0.01	−0.85±0.01	−0.41±0.00	−0.40±0.00
		Sub	−2.93±0.01	−2.11±0.02	−1.83±0.01	−1.43±0.01	−0.60±0.00	−0.52±0.00
		FA	−1.08±0.02	−1.06±0.02	−0.89±0.01	−0.87±0.01	−0.40±0.00	−0.39±0.00
	VB		–±–	–±–	–±–	–±–	–±–	–±–
$\|\mathbf{m} - \mathbf{w}_{tr}\|_2/D$	G-KL	Chev	1.48±0.01	1.48±0.01	1.38±0.01	1.38±0.01	1.11±0.01	1.11±0.01
		Band	1.48±0.01	1.48±0.01	1.38±0.01	1.38±0.01	1.11±0.01	1.11±0.01
		Sub	1.49±0.01	1.48±0.01	1.43±0.01	1.41±0.01	1.20±0.01	1.18±0.01
		FA	1.48±0.01	1.48±0.01	1.38±0.01	1.38±0.01	1.11±0.01	1.11±0.01
	VB		1.52±0.01	1.51±0.01	1.45±0.01	1.45±0.02	1.21±0.01	1.21±0.01
$\log p(\mathbf{y}^*\|\mathbf{X}^*)/N_{tst}$	G-KL	Chev	−0.57±0.01	−0.56±0.01	−0.47±0.01	−0.47±0.01	−0.19±0.00	−0.19±0.00
		Band	−0.56±0.01	−0.56±0.01	−0.47±0.01	−0.46±0.01	−0.19±0.00	−0.19±0.00
		Sub	−0.67±0.02	−0.63±0.02	−0.57±0.02	−0.54±0.02	−0.21±0.01	−0.20±0.01
		FA	−0.57±0.01	−0.57±0.01	−0.48±0.01	−0.47±0.01	−0.19±0.00	−0.19±0.00
	VB		−0.68±0.02	−0.68±0.02	−0.57±0.01	−0.56±0.01	−0.21±0.01	−0.21±0.01

Table 3: Bayesian logistic regression results for a unit variance Gaussian prior, with parameter dimension $D = 250$ and number of test points $N_{tst} = 2500$. Experimental setup and metrics are described in Section 6.2.

References

D. Barber. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 2012.

D. Barber and C. Bishop. Ensemble learning in bayesian neural networks. In *Neural Networks and Machine Learning*, pages 215–237. Springer, 1998.

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

E. Challis and D. Barber. Concave gaussian variational approximations for inference in large-scale bayesian linear models. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 2011.

			$N_{trn}=500$		$N_{trn}=1000$		$N_{trn}=5000$	
			$K=50$	$K=100$	$K=50$	$K=100$	$K=50$	$K=100$
Time (s)	G-KL	Chev	2.68±0.04	3.37±0.05	6.41±0.11	7.28±0.14	75.23±1.51	78.38±2.10
		Band	6.66±0.58	8.97±0.14	12.81±0.15	20.59±0.26	127.69±2.36	190.65±3.47
		Sub	1.59±0.07	2.58±0.12	3.24±0.03	7.71±0.20	56.67±1.80	75.35±1.63
		FA	9.94±1.00	12.24±0.63	16.21±0.74	18.64±1.09	70.87±3.92	82.13±5.83
	VB		1.78±0.03	2.65±0.05	4.12±0.04	6.17±0.07	21.88±0.03	33.87±0.02
$\tilde{\mathcal{B}}$	G-KL	Chev	−1.28±0.01	−1.24±0.01	−0.99±0.00	−0.96±0.00	−0.42±0.00	−0.41±0.00
		Band	−1.24±0.01	−1.17±0.01	−0.98±0.00	−0.94±0.00	−0.42±0.00	−0.42±0.00
		Sub	−5.40±0.23	−4.54±0.25	−7.56±0.00	−1.52±0.00	−0.62±0.00	−0.54±0.00
		FA	−1.29±0.01	−1.26±0.01	−1.00±0.00	−0.97±0.00	−0.42±0.00	−0.41±0.00
	VB		–±–	–±–	–±–	–±–	–±–	–±–
$\|\mathbf{w}-\mathbf{w}_{tr}\|_2/D$	G-KL	Chev	0.53±0.00	0.53±0.00	0.49±0.00	0.49±0.00	0.38±0.00	0.38±0.00
		Band	0.53±0.00	0.53±0.00	0.49±0.00	0.49±0.00	0.38±0.00	0.38±0.00
		Sub	0.56±0.00	0.55±0.00	0.56±0.00	0.50±0.00	0.44±0.00	0.43±0.00
		FA	0.53±0.00	0.53±0.00	0.49±0.00	0.49±0.00	0.38±0.00	0.38±0.00
	VB		0.54±0.00	0.54±0.00	0.52±0.00	0.52±0.00	0.45±0.00	0.45±0.00
$\log p(\mathbf{y}^*\|\mathbf{X}^*)/N_{tst}$	G-KL	Chev	−0.62±0.01	−0.61±0.01	−0.51±0.01	−0.49±0.01	−0.18±0.00	−0.18±0.00
		Band	−0.61±0.01	−0.59±0.01	−0.50±0.01	−0.49±0.01	−0.18±0.00	−0.18±0.00
		Sub	−0.62±0.01	−0.61±0.01	−0.69±0.00	−0.61±0.01	−0.21±0.00	−0.21±0.00
		FA	−0.62±0.01	−0.61±0.01	−0.52±0.01	−0.51±0.01	−0.18±0.00	−0.18±0.00
	VB		−0.88±0.01	−0.95±0.02	−0.68±0.01	−0.70±0.01	−0.21±0.00	−0.21±0.00

Table 4: Bayesian logistic regression results for a unit variance Gaussian prior, with parameter dimension $D=1000$ and number of test points $N_{tst}=5000$. Experimental setup and metrics are described in Section 6.2.

R. Fergus, B. Singh, A. Hertzmann, S. Roweis, and W. Freeman. Removing camera shake from a single photograph. In *ACM Transactions on Graphics*.

J. Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–67, 1991.

M. Gibbs and D. MacKay. Variational gaussian process classifiers. *IEEE Transactions on Neural Networks*, 11(6):1458–1464, 2000.

M. Girolami. A variational method for learning sparse and overcomplete representations. *Neural Computation*, 13(11):2517–2532, 2001.

G. Golub and C. Van Loan. *Matrix Computations*. John Hopkins University Press, 1996.

A. Graves. Practical variational inference for neural networks. In *Advances in Neural Information Processing Systems 24*, 2011.

R. Herbrich. On gaussian expectation propagation. Technical report, Microsoft Research Cambridge, `research.microsoft.com/pubs/74554/EP.pdf`, 2005.

A. Honkela, T. Raiko, M. Kuusela, M. Tornio, and J. Karhunen. Approximate riemannian conjugate gradient learning for fixed-form variational bayes. *Journal of Machine Learning Research*, 11: 3235–3268, 2010.

T. Jaakkola and M. Jordan. A variational approach to bayesian logistic regression problems and their extensions. In *Proceedings of the Sixth International Workshop on Artificial Intelligence and Statistics*, 1997.

P. Jylanki, J. Vanhatalo, and A. Vehtrari. Robust Gaussian Process Regression with a Student-t Likelihood. *Journal of Machine Learning Research*, 12:3187–3225, 2011.

K. Ko and M. Seeger. Large scale variational bayesian inference for structured scale mixture models. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

M. Kuss. *Gaussian Process Models for Robust Regression, Classification, and Reinforcement Learning*. PhD thesis, Technischen Universität Darmstadt, Darmstadt, Germany, 2006.

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

D. MacKay and J. Oldfield. Generalization error and the number of hidden units in a multilayer perceptron. Technical report, Cambridge University, `www.inference.phy.cam.ac.uk/mackay/gen.ps.gz`, 1995.

B. Marlin, M. Khan, and K. Murphy. Piecewise bounds for estimating bernoulli-logistic latent gaussian models. In *Proceedings of the 28th International Conference on Machine Learning*, 2011.

T. Minka. Expectation propagation for approximate bayesian inference. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, 2001.

T. Minka. Power EP. Technical report, Department of Statistics, Carnegie Mellon University, `research.microsoft.com/pubs/67427/tr-2004-149.pdf`, 2004.

R. Neal. Monte carlo implementation of gaussian process models for bayesian regression and classification. Technical report, Department of Statistics and Department of Computer Science, University of Toronto, `arxiv.org/abs/physics/9701026v2`, 1997.

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

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

H. Nickisch and C. Rasmussen. Approximations for binary gaussian process classification. *Journal of Machine Learning Research*, 9:2035–2078, 10 2008.

H. Nickisch and M. Seeger. Convex variational bayesian inference for large scale generalized linear models. In *Proceedings of the 26th International Conference on Machine Learning*, 2009.

J. Nocedal and S. Wright. *Numerical Optimization*. Springer, 2006.

B. Olshausen and D. Field. Natural image statistics and efficient coding. *Network: Computation in Neural Systems*, 7:333–339, 2 1996.

M. Opper and C. Archambeau. The variational gaussian approximation revisited. *Neural Computation*, 21(3):786–792, 2009.

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

J. Ormerod and M. Wand. Gaussian variational approximate inference for generalized linear mixed models. *Journal of Computational and Graphical Statistics*, 21(1):2–17, 2012.

A. Palmer, D. Wipf, K. Kreutz-Delgado, and B. Rao. Variational EM algorithms for non-Gaussian latent variable models. In *Advances in Neural Information Processing Systems 20*, 2006.

G. Papandreou and A. Yuille. Gaussian sampling by local perturbations. In *Advances in Neural Information Processing Systems 11*, 2010.

T. Park and G. Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103: 681–686, 2008.

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

C. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

R. Salakhutdinov, S. Roweis, and Z. Ghahramani. Optimization with EM and expectation-conjugate-gradient. In *Proceedings of the 20th International Conference on Machine Learning*, 2003.

L. Saul, T. Jaakkola, and M. Jordan. Mean field theory for sigmoid belief networks. *Journal of Artificial Intelligence Research*, 4:61–76, 1996.

M. Schmidt, G. Fung, and R. Rosales. Fast optimization methods for l1 regularization: A comparative study and two new approaches. In *The 18th European Conference on Machine Learning*, 2007.

M. Seeger. Bayesian methods for support vector machines and gaussian processes. Master's thesis, University of Karlsruhe, 1999a.

M. Seeger. Bayesian model selection for support vector machines, gaussian processes and other kernel classifiers. In *Advances in Neural Information Processing Systems 12*. 1999b.

M. Seeger. Low rank updates for the cholsky decomposition. Technical report, University of California at Berkeley, `infoscience.epfl.ch/record/161468/files/cholupdate.pdf`, 2007.

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

M. Seeger. Sparse linear models: Variational approximate inference and bayesian experimental design. *Journal of Physics: Conference Series*, 197(1), 2009.

M. Seeger. Gaussian covariance and scalable variational inference. In *Proceedings of the 27th International Conference on Machine Learning*, 2010.

M. Seeger and H. Nickisch. Compressed sensing and bayesian experimental design. In *Proceedings of the 25th International Conference on Machine Learning*, pages 912–919, 2008.

M. Seeger and H. Nickisch. Large scale variational inference and experimental design for sparse generalized linear models. Technical report, Max Planck Institute for Biological Cybernetics, //arxiv.org/abs/0810.0901, 2010.

M. Seeger and H. Nickisch. Fast convergent algorithms for expectation propagation approximate inference. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 2011a.

M. Seeger and H. Nickisch. Large scale bayesian inference and experimental design for sparse linear models. *SIAM Journal on Imaging Sciences*, 4(1):166–199, 2011b.

M. Seeger, S. Gerwinn, and M. Bethge. Bayesian inference for sparse generalized linear models. In *The 18th European Conference on Machine Learning*, 2007.

M. Tipping. Probabilistic visualisation of high-dimensional binary data. In *Advances in Neural Information Processing Systems 11*, 1999.

J. Vanhatalo, P. Jylnki, and A. Vehtari. Gaussian process regression with a student-t likelihood. In *Advances in Neural Information Processing Systems 22*, 2009.

D. Wipf. Sparse bayesian learning for basis selection. *IEEE Transactions on Signal Processing*, 52 (8):2153–2164, 2004.