Supplementary Material for "A Geometric Variational Approach to Bayesian Inference"

Abhijoy Saha[1], Karthik Bharath[2], Sebastian Kurtek[1]

[1]Department of Statistics, The Ohio State University

[2]School of Mathematical Sciences, University of Nottingham

1 Proofs

Proof of Proposition 1.

Denote by $J(\boldsymbol{\eta}) := \left|\det\Big(d\theta_i/d\eta_j\Big)\right|$ the absolute value of the determinant of the Jacobian matrix corresponding to the transformation $(\theta_1, \ldots, \theta_d) \mapsto (\phi_1(\theta_1), \ldots, \phi_d(\theta_d)) = (\eta_1, \ldots, \eta_d)$. For $i = 1, \ldots, d$ since each ϕ_i transforms θ_i individually, $J(\boldsymbol{\eta}) = \prod_{i=1}^d d\theta_i/d\eta_i$. Then,

$$\begin{aligned}\mathcal{E}_{1/2}(\psi_q; \boldsymbol{\eta}) &= \int_{\boldsymbol{\Theta}} \psi_f(x, \boldsymbol{\eta}) \prod_{i=1}^d \psi_{q_i}(\eta_i) d\boldsymbol{\eta} = \int_{\boldsymbol{\Theta}} \psi_f(x, \boldsymbol{\theta}) \prod_{i=1}^d \psi_{q_i}(\theta_i) J(\boldsymbol{\eta}) d\boldsymbol{\eta} \\ &= \int_{\boldsymbol{\Theta}} \psi_f(x, \boldsymbol{\theta}) \prod_{i=1}^d \psi_{q_i}(\theta_i) d\boldsymbol{\theta} = \mathcal{E}_{1/2}(\psi_q; \boldsymbol{\theta}).\end{aligned}$$

Proof of Proposition 2.

$$m(x)^\alpha = \Big[\int_{\boldsymbol{\Theta}} f(x, \boldsymbol{\theta}) d\boldsymbol{\theta}\Big]^\alpha = \Big[\int_{\boldsymbol{\Theta}} \frac{f(x, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} q(\boldsymbol{\theta}) d\boldsymbol{\theta}\Big]^\alpha = g\left(\mathbb{E}_q\Big[\frac{f(x, \boldsymbol{\theta})}{q(\boldsymbol{\theta})}\Big]\right),$$

where $g(y) = y^\alpha$ is concave on $y > 0$ and $0 < \alpha < 1$. Using Jensen's inequality, we obtain:

$$m(x)^\alpha \geq \mathbb{E}_q\left[g\Big(\frac{f(x,\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\Big)\right] = \int_{\boldsymbol{\Theta}} \Big(\frac{f(x,\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\Big)^\alpha q(\boldsymbol{\theta})d\boldsymbol{\theta} = \int_{\boldsymbol{\Theta}} f(x,\boldsymbol{\theta})^\alpha q(\boldsymbol{\theta})^{1-\alpha} d\boldsymbol{\theta},$$

which is equivalent to $\mathcal{E}_\alpha$ since $f = \psi_f^2$ and $q = \psi_q^2$. Taking the natural logarithm on both sides, we obtain $\alpha \ln m(x) \geq \ln \mathcal{E}_\alpha(\psi_q; \cdot)$. Further, consider $\ln \mathcal{E}_\alpha(\psi_q; \cdot) = \ln \mathbb{E}_q\left[\Big(\frac{f(x,\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\Big)^\alpha\right]$. Since ln is a concave function, using Jensen's inequality again we obtain:

$$\ln \mathbb{E}_q\left[\Big(\frac{f(x,\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\Big)^\alpha\right] \geq \mathbb{E}_q\left[\ln\Big(\frac{f(x,\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\Big)^\alpha\right] = \alpha\mathbb{E}_q\left[\ln\Big(\frac{f(x,\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\Big)\right] = \alpha\mathcal{H}(f,q).$$

Thus, for $0 < \alpha < 1$ we have

$$\alpha \ln m(x) \geq \ln \mathcal{E}_\alpha(\psi_q; \cdot) \geq \alpha\mathcal{H}(f,q).$$

Similarly, for part (ii) of the proposition, noting that $g(y) = y^\alpha$ is convex for $y > 0$ and $\alpha > 1$, using Jensen's inequality we obtain:

$$\alpha \ln m(x) \leq \ln \mathcal{E}_\alpha(\psi_q; \cdot).$$

Proof of Proposition 3.

The directional or Fréchet derivative $D^i\mathcal{E}_\alpha$ at a point $\psi_{q_i} \in \Psi$ is defined as a linear functional $T_{\psi_{q_i}}(\Psi) \to \mathbb{R}$, the element of the dual space of $T_{\psi_{q_i}}(\Psi)$, via the relation $D^i\mathcal{E}_\alpha(v_i) = \langle D^i\mathcal{E}_\alpha, v_i\rangle$ for all $v_i \in T_{\psi_{q_i}}(\Psi)$ since $T_{\psi_{q_i}}(\Psi)$ is a linear subspace of $\mathbb{L}^2(\Theta_i)$, the Hilbert space of square-integrable functions on Θ_i, and inherits the usual inner product. The Riesz representation theorem implies that the gradient $\nabla\mathcal{E}_\alpha^i$ exists as an element of $T_{\psi_{q_i}}(\Psi)$ and is defined such that $D^i\mathcal{E}_\alpha(v_i) = \langle\nabla\mathcal{E}_\alpha^i, v_i\rangle, \quad v_i \in T_{\psi_{q_i}}(\Psi)$. Along a basis

direction b_i^k, we can therefore express the gradient as

$$\nabla \mathcal{E}_\alpha^i = \sum_{k=1}^{\infty} D^i \mathcal{E}_\alpha(b_i^k) b_i^k.$$

Therefore, on the restriction $\mathcal{E}_{\alpha|\Psi} : \Psi \to \mathbb{R}_{>0}$ to Ψ for each $i = 1, \ldots, d$, the directional derivative $D^i \mathcal{E}_\alpha$ along b_i^k can be computed as

$$\begin{aligned} D^i \mathcal{E}_\alpha(b_i^k) &= \lim_{t \to 0} \frac{1}{t} \Big[\mathcal{E}_{\alpha|\Psi}(\psi_{q_i} + t b_i^k) - \mathcal{E}_{\alpha|\Psi}(\psi_{q_i}) \Big] \\ &= \lim_{t \to 0} \frac{1}{t} \Big[\int_{\boldsymbol{\Theta}} \psi_f(x, \boldsymbol{\theta})^{2\alpha} \prod_{j \neq i} (\psi_{q_j}(\theta_j))^{2(1-\alpha)} \left\{ (\psi_{q_i}(\theta_i) + t b_i^k(\theta_i))^{2(1-\alpha)} - (\psi_{q_i}(\theta_i))^{2(1-\alpha)} \right\} d\boldsymbol{\theta} \Big]. \end{aligned}$$

With $p = 2(1-\alpha)$, using the binomial expansion for real powers we obtain

$$\frac{1}{t} \Big[(\psi_{q_i}(\theta_i) + t b_i^k(\theta_i))^p \Big] = \frac{1}{t} \Big[\psi_{q_i}^p(\theta_i) + p \psi_{q_i}^{p-1}(\theta_i) t b_i^k(\theta_i) + R(\theta_i, \alpha, t) \Big],$$

where $R(\theta_i, \alpha, t) = O(m(\theta_i, \alpha) t^2)$ with $m(\cdot, \cdot)$ a function of only θ_i and $2(1-\alpha)$. Therefore,

$$\begin{aligned} D^i \mathcal{E}_\alpha(b_i^k) &= \lim_{t \to 0} \Big[\int_{\boldsymbol{\Theta}} \psi_f(x, \boldsymbol{\theta})^{2\alpha} \prod_{j \neq i} (\psi_{q_j}(\theta_j))^{2(1-\alpha)} \Big\{ 2(\alpha - 1) \psi_{q_i}^{1-2\alpha}(\theta_i) b_i^k(\theta_i) + \frac{1}{t} R(\theta_i, \alpha, t) \Big\} d\boldsymbol{\theta} \Big] \\ &= 2(\alpha - 1) \int_{\boldsymbol{\Theta}} \psi_f(x, \boldsymbol{\theta})^{2\alpha} \prod_{j \neq i} (\psi_{q_j}(\theta_j))^{2(1-\alpha)} \psi_{q_i}^{1-2\alpha}(\theta_i) b_i^k(\theta_i) d\boldsymbol{\theta} \\ &\quad + \lim_{t \to 0} \int_{\boldsymbol{\Theta}} \psi_f(x, \boldsymbol{\theta})^{2\alpha} \prod_{j \neq i} (\psi_{q_j}(\theta_j))^{2(1-\alpha)} \frac{1}{t} R(\theta_i, \alpha, t) \Big\} d\boldsymbol{\theta}. \end{aligned} \tag{1}$$

For fixed $(x, \alpha) \in \mathcal{X} \times (0, \infty) \backslash \{1\}$, the sequence of functions

$$\boldsymbol{\Theta} \ni \boldsymbol{\theta} \mapsto H_t(x, \boldsymbol{\theta}, \alpha) := \psi_f(x, \boldsymbol{\theta})^{2\alpha} \prod_{j \neq i} (\psi_{q_j}(\theta_j))^{2(1-\alpha)} \frac{1}{t} R(\theta_i, \alpha, t)$$

converges to $H(x, \boldsymbol{\theta}, \alpha) := \psi_f(x, \boldsymbol{\theta})^{2\alpha} \prod_{j \neq i} (\psi_{q_j}(\theta_j))^{2(1-\alpha)}$ as $t \to 0$ with $|H(x, \boldsymbol{\theta}, \alpha)| < \infty$ since $q(\boldsymbol{\theta})$ is strictly positive based on our assumptions. The result follows by an application of the dominated convergence theorem to the second integral in Equation 1.

Proof of Proposition 4.

We prove the result for the case when $\alpha = 1/2$ to avoid cumbersome notation; the case of a general α can be worked out along the lines of the proof of Proposition 3 using the binomial expansion with real-valued powers. We use the result provided in Proposition 3.2 of Ring and Wirth (2012) to prove our claim. They state that a sufficient condition to ensure the existence such an ϵ is that the restriction of $\tilde{\mathcal{E}}^i_\alpha$ to the linear span of the set $\{v_i\}$ be continuously differentiable.

Suppose we choose basis functions $\mathcal{B}_i$ for $T_{\psi_{q_i}}(\Psi)$ that are bounded (in $\mathbb{L}^2$) and smooth (e.g., Fourier basis). Observe then that from the expression of the gradient $\nabla\mathcal{E}^i_\alpha$ in Proposition 3, differentiability of $\tilde{\mathcal{E}}^i_\alpha$ depends solely on the existence and continuity of the directional derivative at $v_i \in T_{\psi_{q_i}}(\Psi)$ along a direction b_i, defined as

$$D^i\mathcal{E}_\alpha \circ \exp(v_i)(b_i) := \lim_{t\to 0} \frac{1}{t}\Big[\tilde{\mathcal{E}}_{\alpha|\Psi}(\psi_{q_i} + tb_i) - \tilde{\mathcal{E}}_{\alpha|\Psi}(\psi_{q_i})\Big].$$

For convenience, denoting a function $h(\cdot)$ as h, similar calculations as in the proof of Proposition 3 results in

$$D^i\mathcal{E}_\alpha \circ \exp(v_i)(b_i) = \int_{\boldsymbol{\Theta}} \frac{\psi_f}{2\exp_{\psi_{q_i}}(v_i)} \frac{A(v_i)}{\|v_i\|}\left[\frac{v_i}{2\|v_i\|^2} - \frac{\psi_{q_i}\sin(\|v_i\|)}{2} + \frac{b_i\sin(\|v_i\|)}{A(v_i)} + \frac{v_i\cos(\|v_i\|)}{2\|v_i\|}\right]d\boldsymbol{\theta},$$

where $A(v_i) := \|b_i\|^2 + 2\|v_i\|\|b_i\|$. It's easy to verify that $D^i\mathcal{E}_\alpha \circ \exp(v_i)(b_i)$ is finite for every choice of b_i and v_i as long as $v_i \neq 0$.

The directional derivative $D^i\mathcal{E}_\alpha \circ \exp(v_i)(b_i)$ is clearly a linear operator from $T_{\psi_{q_i}}(\Psi) \to \mathbb{R}$. The space Ψ, which is the positive orthant of the unit sphere S^∞ in $\mathbb{L}^2(\Theta_i)$, is an open subset of S^∞ (Theorem 3.2 of Khesin et al. (2013)) and hence a Hilbert submanifold of S^∞. The domain of the exponential map $\exp_{\psi_{q_i}}$ is all of $T_{\psi_{q_i}}(\Psi)$ and hence maps every element of $T_{\psi_{q_i}}(\Psi)$ to Ψ. Therefore it is a diffeomorphism from $T_{\psi_{q_i}}(\Psi)$ to Ψ and clearly continuous.

Consider a sequence $v_{i,n}$ such that $\|v_{i,n} - v_i\| \to 0$. Noting that the norm function $x \mapsto \|x\|$ is continuous in $\mathbb{L}^2$, it is easy to see that every term inside the integral is a continuous function of v_i. This implies that $D^i\mathcal{E}_\alpha \circ \exp(v_{i,n})(b_i) \to D^i\mathcal{E}_\alpha \circ \exp(v_i)(b_i)$ as $n \to \infty$. For a finite N chosen in the algorithm that determines the number of iterations, the argument can be extended to any linear combination of possible directions $\sum_{j=1}^N \alpha_{i,j} v_{i,j}$ on each tangent space $T_{\psi_{q_i}}(\Psi)$, and thus to the span under consideration. This completes the proof.

2 Bayesian Linear Regression

All of the following run time experiments were performed in Matlab on an Intel Core i7 processor (3.40 GHz) with 8 GB of RAM.

2.1 Run Time Comparison for Gibbs Sampler and Proposed Method

d	n	MSE Gibbs (*Time*) *iter*: 5000, *burn-in*: 1000	MSE PM (*Time*)	MSE Gibbs (*Time*) *iter*: 30000, *burn-in*: 5000	MSE PM (*Time*)
25	100	1.0034e-05 (*6.9151*)	9.5201e-06 (*1.3692*)	7.0139e-07 (*40.8964*)	5.2551e-07 (*1.8097*)
50	100	1.7488e-05 (*16.2920*)	1.7234e-05 (*9.9007*)	4.9125e-06 (*96.8816*)	4.7913e-06 (*11.2436*)
75	100	2.2089e-04 (*27.9082*)	2.1717e-04 (*24.2695*)	1.8478e-05 (*165.5516*)	1.8268e-05 (*39.1568*)
100	200	7.0274e-06 (*60.0844*)	6.8176e-06 (*37.2117*)	1.0505e-06 (*381.3203*)	9.9089e-07 (*47.4571*)

Table 1: Comparison of running times (in seconds) between the Gibbs sampler and the proposed method (PM) with $\alpha = 0.5$.

In Table 1, we compare the running times of our method to the Gibbs sampling approach. For each choice of d, a dataset of size n is generated. First, we apply the Gibbs sampler with two choices of the total number of iterations and burn-in. The MSE in each case is noted along with the respective running time. Next, we fix $\alpha = 0.5$ and apply our method to the same dataset. We continue to update the corresponding q_is in each case and run the algorithm as long as the MSE for the proposed method is larger than the one obtained using the Gibbs sampling technique. We stop the algorithm as soon as the MSE for our method becomes smaller than the Gibbs MSE and note the running time. It can

be easily seen that the proposed variational method attains the MSE obtained by Gibbs sampling much faster in each case.

2.2 Effect of Basis Size on Run Time

	$d = 10$	$d = 25$	$d = 50$	$d = 100$	$d = 200$
	$n = 100$	$n = 100$	$n = 100$	$n = 500$	$n = 500$
$N = 49$	0.1151	0.5717	4.2978	12.6686	76.2938
$N = 99$	0.1748	0.9445	6.3781	14.1433	106.5231
$N = 199$	0.2892	1.4250	8.8313	21.6562	138.5783
$N = 499$	1.5809	7.7542	39.7014	72.8446	377.8602

Table 2: Comparison of running times (in seconds) for varying number of basis elements (N), sample size (n) and dimensionality (d) with $\alpha = 0.5$.

In our setup, the tangent space for each q_i is spanned by a finite collection of pre-specified orthonormal basis functions. We indicated in the main paper that increasing the number of basis elements leads to better approximations of the posterior. Table 2 shows the running times for our method for different choices of dimensionality d and sample size n as we vary the number of basis functions N. In this simulation, we use $\alpha = 0.5$. The proposed algorithm is fast for a moderately sized dataset even with $N = 199$ basis elements, i.e., run time is under 10 seconds for a dataset of dimension $d = 50$, and sample size $n = 100$.

3 Bayesian Density Estimation

3.1 Effect of Sample Size on Density Estimation

Figure 1 shows the effect of the sample size n on density estimation under the proposed approach. As is common with other nonparametric density estimation techniques, we can see that with the increase in sample size, we obtain a better estimate of the true density function. Again, the proposed method with $\alpha = 0.9$ performs favorably compared to standard kernel density estimation.

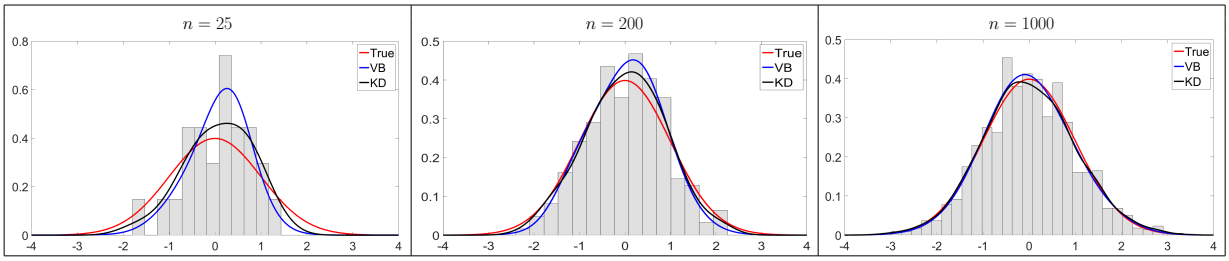

Figure 1: Effect of increasing the sample size on density estimation. Data was generated from a N(0,1).

3.2 Effect of α on Density Estimation

Next, we assess the effect of α on density estimation in Figure 2. The plot in the left panel of Figure 2 shows the results for standard VB (i.e., $\alpha \to 0$) and EP (i.e., $\alpha \to 1$) in the limiting case, as noted in Section 3.1 in the main paper. We note that, using the proposed D_α-based approach, we are able to explore a richer class of divergences with minor adjustments by simply changing the value of α accordingly. The right panel shows similar results for $\alpha = 1.1$ and $\alpha = 2$.

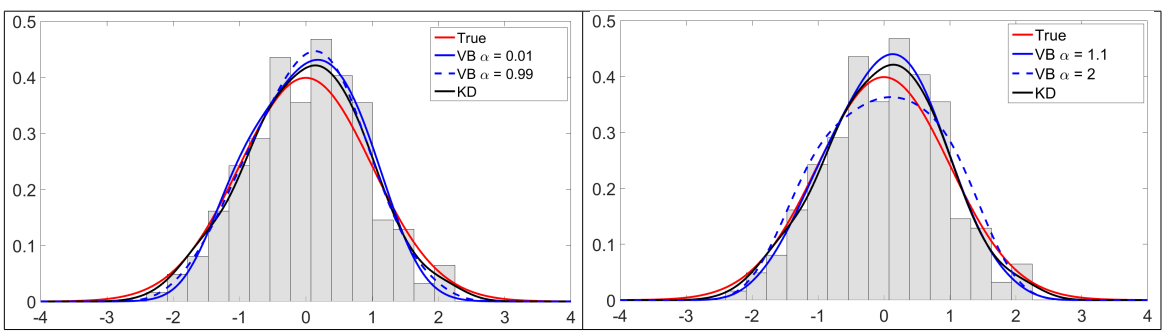

Figure 2: Effect of different choices of α on density estimation. Data was generated from a N(0,1).

4 Bayesian Logistic Regression for Real Data Applications

We report the accuracy and average log predictive likelihood (ALPL) of the proposed method on two multi-label datasets [1] considered in Wang and Blei (2013). As in the main

[1] `http://mulan.sourceforge.net/datasets.html`

paper MAP, PMEA, PMED and PPRED represent the maximum a posteriori, posterior mean, posterior median and posterior predictive, respectively, under the proposed approach. The results based on these summaries for KLD-based VB are also reported under KLMAP, KLMEA, KLMED and KLPRED.

4.1 Yeast Data

	MAP	PMEA	PMED	PPRED	KLMAP	KLPMEA	KLPMED	KLPPRED
Accuracy (in %)	79.5	79.5	79.4	79.6	79.3	79.2	79.3	79.2
ALPL	-0.6587	-0.6579	-0.6595	-0.6576	-0.6718	-0.6731	-0.6734	-0.6723

Table 3: Classification results for the the yeast dataset.

The yeast dataset (Elisseeff and Weston, 2001) is composed of micro-array expression data and phylogenetic profiles with 1500 genes in the training set and 917 in the test set. For each gene, we have 103 covariates and up to 14 different gene functional classes, making it a multi-label problem. This can be reformulated into 14 separate binary classification tasks, similar to the problem studied in Wang and Blei (2013).

We use 299 basis elements for this dataset, and choose $\alpha = 0.9$. Table 3 presents the classification results, averaged over the 14 binary problems, obtained under this setup. The results for all of the methods are very similar in this case. The ALPL values based on all of the D_α summaries in Table 3 are greater than the KLD-based summaries, indicating better fit. Wang and Blei (2013) report accuracy rates of 80.1% and 80.2%, and ALPL values of -0.449 and -0.450 using Laplace inference and delta method inference, respectively. The method of Jaakkola and Jordan (1997) gives accuracy of 79.7% and an ALPL value of -0.678. However, note that both of those papers used cross-validation to evaluate the performance of their algorithm, which is different from the training-testing split considered here. Nonetheless, the proposed method produces results that are comparable to the previously reported classification results.

	MAP	PMEA	PMED	PPRED	KLMAP	KLPMEA	KLPMED	KLPPRED
Accuracy (in %)	88.3	88.6	84.1	88.6	82.7	85.7	83.1	85.9
ALPL	-0.5649	-0.5596	-0.6375	-0.5584	-0.8375	-0.7653	-0.8425	-0.7581

Table 4: Classification results for the scene dataset.

4.2 Scene Data

The scene dataset was used for the problem of semantic scene classification in Boutell et al. (2004), where a scene might contain different objects such that it can be described by multiple class labels. The dataset contains 1,211 images in the training set and 1,196 images in the test set. The images consist of 294 image features that can be used to predict scene labels. There are up to six scene labels per image. Analogous to the previous example, this corresponds to six separate binary classification problems.

We use 299 basis elements to estimate the energy gradient in this classification problem, and choose $\alpha = 1.1$. The classification results, averaged over the six binary problems, obtained using the summary measures based on the proposed D_α-based VB and KLD-based VB are presented in Table 4. With the exception of the posterior median, summaries based on D_α perform better than those of Jaakkola and Jordan (1997) (they achieve 87.4% accuracy). Moreover, the ALPL values based on D_α summaries are significantly higher as well (they obtain a an ALPL value of -0.670). Our classification accuracy results are comparable to the results obtained using Laplace inference and delta method inference introduced in Wang and Blei (2013). Similar to the yeast example considered in Section 4.1, the results in those papers were generated via cross-validation, which is different from our training-testing split setup.

References

Boutell, M. R., J. Luo, X. Shen, and C. M. Brown (2004). Learning multi-label scene classification. *Pattern Recognition 37*(9), 1757–1771.

Elisseeff, A. and J. Weston (2001). A kernel method for multi-labelled classification. In *Neural Information Processing Systems*, pp. 681–687.

Jaakkola, T. and M. I. Jordan (1997). A variational approach to Bayesian logistic regression models and their extensions. In *International Workshop on Artificial Intelligence and Statistics*, Volume 82.

Khesin, B., J. Lenells, G. Misiolek, and S. C. Preston (2013). Geometry of diffeomorphism groups, complete integrability and geometric statistics. *Geometric Functional Analysis 23*(2), 334–366.

Ring, W. and B. Wirth (2012). Optimization methods on Riemannian manifolds and their applications to shape space. *SIAM Journal of Optimization 22*(2), 596–627.

Wang, C. and D. M. Blei (2013). Variational inference in nonconjugate models. *Journal of Machine Learning Research 14*, 1005–1031.