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

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

0.1 Alternative Concavity Proof

We present an alternative derivation showing that for log-concave potentials $\phi(\mathbf{w})$ the Gaussian expectation $\langle \log \phi(\mathbf{w}) \rangle_{\mathcal{N}(\mathbf{w}|\mathbf{m},\mathbf{C}^\intercal\mathbf{C})}$ is concave *w.r.t.* $\mathbf{m}$ and $\mathbf{C}$. This derivation is due to Michalis K. Titsias at the University of Manchester.

Since $\log \phi(\mathbf{w})$ is log concave we have that for all $\theta \in [0,1]$ and for any $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^D$

$$\log \phi(\theta \mathbf{w}_1 + (1-\theta)\mathbf{w}_2) \geq \theta \log \phi(\mathbf{w}_1) + (1-\theta) \log \phi(\mathbf{w}_2). \tag{0.1}$$

To show that $E(\mathbf{m}, \mathbf{C}) = \langle \log \phi(\mathbf{w}) \rangle_{\mathcal{N}(\mathbf{w}|\mathbf{m},\mathbf{C}^\intercal\mathbf{C})}$ is concave it suffices to show that

$$E(\theta \mathbf{m}_1 + (1-\theta)\mathbf{m}_2, \theta \mathbf{C}_1 + (1-\theta)\mathbf{C}_2) \geq \theta E(\mathbf{m}_1, \mathbf{C}_1) + (1-\theta) E(\mathbf{m}_2, \mathbf{C}_2). \tag{0.2}$$

This can be done by making use of the substitution $\mathbf{w} = \theta \mathbf{m}_1 + (1-\theta)\mathbf{m}_2 + (\theta \mathbf{C}_1 + (1-\theta)\mathbf{C}_2)\,\mathbf{z}$. Which gives us the following expression

$$E(\theta \mathbf{m}_1 + (1-\theta)\mathbf{m}_2, \theta \mathbf{C}_1 + (1-\theta)\mathbf{C}_2) = \int \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I}) \times \log \phi\left(\theta(\mathbf{m}_1 + \mathbf{C}_1\mathbf{z}) + (1-\theta)(\mathbf{m}_2 + \mathbf{C}_2\mathbf{z})\right) d\mathbf{z}$$

Using then the concavity of $\log \phi(\mathbf{w})$ *w.r.t.* $\mathbf{w}$ and using equation (0.1) where $\mathbf{w}_1 = \mathbf{m}_1 + \mathbf{C}_1\mathbf{z}$ and $\mathbf{w}_2 = \mathbf{m}_2 + \mathbf{C}_2\mathbf{z}$ we have that

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

Where the last line is obtained by substituting back to the original coordinate system.

0.1.1 Subspace Covariance Decomposition

For generalized linear models with isotropic $\mathbf{\Sigma} = s^2\mathbf{I}$ then the form for the optimal inverse covariance, equation (2.6), simplifies to

$$\mathbf{S}^{-1} = \frac{1}{s^2}\mathbf{I} + \mathbf{H}\mathbf{\Gamma}\mathbf{H}^\intercal \tag{0.3}$$

Thus to approximate the K leading eigenvectors of $\mathbf{S}$ we wish to evaluate the K smallest eigenvectors of $\mathbf{\Sigma}^{-1} + \mathbf{H}\mathbf{\Gamma}\mathbf{H}^\intercal$. To do so we note that $\mathbf{H}\mathbf{\Gamma}\mathbf{H}^\intercal \approx \mathbf{H}\mathbf{\Gamma}'\mathbf{H}^\intercal$ where $\Gamma'_{nn} = \Gamma_{nn}$ if $\Gamma_{nn} > \delta$ and zero otherwise - we set δ such that there are K diagonal terms Γ_{nn} where $\Gamma_{nn} > \delta$. If we now calculate the eigen decomposition to $\mathbf{H}\mathbf{\Gamma}'\mathbf{H}^\intercal = \mathbf{E}\mathbf{\Lambda}\mathbf{E}^\intercal$ we see that

$$\left[\frac{1}{s^2}\mathbf{I} + \mathbf{H}\mathbf{\Gamma}'\mathbf{H}^\intercal\right]^{-1} = \mathbf{E}\mathrm{diag}\left(\frac{s^2}{1 + \lambda'_{nn}s^2}\right)\mathbf{E}^\intercal \tag{0.4}$$

For $L \ll D$ we can evaluate the L eigenvectors of $\mathbf{H}\mathbf{\Gamma}'\mathbf{H}^\intercal$ cheaply since the eigenvalues of $\mathbf{X}\mathbf{X}^\intercal$ coincide with the eigenvalues of $\mathbf{X}^\intercal\mathbf{X}$[1]. And so approximating K subspace eigen decomposition reduces to the complexity of decomposing a $K \times K$ matrix. If δ is small this method can often outperform approximate iterative decompositions provided the data is non-sparse and of moderate dimensionality.

0.2 VG Bound Gradients

0.2.1 Analytic forms for Laplace Priors

The expectation of Laplace site functions can be evaluated analytically using the following form

$$\langle \log p(w_i) \rangle_{\mathcal{N}(\mathbf{w}|\mathbf{m},\mathbf{S})} = -\log(2\tau_i) - \frac{1}{\tau_i}\langle |w_i| \rangle_{\mathcal{N}(w_i|m_i,S_{ii})} \tag{0.5}$$

where

$$\langle |w_i| \rangle_{N(w_i|m_i,S_{ii})} = \left(\frac{2S_{ii}}{\pi}\right)^{\frac{1}{2}} e^{-\frac{1}{2}a_i^2} + m_i\left[1 - 2\Phi\left(-a_i\right)\right]$$

[1] to see this consider the eigen equation for $\mathbf{X}^\intercal\mathbf{X}\mathbf{E} = \mathbf{E}\mathbf{\Lambda}$ thus $\mathbf{X}\mathbf{X}^\intercal\mathbf{X}\mathbf{E} = \mathbf{X}\mathbf{E}\mathbf{\Lambda}$

(0.6)

where $\Phi(x) = \int_{-\infty}^{x} \mathcal{N}(t|0,1)\,dt$ and $a_i = m_i/\sqrt{S_{ii}}$. The required derivatives are

$$\frac{\partial}{\partial m_j}\langle|w_j|\rangle = 1-\sqrt{\frac{2}{\pi}}a_j e^{-\frac{1}{2}a_j^2}-2\Phi(-a_j)+2a_j\mathcal{N}(a_j|0,1) \tag{0.7}$$

and

$$\frac{\partial}{\partial S_{jj}}\langle|w_j|\rangle = \frac{a_j^2+1}{\sqrt{2\pi S_{jj}}}e^{-\frac{1}{2}a_j^2}-\frac{a_j^2}{\sqrt{S_{jj}}}\mathcal{N}(a_j|0,1) \tag{0.8}$$

Thus in terms of the Cholesky factorisation we have

$$\frac{\partial}{\partial \mathbf{C}}\langle|w_j|\rangle = 2\mathbf{C}^{\mathsf{T}}\mathrm{diag}\left(\frac{\partial}{\partial S_{jj}}\langle|w_j|\rangle\right) \tag{0.9}$$

0.2.2 Inverse Modelling Gradients

Unconstrained Cholesky factorisations of the covariance results in VG bound gradients of the form:

$$\frac{\partial \mathcal{B}_{KL}}{\partial \mathbf{m}} = \frac{1}{s^2}\left(\mathbf{y}^{\mathsf{T}}\mathbf{M}-\mathbf{M}^{\mathsf{T}}\mathbf{M}\mathbf{m}\right)+\frac{\partial}{\partial \mathbf{m}}\langle\log p(\mathbf{w})\rangle \tag{0.10}$$

$$\frac{\partial \mathcal{B}_{KL}}{\partial \mathbf{C}} = -\frac{1}{2s^2}\mathrm{tril}\left(\mathbf{M}^{\mathsf{T}}\mathbf{M}\mathbf{C}\right)+\mathrm{diag}\left(1/C_{ii}\right) + \frac{\partial}{\partial \mathbf{C}}\langle\log p(\mathbf{w})\rangle \tag{0.11}$$

where $\mathrm{tril}(\mathbf{X})_{ij} = X_{ij}$ for $i \leq j$ and 0 otherwise.

For Chevron parameterised Cholesky covariances we have that

$$\frac{\partial \mathcal{B}_{KL}}{\partial \mathbf{\Theta}} = \frac{1}{2s^2}\mathrm{tril}\left(\mathbf{M}^{\mathsf{T}}\mathbf{M}\mathbf{\Theta}\right)+\mathrm{diag}\left(1/\Theta_{ii}\right) + \frac{\partial}{\partial \mathbf{\Theta}}\langle\log p(\mathbf{w})\rangle \tag{0.12}$$

and with respect to the diagonal elements $\mathbf{d}$

$$\frac{\partial \mathcal{B}_{KL}}{\partial d_i} = \frac{1}{2s^2}d_i\left[\mathbf{M}^{\mathsf{T}}\mathbf{M}\right]_{ii}+\frac{1}{d_i} + 2d_i\left(\frac{\partial}{\partial S_{ii}}\langle\log p(\mathbf{w})\rangle\right) \tag{0.13}$$

where $[\mathbf{X}]_{ii}$ refers to the i^{th} diagonal element of $\mathbf{X}$.

0.2.3 Logistic Regression VG Gradients

This bound admits the following gradients for an unconstrained Cholesky factorisation of the covariance.

$$\frac{\partial \mathcal{B}_{KL}}{\partial \mathbf{m}} = -\mathbf{\Sigma}^{-1}\mathbf{m}+\sum_{n=1}^{N}s_n\mathbf{x}_n\left(1-\langle f(\mu_n+z\sigma_n)\rangle_z\right)$$

In problems of sufficiently low dimensionality, it is useful to have the bound's Hessian with respect to $\mathbf{m}$ which is

$$\frac{\partial^2 \mathcal{B}_{KL}}{\partial m_i \partial m_j} = -\left[\mathbf{\Sigma}^{-1}\right]_{ij} - \sum_{n=1}^{N}x_i^n x_j^n\left\langle f(\mu_n+z\sigma_n)\left(1-f(\mu_n+z\sigma_n)\right)\right\rangle_z$$

The gradient with respect the Cholesky parameterisations of the variational covariance we have

$$\frac{\partial \mathcal{B}_{KL}}{\partial \mathbf{C}} = \mathrm{diag}\left(\frac{1}{C_{ii}}\right)-\mathrm{tril}\left(\mathbf{\Sigma}^{-1}\mathbf{C}+\mathbf{X}\mathbf{\Gamma}\mathbf{X}^{\mathsf{T}}\mathbf{C}\right) \tag{0.14}$$

where $\mathbf{X}\mathbf{\Gamma}\mathbf{X} = \left[\sum_{n=1}^{N}\frac{\mathbf{x}_n\mathbf{x}_n^{\mathsf{T}}}{\sigma_n}\langle zf(\mu_n+z\sigma_n)\rangle_z\right]$. For banded Cholesky parameterisations of the covariance the above algebraic form is accurate but elements indexed out of band width should be fixed to zero.

Chevron Chevron parameterisations of covariance admit the following gradients with respect to $\mathbf{\Theta}$

$$\frac{\partial \mathcal{B}_{KL}}{\partial \mathbf{\Theta}} = \mathrm{diag}\left(\frac{1}{\Theta_{ii}}\right)-\mathrm{tril}\left(\mathbf{\Sigma}^{-1}\mathbf{\Theta}+\mathbf{X}\mathbf{\Gamma}\mathbf{X}^{\mathsf{T}}\mathbf{\Theta}\right) \tag{0.15}$$

and with respect to the diagonal $\mathbf{d}$ we have

$$\frac{\partial \mathcal{B}_{KL}}{\partial d_i} = \frac{1}{d_i}-\left[\mathbf{\Sigma}^{-1}\right]_{ii}d_i - d_i\left[\mathbf{X}\mathbf{\Gamma}\mathbf{X}^{\mathsf{T}}\mathbf{C}\right]_{ii} \tag{0.16}$$

Subspace Letting $\mathbf{C} = \mathrm{blkdiag}\left(\mathbf{C}^{sub}, c\mathbf{I}_{L\times L}\right)$ where $\mathbf{C}^{sub}$ is the $K \times K$ subspace Cholesky matrix, $L = D-K$, and spherical prior variance $\mathbf{\Sigma} = s^2\mathbf{I}$,

$$\frac{\partial \mathcal{B}_{KL}}{\partial \mathbf{C}^{sub}} = \mathrm{diag}\left(\frac{1}{C_{ii}^{sub}}\right) - \mathrm{tril}\left(\mathbf{\Sigma}^{-1}\mathbf{C}^{sub}+\mathbf{X}'\mathbf{\Gamma}'\mathbf{X}'^{\mathsf{T}}\mathbf{C}^{sub}\right) \tag{0.17}$$

where $\mathbf{X}' = \mathbf{E}_1^{\mathsf{T}}\mathbf{X}$, $\Gamma'_{nn} = \frac{\langle zf(\mu_n+z\sigma'_n)\rangle_z}{\sigma'_n}$ where $\sigma_n'^2 = \|\mathbf{C}^{sub\mathsf{T}}\mathbf{E}_1^{\mathsf{T}}\mathbf{x}_n\|^2 + c^2(\|\mathbf{x}_n\|^2 - \|\mathbf{E}_1^{\mathsf{T}}\mathbf{x}_n\|^2)$. The gradient with respect to c is then

$$\frac{\partial \mathcal{B}_{KL}}{\partial c} = \frac{L}{c}-\frac{cL}{s^2}- c\sum_n \Gamma'_{nn}\left(\|\mathbf{x}_n\|^2-\|\mathbf{C}^{sub\mathsf{T}}\mathbf{E}_1^{\mathsf{T}}\mathbf{x}_n\|^2\right) \tag{0.18}$$