Sparse Approximations for Non-Conjugate Gaussian Process Regression

Thang Bui and Richard Turner
Computational and Biological Learning lab
Department of Engineering
University of Cambridge

November 11, 2014

Abstract

Notes: This report only shows some preliminary work on scaling Gaussian process models that use non-Gaussian likelihoods. As there are recently arxived papers on the similar idea [1, 2], this report will stay as is, please consult the two papers above for a proper discussion and experiments.

1 Introduction

Gaussian Processes (GPs) is an important tool for probabilistic models in both supervised and unsupervised settings. Typically, GPs are used as priors over unknown continuous functions which govern the trajectory of the observed variables. When the likelihood term is conjugate to this prior, e.g. a Gaussian likelihood, the posterior distribution is analytically tractable. It is however, not tractable for non-conjugate likelihoods, such as a Poisson likelihood in Poisson count regression or a logistic likelihood in binary classification. Approximation inference schemes such as Laplace approximation or Expectation Propagation can be applied in these cases [3]. Critically, these GP based models suffer from a high complexity, typically $\mathcal{O}(N^3)$ and $\mathcal{O}(N^2)$ for learning and prediction respectively. When dealing with large datasets, sparse approximation techniques that can reduce the computation demand are preferred. However, most of the approximations were developed for regression problems that use a Gaussian likelihood, and hence often cannot be applied to problems that deal with other forms of likelihood in a straightforward way.

This report introduces a variational inference formulation that 1. uses inducing points to augment the model as in current approximations for Gaussian likelihood, 2. uses variational free enery approach as a basis to find a generic lowerbound for many non-Gaussian likelihoods, 3. can perform joint optimization of variational parameters and model hyperparameters, 4. offers a small computational complexity.

2 Review of GPs for regression

$$p(\mathbf{f}) = \mathcal{N}(\mathbf{f}; \mathbf{0}, \mathbf{K_{ff}}) \tag{1}$$

$$p(\mathbf{y}|\mathbf{f}) = \prod_{n=1}^{N} p(y_n|f_n) \tag{2}$$

Type	Distribution	$p(y_n\|f_n)$
Continuous	Gaussian	$p(y_n\|f_n) = \mathcal{N}(y_n; f_n, \sigma_n^2)$
Binary	Bernoulli logistic	$p(y_n = 1\|f_n) = \sigma(f_n)$
Categorical	Multinomial logistic	$p(y_n = k\|\mathbf{f}_i) = \exp(f_{i,k} - \text{lse}(\mathbf{f}_i))$
Ordinal	Cumulative logistic	$p(y_n \le k\|f_n) = \sigma(\phi_k - f_n)$
Count	Poisson	$p(y_n = k\|f_n) = \frac{\exp(kf_n - \exp(f_n))}{k!}$
Continuous	Laplace	$p(y_n\|f_n) = \frac{1}{2b}\exp(-\frac{\|y_n - f_n\|}{b}), b = \frac{\sigma}{\sqrt{2}}$
Positive, Real	Exponential	$p(y_n\|f_n) = \frac{\exp(-y_n/\exp(f_n))}{\exp(f_n)}$

Table 1: Type of observed variables, possible likelihood functions and their forms.

3 Variational inference and learning

3.1 Formulation

Augment the model [4–7]
Choose variational distribution for the joint: $q(\mathbf{f}, \mathbf{u}) = q(\mathbf{f}|\mathbf{u})q(\mathbf{u})$ and $q(\mathbf{f}|\mathbf{u}) = p(\mathbf{f}|\mathbf{u})$ [6]. It could be shown that the lower

bound using the augmented variational distribution is smaller than the bound obtained if there is no auxiliary variables $\mathbf{u}$ [7]. The log marginal likelihood:

$$\begin{aligned}
\mathcal{L} &= \log p(\mathbf{y}) && (3)\\
&= \log \int \mathrm{d}\mathbf{u}\mathrm{d}\mathbf{f} p(\mathbf{f},\mathbf{u}) p(\mathbf{y}|\mathbf{f}) && (4)\\
&= \log \int \mathrm{d}\mathbf{u}\mathrm{d}\mathbf{f} q(\mathbf{f},\mathbf{u}) \frac{p(\mathbf{f},\mathbf{u}) p(\mathbf{y}|\mathbf{f})}{q(\mathbf{f},\mathbf{u})} && (5)\\
&\geq \int \mathrm{d}\mathbf{u}\mathrm{d}\mathbf{f} q(\mathbf{f},\mathbf{u}) \log \frac{p(\mathbf{f},\mathbf{u}) p(\mathbf{y}|\mathbf{f})}{q(\mathbf{f},\mathbf{u})} && (6)\\
&= \int \mathrm{d}\mathbf{u}\mathrm{d}\mathbf{f} q(\mathbf{f},\mathbf{u}) \log \frac{\cancel{p(\mathbf{f}|\mathbf{u})} p(\mathbf{u}) p(\mathbf{y}|\mathbf{f})}{\cancel{p(\mathbf{f}|\mathbf{u})} q(\mathbf{u})} && (7)\\
&= \int \mathrm{d}\mathbf{u} q(\mathbf{u}) \log \frac{p(\mathbf{u})}{q(\mathbf{u})} + \int \mathrm{d}\mathbf{u}\mathrm{d}\mathbf{f} q(\mathbf{f},\mathbf{u}) \log p(\mathbf{y}|\mathbf{f}) = \mathcal{F} && (8)
\end{aligned}$$

The first term in the lower bound above is the negative KL divergence between distributions $q(\mathbf{u})$ and $p(\mathbf{u})$. The second term can be expressed as a sum of N terms, $\mathcal{F}_2 = \sum_{n=1}^{N} \mathcal{F}_{2,n}$, where

$$\begin{aligned}
\mathcal{F}_{2,n} &= \int \mathrm{d}\mathbf{u} q(\mathbf{u}) \int \mathrm{d}f_n p(f_n|\mathbf{u}) \log p(y_n|f_n) && (9)\\
&= \int \mathrm{d}f_n \log p(y_n|f_n) \int \mathrm{d}\mathbf{u} q(\mathbf{u}) p(f_n|\mathbf{u}). && (10)
\end{aligned}$$

Therefore, the lower bound to the log marginal likelihood can be expressed as,

$$\mathcal{F} = -\mathrm{KL}(q(\mathbf{u})||p(\mathbf{u})) + \sum_{n=1}^{N} \int \mathrm{d}f_n q(f_n) \log p(y_n|f_n), \tag{11}$$

where $q(f_n) = \int \mathrm{d}\mathbf{u} q(\mathbf{u}) p(f_n|\mathbf{u})$. In the next sections, we will discuss the choice of the variational distribution $q(\mathbf{u})$ such that the KL term and $q(f_n)$ can be computed analytically, and how to find a tight or exact bounds of the expectation of the local likelihood under $q(f_n)$. We discuss several likelihood classes for which this approximation can be used.

3.2 Choice of variational distribution

If we choose a Gaussian variational distribution, $q(\mathbf{u}) = \mathcal{N}(\mathbf{u}; \mathbf{m}, \mathbf{C})$, the first term in the lower bound, that is a negative KL divergence, can be computed exactly,

$$\mathcal{F}_1 = -\mathrm{KL}(q(\mathbf{u})||p(\mathbf{u})) = -\frac{1}{2}\left[\mathrm{tr}(\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{C}) + (\mathbf{m}-\boldsymbol{\mu}_{\mathbf{u}})^{\intercal}\mathbf{K}_{\mathbf{uu}}^{-1}(\mathbf{m}-\boldsymbol{\mu}_{\mathbf{u}}) - M - \log\frac{\det\mathbf{C}}{\det\mathbf{K}_{\mathbf{uu}}}\right], \tag{12}$$

where $p(\mathbf{u}) = \mathcal{N}(\mathbf{u}; \boldsymbol{\mu}_{\mathbf{u}}, \mathbf{K}_{\mathbf{uu}})$ is the prior on the inducing variables. Typically, $\boldsymbol{\mu}_{\mathbf{u}} = \mathbf{0}$ which leads to a trivial simplication in the above equation,

$$\mathcal{F}_1 = -\mathrm{KL}(q(\mathbf{u})||p(\mathbf{u})) = -\frac{1}{2}\left[\mathrm{tr}(\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{C}) + \mathbf{m}^{\intercal}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{m} - M - \log\frac{\det\mathbf{C}}{\det\mathbf{K}_{\mathbf{uu}}}\right]. \tag{13}$$

The distribution $q(\mathbf{u})$ is parameterised by its mean $\mathbf{m}$ and covariance matrix $\mathbf{C}$, which means there are $M + M^2$ parameters in total. For a large M, this may not be desirable and could lead to problems with the optimisation of the lower bound. TODO: 1. Is there any transformation we could do here to reduce the number of parameters? 2. Can use a sparse precision parameterisation as in [8].

Choosing a single Gaussian distribution to represent the posterior of $\mathbf{u}$ may be too coarse and hence a multimodal distribution such as a mixture of Gaussian distributions may be a better choice for a multimodal posterior distribution [9, 10](TODO: this reduces significantly the number of variational parameters.).We could use a mixture of (uniformly) weighted Gaussians with isotropic covariances,

$$q(\mathbf{u}) = \frac{1}{K}\sum_{k=1}^{K} \mathcal{N}(\mathbf{u}; \mathbf{m}_k, \sigma_k^2\mathrm{I}). \tag{14}$$

However, this distribution leads to an intractabily in computing the KL term, as it involves a log of a sum of distributions. Fortunately, we can bound this using Jensen's inequality [9, 11] as follows,

$$\begin{aligned}
\mathcal{F}_1 &= -\mathrm{KL}(q(\mathbf{u})||p(\mathbf{u})) && (15)\\
&= -\int \mathrm{d}\mathbf{u} q(\mathbf{u}) \log\frac{q(\mathbf{u})}{p(\mathbf{u})} && (16)
\end{aligned}$$

$$= -\int \mathrm{d}\mathbf{u} q(\mathbf{u}) \log q(\mathbf{u}) + \int \mathrm{d}\mathbf{u} q(\mathbf{u}) \log p(\mathbf{u}) \tag{17}$$

$$\geq -\frac{1}{K}\sum_{k=1}^{K} \log \int \mathrm{d}\mathbf{u} q(\mathbf{u}) \mathcal{N}(\mathbf{u}; \mathbf{m}_k, \sigma_k^2 \mathrm{I}) + \frac{1}{K}\sum_{k=1}^{K} \int \mathrm{d}\mathbf{u} \mathcal{N}(\mathbf{u}; \mathbf{m}_k, \sigma_k^2 \mathrm{I}) \log p(\mathbf{u}), \tag{18}$$

where we have used the inequality to bound the first term in the equations above.
TODO: R. Turner's comment: there's been limited success using mixture of gaussians to parameterise the variational distribution, perhaps because of this bound is not tight!
The first term in the bound above involves a convolution of two Gaussian distributions and hence,

$$\mathcal{F}_{1a} = -\frac{1}{K}\sum_{k=1}^{K} \log \int \mathrm{d}\mathbf{u} q(\mathbf{u}) \mathcal{N}(\mathbf{u}; \mathbf{m}_k, \sigma_k^2 \mathrm{I}) \tag{19}$$

$$= -\frac{1}{K}\sum_{k=1}^{K} \log \frac{1}{K}\sum_{k'=1}^{K} \mathcal{N}(\mathbf{m}_k; \mathbf{m}_{k'}, (\sigma_k^2 + \sigma_{k'}^2)\mathrm{I}) \tag{20}$$

The second term is due to the assumption about the variational distribution,

$$\mathcal{F}_{1b} = \frac{1}{K}\sum_{k=1}^{K} \left[-\frac{M}{2} - \log \det \mathbf{K}_{\mathbf{uu}} - \frac{1}{2}(\boldsymbol{\mu}_{\mathbf{u}} - \mathbf{m}_k)^\intercal \mathbf{K}_{\mathbf{uu}}^{-1}(\boldsymbol{\mu}_{\mathbf{u}} - \mathbf{m}_k) + \mathrm{tr}(\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{C}_k) \right], \tag{21}$$

where $\mathbf{C}_k = \sigma_k^2 \mathrm{I}$. The expressions presented above are tractable to compute, and its gradient with respect to the variational parameters $\{\mathbf{m}_k, \sigma_k^2\}_{k=1}^K$ can be computed in closed-form. Note that when $K = 1$, the bound above is exact, and we obtain the result of the single Gaussian case presented at the beginning of this section. In what follows, we will present the result for the single Gaussian case, but the derivation for the mixture is straightforward.

3.3 Choice of local bounds

We have discussed the choice of $q(\mathbf{u})$ and how to deal with the first term of the bound $\mathcal{F}$. In this section, we provide a treatment for the second term,

$$\mathcal{F}_2 = \sum_{n=1}^{N} \int \mathrm{d}f_n q(f_n) \log p(y_n | f_n), \tag{22}$$

where

$$q(f_n) = \int \mathrm{d}\mathbf{u} q(\mathbf{u}) p(f_n | \mathbf{u}) \tag{23}$$

$$= \int \mathrm{d}\mathbf{u} \mathcal{N}(\mathbf{u}; \mathbf{m}, \mathbf{C}) \mathcal{N}(f_n; a_n \mathbf{u}, b_n) \tag{24}$$

$$= \mathcal{N}(f_n; a_n \mathbf{m}, b_n + a_n \mathbf{C} a_n^\intercal), \tag{25}$$

and $a_n = \mathbf{K}_{f_n \mathbf{u}} \mathbf{K}_{\mathbf{uu}}^{-1}$, $b_n = k_{f_n, f_n} - \mathbf{K}_{f_n \mathbf{u}} \mathbf{K}_{\mathbf{uu}}^{-1} \mathbf{K}_{\mathbf{u} f_n}$ in the equations above. Let denote $\mu_n = a_n \mathbf{m}$ and $v_n = b_n + a_n \mathbf{C} a_n^\intercal$. We note that $\mathcal{F}_2$ involves one dimensional integrals which are the expectation of local likelihood terms wrt the distribution $q(f_n)$. Since $q(f_n)$ is a Gaussian distribution, the local integral can be computed exactly or can be bounded so that a tractable lower bound of $\mathcal{F}_2$ can be used instead of the original intractable integrals.

3.3.1 Poisson likelihood for count data regression

Consider,

$$p(y_n = k | f_n) = \frac{\exp(y_n f_n - \exp(f_n))}{y_n!}, \tag{26}$$

which means,

$$\log p(y_n | f_n) = y_n f_n - \exp(f_n) - \log y_n!. \tag{27}$$

Therefore,

$$\mathcal{F}_{2,n} = \int \mathrm{d}f_n q_(f_n) \log p(y_n | f_n) \tag{28}$$

$$= \int \mathrm{d}f_n \mathcal{N}(f_n; \mu_n, v_n) \left[y_n f_n - \exp(f_n) - \log y_n! \right] \tag{29}$$

$$= y_n \mu_n - \exp(\mu_n + \frac{v_n}{2}) - \log y_n! \tag{30}$$

$$= y_n a_n \mathbf{m} - \exp(\frac{b_n}{2} + a_n \mathbf{m} + \frac{a_n \mathbf{C} a_n'}{2}) - \log y_n! \tag{31}$$

In words, $\mathcal{F}_2$ can be computed analytically when the likelihood is a Poisson distribution. We have used the exponential reverse-link function here. See [2] for a treatment for a different reverse-link function.

3.3.2 Bernoulli logistic likelihood for logit regression or binary classification

The bound of the log partition function can be found using Taylor's approximation [12] or the Fenchel inequality [13]. These bounds have been used extensively, but recently shown that they are not tight, especially, when being converted back to the logistic function [14]. Motivated by the performance of the piecewise linear/quadratic bound in [14, 15], we apply this technique for Bernoulli logistic and ordinal regression problems, in this section and next section respectively.

Consider the logistic function for binary classification or logit regression,

$$p(y_n = 1|f_n) = \frac{1}{1 + \exp(-f_n)} = \frac{\exp(y_n f_n)}{1 + \exp(f_n)}. \tag{32}$$

The log likelihood is,

$$\log p(y_n|f_n) = y_n f_n - \log(1 + \exp(f_n)) \tag{33}$$

Using the piecewise bound in [14], we obtain,

$$\mathcal{F}_{2,n} = \int \mathrm{d}f_n q(f_n) \left[y_n f_n - \log(1 + \exp(f_n)\right] \tag{34}$$

$$\geq y_n \mu_n - \int \mathrm{d}f_n q(f_n) \mathrm{bound}(\log(1 + \exp(f_n)) \tag{35}$$

$$= y_n \mu_n - \sum_{r=1}^{R} \int_{l_r}^{h^r} \mathrm{d}f_n q(f_n)(a_{n,r} f_n^2 + b_{n,r} f_n + c_{n,r}) \tag{36}$$

$$= y_n \mu_n - \sum_{r=1}^{R} a_{n,r} \mathrm{E}_{l_r}^{h_r}[f_n^2|\mu_n, v_n] + b_{n,r} \mathrm{E}_{l_r}^{h_r}[f_n^2|\mu_n, v_n] + c_{n,r} \mathrm{E}_{l_r}^{h_r}[f_n^0|\mu_n, v_n], \tag{37}$$

where $\mathrm{E}_l^h[f^m|\mu, \sigma^2] = \int_l^h \mathrm{d}f \mathcal{N}(f; \mu, \sigma^2) f^m$. Importantly, the gradients of the expectations in the equations above wrt the mean and variance of the integrating distribution $\{\mu_n, v_n\}$, can be computed, which means the gradients wrt the variational parameters and the inducing inputs $\mathbf{x_u}$ can be computed. The forms of the truncated Gaussian moments can be found in, e.g. [16, pp. 144-145].

3.3.3 Cumulative logistic likelihood for ordinal regression

The following is an extension of [16, Ch. 6, pp. 116-117]. Consider the cumulative logistic likelihood,

$$p(y_n \leq k|f_n) = \sigma(\phi_k - f_n), \tag{38}$$

which means

$$p(y_n = k|f_n) = \sigma(\phi_k - f_n) - \sigma(\phi_{k-1} - f_n) \tag{39}$$

$$= \frac{e^{f_n}(e^{-\phi_{k-1}} - e^{-\phi_k})}{[1 + e^{-(\phi_{k-1} - f_n)}][1 + e^{-(\phi_{k-1} - f_n)}]} \tag{40}$$

hence its log,

$$\log p(y_n = k|f_n) = f_n - \log(e^{-\phi_{k-1}} - e^{-\phi_k}) - \mathrm{llp}(f_n - \phi_k) - \mathrm{llp}(f_n - \phi_{k-1}). \tag{41}$$

Similar to the previous section, we can obtain the bound on the llp function using R linear/quadratic pieces. The expectation of these pieces under a one dimensional Gaussian and their corresponding gradients can also be obtained as in the above sections.

3.3.4 Laplace likelihood

Consider the Laplace likelihood,

$$p(y_n|f_n) = \frac{1}{2b}\exp(-\frac{|y_n - f_n|}{b}) \tag{42}$$

Taking the log gives,

$$\log p(y_n|f_n) = -\log(2b) - \frac{|y_n - f_n|}{b} \tag{43}$$

$$= \begin{cases} \frac{-y_n+f_n}{b} - \log(2b) & \text{if } y_n \geq f_n \\ \frac{y_n-f_n}{b} - \log(2b) & \text{if } y_n < f_n \end{cases} \tag{44}$$

Hence the intergral of $\log p(y_n|f_n)$ is analytically tractable and involves evaluating the Gaussian cummulative distribution ... TODO

3.3.5 Exponential likelihood

Consider the exponential likelihood,

$$p(y_n|f_n) = \frac{\exp(-y_n/\exp(f_n))}{\exp(f_n)} \tag{45}$$

Taking the log gives,

$$\log p(y_n|f_n) = \frac{-y_n}{\exp(f_n)} - f_n, \tag{46}$$

which means the intergral will be tractable as in the case of the Poisson likelihood with the exponential inverse-link function above.

3.3.6 Multinomial logistic likelihood for categorical regression or multiclass classification

Consider the likelihood for multinomial logistic regression,

$$p(y_n = k|\mathbf{f}_i) = \exp(f_{i,k} - \text{lse}(\mathbf{f}_i)) \tag{47}$$

TODO: [15–18]

3.3.7 Other likelihood classes that cannot be analytically bounded

If the bound on the log likelihood term $g(f_n) = \log(y_n|f_n)$ cannot be found analytically, we can use stochastic optimisation to obtain the unbiased estimate of $\mathcal{F}_{2,n,k}$ and its gradients wrt the mean and variance of the distribution $q_k(f_n)$. That is,

$$\mathcal{F}_{2,n,k} \approx \frac{1}{T}\sum_{t=1}^{T} g(f_t) \quad , f_t \sim q_k(f_n) \tag{48}$$

Furthermore, if $q_k(f_n)$ is parameterised by the mean μ_n and covariance v_n, the gradient of the expectation can be expressed as an expectation of a gradient. This is due to Price's theorem [19] and Bonnet's theorem [20], which have been employed recently for deep models [8]. That is,

$$\frac{\mathrm{d}}{\mathrm{d}\mu_{n,k}}\mathcal{F}_{2,n,k} = \frac{\mathrm{d}}{\mathrm{d}\mu_{n,k}}\mathrm{E}_{q_k(f_n)}[g(f_n)] = \mathrm{E}_{q_k(f_n)}\left[\frac{\mathrm{d}}{\mathrm{d}f_n}g(f_n)\right] \tag{49}$$

$$\frac{\mathrm{d}}{\mathrm{d}v_{n,k}}\mathcal{F}_{2,n,k} = \frac{\mathrm{d}}{\mathrm{d}v_{n,k}}\mathrm{E}_{q_k(f_n)}[g(f_n)] = \mathrm{E}_{q_k(f_n)}\left[\frac{\mathrm{d}^2}{\mathrm{d}f_n^2}g(f_n)\right] \tag{50}$$

$$\tag{51}$$

The above equation means that the gradients can be estimated using a Monte Carlo summation. The condition for the following theorem is that $\log p(y_n|f_n)$ needs to be twice differentiable wrt f_n.

Two potential issues with this approach is:

- Samples drawn from $q_k(\mathbf{u})$ may not be important as they fall into regions in which $g_n(\mathbf{u})$ is too small.
- Variance of the unbiased estimates may be large, TODO: variance reduction techniques?: Rao-Blackwellisation, change of variables?, read [8, 21].

However, the problem may not be as worse as it sounds, as the expectation above is only one dimensional and therefore, the number of samples needed is much smaller compared to similar approaches in high dimensions [8, 22].

3.4 Complexity

$\mathcal{O}(NM^2)$

4 Can this be applied to non-conjugate GPLVMs?

The answer is no if you are looking for an analytic solution because of the non-linear dependencies of $\mathbf{X}$ in $p(\mathbf{F}|\mathbf{U})$, but stochastic approximation could be used. Again, getting this to work in high dimensions could be challenging.

References

[1] J. Hensman, A. Matthews, and Z. Ghahramani, "Scalable variational Gaussian process classification," 2014.

[2] C. Lloyd, T. Gunter, M. A. Osborne, and S. J. Roberts, "Variational inference for Gaussian process modulated Poisson processes," 2014.

[3] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[4] J. Quiñonero-Candela and C. E. Rasmussen, "A unifying view of sparse approximate Gaussian process regression," *The Journal of Machine Learning Research*, vol. 6, pp. 1939–1959, 2005.

[5] E. Snelson and Z. Ghahramani, "Sparse Gaussian processes using pseudo-inputs," in *Advances in Neural Information Processing Systems 19*, pp. 1257–1264, MIT press, 2006.

[6] M. K. Titsias, "Variational learning of inducing variables in sparse Gaussian processes," in *International Conference on Artificial Intelligence and Statistics*, pp. 567–574, 2009.

[7] T. Salimans, "Markov chain Monte Carlo and variational inference: Bridging the gap," 2014.

[8] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *Proceedings of the 31st International Conference on Machine Learning*, pp. 1278–1286, 2014.

[9] S. Gershman, M. Hoffman, and D. M. Blei, "Nonparametric variational inference," in *Proceedings of the 29th International Conference on Machine Learning*, pp. 663–670, 2012.

[10] T. Nguyen and E. Bonilla, "Automated variational inference for Gaussian process models," in *Advances in Neural Information Processing Systems* (N. Lawrence and C. Cortes, eds.), vol. 26, 2014.

[11] M. Huber, T. Bailey, H. Durrant-Whyte, and U. Hanebeck, "On entropy approximation for Gaussian mixture random vectors," in *Multisensor Fusion and Integration for Intelligent Systems, 2008. MFI 2008. IEEE International Conference on*, pp. 181–188, Aug 2008.

[12] D. Böhning, "Multinomial logistic regression algorithm," *Annals of the Institute of Statistical Mathematics*, vol. 44, no. 1, pp. 197–200, 1992.

[13] T. S. Jaakkola and M. I. Jordan, "Bayesian parameter estimation via variational methods," *Statistics and Computing*, vol. 10, no. 1, pp. 25–37, 2000.

[14] B. M. Marlin, M. E. Khan, and K. P. Murphy, "Piecewise bounds for estimating Bernoulli-logistic latent Gaussian models," in *Proceedings of the 28th International Conference on Machine Learning*, 2011.

[15] M. E. Khan, S. Mohamed, and K. P. Murphy, "Fast Bayesian inference for non-conjugate Gaussian process regression," in *Advances in Neural Information Processing Systems*, pp. 3140–3148, 2012.

[16] M. E. Khan, *Variational learning for latent Gaussian models of discrete data*. PhD thesis, The University of British Columbia, 2012.

[17] G. Bouchard, "Efficient bounds for the softmax function and applications to inference in hybrid models," in *NIPS 2007 Workshop on Approximate Bayesian Inference in Continuous/Hybrid Systems*, 2007.

[18] D. Blei and J. Lafferty, "Correlated topic models," *Advances in neural information processing systems*, vol. 18, p. 147, 2006.

[19] R. Price, "A useful theorem for nonlinear devices having Gaussian inputs," *Information Theory, IRE Transactions on*, vol. 4, pp. 69–72, June 1958.

[20] G. Bonnet, "Transformations des signaux aléatoires a travers les systèmes non linéaires sans mémoire," *Annales des Télécommunications*, vol. 19, no. 9-10, pp. 203–220, 1964.

[21] R. Ranganath, S. Gerrish, and D. Blei, "Black box variational inference," in *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, pp. 814–822, 2014.

[22] M. Titsias and M. Lázaro-Gredilla, "Doubly stochastic variational bayes for non-conjugate inference," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 1971–1979, 2014.