Fast Dual Variational Inference for Non-Conjugate Latent Gaussian Models

Mohammad Emtiyaz Khan EMTIYAZ.KHAN@EPFL.CH
School of Computer and Communication Sciences, Ecole Polytechnique Fédérale de Lausanne, Switzerland

Aleksandr Y. Aravkin SARAVKIN@US.IBM.COM
Numerical Analysis and Optimization, IBM T.J. Watson Research Center, Yorktown Heights, NY, USA

Michael P. Friedlander MPF@CS.UBC.CA
Department of Computer Science, University of British Columbia, Vancouver, Canada

Matthias Seeger MATTHIAS.SEEGER@EPFL.CH
School of Computer and Communication Sciences, Ecole Polytechnique Fédérale de Lausanne, Switzerland

Abstract

Latent Gaussian models (LGMs) are widely used in statistics and machine learning. Bayesian inference in non-conjugate LGMs is difficult due to intractable integrals involving the Gaussian prior and non-conjugate likelihoods. Algorithms based on variational Gaussian (VG) approximations are widely employed since they strike a favorable balance between accuracy, generality, speed, and ease of use. However, the structure of the optimization problems associated with these approximations remains poorly understood, and standard solvers take too long to converge. We derive a novel dual variational inference approach that exploits the convexity property of the VG approximations. We obtain an algorithm that solves a convex optimization problem, reduces the number of variational parameters, and converges much faster than previous methods. Using real-world data, we demonstrate these advantages on a variety of LGMs, including Gaussian process classification, and latent Gaussian Markov random fields.

Proceedings of the 30^{th} International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

1. Introduction

Latent Gaussian models (LGM) are ubiquitous in machine learning and statistics (e.g., Gaussian process models, Bayesian generalized linear models, dynamical systems with non-Gaussian observations, robust PCA, and non-conjugate matrix factorization). In many real-world applications, the likelihood is not conjugate to the Gaussian distribution, making exact Bayesian inference intractable. These modern applications, especially those with large latent dimensionality and number of observations, require fast, robust, and reliable algorithms for approximate inference.

In this context, algorithms based on variational Gaussian (VG) approximations are growing in popularity (Opper & Archambeau, 2009; Challis & Barber, 2011; Lázaro-Gredilla & Titsias, 2011; Honkela et al., 2011), since they strike a favorable balance between accuracy, generality, speed, and ease of use. However, compared to other approximations such as that of Seeger & Nickisch (2011), the structure of optimization problems associated with VG approximations remains poorly understood, and standard solvers for optimization take too long to converge.

While some variants of VG inference are convex (Khan et al., 2012b), they require $O(L^2)$ variational parameters to be optimized, where L is the dimensionality of the latent Gaussian vector. This slows down the optimization dramatically. One approach is to restrict the covariance representations up front, whether by naive mean field (Braun & McAuliffe, 2010; Knowles & Minka, 2011) or restricted Cholesky assumptions (Challis & Barber, 2011). Unfortunately, this can

result in considerable loss in accuracy, since typical LGMs, such as Gaussian processes, are tightly coupled. Another approach is to reduce the number of parameters to $O(N)$, where N is the dimension of the observation vector, using an exact covariance parameterization (Opper & Archambeau, 2009). This reparameterization destroys the convexity of the original problem, and very slow convergence is typically observed (Khan et al., 2012b). A recent coordinate-ascent method improves upon the state of the art (Khan et al., 2012b), but is restricted to Gaussian process models only and uses inefficient low-rank matrix updates.

We propose a dual decomposition approach that allows us to reduce the number of parameters to $O(N)$ while retaining convexity. The new dual optimization problem can be solved very rapidly with standard methods for smooth optimization. Using real-world data, we demonstrate that our algorithm converges much faster than the state of the art on a variety of LGMs. Unlike the approach of Khan et al. (2012b), our algorithm is generic and is not restricted to Gaussian processes.

2. Latent Gaussian Models

Given a vector of observations $\boldsymbol{y} \in \mathcal{Y}^N$, the dependencies among its components can be modeled using a latent vector $\boldsymbol{z} \in \mathbb{R}^L$. Here, the set $\mathcal{Y}$ is the domain of each observation, e.g., for binary observations, $\mathcal{Y} = \{0, 1\}$. The latent vector $\boldsymbol{z}$ is assumed to follow a Gaussian distribution $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The likelihood has the general form

$$p(\boldsymbol{y}|\boldsymbol{z}) = \prod_{n=1}^{N} p(y_n|\eta_n), \quad \boldsymbol{\eta} = \boldsymbol{W}\boldsymbol{z}, \tag{1}$$

where $\boldsymbol{W} \in \mathbb{R}^{N \times L}$. Model parameters $\boldsymbol{\theta}$ consist of parameters required to specify $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, $\boldsymbol{W}$, as well as parameters of the distribution $p(y_n|\eta_n)$. All densities are implicitly conditioned on $\boldsymbol{\theta}$, which we suppress from the notation. Also note that η_n can be a vector but we restrict ourselves to scalar η_n. Our results can be easily extended to the vector case.

Many models used in statistics and machine learning are instances of LGMs. Several examples are listed in Table 1, and an extensive list can be found in Khan (2012, Chapter 1). Bayesian generalized linear models constitute one such example, where we assume a latent Gaussian weight vector and use exponential family likelihoods with natural parameter η_n. Similarly, latent Gaussian Markov random fields (GMRF) model spatial correlations by using a GMRF with a sparse inverse covariance matrix $\boldsymbol{\Sigma}^{-1}$, along with an exponential family likelihood to model non-normal observations (Rue & Held, 2005). For example, count data with spatial dependence (e.g., incidences of a disease in different regions of a country) can be modeled using a Poisson likelihood with rate $r_n = \exp(\eta_n)$. The log-Gaussian Cox process is a non-parameteric generalization of this setting (Rue et al., 2009). Other non-parameteric examples are Gaussian process (GP) models, where observation pairs $\{y_n, \boldsymbol{x}_n\}$ are modelled via a latent Gaussian process $z(\boldsymbol{x})$ with the prior specified by mean and covariance functions.

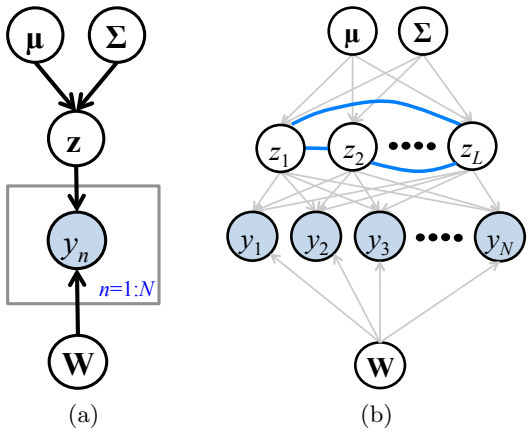


Figure 1. The graphical model for latent Gaussian models shown in left figure, and expanded in the right figure to explicitly show the correlation in the latent vector $\boldsymbol{z}$ induced due to a non-diagonal $\boldsymbol{\Sigma}$.

In Bayesian inference, we wish to compute expectations with respect to the posterior distribution

$$p(\boldsymbol{z}|\boldsymbol{y}) \propto \prod_{n=1}^{N} p(y_n|\eta_n)\mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{2}$$

For example, prediction of a new observation y_* can be obtained by computing the expectation $p(y_*|\boldsymbol{y}) = \int p(y_*|\boldsymbol{\eta})p(\boldsymbol{z}|\boldsymbol{y})\, d\boldsymbol{z}$. Another important task is computation of the marginal likelihood

$$p(\boldsymbol{y}) = \int \prod_{n=1}^{N} p(y_n|\eta_n)\mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\, d\boldsymbol{z}. \tag{3}$$

For example, parameters $\boldsymbol{\theta}$ can be learned by maximizing the log of the marginal likelihood, $\log p(\boldsymbol{y})$. This is also referred to as empirical Bayes or automatic relevance determination (ARD) (Tipping, 2001; Rasmussen & Williams, 2006).

For non-Gaussian likelihoods, both of these tasks are intractable. Applications in practice demand good approximations that scale favorably in N and L.

Model	Data	$\boldsymbol{z}$	$\boldsymbol{\theta}$	N	L	Remarks
Bayesian Logistic Regression	$\{y_n, \boldsymbol{x}_n\}$	Regression weights $y_n \leftarrow f(\boldsymbol{z}^T\boldsymbol{x}_n)$	$\boldsymbol{\mu}, \boldsymbol{\Sigma}$	#Obs	#Features	Row of $\boldsymbol{W}$ $\boldsymbol{w}_n = \boldsymbol{x}_n$
Gaussian Process Classification	$\{y_n, \boldsymbol{x}_n\}$	Regression function $y_n \leftarrow f(z_n)$	s, σ	#Obs	#Features	$\boldsymbol{W} = \boldsymbol{I}$ $N = L$
Gaussian Markov Random Field	$\{y_n\}$	Latent Gaussian field $y_n \leftarrow f(\boldsymbol{z}_n)$	k_v, k_u	#Obs	# Latent dims	
Probabilistic PCA	$\{y_{ni}\}$	Latent factors $y_n \leftarrow f(\boldsymbol{w}_i^T\boldsymbol{z}_n)$	$\boldsymbol{W}$	#Obs dims	#Latent factors	$N > L$ $\boldsymbol{\mu} = \boldsymbol{0}, \boldsymbol{\Sigma} = \boldsymbol{I}$

Table 1. Examples of LGM. Each column is a quantity from our generic LGM definition. Each row shows corresponding quantities for a model. First two models are supervised and the last two are unsupervised. For columns 2 and 3, n ranges over 1 to N and $\{a_n\}$ denotes the set of variables indexed by all values of n. $y \leftarrow f(z)$ implies that y can be generated using some function f of z. In last three columns, 'Obs' means observations, 'Dims' means dimensions, and '#' represents the number of a quantity. For GP, s and σ are hyperparameters of the covariance function. Similarly, k_u and k_v are hyperparameters for the latent field. See Section 6 for details. For PPCA, the subscript i indexes the observation vector.

3. Variational Gaussian Inference

In the *variational Gaussian* approximation (Opper & Archambeau, 2009), we assume the posterior to be a Gaussian $q(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{m}, \boldsymbol{V})$. The posterior mean $\boldsymbol{m}$ and covariance $\boldsymbol{V}$ form the set of variational parameters, and are chosen to maximize the variational lower bound to the log marginal likelihood shown in Eq. 5. To get this lower bound, we first multiply and divide by $q(\boldsymbol{z})$ in Eq. 4, and then use Jensen's inequality and the concavity of log (we denote the expectation with respect to $q(\boldsymbol{z})$ by $\mathrm{E}_{q(z)}$):

$$\log p(\boldsymbol{y}) = \log \int q(\boldsymbol{z}) \frac{\prod_n p(y_n|\eta_n)p(\boldsymbol{z})}{q(\boldsymbol{z})}\, d\boldsymbol{z} \tag{4}$$

$$\geq \mathrm{E}_{q(z)}\left[\log \frac{\prod_n p(y_n|\eta_n)p(\boldsymbol{z})}{q(\boldsymbol{z})}\right]. \tag{5}$$

The lower bound can be simplified further, and variational parameters $\boldsymbol{m}$ and $\boldsymbol{V}$ can be obtained by maximizing it:

$$\max_{\boldsymbol{m}, \boldsymbol{V} \succ \boldsymbol{0}} -\mathrm{D}[q(\boldsymbol{z}) \,\|\, p(\boldsymbol{z})] - \sum_{n=1}^{N} \mathrm{E}_{q(\eta_n)}[-\log p(y_n|\eta_n)], \tag{6}$$

where

$$\mathrm{D}[q \,\|\, p] = \mathrm{E}_q[\log q(\boldsymbol{z}) - \log p(\boldsymbol{z})] \tag{7}$$

$$q(\eta_n) = \mathcal{N}(\bar{m}_n, \bar{v}_n) \tag{8}$$

$$\bar{\boldsymbol{m}} = \boldsymbol{W}\boldsymbol{m}, \quad \bar{\boldsymbol{v}} = \mathrm{diag}(\boldsymbol{W}\boldsymbol{V}\boldsymbol{W}^T). \tag{9}$$

See Eqs. 4–7 in Khan et al. (2012a) for details of this derivation.

The first term in Eq. 6 is the relative entropy, and is jointly concave in $(\boldsymbol{m}, \boldsymbol{V})$. The second term $\mathrm{E}_{q(\eta_n)}[-\log p(y_n|\eta_n)]$ is not always available in closed form. We assume in this paper that, in such cases, we can evaluate an upper bound f_n to this term, i.e.,

$$\mathrm{E}_{q(\eta_n)}[-\log p(y_n|\eta_n)] \leq f_n(\bar{m}_n, \bar{v}_n). \tag{10}$$

This is also known as the local variational bound (LVB). We assume that f_n is differentiable and—most importantly—convex. We discuss a few such LVBs in Section 5; see Khan (2012) for an extensive list.

The resulting optimization problem is shown below in Eq. 11 and is expanded in Eq. 12:

$$\max_{\boldsymbol{m}, \boldsymbol{V} \succ \boldsymbol{0}} -\mathrm{D}[q(\boldsymbol{z}) \,\|\, p(\boldsymbol{z})] - \sum_{n=1}^{N} f_n(\bar{m}_n, \bar{v}_n) \tag{11}$$

$$:= \tfrac{1}{2}[\log|\boldsymbol{V}| - \mathrm{tr}(\boldsymbol{V}\boldsymbol{\Sigma}^{-1}) - (\boldsymbol{m}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{m}-\boldsymbol{\mu})] - \sum_{n=1}^{N} f_n(\bar{m}_n, \bar{v}_n) + \mathrm{cnst}. \tag{12}$$

The above lower bound is strictly concave (Braun & McAuliffe, 2010; Challis & Barber, 2011; Khan, 2012).

3.1. Related Work

A straight-forward approach is to solve Eq. 11 directly in $(\boldsymbol{m}, \boldsymbol{V})$ (Braun & McAuliffe, 2010; Challis & Barber, 2011; Marlin et al., 2011; Khan et al., 2012a). In practice, direct methods are slow and memory-intensive because of the very large number $L + L(L+1)/2$ of primal variables. Challis & Barber (2011) show that for log-concave likelihoods $p(y_n|\eta_n)$, the original problem Eq. 6 is jointly concave in $\boldsymbol{m}$ and the Cholesky factor of $\boldsymbol{V}$, and additional LVBs are not required. This fact, however, does not result in any reduction in number of parameters, and they propose to use factorizations of a restricted form, which negatively affects the approximation accuracy.

Opper & Archambeau (2009) and Nickisch & Rasmussen (2008) note that the optimal $\boldsymbol{V}_*$ must be of the form

$$\boldsymbol{V}_* = (\boldsymbol{\Sigma}^{-1} + \boldsymbol{W}^T(\operatorname{diag}\boldsymbol{\lambda})\boldsymbol{W})^{-1}, \tag{13}$$

which suggests reparameterizing Eq. 11 in terms of $L+N$ parameters $(\boldsymbol{m}, \boldsymbol{\lambda})$, where $\boldsymbol{\lambda}$ is the new variable. However, the problem is non-concave in this alternative parameterization (Khan et al., 2012b). Moreover, as shown in (Khan et al., 2012b) and our experiments here, convergence can be exceedingly slow. The coordinate-ascent algorithm proposed in (Khan et al., 2012b) solves the problem of convergence, but seems limited to the case $N = L$ and $\boldsymbol{W} = \boldsymbol{I}$. In addition, it requires N rank-one updates of $\boldsymbol{V}$ per iteration, which is slow on modern architectures optimized for block-matrix computations.

A range of different deterministic inference approximations apply to latent Gaussian models. The local variational method is convex for log-concave potentials and can be solved at very large scales (Seeger & Nickisch, 2011). However, it applies to super-Gaussian[1] potentials only. The bound it maximizes is provably less tight than Eq. 6 (Seeger, 2009; Challis & Barber, 2011), and it leads to worse results than the variational Gaussian approximation in general (Nickisch & Rasmussen, 2008; Khan, 2012). A key interpretation of this method is that it can be seen as one way to generate LVBs (for super-Gaussian potentials), which can be used in our VG setup (Seeger, 2009). Expectation propagation (Minka, 2001; Seeger, 2008) is more general and can be more accurate than most other approximations mentioned here. Based on a saddlepoint rather than an optimization problem, the standard EP algorithm does not always converge and can be numerically unstable. Among these alternatives, the variational Gaussian approximation stands out as a compromise between accuracy and good algorithmic properties, which is widely used beyond latent Gaussian model applications as well (Lázaro-Gredilla & Titsias, 2011; Honkela et al., 2011).

4. Dual Variational Inference

In this section, we show how Eq. 11 can be solved using a convex dual formulation in only N variational parameters. As shown in our experiments, the novel formulation admits simple algorithms which converge much more rapidly and have a lower per-iteration cost than previous methods reviewed above.

We achieve this by dual decomposition: decoupling the two terms in Eq. 11 by equality constraints, and then forming the Lagrangian dual. To be precise, we first introduce two new variables $h_n, \rho_n \in \mathbb{R}$ for each n and introduce constraints $h_n = \bar{m}_n$ and $\rho_n = \bar{v}_n$. The resulting (equivalent) optimization problem can be written as

$$\max_{\boldsymbol{m},\boldsymbol{V},\boldsymbol{h},\boldsymbol{\rho}} -\mathrm{D}[q(\boldsymbol{z}) \,\|\, p(\boldsymbol{z})] - \sum\nolimits_{n=1}^{N} f_n(h_n, \rho_n) \quad \text{s.t. } \boldsymbol{h} = \boldsymbol{W}\boldsymbol{m}, \quad \boldsymbol{\rho} = \operatorname{diag}(\boldsymbol{W}\boldsymbol{V}\boldsymbol{W}^T). \tag{14}$$

Next, we introduce dual variables $\boldsymbol{\alpha}, \boldsymbol{\lambda} \in \mathbb{R}^N$ associated to these constraints, and form the corresponding Lagrangian

$$\mathcal{L} = -\mathrm{D}[q(\boldsymbol{z}) \,\|\, p(\boldsymbol{z})] - \sum\nolimits_{n=1}^{N} f_n(h_n, \rho_n) + \boldsymbol{\alpha}^T(\boldsymbol{h} - \boldsymbol{W}\boldsymbol{m}) + \tfrac{1}{2}\boldsymbol{\lambda}^T(\boldsymbol{\rho} - \operatorname{diag}(\boldsymbol{W}\boldsymbol{V}\boldsymbol{W}^T)). \tag{15}$$

Strong duality holds because the constraints are affine, and so the solution to the original problem can be found by minimizing the Lagrangian dual with respect to $(\boldsymbol{\alpha}, \boldsymbol{\lambda})$, i.e.,

$$\min_{\boldsymbol{\alpha},\boldsymbol{\lambda}} \mathcal{D}(\boldsymbol{\alpha}, \boldsymbol{\lambda}) = \min_{\boldsymbol{\alpha},\boldsymbol{\lambda}} \max_{\boldsymbol{m},\boldsymbol{V},\boldsymbol{h},\boldsymbol{\rho}} \mathcal{L}. \tag{16}$$

The advantage of this formulation is that we can solve analytically for $(\boldsymbol{m}, \boldsymbol{V})$ and $(\boldsymbol{h}, \boldsymbol{\rho})$, and the resulting dual $\mathcal{D}(\boldsymbol{\alpha}, \boldsymbol{\lambda})$ is available in closed form. Since $\boldsymbol{\alpha}$ and $\boldsymbol{\lambda}$ are length N vector, the dual minimization involves only $O(N)$ parameters.

Derivations of the following statements are given in the Appendix. The unique maximizer with respect to $(\boldsymbol{m}, \boldsymbol{V})$ is given by

$$\boldsymbol{m}_* = \boldsymbol{\mu} - \boldsymbol{\Sigma}\boldsymbol{W}^T\boldsymbol{\alpha} \tag{17}$$

$$\boldsymbol{V}_* = \boldsymbol{A}_{\boldsymbol{\lambda}}^{-1} := (\boldsymbol{\Sigma}^{-1} + \boldsymbol{W}^T(\operatorname{diag}\boldsymbol{\lambda})\boldsymbol{W})^{-1}. \tag{18}$$

Importantly, $\boldsymbol{V}_*$ has precisely the economical form pointed out by Opper & Archambeau (2009).

Maximization over $(\boldsymbol{h}, \boldsymbol{\rho})$ is also available in closed form. Collecting the terms involving (h_n, ρ_n) in Eq. 15, we get the following optimization problem,

$$f_n^*(\alpha_n, \lambda_n) := \max_{h_n,\rho_n} \alpha_n h_n + \lambda_n \rho_n/2 - f_n(h_n, \rho_n), \tag{19}$$

which is in fact the f_n^* the *Fenchel conjugate* of f_n (Rockafellar, 1970), and is convex and well-defined due to the convexity of f_n. For many likelihoods (and LVBs), f_n^* is available in closed form. We give several examples in Section 5, summarized in Table 2.

Note that the effective domain of f_n^* (i.e., values of (α, λ) for which f_n^* is finite) may be restricted. We

[1] Neither the Poisson, nor the stochastic volatility likelihood are super-Gaussian (Section 5).

give details of this and show the effective domain of $f_n^*(\alpha_n, \lambda_n)$ for several commonly used likelihoods in Section 5. We denote the effective domain of f_n^* by $\mathcal{S}$.

Plugging in Eq. 17, 18, and 19 into Eq. 15 and ignoring the constants, directly gives us the optimization problem

$$\min_{\boldsymbol{\alpha},\boldsymbol{\lambda}\in\mathcal{S}} \tfrac{1}{2}\boldsymbol{\alpha}^T\widetilde{\boldsymbol{\Sigma}}\boldsymbol{\alpha} - \widetilde{\boldsymbol{\mu}}^T\boldsymbol{\alpha} - \tfrac{1}{2}\log|\boldsymbol{A}_{\boldsymbol{\lambda}}| + \sum_{n=1}^{N} f_n^*(\alpha_n, \lambda_n), \tag{20}$$

where $\widetilde{\boldsymbol{\mu}} = \boldsymbol{W}\boldsymbol{\mu}$ and $\widetilde{\boldsymbol{\Sigma}} = \boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^T$.

This is a strictly convex optimization problem involving $2N$ parameters, in contrast to Eq. 11, which involves $O(L^2)$ number of parameters. Given $(\boldsymbol{\alpha}_*, \boldsymbol{\lambda}_*)$ that minimizes the dual, the primal solution $(\boldsymbol{m}_*, \boldsymbol{V}_*)$ is obtained using Eq. 17 and 18. It might appear that minimizing the dual might be a difficult problem due to the constraints, but as we show later f_n^*, act as barrier functions, which simplify the optimization.

5. Algorithmic Details

Here we give details on the function f_n and its conjugate f_n^*. We also provide computational details about our algorithm for solving the dual problem Eq. 20.

5.1. Fenchel conjugates

We give an illustrative example to show the derivation of Fenchel conjugates. For simplicity, we drop the subscript n. Consider the Poisson likelihood $\log p(y|\eta) = y\eta - \exp(\eta) + \text{cnst}$:

$$f(h, \rho) = \mathrm{E}[-\log p(y|\eta)] = -yh + e^{h+\rho/2} + \text{cnst} \tag{21}$$

This function is convex. To determine the Fenchel conjugate $f^*(\alpha, \lambda)$, we use Eq. 19 and first maximize over ρ, obtaining $\lambda = e^{h+\rho_*/2}$. This implies $\lambda > 0$, since otherwise the conjugate takes the value $+\infty$. Then,

$$f^*(\alpha, \lambda) = \max_h \lambda(\log\lambda - 1) + (\alpha + y - \lambda)h \tag{22}$$

$$= \lambda(\log\lambda - 1) + \delta_0(\alpha - \lambda + y), \tag{23}$$

where $\delta_0(\cdot)$ is the convex indicator function, which equals 0 if the argument is 0, and $+\infty$ otherwise; the indicator term enforces the constraint $\alpha = \lambda - y$. Note that λ is constrained to lie in $S = \{\lambda > 0\}$.

Examples of f_n and f_n^* for a range of other likelihood functions are given in Table 2. Detailed derivation of these is available in an online appendix to the paper. In all the cases, $\alpha_n = \lambda_n - y_n$ applies, except for the stochastic volatility where $\alpha_n = \frac{1}{2} - \lambda_n$.

5.2. Reduced dual

As discussed in previous section, for all likelihoods discussed in this paper, we have a restriction on $\boldsymbol{\alpha}$. For example, for the first three likelihoods $\boldsymbol{\alpha} = \boldsymbol{\lambda} - \boldsymbol{y}$. Plugging this in Eq. 20, we get the reduced dual

$$\min_{\boldsymbol{\lambda}\in\mathcal{S}} \tfrac{1}{2}(\boldsymbol{\lambda}-\boldsymbol{y})^T\widetilde{\boldsymbol{\Sigma}}(\boldsymbol{\lambda}-\boldsymbol{y}) - \widetilde{\boldsymbol{\mu}}^T(\boldsymbol{\lambda}-\boldsymbol{y}) - \tfrac{1}{2}\log|\boldsymbol{A}_{\boldsymbol{\lambda}}| + \sum_{n=1}^{N} f_n^*(\lambda_n). \tag{24}$$

In other words, the equality constraints $\alpha_n = \lambda_n - y_n$ are enforced by the domain of the conjugate $f_n^*(\alpha_n, \lambda_n)$, which allows us to eliminate $\boldsymbol{\alpha}$ altogether using an affine substitution.

5.3. Algorithm Details

In this section, we show that the constrained problem of Eq. 24 can be optimized efficiently using quasi-Newton methods. We make use of the fact that the Fenchel conjugates act as barrier functions, thereby allowing us to limit the line search within the feasible set. This way, we avoid any unnecessary function evaluations to get an efficient implementation, treating the problem as if it was unconstrained.

First of all, note that the gradient of Eq. 24 with respect to $\boldsymbol{\lambda}$ is given by

$$\widetilde{\boldsymbol{\Sigma}}(\boldsymbol{\lambda}-\boldsymbol{y}) - \widetilde{\boldsymbol{\mu}} - \tfrac{1}{2}\operatorname{diag}(\boldsymbol{W}^T\boldsymbol{A}_{\boldsymbol{\lambda}}^{-1}\boldsymbol{W}) + \boldsymbol{g}_{\boldsymbol{\lambda}}^*, \tag{25}$$

where $\boldsymbol{g}_{\boldsymbol{\lambda}}^*$ is the vector of gradients of f_n^* with respect to λ_n. This gradient is used to obtain a descent descent direction $\boldsymbol{d}$.

Given the descent direction $\boldsymbol{d}$ and an initial step size δ_0, our goal is find a new step size δ while keeping $\boldsymbol{\lambda}$ feasible. We do this by restricting the linesearch to the feasible set only, and then using Armijo or Wolfe condition in exactly the same way as in the unconstrained case. We illustrate this for the constraint $\lambda_n > 0$, which arise when Fenchel conjugate contains terms such as $\log(\lambda_n)$. Other constraints can be implemented in a similar way. Assume that the current $\boldsymbol{\lambda}$ is in the feasible set, i.e., $\lambda_n > 0$ for all n. We find the indices $\mathcal{I}$ where $\lambda_n + \delta_0 d_n < 0$. Since $\lambda_n > 0$, we have $d_n < 0$ for all $n \in \mathcal{I}$. To keep the next $\lambda_n > 0$, the largest step should be less than the minimum $\lambda_n/|d_n|$ of all $n \in \mathcal{I}$. Hence, we restrict the search to the set

$$\delta = (1-\epsilon)\min\left\{\min_{n\in\mathcal{I}} \frac{\lambda_n}{|d_n|}, \delta_0\right\}, \tag{26}$$

where $\epsilon > 0$ ensures strict feasibility. Other constraints can be dealt with in a similar way.

	Poisson	Bernoulli-logit	Multi-Logit	Stochastic Volatility
$p(y\|\eta)$	$\exp(y\eta - e^\eta)/y!$	$e^{y\eta}/(1+e^\eta)$	$\exp(\boldsymbol{y}^T\boldsymbol{\eta})/\sum_k \exp(\eta_k)$	$\mathcal{N}(y\|0, e^\eta)$
LVB	Not required	Yes	Yes	Not required
$f(h,\rho)$	$-yh + e^{h+\rho/2}$	$-yh + \log(1+e^{h+\rho/2})$	$-\boldsymbol{y}^T\boldsymbol{h} + \text{lse}(\boldsymbol{h} + \frac{1}{2}\boldsymbol{\rho})$	$\frac{1}{2}h + \frac{1}{2}y^2 e^{-h+\rho/2}$
$f^*(\lambda)$	$\lambda(\log\lambda - 1)$	$\lambda\log\lambda$ $+(1-\lambda)\log(1-\lambda)$	$\sum_{k=1}^{K-1}\lambda_k\log\lambda_k + t\log t$ where $t := \sum_{k=1}^{K-1}\lambda_k$	$\lambda\log(2\lambda/y^2) - \lambda$
Range $\mathcal{S}$	$\lambda > 0$	$\lambda \in (0,1)$	$\lambda_k > 0$, $t < 1$	$\lambda > 0$

Table 2. This table summarizes LVBs (or exact expressions) and Fenchel conjugates for a number of likelihoods. Stochastic volatility is from (Rue et al., 2009), the Bernoulli-logit multi-logit bound from (Blei & Lafferty, 2006). Here, $\text{lse}(\boldsymbol{v}) = \log(1 + \sum_{k=1}^{K-1} e^{v_k})$. For first 3 columns, α is constrained to be equal to $\lambda - y$, and for the last one $\alpha = \frac{1}{2} - \lambda$.

6. Experiments

In this section, we apply our novel dual variational algorithm to a range of real-world Bayesian inference problems. We compare our algorithm to the widely used method of Opper & Archambeau (2009), which plugs the covariance parameterization of 13 into the primal problem Eq. 11 and optimizes it over $(\boldsymbol{m}, \boldsymbol{\lambda})$. We refer to this method as 'Opper-Arch'. We do not present results for the naive method of solving the primal in $(\boldsymbol{m}, \boldsymbol{V})$ directly, since this turns out to be much slower than the alternatives.

6.1. Multi-Way GP Classification

In this section, we consider a multinomial logit K-way Gaussian process classification (mGPC) model, following the experimental setup outlined in Khan et al. (2012a) and (Girolami & Rogers, 2006). See Khan (2012, Chapter 1) for details how GP classification can be formulated as an LGM.

For multinomial logit likelihood, the term f_n is not available in closed form, and we use the LVB proposed by Braun & McAuliffe (2010). Details of this LVB and its Fenchel conjugate are given in Table 2.

We apply the mGPC model to the forensic glass data set (available from the UCI repository) which has $N = 214$ data examples, $K = 6$ categories, and features $\boldsymbol{x}$ of length 8. We use 80% of the dataset for training and the rest for testing. We set $\boldsymbol{\mu} = 0$ and use a squared-exponential kernel, for which the (i,j)th entry of $\boldsymbol{\Sigma}$ is defined as: $\boldsymbol{\Sigma}_{ij} = -\sigma^2 \exp[-\frac{1}{2}||\boldsymbol{x}_i - \boldsymbol{x}_j||^2/s]$. Similar to the setup of Girolami & Rogers (2006), the prior factorizes across classes and we fix the hyperparameters σ and s to be same for all the classes. We find a good setting of these hyperparameters using the approximate marginal likelihood on training set. We compute this on a 11×11 grid, giving us total 121 hyperparameter settings. We also compute the prediction error defined as $-\log_2 \tilde{p}(y_{test}|\boldsymbol{\theta}, \boldsymbol{y}_{train}, \boldsymbol{x}_{train}, \boldsymbol{x}_{test})$, where $(\boldsymbol{y}_{train}, \boldsymbol{x}_{train})$ and $(\boldsymbol{y}_{test}, \boldsymbol{x}_{test})$ are training and testing data, respectively. Here, $\tilde{p}(y_{test}|\cdot)$ is the marginal predictive distribution approximated using the Monte Carlo method (see Rasmussen & Williams (2006, Chapter 3) for details).

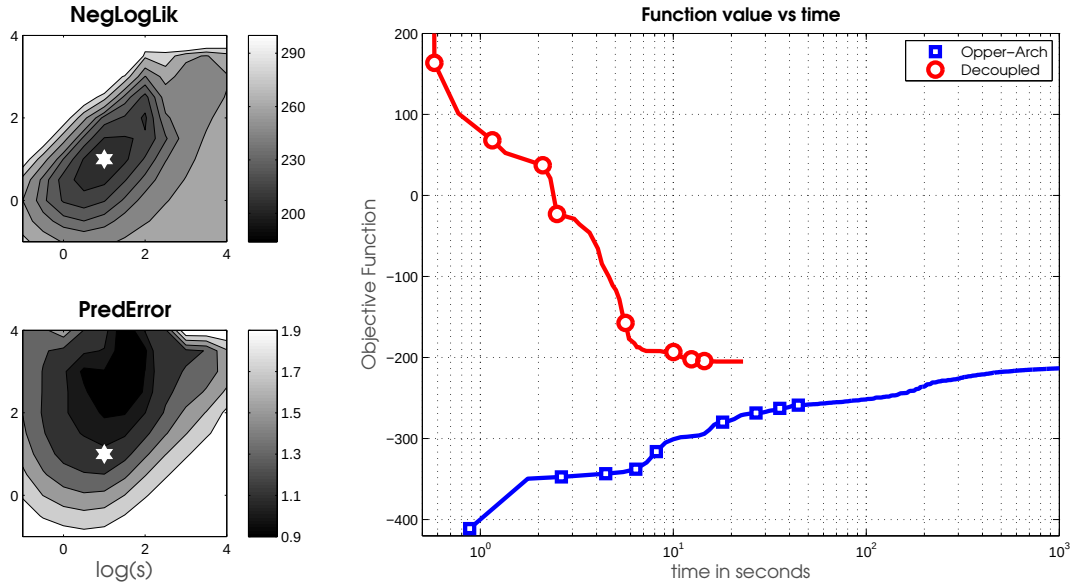
The results are shown in Fig. 2(a), where we plot the two quantities. The star shows the minimum of the negative marginal likelihood. We see that at this setting the algorithm also achieves a reasonable prediction error.

Fig. 2(b) shows the traces of the objective function for the two methods. The Opper-Arch method maximizes the primal objective function while dual variational inference minimizes the dual objective function. We show markers for iterations 1, 3, 5, 7, 9, and then at 20, 30, 40, and 50. We see that the dual inference coverges at least 100 times faster that the existing method (which has not yet converged in the plot). Each gradient step in Opper-Arch is also more expensive than our method since the number of parameters is $2NK$ (where K is the number of categories) in contrast to our algorithm which require only NK parameters. In addition, each function evaluation of Opper-Arch is more expensive than ours. This is due to the additional trace term in the primal problem Eq. 11, which is not present in the dual problem Eq. 16. Hence, our proposed algorithm has advantage in terms of the rate of convergence, cost of function evaluation, and the number of parameters.

We observed similar trends for other hyperparameter settings.

6.2. Latent Gaussian Markov Random Field

We consider the modeling of the oral cancer mortality rates using a latent GMRF, described in Rue & Held (2005). The data consists of mortality counts in 544 regions in Germany during 1986-1990. We model the count y_i in a region i using a Poisson likelihood with the rate $\lambda_i = \exp(\mu + u_i + v_i)$. Here, μ is the offset, v_i is an unstructured component, and u_i a spatially



(a) (b)

Figure 2. Comparisons for multinomial logit GP classification on the glass dataset. Figure (a) shows the negative log marginal likelihood approximations at the top and prediction errors at the bottom for many values of $\log(s)$ and $\log(\sigma)$. The star shows the minimum of the negative marginal likelihood, which achieves a reasonable prediction error. Figure (b) compares the traces of objective function with respect to time in seconds. We choose the hyperparameter setting which minimizes the negative of train log-likelihood. We see that dual variational inference converges much faster than the existing method.

structured component. The prior on the last two terms is shown below in Eq. 27. We assume an independent Gaussian prior over $\boldsymbol{v}$ with hyperparameter k_v, and an intrinsic GMRF of first-order with hyperparameter k_u (see Rue & Held (2005) for details on GMRFs). Here, $i \sim j$ are all unordered pairs (i, j) such that regions i and j are neighbors, i.e.,

$$p(\boldsymbol{u}, \boldsymbol{v}|k_u, k_v) \propto \exp\Big[-\tfrac{1}{2}k_v \sum_i v_i^2 - \tfrac{1}{2}k_u \sum_{i\sim j}(u_i - u_j)^2\Big]. \tag{27}$$

The GMRF prior can be easily written in the form of the LGM discussed in Section 2.

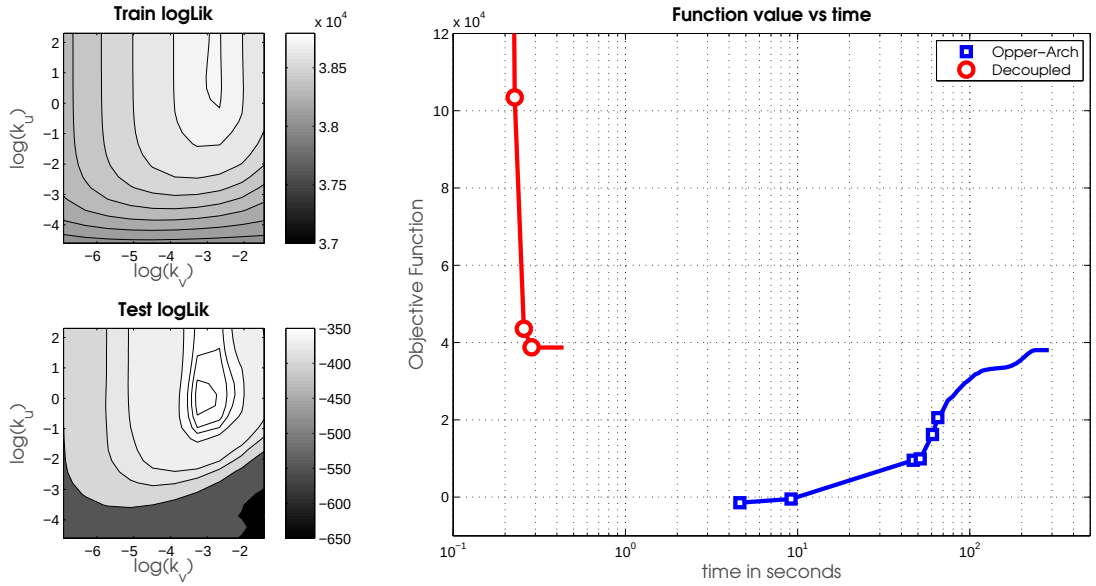
We choose 500 regions at random as training data and keep the rest as testing data. For simplicity, we set μ to 0. To find a good setting of other hyperparameters, we compute train and test log-likelihoods for several (k_u, k_v). The results are shown in Fig. 3(a). We see that the shape of train and test log-likelihoods are similar, justifying the maximization of the train log-likelihood to achieve good test accuracy. The maximum occurs at $k_u = 2.637$ and $k_v = 0.088$.

Fig. 3(b) shows the traces of optimizers for this setting of k_u and k_v. We show markers at iterations 1 to 6. We see that the proposed algorithm converges just in 6 iterations, and is much faster than the Opper-Arch method. Similar to mGPC, our method beats Opper-Arch on the number of iterations to converge, number of parameters, and cost of each function evaluations.

7. Conclusions

We presented a novel dual decomposition viewpoint on the variational Gaussian inference problem for latent Gaussian models. Our approach applies generally to any likelihood potential for which convex local variational bounds are available (e.g., Poisson, Bernoulli-logit, multi-way logistic, super-Gaussian) and is easy to configure to a new setup. Applying standard optimization technology to the dual formulation leads to an algorithm which has lower per iteration cost (time and memory) and can converge in orders of magnitude less iterations than the previous state of the art.

Dual decomposition has been popular for MAP inference in graphical model, for example, see Sontag et al. (2011); Jojic et al. (2010). In this paper, we applied the decomposition to the VG inference problem. We would like point that the coordinate-ascent approach of Khan et al. (2012b) also has a dual interpretation. There, each coordinate update can be interpreted as optimization of an element of the dual variable (see



(a) (b)

Figure 3. Comparisons for latent GMRF model on the glass dataset. Figure (a) shows the train and test log-likelihood approximations for many values of k_v and k_u. Figure (b) compares the traces of objective functions vs time. We choose the hyperparameter setting which maximizes the test log-likelihood.

Appendix of the paper). Our dual approach improves the approach of Khan et al. (2012b) by allowing parallel updates leading to an efficient implementation, while maintaining fast convergence.

A disadvantage of our approach is its restrition to the likelihood potentials with convex local variational bounds. Extension to the non-convex case remains an open problem which we would like to research in the future. We also aim to combine our dual formulation with covariance decoupling ideas from Seeger & Nickisch (2011), in order to break the $O(L^3)$ computational barrier and to make variational Gaussian inference applicable to very large problems.

Appendix

We describe the maximization with respect to $\boldsymbol{m}$ and $\boldsymbol{V}$ to get Eq. 17 and Eq. 18. We substitute the definition of $\mathrm{D}[q(\boldsymbol{z}) \,\|\, p(\boldsymbol{z})]$ from Eq. 12 into the Lagrangian Eq. 15. Derivatives of the Lagrangian with respect to $\boldsymbol{m}$ and $\boldsymbol{V}$ are given by

$$\tfrac{1}{2}(\boldsymbol{V}^{-1} - \boldsymbol{\Sigma}^{-1} - \boldsymbol{W}^T \operatorname{diag}(\boldsymbol{\lambda})\boldsymbol{W}) = 0, \tag{28}$$

$$-\boldsymbol{\Sigma}^{-1}(\boldsymbol{m} - \boldsymbol{\mu}) - \boldsymbol{W}^T\boldsymbol{\alpha} = 0. \tag{29}$$

Simplifying, we get Eq. 17 and Eq. 18.

Acknowledgments

MEK and MS were supported by an ERC starting grant (277815-SCALABIM). MEK would also like to thank Dr. Kevin Murphy for financial support at the beginning of this project.

References

Blei, D. and Lafferty, J. Correlated topic models. In *Advances in Neural Information Processing Systems*, 2006.

Braun, M. and McAuliffe, J. Variational inference for large-scale models of discrete choice. *Journal of the American Statistical Association*, 105(489):324–335, 2010.

Challis, E. and Barber, D. Concave Gaussian variational approximations for inference in large-scale Bayesian linear models. In *International conference on Artificial Intelligence and Statistics*, volume 6, pp. 7, 2011.

Girolami, M. and Rogers, S. Variational Bayesian multinomial probit regression with Gaussian process priors. *Neural Comptuation*, 18(8):1790 – 1817, 2006.

Honkela, A., Raiko, T., Kuusela, M., Tornio, M., and Karhunen, J. Approximate Riemannian conjugate gradient learning for fixed-form variational

Bayes. *Journal of Machine Learning Research*, 11: 3235–3268, 2011.

Jojic, Vladimir, Gould, Stephen, and Koller, Daphne. Accelerated dual decomposition for map inference. In *International Conference on Machine Learning*, 2010.

Khan, Mohammad Emtiyaz. *Variational Learning for Latent Gaussian Models of Discrete Data*. PhD thesis, University of British Columbia, 2012.

Khan, Mohammad Emtiyaz, Mohamed, Shakir, Marlin, Benjamin, and Murphy, Kevin. A stick breaking likelihood for categorical data analysis with latent Gaussian models. In *International conference on Artificial Intelligence and Statistics*, 2012a.

Khan, Mohammad Emtiyaz, Mohamed, Shakir, and Murphy, Kevin. Fast Bayesian inference for non-conjugate Gaussian process regression. In *Advances in Neural Information Processing Systems*, 2012b.

Knowles, D. and Minka, T. Non-conjugate variational message passing for multinomial and binary regression. In *Advances in Neural Information Processing Systems*, 2011.

Lázaro-Gredilla, M. and Titsias, M. Variational heteroscedastic Gaussian process regression. In *International Conference on Machine Learning 28*, 2011.

Marlin, B., Khan, M., and Murphy, K. Piecewise bounds for estimating Bernoulli-logistic latent Gaussian models. In *International Conference on Machine Learning*, 2011.

Minka, T. Expectation propagation for approximate Bayesian inference. In *Uncertainty in Artificial Intelligence 17*, 2001.

Nickisch, H. and Rasmussen, C.E. Approximations for binary Gaussian process classification. *Journal of Machine Learning Research*, 9(10), 2008.

Opper, M. and Archambeau, C. The variational Gaussian approximation revisited. *Neural computation*, 21(3):786–792, 2009.

Rasmussen, Carl Edward and Williams, Christopher K. I. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

Rockafellar, R. *Convex Analysis*. Princeton University Press, 1970.

Rue, H. and Held, L. *Gaussian Markov Random Fields: Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 2005.

Rue, H., Martino, S., and Chopin, N. Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations. *Journal of Royal Statistical Sociecty, Series B*, 71:319–392, 2009.

Seeger, M. Bayesian inference and optimal design for the sparse linear model. *Journal of Machine Learning Research*, 9:759–813, 2008.

Seeger, M. Sparse linear models: Variational approximate inference and Bayesian experimental design. *Journal of Physics: Conference Series*, 197(012001), 2009.

Seeger, M. and Nickisch, H. Large scale Bayesian inference and experimental design for sparse linear models. *SIAM J. Imag. Sciences*, 4(1):166–199, 2011.

Sontag, David, Globerson, Amir, and Jaakkola, Tommi. Introduction to dual decomposition for inference. *Optimization for Machine Learning*, 1, 2011.

Tipping, M. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, 2001.