Concave Gaussian Variational Approximations for Inference in Large-Scale Bayesian Linear Models

Edward Challis **David Barber**
Computer Science Department, University College London, London WC1E 6BT, UK.

Abstract

Two popular approaches to forming bounds in approximate Bayesian inference are local variational methods and minimal Kullback-Leibler divergence methods. For a large class of models we explicitly relate the two approaches, showing that the local variational method is equivalent to a weakened form of Kullback-Leibler Gaussian approximation. This gives a strong motivation to develop efficient methods for KL minimisation. An important and previously unproven property of the KL variational Gaussian bound is that it is a concave function in the parameters of the Gaussian for log concave sites. This observation, along with compact concave parametrisations of the covariance, enables us to develop fast scalable optimisation procedures to obtain lower bounds on the marginal likelihood in large scale Bayesian linear models.

1 BAYESIAN MODELS

For parameter $\mathbf{w}$ and data $\mathcal{D}$, a large class of Bayesian models describe posteriors of the form

$$p(\mathbf{w}|\mathcal{D}) = \frac{1}{Z}\mathcal{N}(\mathbf{w}|\boldsymbol{\mu},\boldsymbol{\Sigma})\,\phi(\mathbf{w}), \qquad (1.1)$$
$$Z = \int \mathcal{N}(\mathbf{w}|\boldsymbol{\mu},\boldsymbol{\Sigma})\,\phi(\mathbf{w})d\mathbf{w}$$

for a Gaussian factor $\mathcal{N}(\mathbf{w}|\boldsymbol{\mu},\boldsymbol{\Sigma})$ and positive potential function $\phi(\mathbf{w})$. This class includes generalised linear models, see *e.g.* Hardin and Hilbe (2007), and Gaussian noise models in inverse modeling, see *e.g.* Wipf and Nagarajan (2009). A classic example is Bayesian logistic regression in which $\mathcal{N}(\mathbf{w}|\boldsymbol{\mu},\boldsymbol{\Sigma})$ is the prior on the weight $\mathbf{w}$, $\phi(\mathbf{w})$ the likelihood $p(\mathcal{D}|\mathbf{w})$ and $Z = p(\mathcal{D})$.

For large parameter dimension, $D = \dim(\mathbf{w})$, the normalisation constant Z in equation (1.1) is computationally intractable, except for limited special cases. Evaluating Z is essential for the purposes of model comparison, hyper-parameter estimation, active learning and experimental design. Indeed, any marginal function of the posterior $p(\mathbf{w}|\mathcal{D})$, such as a moment, also implicitly requires Z.

Due to the importance of this large model class, a great deal of effort has been dedicated to finding accurate approximations to posteriors of the form equation (1.1). Whilst there are many different possible approximation routes, including sampling, consistency methods such as expectation propagation and perturbation techniques such as Laplace, see *e.g.* Barber (2011), our interest here is uniquely in techniques that form a lower bound on Z. Such lower bounds are particularly useful in parameter estimation and provide concrete exact knowledge about Z. Furthermore, lower bounds may be coupled with upper bounds on Z to form bounds on marginal quantities of interest (Gibbs and MacKay, 2000).

A well studied route to forming a lower bound on Z is to use a so-called local variational method that bounds the integrand with a parametric function, see *e.g.* Gibbs and MacKay (2000); Jaakkola and Jordan (1996); Girolami (2001); Nickisch and Seeger (2009); Palmer et al. (2006). Local variational optimisation procedures are, however, computationally demanding, requiring the solution of linear systems of dimension D. For this reason, considerable attention has been paid to characterising the convexity of local bounds and developing fast scalable solvers (Nickisch and Seeger, 2009; Palmer et al., 2006).

In contrast the Variational Gaussian (VG) method directly approximates the posterior by minimising the Kullback-Leibler divergence between a parametrised Gaussian approximation and the posterior. The VG method provides a bound on Z for all positive func-

Appearing in Proceedings of the 14^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2011, Fort Lauderdale, FL, USA. Volume 15 of JMLR: W&CP 15.

tions $\phi(\mathbf{w})$, whilst the local bounding procedure requires ϕ to be super-Gaussian[1]. Whilst such variational Gaussian (VG) approximations are not new (Barber and Bishop, 1998; Seeger, 1999; Kuss and Rasmussen, 2005; Opper and Archambeau, 2009), we contribute several results concerning this procedure:

- For posteriors in the form of equation (1.1) we make clear the relationship between local and VG bounds and show that VG bounds are provably tighter than local ones. Furthermore, this improvement can give rise to differences in the mass they assign to competing models.
- The VG bound has been considered unfavourable compared to local bounds due to their important convexity properties. Here we prove that the VG bound is in fact also concave for log-concave ϕ.
- The VG method is often dismissed as impractical due to the difficulty of specifying covariances in large systems. We provide explicit scalable concave parametrisations for the covariance. To demonstrate the efficacy of our approach, we apply the method to large datasets, outperforming the local method in bound value and matching or exceeding it in terms of computational speed.

1.1 Local variational method

The local variational method replaces $\phi(\mathbf{w})$ in equation (1.1) with a bound that renders the integral analytically tractable. Provided the function ϕ is super-Gaussian, one may bound $\phi(\mathbf{w})$ by an exponential quadratic function (Palmer et al., 2006)

$$\phi(\mathbf{w}) \geq c(\xi) e^{-\frac{1}{2}\mathbf{w}^\mathsf{T}\mathbf{F}(\xi)\mathbf{w}+\mathbf{w}^\mathsf{T}\mathbf{f}(\xi)} \tag{1.2}$$

where the matrix $\mathbf{F}(\xi)$, vector $\mathbf{f}(\xi)$ and scalar $c(\xi)$ depend on the specific function ϕ; ξ is a variational parameter that enables one to find the tightest bound.

Many models of practical utility have super-Gaussian potentials, examples of which are the logistic sigmoid inverse link function $\phi(x) = (1+\exp(-x))^{-1}$, Laplace potentials where $\phi(x) \propto \exp(-|x|)$ and Student's t-distribution. We discuss explicit c, $\mathbf{F}$ and $\mathbf{f}$ functions for such potentials later, but for the moment leave them unspecified.

Bounding $\phi(\mathbf{w})$ with the squared exponential, we obtain

$$Z \geq c(\xi) \frac{e^{-\frac{1}{2}\boldsymbol{\mu}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}}{\sqrt{\det(2\pi\boldsymbol{\Sigma})}} \int e^{-\frac{1}{2}\mathbf{w}^\mathsf{T}\mathbf{A}\mathbf{w}+\mathbf{w}^\mathsf{T}\mathbf{b}} d\mathbf{w} \tag{1.3}$$

[1] A function $\phi(x)$ is super-Gaussian if $\exists b \in \mathbb{R}$ s.t. for $g(x) := \log\phi(x) - bx$ is even, and is convex and decreasing as a function of $y = x^2$ (Seeger and Nickisch, 2010).

where

$$\mathbf{A} \equiv \boldsymbol{\Sigma}^{-1} + \mathbf{F}(\xi), \quad \mathbf{b} \equiv \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \mathbf{f}(\xi) \tag{1.4}$$

Whilst both $\mathbf{A}$ and $\mathbf{b}$ are functions of ξ, we drop this dependency for a more compact notation. One can interpret equation (1.3) as a Gaussian approximation to the posterior where $p(\mathbf{w}|\mathcal{D}) \approx \mathcal{N}\left(\mathbf{w}|\mathbf{A}^{-1}\mathbf{b}, \mathbf{A}^{-1}\right)$. Completing the square in equation (1.3) and integrating, we have $\log Z \geq \mathcal{B}(\xi)$, where

$$\mathcal{B}(\xi) \equiv \log c(\xi) - \frac{1}{2}\boldsymbol{\mu}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \frac{1}{2}\mathbf{b}^\mathsf{T}\mathbf{A}^{-1}\mathbf{b} - \frac{1}{2}\log\det(\boldsymbol{\Sigma}\mathbf{A}) \tag{1.5}$$

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

In many practical problems of interest, including generalised linear models and inverse modelling,

$$\phi(\mathbf{w}) = \prod_{n=1}^{N} \phi_n(\mathbf{w}^\mathsf{T}\mathbf{h}_n) \tag{1.6}$$

for local site functions ϕ_n and fixed vectors $\mathbf{h}_n$. In this case, a local bound is applied to each site factor ϕ_n. Since the local bounds are all positive, this gives a bound dependent on the set of variational parameters

$$\log Z \geq \mathcal{B}(\xi_1, \ldots, \xi_N)$$

The problem of integrating over $\mathbf{w}$ has thus been approximated by the requirement to optimise the bound with respect to the vector $\boldsymbol{\xi}$ of variational parameters. The formal complexity of evaluating the local bound equation (1.5) requires computing $\det(\mathbf{A})$ and thus in general scales $O\left(D^3\right)$. Furthermore, optimising the local bound with respect to $\boldsymbol{\xi}$ requires solving a $D \times D$ linear system N times (Jaakkola and Jordan, 2000; Girolami, 2001).

Such computations are prohibitively expensive when $D \gg 1$, and scalable solvers have recently been developed to address this (Seeger, 2009; Nickisch and Seeger, 2009). The reduced computational burden of these methods is principally derived from making two relaxations. Firstly, double loop algorithms are used that, by decoupling the variational bound, reduce the number of times that the expensive $\log\det(\mathbf{A})$ term and its gradient have to be computed. Further savings are obtained by employing low rank approximate factorisations of $\mathbf{A}$. The computational demand of this procedure shifts then from evaluating the gradient of $\log\det(\mathbf{A})$ to calculating the low K-rank approximate factorisation. This may be achieved using Lanczos codes whose complexity scale super-linearly in K. The overall resulting complexity of this 'relaxed' local method is both problem and user dependent although, roughly speaking, the dimensionality

contributes $O\left(D^2\right)$ to the complexity of each update. We refer the reader to Nickisch and Seeger (2009) for a detailed discussion. An unfortunate aspect of this approximate decomposition is that it does not retain a bound on Z and indeed the quality of the resulting approximation to Z can be poor.

2 VARIATIONAL GAUSSIAN APPROXIMATION

An alternative to the local bounding method is to fit a Gaussian $q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}, \mathbf{S})$ based on minimising $\mathrm{KL}(q(\mathbf{w})|p(\mathbf{w}|\mathcal{D}))$. Due to the non-negativity of the KL divergence, we immediately obtain the bound $\log Z \geq \mathcal{B}_{KL}(\mathbf{m}, \mathbf{S})$, where

$$\mathcal{B}_{KL}(\mathbf{m}, \mathbf{S}) \equiv -\langle \log q(\mathbf{w}) \rangle + \langle \log \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \rangle + \langle \log \phi(\mathbf{w}) \rangle \quad (2.1)$$

and $\langle \cdot \rangle$ denotes expectation with respect to $q(\mathbf{w})$. This VG bound holds for any ϕ, compared with the local method which requires ϕ to be super-Gaussian. An important question, however, is whether the local bound equation (1.5) or the VG bound equation (2.1) is tighter and, furthermore, how these bounds are related. The VG bound has been noted before both empirically in the case of logistic regression (Nickisch and Rasmussen, 2008) and analytically for the special case of symmetric potentials (Seeger, 2009) to be tighter than the local bound. It is also tempting to presume that the VG bound is to be expected to be tighter due to the potentially unrestricted covariance $\mathbf{S}$. In this, however, one needs to bear in mind that for local site functions ϕ_n, $n = 1, \ldots, N$, provided $N > \frac{1}{2}D(D+2)$, the number of variational parameters in the local method actually exceeds the number of parameters in the VG method, and such intuition breaks down.

2.1 Relating the local and VG bounds

We derive a relationship between the local and VG bounds based on a generic super-Gaussian function $\phi(\mathbf{w})$. We first use the local bound on $\phi(\mathbf{w})$, equation (1.2), in equation (2.1) to obtain a new bound

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

where

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

Using equation (1.4) this can be written as

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

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

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

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 $\tilde{\mathcal{B}}_{KL}$ disappears and $\mathbf{m}$ and $\mathbf{S}$ are given by

$$\mathbf{S}_\xi = \left(\boldsymbol{\Sigma}^{-1} + \mathbf{F}(\xi)\right)^{-1}, \ \mathbf{m}_\xi = \mathbf{S}_\xi\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \mathbf{f}(\xi)\right) \quad (2.2)$$

Since $\mathcal{B}(\mathbf{m}, \mathbf{S}) \geq \tilde{\mathcal{B}}(\mathbf{m}, \mathbf{S}, \xi)$ we have that,

$$\mathcal{B}_{KL}(\mathbf{m}_\xi, \mathbf{S}_\xi) \geq \tilde{\mathcal{B}}_{KL}(\mathbf{m}_\xi, \mathbf{S}_\xi, \xi) = \mathcal{B}(\xi) \quad (2.3)$$

Importantly, the VG bound can be tightened beyond this setting:

$$\max_{\mathbf{m},\mathbf{S}} \mathcal{B}_{KL}(\mathbf{m}, \mathbf{S}) \geq \mathcal{B}_{KL}(\mathbf{m}_\xi, \mathbf{S}_\xi) \quad (2.4)$$

Thus optimal VG bounds are provably tighter than both the local variational bound and the VG bound calculated using the optimal local moments $\mathbf{m}_\xi$ and $\mathbf{S}_\xi$. The experiments in section(5) show that the improvement in VG bound values can be significant. Furthermore, constrained parametrisations of covariance, which are required when $D \gg 1$, are also frequently observed to outperform local variational solutions.

2.2 Tractable VG Approximations

The bound equation (2.1) assumes we can compute $\langle \log \phi(\mathbf{w}) \rangle$. For generic functions ϕ, this may not be practical. However, for the product of site projections form, equation (1.6), each projection $\mathbf{w}^\mathsf{T}\mathbf{h}_n$ is Gaussian distributed and

$$I \equiv \langle \log \phi(\mathbf{w}) \rangle = \sum_n \langle \log \phi_n(\mu_n + z\sigma_n) \rangle_{\mathcal{N}(z|0,1)} \quad (2.5)$$

with $\mu_n \equiv \mathbf{m}^\mathsf{T}\mathbf{h}_n$ and $\sigma_n^2 \equiv \mathbf{h}_n^\mathsf{T}\mathbf{S}\mathbf{h}_n$. Hence I can be readily computed either analytically (for example for $\phi(x) \propto e^{-|x|}$) or more generally using one-dimensional numerical integration (Barber and Bishop, 1998; Kuss and Rasmussen, 2005).

A further point of interest is that for local sites, equation (1.6), by differentiating the VG bound with respect to $\mathbf{S}$ and equating to zero, the optimal form for the covariance satisfies

$$\mathbf{S}^{-1} = \boldsymbol{\Sigma}^{-1} + \mathbf{H}\boldsymbol{\Gamma}\mathbf{H}^\mathsf{T} \quad (2.6)$$

where $\boldsymbol{\Gamma}$ is diagonal such that

$$\Gamma_{nn} = \left\langle \frac{z\phi_n'(\mu_n + z\sigma_n)}{2\sigma_n\phi_n(\mu_n + z\sigma_n)} \right\rangle \quad (2.7)$$

and $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_N]$. Here $\mathbf{\Gamma}$ is dependent on $\mathbf{S}$ through the projected variance terms σ_n and we do not have a closed expression or fixed point procedure for optimising the bound. Unfortunately, optimising the bound directly with respect to Γ_{nn} is infeasible due the computational cost of storing and inverting $\mathbf{S}$ when D and $N \gg 1$. Furthermore, $\mathbf{S}$ parametrised in the form of equation (2.6) renders the bound non-concave in Γ_{nn}. We shall return to the issue of scalable alternative parametrisations of $\mathbf{S}$ in section(4).

3 VG BOUND CONCAVITY

An attractive property of local variational methods is that they have been proved to be convex problems when ϕ is log-concave (Seeger, 2009). Here we show that this important property is not restricted to local variational methods. For log-concave potentials $\phi(\mathbf{w})$ of the form equation (1.6), the VG bound $\mathcal{B}_{KL}(\mathbf{m}, \mathbf{S})$, equation (2.1), is jointly concave with respect to the variational Gaussian parameters $\mathbf{m}$ and $\mathbf{S}$.

In order to show this we parametrise the covariance $\mathbf{S}$ of the variational Gaussian distribution $q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}, \mathbf{S})$ using the Cholesky decomposition $\mathbf{S} = \mathbf{C}\mathbf{C}^\mathsf{T}$ where $\mathbf{C}$ is a square lower triangular matrix of dimension D. Since the bound depends on the logarithm of ϕ, without loss of generality we may take $N = 1$, and on ignoring constants with respect to $\mathbf{m}$ and $\mathbf{S}$, we have that

$$\mathcal{B}_{KL}(\mathbf{m}, \mathbf{C}) \overset{c.}{=} \sum_i \log C_{ii} - \frac{1}{2}\mathbf{m}^\mathsf{T}\mathbf{\Sigma}^{-1}\mathbf{m} + \boldsymbol{\mu}^\mathsf{T}\mathbf{\Sigma}^{-1}\mathbf{m} - \frac{1}{2}\mathrm{trace}\left(\mathbf{\Sigma}^{-1}\mathbf{C}\mathbf{C}^\mathsf{T}\right) + \left\langle \log \phi(\mathbf{w}^\mathsf{T}\mathbf{h}) \right\rangle \quad (3.1)$$

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, these terms represent a jointly concave contribution. To complete the proof we therefore need to show that $\left\langle \log \phi(\mathbf{w}^\mathsf{T}\mathbf{h}) \right\rangle$ is jointly concave in $\mathbf{m}$ and $\mathbf{C}$. We first transform variables to write $\left\langle \log \phi(\mathbf{w}^\mathsf{T}\mathbf{h}) \right\rangle$ as

$$\langle \log \phi(a) \rangle_{\mathcal{N}(a|\mathbf{m}^\mathsf{T}\mathbf{h}, \mathbf{h}^\mathsf{T}\mathbf{S}\mathbf{h})} = \langle \psi(\mu(\mathbf{m}) + z\sigma(\mathbf{C})) \rangle_z \quad (3.2)$$

where $\langle \cdot \rangle_z$ refers to taking the expectation with respect to the standard normal $\mathcal{N}(z|0, 1)$ and,

$$\mu(\mathbf{m}) \equiv \mathbf{m}^\mathsf{T}\mathbf{h}, \quad \sigma(\mathbf{C}) \equiv \sqrt{\mathbf{h}^\mathsf{T}\mathbf{C}\mathbf{C}^\mathsf{T}\mathbf{h}}, \quad \psi \equiv \log \phi$$

Note that establishing the concavity of equation (3.2) is non-trivial since the function $\psi(\mu(\mathbf{m}) + z\sigma(\mathbf{C}))$ is itself *not* jointly concave in $\mathbf{C}$ and $\mathbf{m}$.

For ease of notation we let $\sigma' \equiv \mathrm{vec}\left(\frac{\partial \sigma(\mathbf{C})}{\partial \mathbf{C}}\right)$, where $\mathrm{vec}(\mathbf{X})$ is the vector obtained by concatenating the columns of $\mathbf{X}$, with dimension D^2; $\sigma'' \equiv \frac{\partial^2 \sigma(\mathbf{C})}{\partial \mathbf{C}^2}$ is the Hessian of σ with respect to $\mathbf{C}$ with dimension $D^2 \times D^2$; $\mu' \equiv \frac{\partial \mu(\mathbf{m})}{\partial \mathbf{m}}$ is a column vector with dimension D. Then the Hessian of ψ with respect to $\mathbf{m}$ and $\mathbf{C}$ can be expressed in the following block matrix form

$$H[\psi] = \begin{bmatrix} \frac{\partial^2 \psi}{\partial \mathbf{C}^2} & \frac{\partial^2 \psi}{\partial \mathbf{C} \partial \mathbf{m}} \\ \frac{\partial^2 \psi}{\partial \mathbf{m} \partial \mathbf{C}} & \frac{\partial^2 \psi}{\partial \mathbf{m}^2} \end{bmatrix} = \begin{bmatrix} \psi'' z^2 \sigma' \sigma'^\mathsf{T} + \psi' z \sigma'' & \psi'' z \sigma' \mu'^\mathsf{T} \\ \psi'' z \mu' \sigma'^\mathsf{T} & \psi'' \mu' \mu'^\mathsf{T} \end{bmatrix}$$

The Hessian of $\langle \psi(\mu(\mathbf{m}) + z\sigma(\mathbf{C})) \rangle_z$ is equivalent to $\langle H[\psi(\mu(\mathbf{m}) + z\sigma(\mathbf{C}))] \rangle_z$, which we now show to be negative semi-definite. Since the expectation in $\langle H[\psi(\mu(\mathbf{m}) + z\sigma(\mathbf{C}))] \rangle_z$ is with respect to an even Gaussian density function, provided that for all $\gamma \geq 0$, the combined Hessian is negative definite, *i.e.*

$$H_{z=-\gamma}[\psi] + H_{z=+\gamma}[\psi] \preceq 0 \quad (3.3)$$

then the expectation of $H[g]$ with respect to z is negative definite. To show this we first note that for all $\mathbf{u} \in \mathbb{R}^{D^2}$ and $\mathbf{v} \in \mathbb{R}^D$

$$\begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix}^\mathsf{T} H[\psi] \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix} = \psi'' \left[\mathbf{v}^\mathsf{T}\mu' + z\mathbf{u}^\mathsf{T}\sigma'\right]^2 + \psi' z \mathbf{u}^\mathsf{T}\sigma''\mathbf{u}$$

The first term of the right hand side is negative for all values of z since $\psi''(x) \leq 0$.

To show that equation (3.3) is satisfied it is sufficient to show that

$$\left(\psi'(\mu + \gamma\sigma) - \psi'(\mu - \gamma\sigma)\right) \gamma \mathbf{u}^\mathsf{T}\sigma''\mathbf{u} \leq 0$$

which is true since $\sigma'' \succeq 0$, $\sigma(\mathbf{C}) \geq 0$ and because $\psi'(x)$ is a decreasing function from the assumed log-concavity of ϕ.

To see that $\sigma'' \succeq 0$ we write, $\sigma^2(\mathbf{C}) = \sum_j g_j^2(\mathbf{C})$ where $g_j(\mathbf{C}) = |\sum_i h_i C_{ij}|$ is convex and non-negative for all j. For convex and non-negative functions g_j and $p > 1$, then $\left(\sum_{j=1}^W g_j(x)^p\right)^{1/p}$ is convex (Boyd and Vandenberghe, 2004), which reveals that $\sigma(\mathbf{C})$ is convex on setting $p = 2$.

The supplementary material contains a simpler proof kindly provided to us after the conference by M. K. Titsias.

4 VG BOUND OPTIMISATION

Whilst highly desirable, concavity of the VG bound with respect to $\mathbf{m}$ and $\mathbf{C}$ does not in itself guarantee that optimisation is scalable. Thus an important practical consideration is the numerical complexity of a simple gradient based procedure.

	$\mathbf{C}_{full}$	$\mathbf{C}_{band}$	$\mathbf{C}_{chev}$	$\mathbf{C}_{sub}$	$\mathbf{C}_{sub\&band}$
Comp.	$O\left(ND^2\right)$	$O\left(NDB\right)$	$O\left(NDK\right)$	$O\left(NK^2\right)$	$O\left(NKB\right)$

Table 1: Time complexity to evaluate the VG bound and its gradient for each covariance parametrisation.

Answering this question in full generality is complex and we therefore restrict ourselves to considering the common case in which $\mathbf{\Sigma} = s^2\mathbf{I}$ and $\phi(\mathbf{w})$ factorises such that $\phi(\mathbf{w}) = \prod_{n=1}^{N}\phi(\mathbf{h}_n^\mathsf{T}\mathbf{w})$. Note that many popular models are in this class, such as inverse models with isotropic Gaussian observation noise, and generalised linear models with isotropic Gaussian priors.

For Cholesky factorisations $\mathbf{C}$ of dimension $D \times D$ the computational bottleneck in computing the VG bound arises from the projected variational variances $\sigma_n^2 = ||\mathbf{C}^\mathsf{T}\mathbf{h}_n||^2$ required in the likelihood term, equation (2.5). Computing all such terms is $O\left(ND^2\right)$. Whilst this appears prohibitive, this complexity compares favourably with the local method in which a single $\boldsymbol{\xi}$ update involves solving N linear $D \times D$ systems, thus scaling $O\left(ND^3\right)$. Note also that in many problems of interest, the $\mathbf{h}_n$ are sparse, in which case D should be taken as the number of non-zero elements, often $\ll D$. Nevertheless, $O\left(D^2\right)$ scaling for the VG method can be expensive for very large problems. For such cases, reduced parametrisations of the covariance $\mathbf{S}$ are required, as presented below. These reduced parametrisations increase the problem size to which the VG procedure can be applied.

Factor Analysis parametrisations of the form $\mathbf{S} = \mathbf{\Theta}\mathbf{\Theta}^\mathsf{T} + \text{diag}\left(\mathbf{d}^2\right)$ can capture the K leading directions of variance for a $D \times K$ dimensional loading matrix $\mathbf{\Theta}$. Unfortunately, however, this parametrisation is not of the square Cholesky form in section(3) and indeed renders the VG bound non-concave. Provided one is happy to accept convergence to possibly local optima, this is still a useful parametrisation.

4.1 Reduced Concave Parametrisations

Whilst full rank Cholesky parametrisations are optimal in terms of achieving the tightest possible VG bound, for $D \gg 1$ storing and evaluating the gradient with respect to $\mathbf{C}$ is prohibitive. Below we consider constrained parametrisations which reduce both the space and time complexity, whilst preserving concavity of the bound.

Banded Cholesky. The simplest option is to constrain the Cholesky matrix to be banded, that is $C_{ij} = 0$ for $i > j+B$ where B is the bandwidth. Doing so reduces the cost of a single bound/gradient computation to $O\left(NDB\right)$, see table(1). Such a parametrisation however assumes zero covariance between variables that are indexed out of bandwidth.

Chevron Cholesky. We constrain $\mathbf{C}$ such that $C_{ij} = \Theta_{ij}$ when $i \geq j$ and $j \leq K$, $C_{ii} = d_i$ for $i > K$ and 0 otherwise. Importantly, this reduced parametrisation does not exclude modelling any individual covariates whilst conserving concavity of the bound. For a Cholesky matrix of this form bound/gradient computations scale $O\left(NDK\right)$.

Subspace Cholesky. Another reduced parametrisation of the covariance can be obtained by considering arbitrary rotations of the covariance, $\mathbf{S} = \mathbf{E}\mathbf{C}\mathbf{C}^\mathsf{T}\mathbf{E}^\mathsf{T}$ where $\mathbf{E}$ forms an orthonormal basis over $\mathbb{R}^D$. Substituting this form of the covariance in equation (3.1) and for $\mathbf{\Sigma} = s^2\mathbf{I}$ we obtain, up to a constant,

$$\mathcal{B}_{KL}(\mathbf{m}, \mathbf{C}) \overset{c.}{=} \sum_i \log C_{ii} - \frac{1}{2s^2}\left[||\mathbf{C}||^2 + ||\mathbf{m}||^2\right] + \frac{1}{s^2}\boldsymbol{\mu}^\mathsf{T}\mathbf{m} + \sum_n \left\langle \log\phi(\mu_n + z\sigma_n)\right\rangle_z \quad (4.1)$$

where $\sigma_n = ||\mathbf{C}^\mathsf{T}\mathbf{E}^\mathsf{T}\mathbf{h}_n||$. One may reduce the computational burden by decomposing the square orthonormal matrix $\mathbf{E}$ into two orthonormal matrices 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 $L = (D-K)$. Then for $\mathbf{C} = \text{blkdiag}\left(\mathbf{C}_1, c\mathbf{I}_{L\times L}\right)$, with $\mathbf{C}_1$ a $K \times K$ Cholesky matrix,

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

meaning that only the K eigenvectors in $\mathbf{E}_1$ need to be approximated. This effect is due to the assumed isotropy of $\mathbf{\Sigma}$, which can be generalised for non-isotropic $\mathbf{\Sigma}$ by using a coordinate transformation for which in the new system, the prior covariance is rendered isotropic. Since terms such as $||\mathbf{h}_n||$ need only be computed once the complexity of bound and gradient computations scales linearly with K, not D. Further savings can be made if we use only banded Cholesky matrices: for $\mathbf{C}_1$ having bandwidth B each bound evaluation and associated gradient computation scales $O\left(NBK\right)$.

The success of this factorisation depends on how well $\mathbf{E}_1$ captures the leading directions of variance. One simple approach is to use the leading Principal Components of the 'dataset' $\mathbf{H}$. Another option is to use a two stage procedure in which we first assume $\mathbf{C}$ is low bandwidth. After optimisation of the VG bound *w.r.t.* $\mathbf{m}$ and $\mathbf{C}$ we then obtain an estimate of the projections μ_n and σ_n. These can be used to approximate the diagonal matrix $\mathbf{\Gamma}$ in equation (2.7). We then seek a rank K approximation to this $\mathbf{S}$. The best rank K approximation is given by evaluating the smallest K eigenvectors of $\mathbf{\Sigma}^{-1} + \mathbf{H}\mathbf{\Gamma}\mathbf{H}^\mathsf{T}$. For very large sparse problems $D \gg 1$ we use iterative Lanczos methods to approximate this, using the methods described in

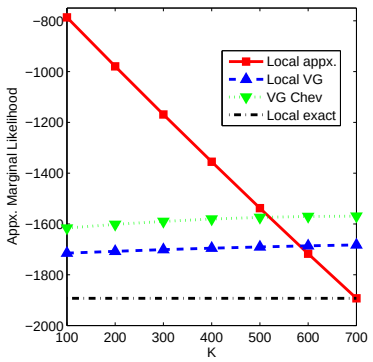

Figure 1: Bound values vs. K the 'size' of covariance parametrisation: approximate local bound values for a rank K approximation (red); VG bound calculated using the implied local variational moments using equation (2.2) (blue); VG bound optimised with Chevron parametrised covariance of width K (green); the exact local variational bound, independent of K, is plotted for comparison (black). Results obtained on a synthetic binary logistic regression problem: $D = 700$, $N = 4000$.

Seeger and Nickisch (2010). For smaller non-sparse problems more accurate approximations are available – see the supplementary material.

5 EXPERIMENTS

To make more concrete the results presented above and as a numerical validation of them we apply both local and VG bounding techniques to two common Bayesian models: binary logistic regression and inverse modelling. Local bound results were obtained using the publicly available `glm-ie` code[2].

5.1 Bayesian Logistic Regression

Given a dataset, $\mathcal{D} = \{(s_n, \mathbf{x}_n), n = 1, \ldots, N\}$ with each class $s_n \in \{-1, 1\}$ and D-dimensional input vector $\mathbf{x}_n$, Bayesian logistic regression models the class probability as $p(c = 1|\mathbf{w}, \mathbf{x}) = f\left(\mathbf{w}^\mathsf{T}\mathbf{x}\right)$, with $f(x) \equiv 1/(1 + e^{-x})$. Under a Gaussian prior, $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \boldsymbol{\Sigma})$, the posterior is given by

$$p(\mathbf{w}|\mathcal{D}) \propto p(\mathbf{w}) \prod_n f\left(s_n \mathbf{w}^\mathsf{T}\mathbf{x}_n\right)$$

This is of the form equation (1.1) and equation (1.6) under log-concave sites $\phi_n(x) \equiv f(x)$ and $\mathbf{h}_n \equiv s_n\mathbf{x}_n$.

VG Bound. Following the procedure outlined in section(2), we obtain a bound on the log marginal likelihood of the form

$$2\mathcal{B}_{KL}(\mathbf{m}, \mathbf{S}) = \log\det\left(\mathbf{S}\boldsymbol{\Sigma}^{-1}\right) + D - \mathrm{trace}\left(\boldsymbol{\Sigma}^{-1}\mathbf{S}\right) - \mathbf{m}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\mathbf{m} + 2\sum_n I_n \quad (5.1)$$

where $I_n \equiv \langle \log f(\mu_n + z\sigma_n)\rangle_z$; $\langle\cdot\rangle_z$ is the expectation with respect to the standard normal, $\mu_n = s_n\mathbf{x}_n^\top\mathbf{m}$ and $\sigma_n^2 = \mathbf{x}_n^\top\mathbf{S}\mathbf{x}_n$. I_n can be computed by any standard one-dimensional numerical integration method – for the results presented, quadrature was used. The gradients for this model are provided in the supplementary material.

Local Bound. Following Jaakkola and Jordan (1996), the logistic function is bounded by

$$f(x) \geq f(\xi)\left[\frac{1}{2}(x - \xi) - \lambda(\xi)\left(x^2 - \xi^2\right)\right] \quad (5.2)$$

where $\lambda(\xi) \equiv \frac{1}{2\xi}\left(f(\xi) - \frac{1}{2}\right)$. Integrating over $\mathbf{w}$ gives

$$\log p(\mathcal{D}) \geq \frac{1}{2}\log\det\left(\mathbf{S}\boldsymbol{\Sigma}^{-1}\right) + \frac{1}{2}\mathbf{m}^\mathsf{T}\mathbf{S}^{-1}\mathbf{m} + \sum_{n=1}^{N}\left[\log f(\xi_n) - \frac{\xi_n}{2} + \lambda(\xi_n)\xi_n^2\right] \quad (5.3)$$

for $\mathbf{m} = \mathbf{S}\sum_{n=1}^N \frac{1}{2}s_n\mathbf{x}_n$, and covariance $\mathbf{S}^{-1} = \boldsymbol{\Sigma}^{-1} + 2\sum_{n=1}^N \lambda(\xi_n)\mathbf{x}_n\mathbf{x}_n^\mathsf{T}$. The bound (5.3), can be optimised *w.r.t.* ξ_n using the variational update $\xi_n^2 = \mathbf{x}_n^\mathsf{T}\left(\mathbf{S} + \mathbf{m}\mathbf{m}^\mathsf{T}\right)\mathbf{x}_n$.

5.1.1 Approximate Bound Comparison

It is important to note that VG methods always return a lower bound to the marginal likelihood, and that computing the bound at most scales $O\left(ND^2\right)$ with reduced parametrisations scaling according to table(1).

In contrast, exact evaluation of the local variational lower bound scales $O\left(D^3\right)$. As discussed in section(1.1), fast scalable approximate solvers that are required when $D \gg 1$ utilise low rank approximations to $\mathbf{A}$ and give only an approximation to $\log Z$.

Figure 1 plots the approximate and exact local variational bound versus the VG bound using a Chevron parametrised covariance matrix for a synthetic binary logistic regression problem. Since the dimensionality of this problem is sufficiently small, the exact local bound and the VG bound implied by the optimal local variational parameters can also be computed. The figure shows that the VG bound improves as more parameters are used to specify the covariance. However, for $K \ll D$ the approximate local bound value can significantly overestimate the exact local bound value. Principally this is due to Lanczos codes ignoring the centre of the eigen-spectrum of $\mathbf{A}$ and thus underestimating the magnitude of the $\log\det(\mathbf{A})$ term. Thus,

[2] `http://mloss.org/software/view/269`

	a9a				realsim				rcv1			
	VG Full	VG Chev	VG Sub	Local	VG diag	VG Chev	VG Sub	Local	VG diag	VG Chev	VG Sub	Local
K	–	80	80	80	–	100	750	750	–	50	750	750
Bound	$-5,374$	$-5,375$	$-5,379$	$-5,383$	$-5,564$	$-5,551$	$-5,723$	–	$-6,981$	$-6,979$	$-7,286$	–
CPU(s)	85	91	68	5	180	350	575	583	176	424	955	436
Acc. %	15.12	15.10	15.12	15.10	2.86	2.86	2.86	2.87	2.90	2.89	2.94	2.94

Table 2: Approximate local and VG results for the `a9a`, `realsm` and `rcv1` binary classification tasks. Bound values in all cases are evaluated using the VG form. K refers to the 'rank' of the approximation: VG Chev $\mathbf{\Theta}$ is $D \times K$; VG Sub parametrisation uses K dimensional subspace and a diagonal Cholesky matrix; approximate local method with K Lanczos vectors. VG diag refers to a bandwidth 1 Cholesky matrix. CPU times were recorded in MATLAB using an Intel 2.5Ghz Core 2 Quad processor.

in problems of sufficiently large dimensionality, where necessarily $K \ll D$, local approximate bound values cannot be relied upon to assess model fidelity. In contrast the VG procedure offers an exact lower bound on the marginal likelihood that scales according to table(1).

5.1.2 Large Scale Numerical Results

To demonstrate the scalability of the VG method we compare the performance against fast local methods on three large scale binary classification tasks[3], `a9a`, `realsim` and `rcv1`.

Training (tr) and test sets (tst) were randomly partitioned such that: `a9a` $D = 123$, $tr = 16,000$, $tst = 16,561$ with the number of non zero elements (nnz) totalling $nnz = 451,592$; `realsim` $D = 20,958$, $tr = 36,000$, $tst = 36,309$ and $nnz = 3,709,083$; `rcv1` $D = 42,736$, $tr = 50,000$, $tst = 50,000$ and $nnz = 7,349,450$.

Model parameters and local optimisation procedures were, for the purposes of comparison, fixed to the values stated in Nickisch and Seeger (2009): τ, a scaling on the likelihood term $p(c_n|\mathbf{x}_n) = f(\tau \mathbf{w}^\mathsf{T}\mathbf{x}_n)$, was set to 1 in the `a9a` dataset and $\tau = 3$ for `realsim` and `rcv1`; the prior variance was fixed $\mathbf{\Sigma} = s^2\mathbf{I}$ with $s^2 = 1$. The VG bound was optimised using a conjugate gradients procedure.

Results for various concave parametrisations of covariance are presented in table(2). Local VG bound values are not presented for the larger problems since evaluating $\mathbf{m}$ and $\mathbf{S}$ using equation (2.2) and the optimal local variational parameters is not computationally feasible, whilst the approximate local bound values are too inaccurate to make meaningful comparisons (see fig(1)). The local bound value for the `a9a` data set was obtained by translating the returned optimal local parameter $\boldsymbol{\xi}$ to a VG Gaussian using equation (2.2). This VG Gaussian is guaranteed to provide a higher bound than the local method itself.

Whilst the local method is significantly faster for the small D problem `a9a` (albeit with a worse bound) than the VG method, this is due only to different overheads in the corresponding implementations. In the larger problems, the results show that a simple VG gradient based optimisation procedure, coupled with constrained parametrisations of the covariance, can achieve results in terms of speed and accuracy on a par with or in excess of fast local solvers.

5.2 Bayesian Inverse Modelling

Bayesian inverse modelling assumes an observed real vector $\mathbf{y} \in \mathbb{R}^N$ is drawn from the generative model $\mathbf{y} = \mathbf{M}\mathbf{w} + \boldsymbol{\eta}$ for $\mathbf{w} \in \mathbb{R}^D$, for a model matrix $\mathbf{M}$ of dimension $N \times D$ with $N \ll D$ and additive spherical Gaussian noise $\boldsymbol{\eta} \sim \mathcal{N}\left(\mathbf{0}, s^2\mathbf{I}\right)$. Sparsity can be imposed on the inferred object by using a prior of the form $\phi(\mathbf{w}) \equiv \prod_i p(w_i)$, where each $p(w_i)$ is super-Gaussian. Given $\mathbf{y}$, the posterior is then of the form

$$p(\mathbf{w}|\mathbf{y}) = \frac{1}{Z}\mathcal{N}\left(\mathbf{y}|\mathbf{M}\mathbf{w}, s^2\mathbf{I}\right)\phi(\mathbf{w}) \tag{5.4}$$

Due to the symmetry $\mathcal{N}(\mathbf{w}|\boldsymbol{\mu}, \mathbf{\Sigma}) \equiv \mathcal{N}(\boldsymbol{\mu}|\mathbf{w}, \mathbf{\Sigma})$, and since $\mathbf{M}$ is a linear operator, then provided that $\phi(\mathbf{w})$ is log-concave, the arguments of section(3) apply. For Laplace sparsity priors $\phi(\mathbf{w}) = \prod_i \frac{1}{2\tau_i} e^{-|w_i|/\tau_i}$ both the VG and local bounds are concave.

VG bound. Following the procedure outlined in section(2) we obtain the following bound on the marginal likelihood,

$$\begin{aligned} 2\mathcal{B}_{KL}(\mathbf{m}, \mathbf{S}) \equiv -N\log(2\pi s^2) + 2\sum_{i=1}^{D}\langle \log p(w_i)\rangle \\ - \frac{1}{s^2}\left[||\mathbf{y}||^2 - 2\mathbf{y}^\mathsf{T}\mathbf{M}\mathbf{m} + ||\mathbf{C}^\mathsf{T}\mathbf{M}||^2 + ||\mathbf{M}\mathbf{m}||^2\right] \\ + 2\log\det(2\pi\mathbf{C}) + D \end{aligned} \tag{5.5}$$

Evaluating the integral $\langle \log p(\mathbf{w})\rangle_{q(\mathbf{w})}$ in this case is analytic, the precise form of which, including variational gradients, is presented in the supplementary

[3] `www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets`

material. Inspecting the form of equation (5.5), it is important to point out that the computational bottleneck is equivalent to that for the generalised linear model case discussed previously, namely computing $||\mathbf{C}^\mathsf{T}\mathbf{m}_n||^2$ where $\mathbf{m}_n$ is the n^{th} column of $\mathbf{M}$.

Local bound. As originally presented by Girolami (2001), a lower bound on $\log Z$ can be obtained by bounding each Laplace site using,

$$e^{-|x|} \geq \lambda(\xi)\mathcal{N}\left(x|0,|\xi|\right), \text{where } \lambda(\xi) \equiv \sqrt{2\pi|\xi|}e^{-\frac{1}{2}|\xi|}$$

Bounding the prior terms with this form gives,

$$\log p(\mathbf{y}|\mathbf{M},s^2) \geq \sum_{i=1}^{D}\log\lambda(\xi_i) - \frac{N}{2}\log(2\pi) - \frac{1}{2}\log\det\left(\mathbf{A}\right) - \frac{1}{2}\mathbf{y}^\mathsf{T}\mathbf{A}^{-1}\mathbf{y} \quad (5.6)$$

where $\mathbf{A} = \mathbf{M}\mathbf{\Gamma}\mathbf{M}^\mathsf{T} + s^2\mathbf{I}$, $\mathbf{\Gamma} = \text{diag}\left(|\xi_1|,...|\xi_D|\right)$.

5.2.1 Numerical Comparison

As a numerical comparison of the local and VG approximate methods for the inverse modelling problem we simulate a hyperparameter selection task, assuming identical priors $\tau_i = \tau$, in which the true value of the hyperparameter τ used to generate the data is known.

The data vector $\mathbf{y}$ was sampled according to the generative model with parameters set such that $\boldsymbol{\tau}_{true} = 0.05 \times \mathbf{1}$ and $s^2 = 10^{-3}$. The model matrix $\mathbf{M}$ has dimension $N \times D$ where $N = 100$ and $D = 200$, with each element generated by sampling $U[-1,1]$. VG bound values are presented in fig(2). Local results were obtained using using both the exact and approximate optimisation procedures with $K = 50$ Lanczos vectors. VG results are presented for a full Cholesky and a constrained Chevron parametrisation with $K = 50$.

Whilst the fully optimal VG parametrisation is guaranteed to outperform the local variational bound, the results presented in fig(2) show that the *constrained* VG parametrisation can also outperform both the exact and approximated local solutions.

As the results in section(5.1.2) testify, the optimal reduced concave covariance parametrisation (Chevron versus Banded Cholesky *etc.*) is problem dependent. Characterising the effects of covariance parametrisation, including those implied by local approximate optimisation procedures, is an important topic for future research. However, until such issues are better understood, the VG bound to the marginal likelihood provides a practical means by which to assess the accuracy of different covariance parametrisations in large systems.

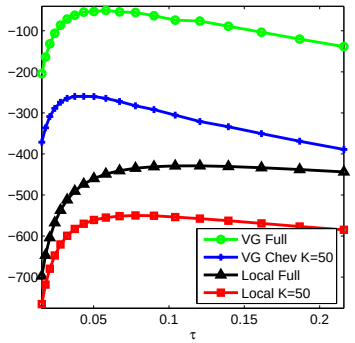

Figure 2: Hyperparameter selection for a sparse linear model. Bound values are calculated using the VG form with posterior moments obtained under each approximation $\boldsymbol{\tau}_{true} = 0.05$. Inferred optimal hyperparameter values under each approximate method are: VG full $\tau^* = 0.058$, VG Chev $\tau^* = 0.043$, local full $\tau^* = 0.12$ and local approx $\tau^* = 0.043$.

6 DISCUSSION

We have presented several novel theoretical and practical developments regarding the application of variational Gaussian KL approximations:

For posteriors of the form in equation (1.1) optimal variational Gaussian bounds are always tighter than local bounds that use squared exponential site bounds.

An important practical issue in finding the optimum of the variational Gaussian bound is its concavity. We have proved that the VG bound is concave in terms of the mean and covariance of the approximating Gaussian.

To enhance scalability and optimisation we have presented constrained covariance parametrisations which retain concavity of the bound. An important practical point is the ease with which VG methods can be implemented; off the shelf gradient based optimisers and constrained concave parametrisations of covariance can achieve fast and scalable approximate inference for a large class of Bayesian generalised linear models.

These observations on the variational Gaussian bound make it, to our minds, an attractive alternative to the more recent focus on local bounding methods.

Code is available at `mloss.org/software/view/308`

Acknowledgements

We are grateful to Peter Sollich and the referees of a previous version of the paper for their technical insights.

References

D. Barber. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 2011.

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.

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.

J. Hardin and J. Hilbe. *Generalized Linear Models and Extensions*. Stata Press, 2007.

T. Jaakkola and M. Jordan. A variational approach to Bayesian logistic regression problems and their extensions. In *Artificial Intelligence and Statistics*, 1996.

T. Jaakkola and M. Jordan. Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1):25–37, 2000.

M. Kuss and C. Rasmussen. Assessing Approximate Inference for Binary Gaussian Process Classification. *Journal of Machine Learning Research*, 6: 1679–1704, 2005.

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. *International Conference on Machine Learning*, 26:761–768, 2009.

M. Opper and C. Archambeau. The Variational Gaussian Approximation Revisited. *Neural Computation*, 21(3):786–792, 2009.

A. Palmer, D. Wipf, K. Kreutz-Delgado, and B. Rao. Variational EM algorithms for non-Gaussian latent variable models. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems (NIPS)*, number 19, pages 1059–1066, Cambridge, MA, 2006. MIT Press.

M. Seeger. Bayesian Model Selection for Support Vector Machines, Gaussian Processes and other Kernel Classifiers. In S. Solla, T. Leen, and Müller, editors, *Advances in Neural Information Processing Systems (NIPS)*, number 12, pages 603–609, Cambridge, MA, 1999. MIT Press.

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

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, http://arxiv.org/abs/0810.0901, 2010.

D. Wipf and S. Nagarajan. A unified Bayesian framework for MEG/EEG source imaging. *NeuroImage*, 44(3):947–966, 2009.