Appendix to

"Variational Bayes with Intractable Likelihood"

Minh-Ngoc Tran, David J. Nott, Robert Kohn

Proof of Theorem 1. (i) Under Assumptions 1 and 2, we have that

$$\mathrm{KL}(\lambda) = \mathrm{KL}(q_\lambda \| \pi) - \int q_\lambda(\theta)\mathbb{E}(z|\theta)d\theta = \mathrm{KL}(q_\lambda \| \pi) + \frac{\sigma^2}{2}, \tag{1}$$

where $\mathrm{KL}(q_\lambda\|\pi)$ is the Kullback-Leibler divergence between the variational distribution $q_\lambda(\theta)$ and the posterior $\pi(\theta)$. So, $\nabla_\lambda \mathrm{KL}(\lambda) = \nabla_\lambda \mathrm{KL}(q_\lambda\|\pi)$ is independent of σ^2, and minimizing $\mathrm{KL}(\lambda)$ with respect to λ is equivalent to minimizing $\mathrm{KL}(q_\lambda\|\pi)$. Algorithm 1 and 2 are the Robbins-Monro procedure for finding the root λ^* of the equation $\nabla_\lambda \mathrm{KL}(q_\lambda\|\pi) = 0$. Then, the result follows from Theorem 1 of Sacks (1958) with the constant c_{λ^*} independent of σ^2.

(ii) Denote $\widehat{h}(\theta,z) = \log(p(\theta)\widehat{p}_N(y|\theta,z)) = \log(p(\theta)p(y|\theta)) + z = h(\theta) + z$. We consider the case with the noisy traditional gradient; the proof for the other cases is similar. We denote by $\widetilde{\nabla_\lambda \mathrm{KL}}(\lambda^*)$ the noisy gradient obtained when the likelihood is available. Then, noting that $\mathbb{E}_*(\zeta_*(\theta)) = 0$, the constant c in (6) is

$$c = \frac{\mathbb{E}_{\theta,z}\{\zeta_*(\theta)^2(\log q_{\lambda^*}(\theta) - h(\theta) - z)\}}{\mathbb{E}_*\{\zeta_*(\theta)^2\}} = \frac{\mathbb{E}_*\{\zeta_*(\theta)^2(\log q_{\lambda^*}(\theta) - h(\theta))\}}{\mathbb{E}_*\{\zeta_*(\theta)^2\}} + \frac{\sigma^2}{2} = \widetilde{c} + \frac{\sigma^2}{2}.$$

We note that $\widetilde{c}$ is the control variate constant we would use to compute $\widetilde{\nabla_\lambda \mathrm{KL}}(\lambda^*)$ if

the likelihood was known.

$$\begin{aligned}
\mathbb{V}\big(\widehat{\nabla_\lambda \mathrm{KL}}(\lambda^*)\big) &= \frac{1}{S}\mathbb{V}_{\theta,z}\Big\{\zeta_*(\theta)(\log q_{\lambda^*}(\theta) - h(\theta) - z - c)\Big\} \\
&= \frac{\sigma^2}{S}\mathbb{V}_*\Big\{\zeta_*(\theta)\Big\} + \frac{1}{S}\mathbb{V}_*\Big\{\zeta_*(\theta)(\log q_{\lambda^*}(\theta) - h(\theta) + \frac{\sigma^2}{2} - c)\Big\} \\
&= \frac{\sigma^2}{S}\mathbb{V}_*\Big\{\zeta_*(\theta)\Big\} + \frac{1}{S}\mathbb{V}_*\Big\{\zeta_*(\theta)(\log q_{\lambda^*}(\theta) - h(\theta) - \widetilde{c})\Big\} \\
&= \frac{\sigma^2}{S}\mathbb{V}_*\Big\{\zeta_*(\theta)\Big\} + \mathbb{V}\big(\widehat{\nabla_\lambda \mathrm{KL}}(\lambda^*)\big).
\end{aligned}$$

Therefore,

$$\sigma^2_{\mathrm{asym}}(\widehat{\lambda}_M) = c_{\lambda^*}\mathbb{V}\big(\widehat{\nabla_\lambda \mathrm{KL}}(\lambda^*)\big) = \sigma^2_{\mathrm{asym}}(\widetilde{\lambda}_M) + c_{\lambda^*}\frac{\sigma^2}{S}\mathbb{V}_*\Big\{\zeta_*(\theta)\Big\}.$$

□

Derivation for Section 5.1

The density of the $d-$variate normal $\mathcal{N}(\mu, \Sigma)$ is

$$q(\beta) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\Big(-\frac{1}{2}(\beta-\mu)'\Sigma^{-1}(\beta-\mu)\Big).$$

A simplified form of the inverse Fisher matrix for a multivariate normal under the natural parameterization is given in Wand (2014). For a $d\times d$ matrix A, denote by $\mathrm{vec}(A)$ the d^2-vector obtained by stacking the columns of A, by $\mathrm{vech}(A)$ the $\frac{1}{2}d(d+1)$-vector obtained by stacking the columns of the lower triangular part of A. The duplication matrix of order d, D_d, is the $d^2\times\frac{1}{2}d(d+1)$ matrix of zeros and ones such that for any symmetric matrix A

$$D_d\mathrm{vech}(A) = \mathrm{vec}(A).$$

Let $D_d^+ = (D_d' D_d)^{-1} D_d'$ be the Moore-Penrose inverse of D_d, and vec^{-1} be the inverse of the operator vec. Then, the exponential family form of the normal distribution $q(\beta)$ is $q(\beta) = \exp(T(\beta)'\lambda - Z(\lambda))$ with

$$T(\beta) = \begin{bmatrix} \beta \\ \text{vech}(\beta\beta') \end{bmatrix}, \quad \lambda = \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} = \begin{bmatrix} \Sigma^{-1}\mu \\ -\frac{1}{2} D_d' \text{vec}(\Sigma^{-1}) \end{bmatrix}. \tag{2}$$

The usual mean and variance parameterization is

$$\begin{cases} \mu = -\frac{1}{2}\left\{\text{vec}^{-1}(D_d^{+\prime}\lambda_2)\right\}^{-1}\lambda_1 \\ \Sigma = -\frac{1}{2}\left\{\text{vec}^{-1}(D_d^{+\prime}\lambda_2)\right\}^{-1}. \end{cases}$$

Wand (2014) derives the following very useful formula

$$I_F(\lambda)^{-1} = \begin{bmatrix} \Sigma^{-1} + M'S^{-1}M & -M'S^{-1} \\ -S^{-1}M & S^{-1} \end{bmatrix}, \tag{3}$$

with

$$M = 2D_d^+(\mu \otimes I_d) \text{ and } S = 2D_d^+(\Sigma \otimes \Sigma)D_d^{+\prime},$$

where $\otimes$ is the Kronecker product and I_d the identity matrix of order d. The gradient $\nabla_\lambda[\log q(\beta)]$ is

$$\nabla_\lambda[\log q(\beta)] = \begin{bmatrix} \beta - \mu \\ \text{vech}(\beta\beta' - \Sigma - \mu\mu') \end{bmatrix}. \tag{4}$$

For the inverse gamma distribution $q(\tau^2)$ with density

$$q(\tau^2) = \frac{a^b}{\Gamma(a)}(\tau^2)^{-1-a}\exp(-b/\tau^2),$$

the natural parameters are (a,b). The Fisher information matrix for the inverse gamma is

$$I_F(a,b) = \begin{pmatrix} \nabla_{aa}[\log\Gamma(a)] & -1/b \\ -1/b & a/b^2 \end{pmatrix},$$

and the gradient

$$\begin{aligned} \nabla_a[\log q_\lambda(\theta)] &= -\log(\tau^2) + \log(b) - \nabla_a[\log\Gamma(a)] \\ \nabla_b[\log q_\lambda(\theta)] &= -\frac{1}{\tau^2} + \frac{a}{b}. \end{aligned}$$

The importance of the natural gradient

We demonstrate the importance of the natural gradient using a simple example where the likelihood is available. We consider a model where the data $y_i \sim B(1,\theta)$ - the Bernoulli distribution with probability θ. We generate $n=200$ observations y_i from $B(1,\theta=0.3)$ and obtain $k=\sum_i y_i = 57$. We use a uniform prior on θ. Then, the posterior $p(\theta|y)$ is Beta$(k+1,n-k+1)$. The variational distribution $q_\lambda(\theta)$ is chosen to be Beta(α,β) which belongs to the exponential family with the natural parameter $\lambda=(\alpha,\beta)'$. The Fisher information matrix $I_F(\lambda)$ is

$$I_F(\lambda) = \begin{bmatrix} \psi_1(\alpha) - \psi_1(\alpha+\beta) & \psi_1(\alpha+\beta) \\ \psi_1(\alpha+\beta) & \psi_1(\beta) - \psi_1(\alpha+\beta) \end{bmatrix},$$

where $\psi_1(x) = \nabla_{xx}[\log\Gamma(x)]$ is the *trigamma* function. In this simple example, the gradient $\nabla_\lambda \mathrm{KL}(\lambda)$ can be computed analytically

$$\nabla_\lambda \mathrm{KL}(\lambda) = I_F(\lambda)\lambda - H(\lambda)$$

with

$$H(\lambda) = \begin{bmatrix} k\psi_1(\alpha) - n\psi_1(\alpha+\beta) \\ (n-k)\psi_1(\beta) - n\psi_1(\alpha+\beta) \end{bmatrix}.$$

Using the traditional gradient, the update in Algorithm 1 is

$$\lambda^{(t+1)} = \lambda^{(t)} - a_t\Big(I_F(\lambda^{(t)})\lambda^{(t)} - H(\lambda^{(t)})\Big).$$

Using the natural gradient, the update is

$$\lambda^{(t+1)} = (1-a_t)\lambda^{(t)} + a_t I_F(\lambda^{(t)})^{-1} H(\lambda^{(t)}).$$

Figure 1 plots the densities of the exact posterior $\pi(\theta)$ and the variational distributions $q_\lambda(\theta)$ estimated by the VBIL using the traditional gradient and the natural gradient, with two different random initializations. The figure shows that the VBIL algorithm using the natural gradient is superior to that using the traditional gradient. Furthermore, the VBIL algorithm based on the natural gradient is insensitive to the initial $\lambda^{(0)}$.

Using VBIL to improve estimates of the marginal posteriors

We illustrate this application by generating $n = 100$ observations from a univariate mixture of two normals $p(x) = \omega\mathcal{N}(x|\mu_1,\sigma_1^2) + (1-\omega)\mathcal{N}(x|\mu_2,\sigma_2^2)$ with $\omega = 0.3$, $\mu_1 = -3$, $\mu_2 = 3$, $\sigma_1^2 = 2$ and $\sigma_2^2 = 3$. Suppose that we are interested in getting an accurate variational approximation of the posterior $p(\omega|y)$. Getting an accurate estimate of w is often more challenging than the other parameters. We use diffuse priors $\omega \sim U(0,1)$, $\mu_1 \sim \mathcal{N}(0,100)$, $\mu_2 \sim \mathcal{N}(0,100)$, $\sigma_1^2 \sim (\sigma_1^2)^{-1}$ and $\sigma_2^2 \sim (\sigma_2^2)^{-1}$, and run VBIL to approximate $p(\omega|y)$ by a Beta distribution. We use the VB algorithm of

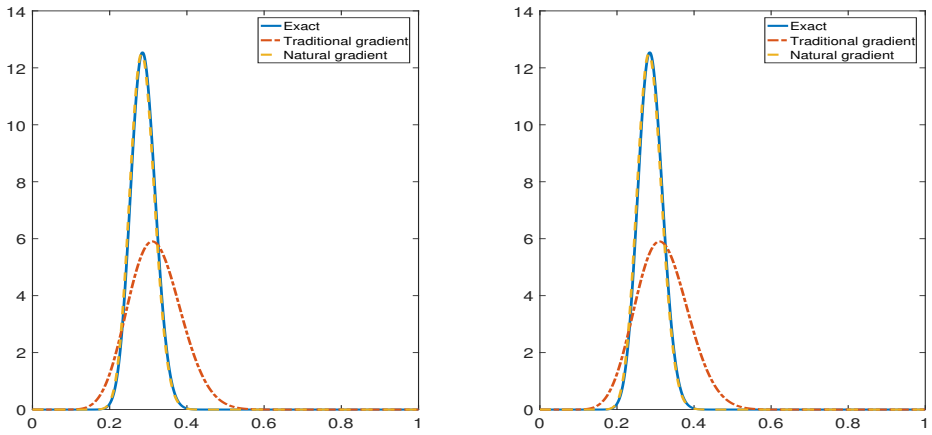

Figure 1: Importance of the natural gradient: Plots of the densities of the exact posterior (solid line) and the variational approximation estimates using the traditional gradient (dotted line) and the natural gradient (dashed line), with two different starting values $\lambda^{(0)}$ at random. VB based on the natural gradient produces highly accurate estimates.

McGrory and Titterington (2007), in which the variational distribution is factorized as $q(\omega)q(\sigma_1^2,\sigma_2^2)q(\mu_1,\mu_2|\sigma_1^2,\sigma_2^2)$, to design the proposal density to obtain an importance sampling estimator of $p(y|\omega)$.

Figure 2 plots the McGrory-Titterington estimate (dashed line) and VBIL estimate (solid line) of the posterior $p(\omega|y)$. As shown, the VBIL estimate has heavier tails than the VB estimate. By (20), it follows that the difference between the two estimates gives an indication of the extent to which the McGrory-Titterington estimate is suboptimal. This example shows that the VBIL method provides an attractive way to obtain accurate approximation of marginal posteriors.

References

McGrory, C. and Titterington, D. (2007). Variational approximations in Bayesian model selection for finite mixture distributions. *Computational Statistics & Data*

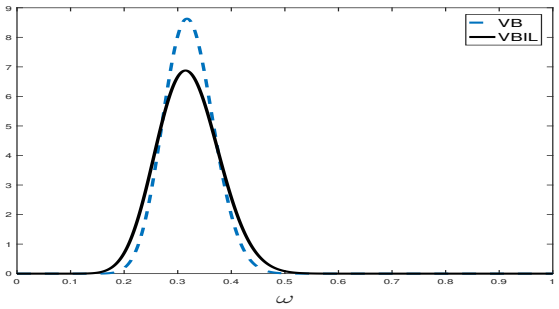

Figure 2: Improving estimates of marginal posteriors: Plots of the VB (dashed line) and VBIL estimates (solid line) of the posterior $p(\omega|y)$. The VBIL estimate has heavier tails than the VB estimate.

Analysis, 51(11):5352 – 5367. Advances in Mixture Models.

Sacks, J. (1958). Asymptotic distribution of stochastic approximation procedures. *The Annals of Mathematical Statistics*, 29(2):373–405.

Wand, M. P. (2014). Fully simplified multivariate normal updates in non-conjugate variational message passing. *Journal of Machine Learning Research*, 15:1351–1369.