Piecewise Bounds for Estimating Bernoulli-Logistic Latent Gaussian Models

Benjamin M. Marlin BMARLIN@CS.UBC.CA
Mohammad Emtiyaz Khan EMTIYAZ@CS.UBC.CA
Kevin P. Murphy MURPHYK@CS.UBC.CA
University of British Columbia, Vancouver, BC, Canada V6T 1Z4

Abstract

Bernoulli-logistic latent Gaussian models (bLGMs) are a useful model class, but accurate parameter estimation is complicated by the fact that the marginal likelihood contains an intractable logistic-Gaussian integral. In this work, we propose the use of fixed piecewise linear and quadratic upper bounds to the logistic-log-partition (LLP) function as a way of circumventing this intractable integral. We describe a framework for approximately computing minimax optimal piecewise quadratic bounds, as well a generalized expectation maximization algorithm based on using piecewise bounds to estimate bLGMs. We prove a theoretical result relating the maximum error in the LLP bound to the maximum error in the marginal likelihood estimate. Finally, we present empirical results showing that piecewise bounds can be significantly more accurate than previously proposed variational bounds.

1. Introduction

Latent Gaussian Models (LGMs) are an important class of probabilistic models that includes factor analysis and probabilistic principal components analysis for continuous data (Tipping & Bishop, 1999), as well as binary and multinomial factor analysis for discrete data (Wedel & Kamakura, 2001; Collins et al., 2002; Mohamed et al., 2008; Khan et al., 2010). The generative process for such models begins by sampling a latent vector $\mathbf{z}$ from the latent Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The canonical parameters of the distribution over the visible variables $\mathbf{y}$ is then given by a linear function $\mathbf{Wz}+\mathbf{b}$ of the latent variables $\mathbf{z}$. Different LGM models are obtained using different likelihoods for the visible variables and different restrictions on the model parameters.

The main difficulty with the LGM class is that the latent variables $\mathbf{z}$ must be integrated away to obtain the marginal likelihood needed for standard maximum likelihood learning. This integration can be carried out analytically in Gaussian-likelihood LGMs because the model is jointly Gaussian in the latent factors and the visible variables. Other likelihood models lack this property, resulting in intractable integrals in the marginal likelihood. The special case of discrete LGMs based on a Bernoulli-logistic or multinomial-softmax likelihood model has been well studied and several previous estimation approaches have been proposed. Collins et al. (2002) propose an approach based on maximizing over the latent variables instead of marginalizing over them in the case of exponential family factor analysis (eFA). Mohamed et al. (2008) propose sampling from the model posterior using Hamiltonian Monte Carlo, again for eFA. Jaakkola & Jordan (1996) propose a variational approach based on an adjustable quadratic bound to the logistic-log-partition function. Khan et al. (2010) propose a related variational approach for multinomial factor analysis based on a bound due to Bohning (1992).

In this work, we focus on the Bernoulli-logistic LGM class for binary data and adopt a strategy of upper bounding the logistic-log-partition (LLP) function $\text{llp}(x) = \log(1 + \exp(x))$. Our main contribution is the proposal of *fixed, piecewise linear and quadratic bounds* as a more accurate replacement for the variational quadratic bounds proposed by Jaakkola & Jordan (1996) and Bohning (1992). Piecewise bounds have the important property that their maximum error is bounded and can be driven to zero by increasing the number of pieces.

Appearing in *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, WA, USA, 2011.

We use recent results from Hsiung et al. (2008) to compute minimax optimal linear bounds and introduce a novel optimization framework for minimax fitting of piecewise quadratic bounds. We show that piecewise quadratic bounds can be ten times more accurate than linear bounds using the same number of pieces at little additional computational cost. We prove a theoretical result relating the maximum error in the logistic-log-partition bound to the maximum error in the marginal likelihood estimate. Similar theoretical results do not exist for variational quadratic bounds. Finally, we apply the bounds to several Bernoulli-logistic LGM (bLGM) models including Bernoulli-logistic latent Gaussian graphical models (bLGGMs) and Bernoulli-logistic factor analysis (bFA). We find significant improvements over the previous variational quadratic bounds.

2. Bernoulli-Logistic LGMs

In this section, we introduce a general Bernoulli-logistic LGM that subsumes Bernoulli-logistic factor analysis (bFA) and Bernoulli-logistic latent Gaussian graphical models (bLGGMs). We denote the visible data vectors by $\mathbf{y}_n$ and the latent vectors by $\mathbf{z}_n$. In general, $\mathbf{y}_n$ and $\mathbf{z}_n$ will have dimensions D and L respectively with $\mathbf{y}_n \in \{0,1\}^D$ and $\mathbf{z}_n \in \mathbb{R}^L$. $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ denote the mean and covariance of the latent Gaussian as seen in Equation 1. The Bernoulli likelihood is defined through the logistic function in Equation 2, which is in turn defined through the logistic-log-partition function given in Equation 4. The mapping between the latent space and the canonical parameter space for each visible dimension d is specified by a length-L weight vector $\mathbf{W}_d$ and a scalar offset b_d, as in Equation 3. Let $\mathbf{W}$ be the matrix with $\mathbf{W}_d$ as rows. We can see that integrating over the latent variable $\mathbf{z}_n$, which is necessary to compute the marginal likelihood, introduces an intractable logistic-Gaussian integral.

$$p(\mathbf{z}_n|\boldsymbol{\theta}) = \mathcal{N}(\mathbf{z}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{1}$$

$$p(\mathbf{y}_n|\mathbf{z}_n, \boldsymbol{\theta}) = \prod_{d=1}^{D_d} \exp(y_{dn}\eta_{dn} - \mathrm{llp}(\eta_{dn})) \tag{2}$$

$$\eta_{dn} = \mathbf{W}_d \mathbf{z}_n + b_d \tag{3}$$

$$\mathrm{llp}(\eta) = \log(1 + \exp(\eta)) \tag{4}$$

As mentioned in the introduction, different binary models can be obtained by restricting the general model in different ways. The prior mean $\boldsymbol{\mu}$ and the offset $\mathbf{b}$ are interchangeable in all the models we consider so we opt to use the mean only. We obtain the bFA model by assuming that $L \leq D$ and $\boldsymbol{\Sigma}$ is the identity matrix, while $\mathbf{W}$ and $\boldsymbol{\mu}$ are unrestricted. Conversely, we obtain the bLGGM by assuming that $D = L$ and $\mathbf{W}$ is the identity matrix, while $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unrestricted. We obtain a sparse bLGGM (sbLGGM) model by additionally placing a Gaussian graphical lasso prior $p(\boldsymbol{\Sigma}^{-1}|\lambda) \propto \exp(-\lambda||\boldsymbol{\Sigma}^{-1}||_1)$ on the precision matrix $\boldsymbol{\Sigma}^{-1}$. The main difference between the three models is that when $L < D$, the bFA model assumes a low-rank structure on the canonical parameters. The bLGGM instead assumes a graphical structure which is more appropriate when latent variables have sparse interactions. The sbLGGM model can further enforce sparsity in the latent graph.

3. Bounds on the LLP Function

In this section, we briefly review the existing variational upper bounds on the logistic-log-partition function (LLP) due to Jaakkola & Jordan (1996) and Bohning (1992). We then move to the piecewise linear and quadratic bounds that form the focus of this paper. The most important feature of all of these bounds is that their expectations with respect to a univariate Gaussian distribution can be obtained in closed form, providing a lower bound on the log marginal likelihood that can be used for tractable model estimation.

3.1. The Jaakkola Bound

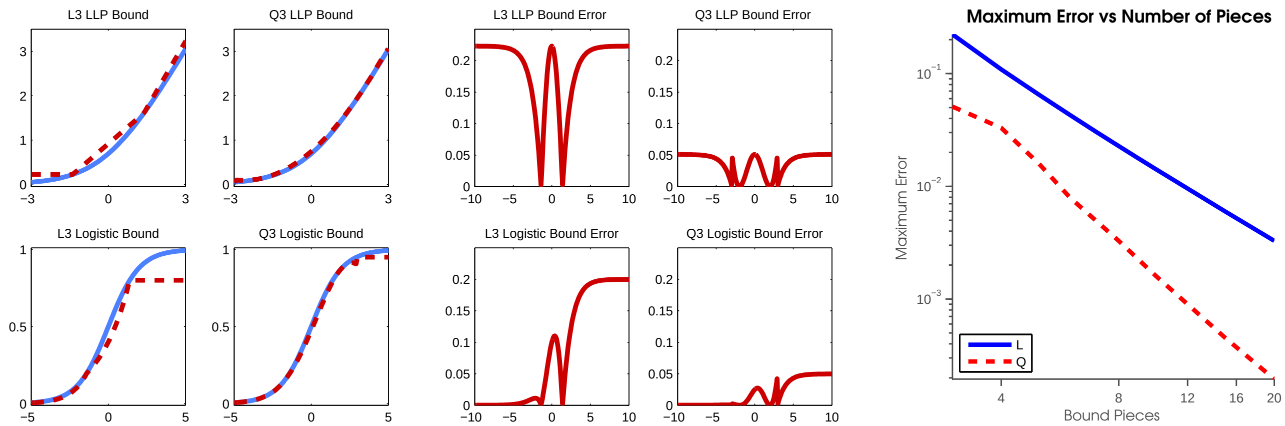
The variational quadratic bound introduced by Jaakkola & Jordan (1996) can be derived through Fenchel duality and has been quite widely used. The bound is given by $\mathrm{llp}(x) \leq a_\xi x^2 + b_\xi x + c_\xi$ where $a_\xi = \lambda_\xi$, $b_\xi = 1/2$, $c_\xi = -\lambda_\xi \xi^2 - \frac{1}{2}\xi + \mathrm{llp}(\xi), \lambda_\xi = \frac{1}{2\xi}(\frac{1}{1+e^{-\xi}} - \frac{1}{2})$. Here ξ is a scalar variational parameter that must be optimized to maximize the approximate marginal likelihood.

3.2. The Bohning Bound

Bohning's bound is a lesser known quadratic variational bound derived from a Taylor series expansion of the LLP function. It is faster to optimize than Jaakkola's bound as it has fixed curvature (Bohning, 1992), but it is less accurate. The bound is given by $\mathrm{llp}(x) \leq x^2/8 + b_\psi x + c_\psi$, where $b_\psi = (1+e^{-\psi})^{-1} - \psi/4$ and $c_\psi = \psi^2/8 - (1+e^{-\psi})^{-1}\psi + \mathrm{llp}(\psi)$. ψ is a scalar variational parameter that must be optimized to maximize the approximate marginal likelihood.

3.3. Piecewise Linear and Quadratic Bounds

The LLP bounds proposed by Jaakkola & Jordan (1996) and Bohning (1992) can be quite accurate locally, however, the induced bounds on the marginal likelihood can be quite inaccurate. This is due to



(a) LLP and Logistic Bounds (b) LLP and Logistic Bound Errors (c) Error vs Number of Pieces

Figure 1. Figure (a) shows a comparison of three-piece linear (L3) and quadratic (Q3) upper bounds on the LLP function (top) and the induced lower bounds on the logistic function (bottom). Figure (b) shows a comparison of the error in the three-piece linear and quadratic bounds on the LLP function (top) and induced bounds on the logistic function (bottom). Figure (c) shows the maximum error in the LLP bounds as a function of the number of pieces in the bound.

the fact that the marginal likelihood integrates over the whole range of the approximation and any single-piece quadratic function will have unbounded error relative to the LLP function. For this reason, we propose the use of piecewise linear and quadratic LLP bounds, which have a finite maximum error that can be driven to zero by increasing the number of pieces.

An R-piece quadratic bound consists of R intervals defined by $R+1$ threshold points $t_0, ..., t_R$ such that $t_r < t_{r+1}$, and R quadratic functions $a_r x^2 + b_r x + c_r$. An R-piece linear bound is a special case where $a_r = 0$ for all r. We fix the first and last threshold points to $-\infty$ and ∞, respectively. For simplicity, we use $\boldsymbol{\alpha}$ to denote the complete set of bound parameters including the threshold points and quadratic coefficients.

The minimax optimal R-piece quadratic upper bound problem for the LLP function is defined in Equation 5. The objective function is simply the maximum gap between the piecewise quadratic bound and the LLP function. The first constraint is required to ensure that each quadratic function is an upper bound over the interval it is defined on. The second constraint ensures that the thresholds are monotonically increasing. The final constraint ensures that the curvature of each quadratic function is non-negative.

$$\min_{\boldsymbol{\alpha}} \max_{r \in \{1,..,R\}} \max_{t_{r-1} \le x < t_r} a_r x^2 + b_r x + c_r - \mathrm{llp}(x) \quad (5)$$
$$\begin{aligned} & a_r x^2 + b_r x + c_r - \mathrm{llp}(x) \ge 0 && \forall\, r, x \in [t_{r-1}, t_r] \\ & t_r - t_{r-1} > 0 && \forall r \in \{1,..,R\} \\ & a_r \ge 0 && \forall r \in \{1,..,R\} \end{aligned}$$

We now reformulate the problem to remove all of the constraints. The second and third constraints can be dealt with using trivial reparameterizations. The first constraint can be replaced with an equality, which can then be solved for c_r yielding $c_r = -\left(\min_{t_{r-1} \le x < t_r} a_r x^2 + b_r x - \mathrm{llp}(x)\right)$. This substitution is essentially finding the minimum gap between the quadratic and the LLP function on each interval and setting it to zero. This converts any quadratic with positive curvature into an upper bound on the LLP function over the corresponding interval. The final unconstrained problem is given below.

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \max_{r \in \{1,..,R\}} & \left(\max_{t_{r-1} \le x < t_r} a_r x^2 + b_r x - \mathrm{llp}(x) \right) \\ & - \left(\min_{t_{r-1} \le x < t_r} a_r x^2 + b_r x - \mathrm{llp}(x) \right) \end{aligned} \quad (6)$$

The main difficulty with this optimization problem comes from the fact that the inner maximization and minimization problems apparently have no closed-form solutions. However, global solutions for both the maximization and minimization problems can be easily found by numerical optimization as the function $ax^2 + bx - \mathrm{llp}(x)$ has at most three critical points for any choice of a and b. However, this means that the outer minimization must be conducted using a derivative-free optimization algorithm since the objective function itself involves solving a non-linear optimization problem. In this work, we use the classical Nelder-Mead method (Nelder & Mead, 1965). In the linear case, Hsiung, Kim, and Boyd (2008) have proposed a constructive search method for determining minimax optimal coefficients and break points. Their work was motivated by the need to obtain linear approximations

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{n=1}^{N}\log\int p(\mathbf{y}_n|\mathbf{z},\boldsymbol{\theta})p(\mathbf{z}|\boldsymbol{\theta})d\mathbf{z} = \frac{1}{N}\sum_{n=1}^{N}\log\int \frac{q_n(\mathbf{z}|\boldsymbol{\gamma}_n)}{q_n(\mathbf{z}|\boldsymbol{\gamma}_n)}p(\mathbf{y}_n|\mathbf{z},\boldsymbol{\theta})p(\mathbf{z}|\boldsymbol{\theta})d\mathbf{z} \tag{7}$$

$$\mathcal{L}_J(\boldsymbol{\theta},\boldsymbol{\gamma}) = \frac{1}{N}\sum_{n=1}^{N} E_{q_n(\mathbf{z}|\boldsymbol{\gamma}_n)}[\log p(\mathbf{y}_n|\mathbf{z},\boldsymbol{\theta})] + E_{q_n(\mathbf{z}|\boldsymbol{\gamma}_n)}[\log p(\mathbf{z}|\boldsymbol{\theta})] - E_{q_n(\mathbf{z}|\boldsymbol{\gamma})}[\log q_n(\mathbf{z}|\boldsymbol{\gamma}_n)] \tag{8}$$

$$\mathcal{L}_{QJ}(\boldsymbol{\theta},\boldsymbol{\gamma}) = \frac{1}{N}\sum_{n=1}^{N}\sum_{d=1}^{D} E_{q_n(\mathbf{z}|\boldsymbol{\gamma}_n)}\left[y_{dn}\mathbf{W}_d^T\mathbf{z} - B_{\boldsymbol{\alpha}}(\mathbf{W}_d^T\mathbf{z})\right] - D_{KL}(q_n(\mathbf{z}|\boldsymbol{\gamma}_n)||p(\mathbf{z}|\boldsymbol{\theta})) \tag{9}$$

$$D_{KL}(q_n(\mathbf{z}|\boldsymbol{\gamma}_n)||p(\mathbf{z}|\boldsymbol{\theta})) = \frac{1}{2}\left(\log|\boldsymbol{\Sigma}| - \log|\mathbf{V}_n| + \mathrm{tr}(\mathbf{V}_n\boldsymbol{\Sigma}^{-1}) + (\mathbf{m}_n-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{m}_n-\boldsymbol{\mu}) - D\right) \tag{10}$$

to LLP constraints in the context of geometric programming. We use their method for computing piecewise linear bounds in this work.

Figure 1 illustrates the gain in accuracy obtained using piecewise quadratic bounds instead of piecewise linear bounds. Figure 1(a) and 1(b) contrast the accuracies obtained using three-piece linear and quadratic bounds while figure 1(c) shows the maximum error of both linear and quadratic bounds as a function of the number of pieces. We see that the piecewise quadratic bounds can be more than an order of magnitude more accurate than the piecewise linear bounds using the same number of pieces. Conversely, it can take more than double the number of pieces for a piecewise linear bound to approach the same accuracy as a piecewise quadratic bound.

4. Learning with Piecewise Bounds

We propose a general expectation maximization (EM) algorithm (Dempster et al., 1977) for bLGMs using piecewise quadratic bounds to overcome the intractable logistic-Gaussian integral. This EM algorithm subsumes both the piecewise linear case and the single variational quadratic bound case.

4.1. Bounding the Marginal Likelihood

We begin with the intractable log marginal likelihood $\mathcal{L}(\boldsymbol{\theta})$ given in Equation 7 and introduce a variational posterior distribution $q_n(\mathbf{z}|\boldsymbol{\gamma}_n)$ for each data case. We use a full covariance Gaussian posterior with mean $\mathbf{m}_n$ and covariance $\mathbf{V}_n$. The full set of variational parameters is thus $\boldsymbol{\gamma}_n = [\mathbf{m}_n, \mathbf{V}_n]$. We apply Jensen's inequality to obtain an initial lower bound $\mathcal{L}_J(\boldsymbol{\theta},\boldsymbol{\gamma})$, as shown in Equation 8. The second and third terms in $\mathcal{L}_J(\boldsymbol{\theta},\boldsymbol{\gamma})$ are easily seen to be the negative of the Kullback−Leibler divergence from the variational Gaussian posterior $q_n(\mathbf{z}|\mathbf{m}_n,\mathbf{V}_n)$ to the Gaussian prior distribution $p(\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\Sigma})$, which has a well-known closed-form expression, as seen in Equation 10. Finally, we apply a piecewise quadratic bound $B_{\boldsymbol{\alpha}}(x)$ to approximate the LLP function, obtaining the final bound $\mathcal{L}_{QJ}(\boldsymbol{\theta},\boldsymbol{\gamma})$ given in Equation 9.

We now expand the likelihood term in $\mathcal{L}_{QJ}(\boldsymbol{\theta},\boldsymbol{\gamma})$. The expectation of the linear term is straightforward. The expectation of the bound requires introducing a change of variables $\eta = \mathbf{W}_d^T\mathbf{z}$ for all d. We then have a tractable expectation of a piecewise quadratic function with respect to a univariate Gaussian distribution with parameters $\tilde{\boldsymbol{\gamma}}_{dn} = \{\tilde{m}_{dn}, \tilde{v}_{dn}\}$ as defined below.

$$\begin{aligned}
&E_{q_n(\mathbf{z}|\boldsymbol{\gamma}_n)}[\log p(\mathbf{y}_n|\mathbf{z},\boldsymbol{\theta})] \\
&\geq \sum_{d=1}^{D}\left(y_{dn}\mathbf{W}_d^T\mathbf{m}_n - E_{q_n(\mathbf{z}|\boldsymbol{\gamma}_n)}[B_{\boldsymbol{\alpha}}(\mathbf{W}_d^T\mathbf{z})]\right) \\
&= \sum_{d=1}^{D}\left(y_{dn}\mathbf{W}_d^T\mathbf{m}_n - E_{q_n(\eta|\tilde{\boldsymbol{\gamma}}_{dn})}[B_{\boldsymbol{\alpha}}(\eta)]\right) \\
&\tilde{\boldsymbol{\gamma}}_{dn} = \{\tilde{m}_{dn}, \tilde{v}_{dn}\},\ \tilde{m}_{dn} = \mathbf{W}_d^T\mathbf{m}_n,\ \tilde{v}_{dn} = \mathbf{W}_d^T\mathbf{V}_n\mathbf{W}_d
\end{aligned}$$

Finally, we apply the definition of the piecewise quadratic bound $B_{\boldsymbol{\alpha}}(x)$ after the change of variables. The result takes the form of a sum of truncated expectations of each piece of the quadratic bound. To simplify the notation, we have introduced the special function $f_r(\mu,\sigma^2,\boldsymbol{\alpha})$ to represent the expectation of the r^{th} piece of the bound with parameters $\boldsymbol{\alpha}$ under a Gaussian with mean μ and variance σ^2. Closed-form expressions for the truncated moments needed to compute the special function $f_r(\mu,\sigma^2,\boldsymbol{\alpha})$ are given in an online appendix to this paper[1].

$$\begin{aligned}
E_{q_n(\eta_{dn}|\tilde{\boldsymbol{\gamma}}_{dn})}[B_{\boldsymbol{\alpha}}(\eta)] &= \sum_{r=1}^{R} f_r(\tilde{m}_{dn},\tilde{v}_{dn},\boldsymbol{\alpha}) \\
&= \sum_{r=1}^{R}\int_{t_{r-1}}^{t_r}(a_r\eta^2 + b_r\eta + c_r)\mathcal{N}(\eta|\tilde{m}_{dn},\tilde{v}_{dn})d\eta
\end{aligned}$$

[1] `http://www.cs.ubc.ca/~bmarlin/research/papers/trunctedGaussianMoments.pdf`

Note that if a piecewise linear bound is used instead of a piecewise quadratic bound, the coefficients a_r will all be zero and the bound on the marginal likelihood will only contain moments of order zero and one. Alternatively, if a single-piece quadratic variational bound is used, the formulas still apply with a single piece defined on $[-\infty, \infty]$.

4.2. A Generalized EM Algorithm

Learning the parameters of a binary LGM model requires optimizing the bound on the marginal likelihood given by $\mathcal{L}_{QJ}(\boldsymbol{\theta}, \boldsymbol{\gamma})$ with respect to the model parameters $\boldsymbol{\theta}$ and the variational posterior parameters $\boldsymbol{\gamma}$. Some of the parameter updates are not available in closed form and require numerical optimization, resulting in a generalized expectation maximization algorithm. The generalized E-Step requires numerically optimizing the variational posterior means and covariances. The generalized M-Step consists of a mix of closed-form updates and numerical optimization. We give the gradients or closed form updates as appropriate in Algorithm 1.

The gradients are given in terms of the gradients of the special function $f_r(\mu, \sigma^2, \boldsymbol{\alpha})$, which are given in the online appendix to this paper. We use limited memory BFGS to perform the updates that require numerical optimization. The piecewise bound $B_{\boldsymbol{\alpha}}(x)$ is computed in advance and fixed during learning and inference. For variational bounds, the free parameters in the LLP bound must be optimized for each data case and each iteration of the EM algorithm. For the sbLGGM model, we compute the maximum likelihood estimate of $\boldsymbol{\Sigma}$ and pass it to a standard convex optimization procedure for the Gaussian graphical Lasso on each iteration. This procedure returns the MAP estimate of $\boldsymbol{\Sigma}$ under the graphical lasso prior.

Algorithm 1 bLGM Generalized EM Algorithm

E-Step:

$$\frac{\partial \mathcal{L}_{QJ}}{\partial \mathbf{m}_{kn}} \leftarrow \sum_{d=1}^{D} y_{dn}\mathbf{W}_{dk} - \sum_{l=1}^{K} (\boldsymbol{\Sigma}^{-1})_{lk}(\mathbf{m}_{ln} - \boldsymbol{\mu}_l) - \sum_{r=1}^{R}\sum_{d=1}^{D} \mathbf{W}_{dk} \frac{\partial f_r(\tilde{m}_{dn}, \tilde{v}_{dn}, \boldsymbol{\alpha})}{\partial \tilde{m}_{dn}}$$

$$\frac{\partial \mathcal{L}_{QJ}}{\partial \mathbf{V}_{kl}} \leftarrow \frac{1}{2}(\boldsymbol{\Sigma}^{-1})_{kl} - \frac{1}{2}(\mathbf{V}_n^{-1})_{kl} - \sum_{r=1}^{R}\sum_{d=1}^{D} \mathbf{W}_{dk}\mathbf{W}_{dl} \frac{\partial f_r(\tilde{m}_{dn}, \tilde{v}_{dn}, \boldsymbol{\alpha})}{\partial \tilde{v}_{dn}}$$

M-Step:

$$\boldsymbol{\mu} \leftarrow \frac{1}{N}\sum_{n=1}^{N} \mathbf{m}_n$$

$$\boldsymbol{\Sigma} \leftarrow \frac{1}{N}\sum_{n=1}^{N} \left(\mathbf{V}_n + (\mathbf{m}_n - \boldsymbol{\mu})(\mathbf{m}_n - \boldsymbol{\mu})^T\right)$$

$$\frac{\partial \mathcal{L}_{QJ}}{\partial \mathbf{W}_{dk}} \leftarrow \sum_{n=1}^{N} \left[\mathbf{m}_{kn} \left(y_{dn} - \sum_{r=1}^{R} \frac{\partial f_r(\tilde{m}_{dn}, \tilde{v}_{dn}, \boldsymbol{\alpha})}{\partial \tilde{m}_{dn}} \right) - \left(2\sum_{l=1}^{K} \mathbf{V}_{kln}\mathbf{W}_{dk} \right) \sum_{r=1}^{R} \frac{\partial f_r(\tilde{m}_{dn}, \tilde{v}_{dn}, \boldsymbol{\alpha})}{\partial \tilde{v}_{dn}} \right]$$

5. Maximum Error Analysis

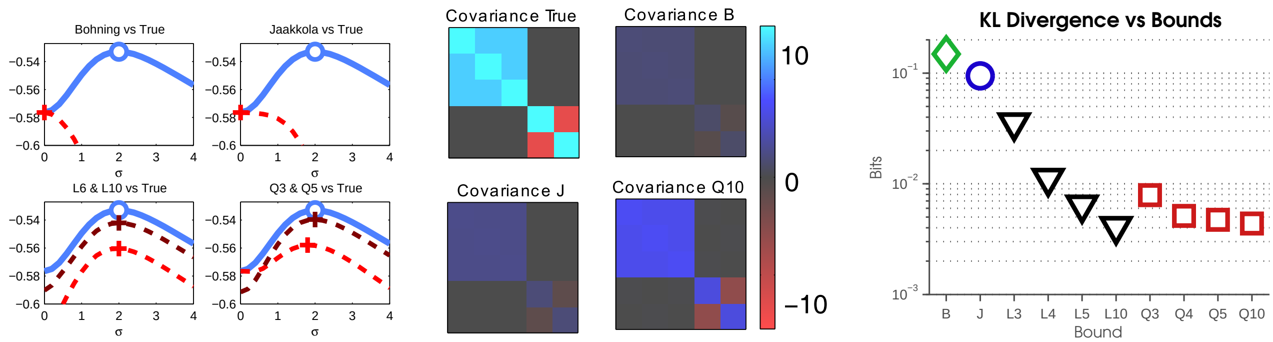
The fact that the piecewise linear and quadratic bounds on the LLP function both have a known finite maximum error ϵ_{max} means that we can easily bound the maximum error in the marginal likelihood or variational free energy due to the application of the LLP bound. Suppose we have a piecewise quadratic bound $B_{\boldsymbol{\alpha}}(x)$ (with a piecewise linear bound being a special case). We can write $B_{\boldsymbol{\alpha}}(x)$ as $\epsilon(x) + \mathrm{llp}(x)$ for any x, where $\epsilon(x) \geq 0$ is the point-wise error function of the bound $B_{\boldsymbol{\alpha}}(x)$. We let ϵ_{max} denote the maximum value of $\epsilon(x)$ over all x.

Theorem 5.1. *The loss in log marginal likelihood incurred by using the piecewise quadratic bound on the LLP function in addition to Jensen's inequality is at most $D\epsilon_{\max}$. In other words, $\mathcal{L}_J(\boldsymbol{\theta}, \boldsymbol{\gamma}) - \mathcal{L}_{QJ}(\boldsymbol{\theta}, \boldsymbol{\gamma}) \leq D\epsilon_{\max}$ for any $\boldsymbol{\theta}, \boldsymbol{\gamma}$. Furthermore, this bound is tight in the sense that a loss arbitrarily close to $D\epsilon_{\max}$ can be realized.*

Proof: Since the maximum error in each piece of the bound is less than or equal to ϵ_{max}, we have that $B_{\boldsymbol{\alpha}}(x) \leq \epsilon_{max} + \mathrm{llp}(x)$ for all x. We can bound the loss in the log marginal likelihood for each data case n as follows:

$$\begin{aligned}
&E_{q_n(\mathbf{z}|\boldsymbol{\gamma}_n)}[\log p(\mathbf{y}_n|\mathbf{z}, \boldsymbol{\theta})] \\
&\geq \sum_{d=1}^{D} E_{q_n(\mathbf{z}_d|\boldsymbol{\gamma})}\left[y_{dn}\mathbf{W}_{dd}\mathbf{z}_d - (\epsilon_{max} + \mathrm{llp}(\mathbf{W}_{dd}\mathbf{z}_d))\right] \\
&= -D\epsilon_{max} + E_{q_n(\mathbf{z}|\boldsymbol{\gamma})}[\log p(\mathbf{y}_n|\mathbf{z}, \boldsymbol{\theta})]
\end{aligned}$$

Since this holds for all n, we have that $\mathcal{L}_J(\boldsymbol{\theta}, \boldsymbol{\gamma}) \geq \mathcal{L}_{QJ}(\boldsymbol{\theta}, \boldsymbol{\gamma}) \geq \mathcal{L}_J(\boldsymbol{\theta}, \boldsymbol{\gamma}) - D\epsilon_{\max}$. A simple rearrangement of these terms yields the desired result that $\mathcal{L}_J(\boldsymbol{\theta}, \boldsymbol{\gamma}) - \mathcal{L}_{QJ}(\boldsymbol{\theta}, \boldsymbol{\gamma}) \leq D\epsilon_{\max}$. To show that the $D\epsilon_{\max}$ bound can be arbitrarily tight, we need only consider a bLGM where the mean on each dimension is located a point that achieves the maximum error



(a) 1D bLGGM Likelihoods (b) 5D bLGGM Covariance (c) 5D bLGGM KL Divergence

Figure 2. Figure (a) shows results for the 1D synthetic bLGGM experiment. We show the Bohning, Jaakkola, 6 and 10 piece linear and 3 and 5 piece quadratic bounds on the marginal likelihood. The bounds are shown in red with darker colors indicating more pieces. The true marginal likelihood is shown in blue. Markers show the true and estimated parameter values. Figure (b) shows the true covariance matrix for the synthetic 5D bLGGM experiment along with the covariance matrices estimated using the Bohning, Jaakkola, and 10 piece quadratic bounds (best viewed in color). Figure (c) shows the KL divergence between the true and estimated distributions for the 5D synthetic bLGGM experiment. We show results for the Bohning and Jaakkola bounds, as well as 3, 4, 5 and 10 piece linear and quadratic bounds.

ϵ_{max}. As the covariance Σ shrinks to 0, $\mathbf{m}_n$ will converge to $\boldsymbol{\mu}$ and $\mathbf{V}_n$ will converge to $\boldsymbol{\Sigma}$ for all n. The result will be an error of exactly $D\epsilon_{\max}$ $\square$.

A simple corollary of the above result is that the maximum difference between $\mathcal{L}_J(\boldsymbol{\theta},\boldsymbol{\gamma})$ and $\mathcal{L}_{QJ}(\boldsymbol{\theta},\boldsymbol{\gamma})$ at their respective optimal parameter values can not exceed $D\epsilon_{\max}$. We also note that the rate at which the error in the LLP bound decreases with number of pieces R is proportional to the rate at which $\mathcal{L}_{QJ}(\boldsymbol{\theta},\boldsymbol{\gamma})$ approaches $\mathcal{L}_J(\boldsymbol{\theta},\boldsymbol{\gamma})$. Hsiung et al. (2008) showed that the error in the optimal piecewise linear bound decreases with the approximate rate $\sqrt{2}/R^2$. The error in the piecewise quadratic bounds decreases at least this fast. This means that $\mathcal{L}_J(\boldsymbol{\theta},\boldsymbol{\gamma}) - \mathcal{L}_{QJ}(\boldsymbol{\theta},\boldsymbol{\gamma})$ approaches zero at a rate that is at least quadratic in the number of pieces. Finally, we note that analogous maximum error results hold if we directly bound the marginal likelihood $\mathcal{L}(\boldsymbol{\theta})$ by introducing the piecewise quadratic bound, obtaining $\mathcal{L}_Q(\boldsymbol{\theta})$ without first applying Jensen's inequality.

6. Experiments and Results

In this section we compare the piecewise linear and quadratic bounds to Jaakkola and Bohning's bounds for bLGM models on several binary data sets. Throughout this section, we use $p(\mathbf{y}|\boldsymbol{\theta})$ to refer to the exact probability of a data vector $\mathbf{y}$ under the distribution with parameters $\boldsymbol{\theta}$. The exact probabilities $p(\mathbf{y}|\boldsymbol{\theta})$ remains intractable, but for small D we can compute them to arbitrary accuracy using numerical integration. We use $\tilde{p}(\mathbf{y}|\boldsymbol{\theta},\boldsymbol{\alpha})$ to refer to the bound on the probability of the data vector computed using the model parameters $\boldsymbol{\theta}$ and bound parameters $\boldsymbol{\alpha}$.

In higher dimensions we use imputation error as a measure of model fit. We hold out exactly one dimension per data case, selected at random. Given the observed data, we use each of the bounds to find the approximate posterior distribution $q(\mathbf{z}_n|\boldsymbol{\gamma}_n)$, and compute the prediction as $\hat{p}(y) = E_{q(\mathbf{z}_n|\boldsymbol{\gamma}_n)}p(y|\mathbf{z})$. To standardize the computation of the final integral, we use the approximation described in (Bishop, 2006) (see Chapter 4, page 218). We use the average cross-entropy of the held-out values as the imputation error measure.

bLGGM 1D-Synthetic: We begin by considering a one-dimensional binary latent Gaussian graphical model parameterized by a scalar mean μ and variance σ^2. The parameter vector is thus $\boldsymbol{\theta} = [\mu, \sigma^2]$. We set the true parameters $\boldsymbol{\theta}^*$ to $\mu^* = 2$ and $\sigma^* = 2$, yielding $p(y = 1|\boldsymbol{\theta}^*) = 0.7752$. We assess the bound on the marginal likelihood in the limit of infinite data by computing $\mathcal{L}_Q(\boldsymbol{\theta}) = \sum_y p(y|\boldsymbol{\theta}^*)\log(\tilde{p}(y|\boldsymbol{\theta},\boldsymbol{\alpha}))$ as we vary σ from 0 to 4 with μ fixed to μ^*. Note that for the variational bounds, we must optimize the free parameters in the LLP bound to maximize $\mathcal{L}_Q(\boldsymbol{\theta})$ for each value of σ.

The results of this experiment are given in Figure 2(a). We plot both the exact marginal likelihood and the bound on the marginal likelihood. We see that the Bohning (B) and Jaakkola (J) bounds fail dramatically, estimating $\sigma = 0$ instead of the correct value $\sigma = 2$. The piecewise bounds do significantly better, converging to the true marginal likelihood and cor-

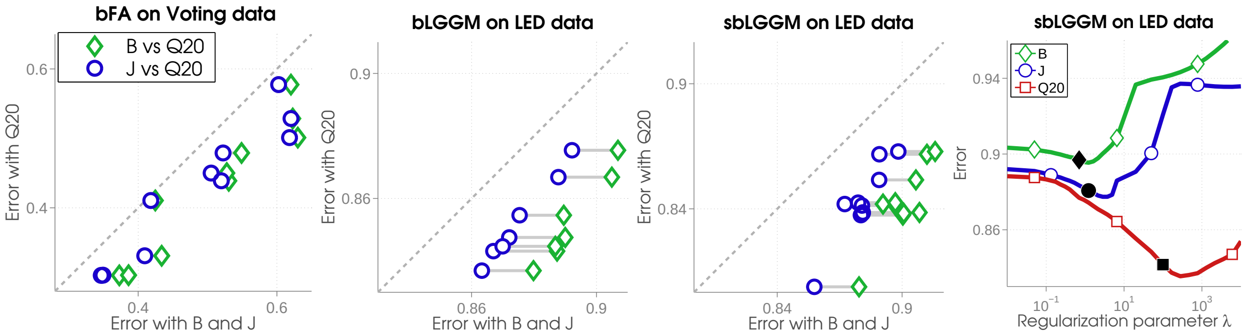


Figure 3. The first three plots show the imputation error of the 20-piece quadratic bound relative to Bohning and Jakkola for the bFA model on the Voting data set, bLGGM on LED and sbLGGM on the LED data set (the piecewise bound has lower error when the marker is *below* the diagonal line). The final plot shows an example of the imputation error versus the regularization parameter setting λ for the sbLGGM experiment.

rect σ value as the number of pieces in the bound increases. The piecewise quadratic bounds (Q3 and Q5) converge significantly faster than the linear bounds (L6 and L10), as predicted by the maximum error analysis in Section 5. Note that the results for Q3 and Q5 match those of L6 and L10, suggesting that the quadratic bound converges twice as fast as a function of the number of pieces.

bLGGM 5D-Synthetic: Next we consider a 5D binary latent Gaussian graphical model. We set the true mean vector $\boldsymbol{\mu}^*$ to 0 and the true covariance matrix $\boldsymbol{\Sigma}^*$ as seen in the top left panel of Figure 2(b). We sample 10^6 data cases from the true model to compute an estimate of the true data distribution. We estimate the model using a data set consisting of all 2^5 data cases $\mathbf{y}$ weighted by $p(\mathbf{y}|\boldsymbol{\theta}^*)$ to again focus on the asymptotic regime. Unlike the previous experiment, in this experiment we estimate the models by optimizing $\mathcal{L}_{QJ}(\boldsymbol{\theta}, \boldsymbol{\gamma})$ using Algorithm 1.

Figure 2(b) shows the covariance matrices estimated using the Jaakkola (J), Bohning (B) and 10 piece quadratic bounds (Q10). We see that both Bohning and Jaakkola shrink the estimated covariance parameters considerably, while the 10 piece quadratic bound results in less biased parameter estimates. Figure 2(c) shows the KL divergence between $p(\mathbf{y}|\boldsymbol{\theta}^*)$ and $p(\mathbf{y}|\hat{\boldsymbol{\theta}})$ for the parameters $\hat{\boldsymbol{\theta}}$ estimated using each bound. This KL divergence is again computed using 10^6 samples from each distribution. We show results for Bohning (B), Jaakkola (J) and 3 to 10 piece linear and quadratic bounds (L3-L10,Q3-Q10). We see that the piecewise bounds have significantly lower KL divergence than the Bohning and Jaakkola bounds when using a sufficient number of pieces. This indicates that they estimate significantly more accurate models, as suggested by the covariance plots in Figure 2(b). We again see

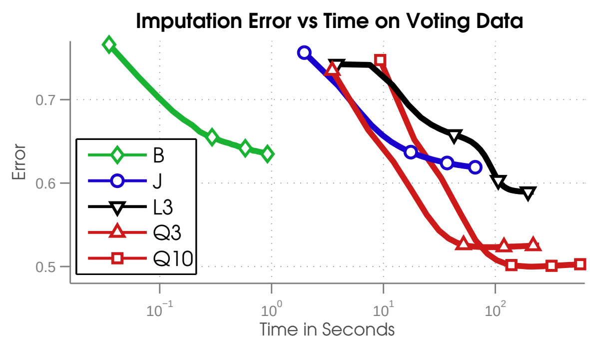


Figure 4. Imputation error versus time on the UCI Voting data. Markers are plotted at iterations $2, 10, 20, 35$.

that the piecewise quadratic bound converges approximately twice as fast as the piecewise linear bound as a function of the number of pieces.

bFA Voting: We fit a three-factor bFA model to the congressional voting records data set (available in the UCI repository) which contains votes of 435 U.S. Congressmen on 16 issues. We remove the data points which contain missing values and 3 issues which only show mild correlation with other issues. This gives us a total of 258 data vectors with 14 variables each. We use 80% of the data for training and 20% for testing. Figure 4 shows traces of the imputation error versus time for Jaakkola (J), Bohning (B), three-piece linear (L3) and three and ten piece quadratic bounds (Q3, Q10) for one training-test split. We see that the piecewise bounds give lower error than the Jaakkola and Bohning bounds, but require more time to converge. We again observe that the quadratic bounds have lower error than the linear bounds and the error decreases as the number of pieces increases. The first plot in Figure 3 shows the final imputation error results for 10 training-test splits. We plot the error of

Q20 against that of B and J. We clearly see that Q20 outperforms both B and J on all splits.

bLGGM and sbLGGM LED: We fit the bLGGM and sbLGGM models to the UCI LED data set (available in the UCI repository). This data set has 2000 data cases and 24 variables. The data is synthetically generated but is out of the model class. It contains 7 highly correlated variables and 17 variables that are marginally independent. In the bLGGM experiment we use 80% of the data for training and 20% for testing. The second plot in Figure 3 shows the results for 10 training-test splits for the bLGGM experiment. As in the Voting data, we see that Q20 outperforms both B and J on all splits.

In the sbLGGM experiment we purposely undersample the training set using 10% of the data for training and 50% for testing. The third plot in Figure 3 shows the results for 10 training-test splits for the sbLGGM experiment. We plot the error of Q20 versus B and J for the optimal choice of the regularization parameter λ found using cross-validation. We again see that Q20 outperforms both B and J on all splits. The final plot in Figure 3 shows traces of the imputation error as a function of the regularization parameter setting for a single split. The optimal value of λ for each bound corresponds to precision matrices that are 82.6%, 83.7% and 80.4% sparse for B, J and Q20, respectively.

7. Discussion

Piecewise quadratic bounds provide a tractable and useful family of estimators with a novel, tunable speed-accuracy trade-off controlled by the number of pieces. The main drawbacks of our approach is its reliance on the assumption of a Gaussian variational posterior. If the Gaussian assumption is strongly violated, the method will likely perform poorly, but in this case no other method based on the same underlying variational posterior can perform well.

An alternative to our approach is expectation-propagation (EP) in probit link-based LGMs (Minka, 2001). For parameter estimation, one can alternate EP inference with optimization of the EP approximation to the marginal likelihood (Minka, 2001, Equation 11). Some authors instead advocate the use of a Jensen's inequality-based approximation to the marginal likelihood (Kuss & Rasmussen, 2005), which is identical to the one we apply. In any case, neither of these EP-based learning schemes are provably convergent, while the variational framework is.

References

Bishop, C. *Pattern recognition and machine learning*. Springer, 2006.

Bohning, D. Multinomial logistic regression algorithm. *Annals of the Institute of Statistical Mathematics*, 44:197–200, 1992.

Collins, M., Dasgupta, S., and Schapire, R.E. A generalization of principal component analysis to the exponential family. In *Advances in neural information processing systems*, pp. 617–624, 2002.

Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 34:1–38, 1977.

Hsiung, K., Kim, S., and Boyd, S. Tractable approximate robust geometric programming. *Optimization and Engineering*, 9:95–118, 2008.

Jaakkola, T. and Jordan, M. A variational approach to Bayesian logistic regression problems and their extensions. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 1996.

Khan, M.E., Marlin, B.M., Bouchard, G., and Murphy, K. Variational bounds for mixed-data factor analysis. In *Advances in Neural Information Proceeding Systems*, 2010.

Kuss and Rasmussen, C. Assessing approximate inference for binary gaussian process classification. *Journal of Machine Learning Research*, 6:1679–1704, 2005.

Minka, T. Expectation propagation for approximate Bayesian inference. In *UAI*, 2001.

Mohamed, S., Heller, K., and Ghahramani, Z. Bayesian Exponential Family PCA. In *Advances in Neural Information Proceeding Systems*, 2008.

Nelder, J.A. and Mead, R. A simplex method for function minimization. *The computer journal*, 7(4):308, 1965.

Tipping, M. and Bishop, C. Probabilistic principal component analysis. *Journal of Royal Statistical Society Series B*, 21(3):611–622, 1999.

Wedel, Michel and Kamakura, Wagner. Factor analysis with (mixed) observed and latent variables in the exponential family. *Psychometrika*, 66(4):515–530, December 2001.