A Geometric Variational Approach to Bayesian Inference

Abstract

We propose a novel Riemannian geometric framework for variational inference in Bayesian models based on the nonparametric Fisher-Rao metric on the manifold of probability density functions. Under the square-root transform representation, the manifold with the Fisher-Rao metric reduces to the unit Hilbert hypersphere with the standard $L^2$ metric. In contrast to existing approaches based on the Kullback-Leibler divergence, we approximate the posterior by a member of an appropriate class closest to the posterior with respect to the $\alpha$-divergence. As a consequence, in comparison with existing methods, our procedure leads to a tighter lower bound on the marginal density of the data. Our procedure also leads to an upper bound on the marginal density, which cannot be obtained from approaches based on Kullback-Leibler divergence. We provide several examples that validate the proposed framework. In particular, we consider classification via Bayesian logistic regression classification on few datasets and show that the performance of our method is comparable to other classification approaches.

1. Introduction

Various algorithms that are based on optimization techniques such as variational inference (VI), variational Bayes (VB) and expectation propagation (EP), have been successfully used to approximate the posterior distribution in the Bayesian setting. Among these approximate inference methods, VB is a popular framework that is often used for complex high-dimensional models since it can be easily used for large scale data analysis. In essence, VB and Markov Chain Monte Carlo (MCMC) sampling techniques are distinct approaches to resolve the same problem of approximating the posterior. VB methods are often preferred as an alternative to standard MCMC due to the two main drawbacks of MCMC methods: high computational complexity and issues with assessing chain convergence [4].

First, let us set up the inference problem. For a given dataset $x$, the objective of the VB framework is to find a distribution $q \in Q$ over the unknown, hidden parameters (or latent variables) $\theta$ that approximates the true posterior distribution $p(\theta|x)$. Usually, the ‘similarity’ or ‘closeness’ between the proposed approximating distribution $q$, and the true posterior $p$, is measured by the Kullback-Leibler divergence (KLD). Typically, the mean-field assumption is used for VB, where the class of approximating distributions is restricted to the class $Q$ where $q$ factorizes into independent partitions and is controlled by its original parameter, $q(\theta) = \prod_i q_i(\theta_i)$. 
In conditionally conjugate models, one assumes that the $q_i$s belong to the same exponential family as the complete conditional distribution $q_i(\theta_i|\theta_{-i},x)$, where $\theta_{-i}$ denotes all of $\theta$ except $\theta_i$. Thus, the inference problem becomes an optimization problem of determining the distribution in the class parameterized by the natural parameter in the exponential family. The solution to this problem is usually obtained by a gradient ascent (or descent) approach along the individual coordinates of $\theta$, where the updates are simple and available as members of the same family [6]. Wang and Blei [7] extended the VB approach for nonconjugate models and proposed two generic methods which use Gaussian approximations: Laplace variational inference and delta method variational inference.

The range of distributions over which the optimization problem is solved depends on the choice of $Q$. Ideally, $Q$ should be a versatile class of distributions chosen in such a manner that it includes distributions close to $p(\theta|x)$. The KLD-based VB framework is appealing and it has computational advantages. However, the geometry of the space of densities is not accounted for when solving the optimization problem. In this paper, we consider a novel nonparametric approach to the optimization problem wherein we do not restrict the class of approximating distributions $Q$ to any particular parametric family. Instead, we utilize the geometry of the space on which probability density functions (PDFs) lie and optimize over this entire class of distributions, equipped with the nonparametric Fisher-Rao Riemannian metric (simply referred to as Fisher-Rao (FR) metric hereafter), on which the parameter space factorizes.

In our proposed framework, we also replace the KLD with the more general $\alpha$-divergence ($\alpha$D) [10] as the dissimilarity measure of the approximating distribution from the exact posterior. We show that the lower bound of the marginal density obtained by using KLD can be improved upon by considering this richer class of divergences. Moreover, we can also derive an upper bound on the marginal density in this scenario, something which is not available with the KLD. Recently, Li and Turner [12] considered a similar approach where they focused on improving the standard variational lower bound with the variational Rényi (VR) bound based on Rényi’s $\alpha$-divergence. To extend the class of models that can be handled in their optimization framework, they implemented Monte Carlo (MC) approximation techniques under certain conditions, which lead to a biased estimate of the bound. As part of future work, they also mentioned the possibility of sandwiching the marginal likelihood with bounds computed using different $\alpha$ values. In this paper, we calculate the bounds for the marginal density and show its usefulness in model selection for a simulated example. Moreover, as opposed to their approach of using MC techniques, or using Gaussian approximations for nonconjugate models [7], we provide a unified nonparametric geometric optimization framework; we consider an optimization problem in which variational approximations to the posterior can be performed in a statistically principled and computationally efficient manner, without imposing severe restrictions on the class of approximating densities. To summarize, the main contributions of this paper
are: (i) We propose a novel geometric framework for VB. Our method utilizes the intrinsic structure of the manifold of PDFs equipped with the nonparametric FR metric. (ii) As the approximating family, we consider the family of all probability densities on the parameter space which factorize. We do not restrict our setup to just parametric families, as is common in the current literature. (iii) We show, theoretically and numerically, that the proposed approach using αD as the dissimilarity measure results in a tighter lower bound on the marginal density than the KL approach. Our approach also provides an upper bound on the marginal density, which cannot be obtained with the KL approach. (iv) We utilize the geometry of the space of PDFs to define a gradient ascent algorithm to solve the VB problem. (v) The proposed method is validated using a simulated example, and multiple logistic regression classification problems. In one of these examples, we consider the task of signature verification. For this purpose, we also define novel shape based signature descriptors.

2. Riemannian Geometric Framework for Variational Bayes and Divergence measures

2.1. Representation Space of PDFs and Riemannian Metric

For simplicity, we restrict our attention to the case of univariate densities on $\mathbb{R}$. We note however, that the framework is equally valid for all finite dimensional distributions. Denote by $\mathcal{P}$, the Banach manifold of PDFs on $\mathbb{R}$, defined as $\mathcal{P} = \{ p : \mathbb{R} \to \mathbb{R}_{\geq} \mid \int_{\mathbb{R}} p(x) dx = 1 \}$. The space $\mathcal{P}$ is not a vector space but a manifold with a boundary because any density function whose value is zero for any $x \in \mathbb{R}$ is a boundary element. Next, for a point $p$ in the interior of $\mathcal{P}$, consider a vector space that contains the set of tangent vectors at this point. This is defined as the tangent space at the point $p$, $T_p(\mathcal{P}) = \{ \delta p : \mathbb{R} \to \mathbb{R} \mid \int_{\mathbb{R}} \delta p(x) p(x) dx = 0 \}$. Intuitively, the tangent space at any point $p$ on the manifold $\mathcal{P}$ contains all possible perturbations of the density function $p$. This tangent space can be used to define a suitable metric between any two probability density functions on the manifold using the FR metric. For any two tangent vectors $\delta p_1, \delta p_2 \in T_p(\mathcal{P})$, the nonparametric version of the FR metric is given by $\langle \langle \delta p_1, \delta p_2 \rangle \rangle_p = \int_{\mathbb{R}} \delta p_1(x) \delta p_2(x) \frac{1}{p(x)} dx$ \[14\].

The FR metric is closely related to the Fisher information matrix, rendering it attractive in various statistical methodologies. However, since the FR metric changes from point to point on the space of PDFs, it leads to cumbersome computations which makes it very difficult to use in practice. Thus, instead of working on the non-linear manifold of PDFs directly under the FR metric, we want to use a suitable transformation that simplifies the Riemannian geometry of this space. The square-root representation proposed by Bhattacharyya \[17\] provides an elegant solution to this problem. We define a continuous mapping $\phi : \mathcal{P} \to \Psi$, where $\phi(p) = \psi = +\sqrt{p}$ is the square root transform (SRT) of a PDF $p$. We omit the $+$ sign from the representation for notational convenience. The inverse mapping is given by $\phi^{-1}(\psi) = p = \psi^2 \[13\]$. The space of all SRT representations of PDFs is $\Psi = \{ \psi : \mathbb{R} \to \mathbb{R}_{\geq} \mid \int_{\mathbb{R}} \psi^2(x) dx = 1 \}$ and represents the positive orthant of the
unit Hilbert sphere. Since the differential geometry of the sphere is well known, one can define standard geometric tools on this space for analyzing PDFs analytically. Thus, our general approach is to represent PDFs using their SRT representation, compute quantities of interest on Ψ, and then map them back to \( P \) using the inverse mapping.

For elements not lying on the boundary, let \( T_\psi(\Psi) = \{ \delta\psi \mid \langle \delta\psi, \psi \rangle = 0 \} \) denote the tangent space at \( \psi \). Thus, with the choice of SRT representation, for any two vectors \( \delta\psi_1, \delta\psi_2 \in T_\psi(\Psi) \), the FR metric becomes the standard \( L^2 \) Riemannian metric. Observe that since we are on the infinite-dimensional unit sphere, the geodesic distance between two PDFs \( p_1, p_2 \in \mathcal{P} \), represented by their SRTs \( \psi_1, \psi_2 \in \Psi \), is defined as the shortest arc connecting them on \( \Psi \). In other words, the geodesic distance between densities \( p_1 \) and \( p_2 \) under the FR metric is given by the angle between them: 

\[
\cos^{-1}(\langle \psi_1, \psi_2 \rangle) = \nu. \quad \text{The corresponding geodesic path is given by the great circle connecting} \ \psi_1 \ \text{and} \ \psi_2.
\]

We will use additional geometric tools for our optimization algorithm. These include the exponential and inverse-exponential maps, and parallel transport. For \( \psi \in \Psi \) and \( \delta\psi \in T_\psi(\Psi) \), the exponential map at \( \psi \), \( \exp : T_\psi(\Psi) \rightarrow \Psi \) is defined as \( \exp_\psi(\delta\psi) = \cos(||\delta\psi||)\psi + \sin(||\delta\psi||)\frac{\delta\psi}{||\delta\psi||} \). The exponential map is used to map points from the tangent space of \( \Psi \) \( (T_\psi(\Psi)) \) to the representation space \( \Psi \). Recall that \( \Psi \) contains all SRTs of PDFs. Similarly for \( \psi_1, \psi_2 \in \Psi \), the inverse-exponential map denoted by \( \exp^{-1}_\psi : \Psi \rightarrow T_\psi(\Psi) \) is given by \( \exp^{-1}_\psi(\nu) = \frac{\nu}{\sin(\nu)} (\psi_2 - \cos(\nu)\psi_1) \). The inverse-exponential map takes a point from the representation space \( \Psi \) to the tangent space \( T_\psi(\Psi) \). Thus, with the help of these two tools from differential geometry, we can travel between the representation space of PDFs and the tangent space with ease.

Finally, we define parallel transport, which is used to map tangent vectors from a tangent space to another while preserving their lengths and the angles between them. We will use the parallel transport along geodesic paths in \( \Psi \). For \( \psi_1, \psi_2 \in \Psi \), and a vector \( \delta\psi_1 \in T_{\psi_1}(\Psi) \), the parallel transport of \( \delta\psi_1 \) from \( \psi_1 \) to \( \psi_2 \) along the geodesic path is defined as \( \delta\psi_2 = \delta\psi_1 = \delta\psi_1 - \frac{2(\delta\psi_1, \delta\psi_2)}{\|\psi_1 + \psi_2\|}(\psi_1 + \psi_2) \). This defines a mapping \( \kappa : T_{\psi_1}(\Psi) \rightarrow T_{\psi_2}(\Psi) \) such that \( \delta\psi_2 = \kappa(\delta\psi_1) \). An important property of parallel transport is that the mapping \( \kappa \) is an isometry between two tangent spaces, i.e., for \( \delta\psi_1, \delta\psi_2 \in T_{\psi_1}(\Psi), \langle \delta\psi_1, \delta\psi_2 \rangle = \langle \kappa(\delta\psi_1), \kappa(\delta\psi_2) \rangle \).

2.2. \( \alpha \)-Divergence

In this section, we briefly review \( \alpha \)-divergences \( (\alpha\text{D}) \) \( [10] \), upon which our proposed VB framework is based. Let us consider two multivariate probability distributions \( p \) and \( q \) of an \( m \)-dimensional random variable \( \theta \in \Theta \). Then, \( \alpha\text{D} \) defined for \( \{ \alpha : \alpha > 0, \alpha \neq 1 \} \) is given by:

\[
D_\alpha[p||q] = \frac{1}{\alpha - 1} \ln \int_\Theta p(\theta)^\alpha q(\theta)^{1-\alpha} d\theta \quad (1)
\]
and is used to measure the ‘closeness’ or ‘similarity’ of the two distributions $p$ and $q$. The class of $\alpha$-divergences has the following properties: (1) $D_\alpha[p||q] \geq 0$, (2) $D_\alpha[p||q] = 0$ when $p = q$ a.e. and (3) $D_\alpha[p||q]$ is convex with respect to both $p$ and $q$. In particular, $\alpha$D is connected to KLD in two ways: (1) $\lim_{\alpha \to 0} D_\alpha[p||q] = KL(q||p)$ and (2) $\lim_{\alpha \to 1} D_\alpha[p||q] = KL(p||q)$. These limiting cases cannot be obtained by plugging the values of $\alpha$ in $[1]$. Instead, the definitions are determined by continuity of $D_\alpha$ $[\Pi]$. With specific regard to approximate inference methods, we note that VB attempts to minimize $KL(q||p)$, whereas EP attempts to minimize $KL(p||q)$. Another special case of $\alpha$D is that for $\alpha = 0.5$, which is very closely related to the aforementioned FR metric. In fact, this is the only choice of $\alpha$, which results in a distance between distributions rather than a divergence.

3. Variational Inference

3.1. Problem Formulation

Let $x$ denote the observed data and $\theta = (\theta_1, \theta_2, \ldots, \theta_m) \in \Theta$ denote the unknown $m$-dimensional parameter, where $\{\Theta = (\Theta_1, \Theta_2, \ldots, \Theta_m) : \theta_i \in \Theta_1\}$. Let $f(\theta, x) = f(x|\theta)\pi(\theta)$ denote the joint distribution of $x$ and $\theta$ where $f(x|\theta)$ is the likelihood function and $\pi(\theta)$ is the prior distribution on $\theta$. The posterior distribution is then given by $p(\theta|x) = \frac{f(x, \theta)}{m(x)}$ where $m(x) = \int_{\Theta} f(x, \theta)d\theta$ denotes the marginal density of $x$. In practice, calculating the posterior is difficult because we cannot evaluate $m(x)$ easily, especially when analytical solutions are not available. To this effect, we propose a VB optimization framework based on $\alpha$D, where we wish to find a density function to approximate the true posterior among the class of all joint PDFs that factorize.

To develop the appropriate optimization problem of interest, we use the SRT defined in Section $2.1$ Suppose the corresponding square-root representations of the joint, marginal and the posterior are denoted by $\psi_f, \psi_m$ and $\psi_p$ respectively. Based on the mean-field assumption described in Section $[1]$ let $Q = \{q \mid q = \prod_{i=1}^m q_i\}$ denote the approximating class of densities. Accordingly, the set $Q_\psi = \{\psi_q \mid \psi_q = \prod_{i=1}^m \psi_{q_i}\}$ consists of elements of the $m$-fold product space $\Psi_m = \Psi \times \Psi \times \cdots \times \Psi$ of SRTs. As mentioned before, we are interested in formulating an optimization problem, which solves for optimal $q_i$s, where the optimality criterion is defined as the $\alpha$D between the true posterior and the product of the $q_i$s:

$$D_\alpha[p||\prod_{i=1}^m q_i] = \frac{1}{\alpha - 1} \ln \int_{\Theta} p(\theta|x)^\alpha \left(\prod_{i=1}^m q_i(\theta)\right)^{1-\alpha} d\theta,$$

(2)

for $\alpha > 0$. Based on $[2]$, let us define the energy function $\tilde{\mathcal{E}}_\alpha : \Psi_m \to \mathbb{R}_{\geq 0}$, as $\tilde{\mathcal{E}}_\alpha(q_1, q_2, \ldots, q_m) = D_\alpha[p||\prod_{i=1}^m q_i]$ for $\alpha > 0$, $\alpha \neq 1$. We wish to minimize $\tilde{\mathcal{E}}_\alpha$ over $Q$; note that for the limiting case of $\alpha = 1$, this gives rise to the optimization problem of minimizing the KLD between $p$ and $q$, i.e., minimizing
\[ \int_{\Theta} \ln \left( \frac{p(\theta|x)}{q(\theta|x)} \right) p(\theta|x) d\theta. \] For high-dimensional models, since the integral in this case is with respect to the computationally intractable posterior distribution \( p \), the optimization problem becomes difficult to handle. Thus, we choose to ignore this limiting case where (2) is not well-defined.

To simplify the problem at hand we observe the following equivalence relationships:

\[
(q_1^*, q_2^*, \ldots, q_m^*) = \arg \min_{q_1, q_2, \ldots, q_m} \mathcal{E}_\alpha(q_1, q_2, \ldots, q_m) = \arg \min_{\Psi_m} D_\alpha[p \mid\mid \prod_{i=1}^{m} q_i] = \arg \min_{\Psi_m} \frac{1}{\alpha - 1} \int_{\Theta} \psi_p(\theta|x)^{2\alpha} \left( \prod_{i=1}^{m} \psi_{q_i}(\theta) \right)^{2-2\alpha} d\theta = \arg \min_{\Psi_m} \frac{1}{\alpha - 1} \int_{\Theta} \psi_f(x, \theta)^{2\alpha} \left( \prod_{i=1}^{m} \psi_{q_i}(\theta) \right)^{2-2\alpha} d\theta.
\]

The third equality holds due to the strict concavity of the natural logarithm, and the fourth equality follows from the fact that the SRT of the marginal density \( m(x) \), \( \psi_m(x) \), is constant in \( \theta \). Furthermore, when \( \alpha < 1 \), the factor \( \frac{1}{\alpha - 1} < 0 \), and thus the minimization problem becomes one of maximization. Consequently, we redefine the energy function as \( \mathcal{E}_\alpha(q_1, q_2, \ldots, q_m) = \int_{\Theta} \psi_f(x, \theta)^{2\alpha} \left( \prod_{i=1}^{m} \psi_{q_i}(\theta) \right)^{2-2\alpha} d\theta \), where we maximize \( \mathcal{E}_\alpha \) if \( \alpha \in (0, 1) \) and minimize \( \mathcal{E}_\alpha \) if \( \alpha \in (1, \infty) \).

Before studying the consequences of using \( \alpha \text{D} \) as a divergence measure in our proposed VB framework, let us consider the traditional VB setup, where one chooses to find the approximating class of densities \( q \) by minimizing the KLD between \( q \) and the true posterior \( p \):

\[
(q_1^{KL}, q_2^{KL}, \ldots, q_m^{KL}) = \arg \min_{q \in \mathcal{Q}} KL(q \mid\mid p) = \arg \min_{q \in \mathcal{Q}} \int_{\Theta} \ln \left( \frac{q(\theta)}{p(\theta|x)} \right) q(\theta) d\theta = \arg \max_{q \in \mathcal{Q}} \int_{\Theta} \ln \left( \frac{f(x, \theta)}{q(\theta)} \right) q(\theta) d\theta = \arg \max_{q \in \mathcal{Q}} \mathcal{H}(f, q),
\]

where the fourth equality again stems from the fact that the marginal does not depend on \( \theta \). Thus, instead of minimizing \( KL(q \mid\mid p) \), one can choose to maximize \( \mathcal{H}(f, q) \) to obtain an equivalent solution to the original optimization problem. We will use the quantity \( \mathcal{H}(f, q) \) in the upcoming Section 3.2 to compare the KLD-based bound on the marginal density with the bound obtained using our proposed \( \alpha \text{D} \) based VB setup.

### 3.2. Bounds on Marginal Density

The two important outcomes of using \( \alpha \text{D} \) as an optimality criterion are: (i) a tighter lower bound on the marginal density than the KLD-based one, and (ii) an upper bound on the marginal density, which is not possible under the KLD setup. For a general variational family (not necessarily one which factorizes), we formally state these two results on the logarithmic scale for ease of comparison with the KLD-based bound.

**Proposition.** The following inequalities hold for the marginal density \( m(x) \):

\[
\int_{\Theta} \ln \left( \frac{p(\theta|x)}{q(\theta|x)} \right) p(\theta|x) d\theta.
\]
3.3. Approximation of the Posterior via Gradient Ascent

We approximate the gradient of the energy \( E \) at the current iteration. We approximate the gradient algorithm. Let \( \psi \) or the novel representation space of PDFs defined in Section 2.1. But, in order to optimize the energy \( E \) in the definition of the variational energy \( \mathcal{E}_\alpha \). This complex optimization problem can be solved using a gradient algorithm. Let \( \psi_0^{(0)}, \psi_1^{(0)}, \ldots, \psi_m^{(0)} \in \mathcal{Q} \) be approximations of the SRT representation of the posterior at the current iteration. We approximate the gradient of the energy \( \mathcal{E}_\alpha \) using directional derivatives under a finite basis. The gradient is defined separately for each \( \psi_\theta^{(0)} \) as:

\[
\nabla \mathcal{E}_\alpha^i \approx \sum_{(k)=1}^{N} \left[ \frac{d}{d\eta} \mathcal{E}_\alpha^i(\psi_\theta^{(0)} + \eta \delta \psi_\theta^{(k)}) \right] \eta=0 \left( 1 - \alpha \right) \sum_{(k)=1}^{N} \left[ \int \psi_f(x, \theta) \prod_{j \neq i} \psi_\theta^{(j)}(\theta_j)^{2\alpha} \psi_\theta^{(i)}(\theta_i)^{1-2\alpha} \delta \psi_\theta^{(k)}(\theta) d\theta \right] \delta \psi_\theta^{(k)} (3)
\]

where \( \delta \psi_\theta \in \mathcal{B} \) is an appropriate orthonormal basis in \( T_{\psi_\theta^{(0)}}(\Psi) \), and \( \eta \) is a small positive constant. We use the geometry of the space \( \Psi \) to define \( \mathcal{B} \). We explain the construction of this basis for \( \theta \in [0, 1] \) and note that it is easily extended to a general compact support. For this purpose, we use the tangent space at the SRT representation of the uniform distribution \( U \) on \( [0, 1] \) defined as \( T_{\psi_U}(\Psi) = \{ \delta \psi_U : [0, 1] \to \mathbb{R} | \int_0^1 \delta \psi_U(\theta) \psi_U(\theta) d\theta = 0 \} \). We define the basis set \( \mathcal{B} = \{ \sin(2\pi n \theta), \cos(2\pi n \theta), 1 - \theta | n \in \mathbb{Z}_+ \} \}. It is easy to verify that all elements of this set are orthogonal to \( \psi_U \). This basis is then orthonormalized using the Gram-Schmidt procedure under the \( L^2 \) metric to result in \( \mathcal{B} \). Note that this is an orthonormal basis for \( T_{\psi_U}(\Psi) \) only. This orthonormal basis can be extended to the entire space using parallel transport. Parallel transport ensures that the basis remains orthonormal and that it remains orthogonal to the representation space. As a result, we can define a one step gradient ascent update using the exponential map as \( \psi_\theta = \exp_{\psi_\theta^{(0)}}(\epsilon \nabla \mathcal{E}_\alpha^i) \), where \( \epsilon \) is a small step size and \( | \cdot | \) denotes the absolute value. The use of the exponential map in this update is needed to ensure that we remain in the space of SRT representations of densities, \( \Psi \). Our approach is then to separately update all \( \psi_\theta \) at each iteration and repeat until convergence. Because this is a gradient ascent approach, we are not guaranteed to arrive at the global maximum, and thus use the local solution for inference. Through extensive simulation, we have found that initialization does not play a crucial role with respect to the convergence of this algorithm. In related work, Minka defined optimization
3.4. Estimation of Integral in Energy Gradient

One major issue that we encounter in computing the approximate gradient is that we must evaluate a potentially high-dimensional integral. This requires further approximation of the integral, for which we use the method introduced in [19]. In short, the approximation method works as follows. Let $Y$ be a random variable with $E(Y) = \mu$. Suppose we are interested in evaluating $E(g(Y))$ for a smooth function $g$. The first-order Taylor expansion of $g$ around $\mu$ is $g(y) = g(\mu) + g'(\mu)(y - \mu) + O(y - \mu)^2$. Taking expectations on both sides, we obtain $E(g(Y)) = g(\mu) + 0 + O(Var(Y))$. Thus, $E(g(Y))$ is approximated with $g(\mu)$. While other approximation strategies can be used for the purpose of evaluating a high-dimensional integral, we have found this one to work well in the examples we considered.

We use this approximation technique in our problem as follows. Consider the integral expression in (3) and rewrite it as follows:

$$\int_{\Theta_m} \cdots \int_{\Theta_2} \int_{\Theta_1} \frac{\psi_f^{2\alpha}(x, \theta_1, \theta_2, \ldots, \theta_m)}{\prod_{j \neq i} \psi_{\alpha j}^{2\alpha}(\theta_j)} \prod_{j \neq i} q_j(\theta_j) \psi_{\alpha i}(\theta_i)^{1-2\alpha} b_i^{(k)}(\theta_i) d\theta_1 d\theta_2 \ldots d\theta_m, \tag{4}$$

since $q_j(\theta_j) = \psi_{\alpha i}(\theta_j)^2$. We first compute the expectations $\mu_j = \int_{\Theta_j} q_j(\theta_j) d\theta_j$, $j \neq i$. We then use these expected values to redefine the high dimensional integral (4) as a one dimensional integral $\int_{\Theta_1} \frac{\psi_f(x, \mu_{-i}, \theta_i)}{\prod_{j \neq i} \psi_{\alpha j}(\mu_j)} b_i^{(k)}(\theta_i) d\theta_i$, which can easily be evaluated using numerical integration.

4. Experimental Results and Applications

4.1. Simulation Study

In this section, we consider the following hierarchical model: $x|\mu, \tau \sim iid N(\mu, \tau^{-1})$, $\mu|\tau \sim N(0, \tau^{-1})$, $\tau \sim Ga(0.01, 0.01)$. Because the posterior in this case is bivariate, we can evaluate our proposed method and its approximation using the ground truth. Additionally, we can compare the estimated marginal computed using our method and that computed under KLD. As described earlier, based on the mean-field approximation, we assume that the posterior distribution factorizes: $q(\mu, \tau) = q(\mu)q(\tau)$. Under this assumption, it is easy to show that under the KLD, the optimal distribution of $\mu$ is $q^{*KL}(\mu) = N(\mu_0^*, \lambda_0^{-1})$ and the optimal distribution of $\tau$ is $q^{*KL}(\tau) = Ga(a^*, b^*)$. Thus, only the parameters of these two distributions need to be updated at each iteration. The updates are given by: $\mu_0^* = \frac{n\bar{x}}{1+n}$, $\lambda_0^* = (1+n)\int_{\mathbb{R}} \tau q(\tau) d\tau$ and $a^* = 0.01 + \frac{n+1}{2}$, $b^* = 0.01 + \frac{1}{2} \int_{\mathbb{R}} 2\mu^2 + (\sum_{i=1}^{n} x_i)^2 - 2\mu \sum_{i=1}^{n} x_i)q(\mu) d\mu$, where $n$ is the sample size and $\bar{x}$ is the sample mean of the data. In this study we utilize only 99 basis elements, to show the efficiency of our algorithms for $\alpha D$, but under the assumption that the approximating class is an exponential family [18]. The proposed approach is more general.
method. Multiple simulation studies reveal that increasing the number of basis elements can lead to better approximations.

We compare KLD-based VB to the proposed method (PM). We also compare the results computed using a numerical integral to evaluate the energy gradient to the approximation of the double integral (PMA) as proposed in Section 3.4. The evaluation is done on three simulated datasets as shown in Figure 1. For each of the simulations, we use $\alpha = 0.9$ for the lower bound (LB) and $\alpha = 1.1$ for the upper bound (UB) on the marginal density. Figure 1 displays the comparison of contour plots of the true posterior and other posterior approximations using the techniques discussed above. For all images, the true posterior is plotted in red and the KLD solution is plotted in green. The top row contains the posterior approximations based on the proposed method without the integral approximation, where $LB_{PM}$ and $UB_{PM}$ are plotted in blue and black, respectively. The bottom row contains the approximated posteriors based on the proposed method with the integral approximation, where $LB_{PMA}$ and $UB_{PMA}$ are plotted in cyan and magenta, respectively.

For improved presentation and for ease of comparison across different simulations, we rescale these values such that the optimal value is 1. In all cases, the different posterior approximations are very close to the true posterior, especially when the sample size is high. We also note that the LB on the marginal computed using PM is always tighter than the KL one. Even when we use an approximation PMA to the double integral in the energy gradient, our method outperforms KLD. Furthermore, the main advantage of PM is that it can also compute an UB on the marginal. Panel (c) shows that the PM is better at estimating the tails of the posterior than KL. Table 1 shows the utility of PM in statistical inference. Here, we use the first dataset (Figure 1(a)). First, we report the LB and UB on the posterior mean of both parameters $\mu$ and $\tau$. Second, we compute the LB and UB for the Bayes factor where Model (1) uses a $N(0, \tau^{-1})$ prior, and Model (2) uses a $N(2, \tau^{-1})$ prior. We note that the bounds on the posterior means and Bayes factor are very tight. In fact, the difference between the bounds is smaller than $1 \times 10^{-5}$ in the posterior mean case. Furthermore, the Bayes factor suggests that Model (1) (prior mean is 0) is better than Model (2), which is in line with our expectation (since the data was sampled from a $N(0, 1)$).

<table>
<thead>
<tr>
<th>Posterior mean of $\mu$</th>
<th>Posterior mean of $\tau$</th>
<th>Bayes factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>$LB_{PM}$</td>
<td>$UB_{PM}$</td>
<td>$LB_{PM}$</td>
</tr>
<tr>
<td>-0.0480</td>
<td>0.0480</td>
<td>1.0205</td>
</tr>
<tr>
<td>$LB_{PMA}$</td>
<td>$UB_{PMA}$</td>
<td>$LB_{PMA}$</td>
</tr>
<tr>
<td>-0.0481</td>
<td>-0.0480</td>
<td>1.0208</td>
</tr>
</tbody>
</table>

Table 1: LB = Lower Bound and UB = Upper Bounds on the Bayes factor and posterior means of $\mu$ and $\tau$. 
4.2. Bayesian Logistic Regression

Bayesian logistic regression is a very popular method for binary classification. Jaakkola and Jordan [20] considered variational methods for such models and extended them to binary belief networks. Before we proceed to the results, we give a brief description of the problem and our classification scheme based on the αD-based VB framework.

Let \( X \) be a \( p \times n \) matrix, where \( p \) is the number of covariates (or features) and \( n \) is the number of observations (or cases). Also, let \( \theta \) be a \( p \)-dimensional coefficient vector and \( y \) be an \( n \)-dimensional vector of class labels corresponding to the observations. The class labels take binary values in \( \{-1, 1\} \) (sometimes also labeled as \( \{0, 1\} \)). Under this setup, the logistic regression model is given by

\[
P(y|X, \theta) = g(\theta^T X),
\]

where \( g(r) = \frac{\exp(r)}{1 + \exp(r)} \). Our final goal is to estimate \( \theta \), the vector of unknown coefficients. For Bayesian inference in this setting, we assume a vague independent Gaussian prior distribution over all of the unknown parameters. Since the posterior does not have a closed form expression, we approximate it using \( q(\theta) \) via the proposed variational approach, where \( q(\theta) = \prod_{i=1}^{p} q_i(\theta_i) \). We use the proposed gradient ascent technique described in Sections 3.3 and 3.4 to obtain the \( q_i(\theta_i) \). Finally, we need to compute the probability \( P(y|X, \theta) \) for classification purposes. There exist various measures based on the posterior that can be used in this scenario. Based on the approximated posterior distribution \( q(\theta) \), we calculate the following summaries: maximum a posteriori (MAP), posterior mean (PMEA), posterior median (PMED) and posterior predictive (PPRED). Note that if the optimality criterion is chosen to be KLD instead of αD, we can still use the same gradient
ascent algorithm that we proposed earlier to approximate the posterior. Thus, all of the aforementioned
summaries (KLMAP, KLPMEA, KLPMED and KLPRED) can be obtained using the proposed algorithm
for a KLD VB framework as well. We present classification results in terms of accuracy (in %) for each of the
summary methods based on $\alpha_D$ and $KLD$.

For each of the examples, we use the training set to approximate the posterior distribution of the coefficient
vector. We then separately use the four summaries mentioned above to predict the binary class label in the
test data and evaluate classification accuracy. When deciding which threshold to choose for the binary
partition, we select the specific value for which the training error rate is minimized based on the posterior
predictive. If the predicted probability is greater than the cutoff, we set $y = 1$, and $y = -1$ otherwise.
Another popular choice for the cutoff in logistic regression is 0.5, which means that we assign a class label
$y = 1$ if the estimated probability of belonging to that specific class is greater than 0.5.

4.2.1. Ionosphere Data

The ionosphere dataset [21] is a standard binary classification benchmark, which we obtained from the
UCI Machine Learning Repository [22]. This data contains 34 continuous predictors corresponding to pulse
numbers of signals received by a radar. We remove the second predictor as it is zero for all cases. The binary
class labels correspond to good ($y = 1$) or bad ($y = -1$) radar returns. There is a total of 351 observations
and no missing values.

For classification, we split the full dataset into 200 training and 151 testing cases. We used the same split as
reported on the following web page: http://www.is.umk.pl/projects/datasets.html#Ionosphere. The
website also lists classification results on the same training-testing split for various classification methods.
We used all of the four summaries listed above, both for $\alpha_D$ (with two choices of $\alpha = 0.9$ and $\alpha = 1.1$),
and KLD VB. Table 2 presents the classification results obtained under this setup. The number of basis
elements used to approximate the energy gradient was fixed at 499. Note that for both choices of $\alpha$, we end
up with the same classification rate using $\alpha_D$-based VB and it also performs better than KLD VB. We can
also compare our results to those listed on the previously mentioned web page. With 6 misclassifications, the
proposed method ranks fifth best in a list of 23 total methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>MAP</th>
<th>PMEA</th>
<th>PMED</th>
<th>PPRED</th>
<th>KLMAP</th>
<th>KLPMEA</th>
<th>KLPMED</th>
<th>KLPPRED</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha = 0.9$</td>
<td>96.0</td>
<td>96.0</td>
<td>96.0</td>
<td>96.0</td>
<td>70</td>
<td>86.1</td>
<td>84.8</td>
<td>85.4</td>
</tr>
<tr>
<td>$\alpha = 1.1$</td>
<td>96.0</td>
<td>96.0</td>
<td>96.0</td>
<td>96.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Classification performance on the ionosphere dataset. Numbers represent classification accuracy in %.
4.2.2. Scene Data

The scene dataset was used for the problem of semantic scene classification in [23], where a scene might contain different objects such that it can be described by multiple class labels. The dataset contains 1,211 images in the training set and 1,196 images in the test set. The images consist of 294 image features that can be used to predict scene labels. There are up to 6 scene labels per image. Analogous to the previous example, this corresponds to six separate binary classification problems.

We use 299 basis elements for this classification problem to estimate the gradient, and choose $\alpha = 1.1$. The classification results obtained using the summary measures based on KLD and $\alpha$D are presented in Table 3. With the exception of the posterior median (PMEA), summaries based on $\alpha$D perform better than Jaakkola and Jordan’s method [20] where they achieve 87.4% accuracy. Our classification results are close to the results obtained by using Laplace inference and delta method inference introduced in [21]. However, they used cross validation techniques to estimate the performance of their algorithm, which is different from our setup since we use the original split to compute the probability threshold used to determine the labels for each binary class. The proposed method still produces results that are comparable to the reported accuracy rates.

<table>
<thead>
<tr>
<th>Method</th>
<th>MAP</th>
<th>PMEA</th>
<th>PMED</th>
<th>PPRED</th>
<th>KLMAP</th>
<th>KLPMEA</th>
<th>KLPMED</th>
<th>KLPPRED</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha = 1.1$</td>
<td>87.9</td>
<td>88.6</td>
<td>71.5</td>
<td>88.6</td>
<td>82.7</td>
<td>85.7</td>
<td>83.1</td>
<td>85.9</td>
</tr>
</tbody>
</table>

Table 3: Classification performance on the scene dataset accumulated over six scene labels. Numbers represent classification accuracy in %.

4.3. Application to Signature Verification

In this section, we consider the problem of signature verification. The signatures used here are a subset of the SVC 2004 signature dataset [24]. The data consists of 40 different signatures, each represented by a planar, open curve. For each signature, 20 genuine writing samples and 20 skilled forgeries are provided. We randomly split the data into half training and half testing. We propose to use novel shape based signature descriptors in conjunction with the proposed VB framework for this binary classification problem. Figure 2 displays four examples of pairs of genuine and forged signatures. The forgeries are extremely difficult to differentiate from the genuine samples making this a difficult classification problem.

To form our descriptors for classification, we use the elastic shape analysis method of Srivastava et al. [15], which provides tools for registering, comparing and averaging shapes of curves. Let $\beta : [0, 1] \to \mathbb{R}^2$ denote a planar, open, parameterized signature curve. In order to analyze its shape, $\beta$ is represented by a spectral function, called the square-root velocity function (SRVF) $q : [0, 1] \to \mathbb{R}^2$, defined as $q(t) = \frac{\dot{\beta}(t)}{\sqrt{\|\dot{\beta}(t)\|}}$. 
where $\dot{\beta} = \frac{d}{dt}\beta$ and $\|\|$ is the standard Euclidean norm. Because the SRVF is defined using the derivative of $\beta$, it is automatically invariant to translation; conversely, $\beta$ can be reconstructed from $q$ up to a translation. In order to achieve invariance to scale, each signature curve is re-scaled to unit length. Because shape is a quantity that is invariant to rotation and re-parameterization, in addition to translation and scale, these variabilities must also be removed from the representation space. This is performed algebraically using equivalence classes. Let $SO(2)$ be the group of $2 \times 2$ rotation matrices (special orthogonal group) and $\Gamma$ be the group of all re-parameterizations (orientation preserving diffeomorphisms of $[0, 1]$). For a curve $\beta$, a rotation $O \in SO(2)$ and a re-parameterization $\gamma \in \Gamma$, the transformed curve is given by $O(\beta \circ \gamma)$. The SRVF of the transformed curve is given by $O(q \circ \gamma)^{\sqrt{\gamma}}$. Using this, one can define equivalence classes of the type: $[q] = \{O(q \circ \gamma)^{\sqrt{\gamma}} | O \in SO(2), \gamma \in \Gamma\}$. Each such equivalence class $[q]$ is associated with a unique shape and vice-versa. Consider two signature curves $\beta_1$ and $\beta_2$, represented by their SRVFs $q_1$ and $q_2$. In order to compare their equivalence classes $[q_1]$ and $[q_2]$, fix $q_1$ and find the optimal rotation and re-parameterization of $q_2$ by solving $(O^*, \gamma^*) = \arg\min_{O \in SO(2), \gamma \in \Gamma} \|q_1 - O(q_2 \circ \gamma)^{\sqrt{\gamma}}\|^2$. This procedure optimally registers these two shapes. Minimization over the rotation group is performed using Procrustes analysis. Optimization over the re-parameterization group requires the dynamic programming algorithm. Once can also compute an average shape in this framework using the Karcher mean (minimizer of the sum of squared distances between shapes).

To form the signature shape descriptors, we begin by separately computing the average shapes for the genuine and forgery training sets. Next, we register each of the signatures in the training and test sets to both the genuine training average shape and the forgery training average. For each signature, this results in two different curves $\beta_{gen}^* = O_{gen}^*(\beta \circ \gamma_{gen}^*)$ and $\beta_{for}^* = O_{for}^*(\beta \circ \gamma_{for}^*)$. We then compute the speed functions (magnitude of tangential velocity) defined as $S_{gen}(t) = |\dot{\beta}_{gen}^*(t)|$ and $S_{for}(t) = |\dot{\beta}_{for}^*(t)|$ for each of these curves and concatenate them. The original signature curves are sampled with 100 points resulting in 200 signature shape descriptors.

For each type of signature, we use the training set to approximate the posterior distribution of the logistic regression model parameters using the proposed VB approach. As before, we use summaries of the posterior
based on $\alpha$D. The results accumulated over all test signatures (total of 800) are presented in Table 4. Note that the proposed shape based signature descriptors perform extremely well on this signature verification task. We used 99 basis elements with the choice of $\alpha = 0.9$. Since the split of the training and test set in this case is very balanced, we also present classification results obtained by using an empirical cutoff of 0.5. Interestingly, this choice of cutoff performs better than the results obtained by using the minimum training error cutoff based on the posterior predictive. Overall, the proposed method is very successful in this application.

<table>
<thead>
<tr>
<th>Method</th>
<th>MAP</th>
<th>PMEA</th>
<th>PMED</th>
<th>PPRED</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>100</td>
<td>91</td>
<td>96.5</td>
<td>83.3</td>
</tr>
<tr>
<td>(b)</td>
<td>100</td>
<td>99.8</td>
<td>99.6</td>
<td>99.8</td>
</tr>
</tbody>
</table>

Table 4: Classification performance on the signature dataset accumulated over 40 different signature types. Numbers represent classification accuracy in % for the probability cutoff based on (a) minimum training error using the posterior predictive and (b) 0.5.

5. Summary

We have proposed a novel variational Bayes approach based on $\alpha$-divergence under the mean-field restriction. Our approach has its roots in Riemannian geometry and uses the Fisher-Rao metric and square root transform representation of densities to define the variational energy. We propose an intrinsic gradient ascent algorithm on the space of PDFs to solve the problem at hand. We show theoretically and numerically that the proposed approach provides a tighter lower bound on the marginal density than the often used KL divergence-based setup. The framework is validated on multiple simulated examples to show the effectiveness of the proposed optimization algorithm. We apply this methodology to Bayesian logistic regression classification problems and show that our method performs favorably. Finally, we also define a novel set of signature shape descriptors and perform classification using the proposed method.

Our current approach has been promising with regard to accuracy and computation time since the geometric quantities are available analytically. However, there are some further issues which need to be addressed in future work: (1) judicious choice of appropriate basis functions in the tangent space, (2) theoretical guarantees for the proposed algorithm, and (3) efficient initialization schemes for different problems of interest. Encouragingly, the SRT representation space is a convex subset of the Hilbert sphere, and this fact might greatly assist us in studying convergence properties of the algorithm. Future directions of study include examining the behavior of the summaries based on the posterior distribution used for Bayesian logistic regression classification, and extending the proposed VB framework to a variety of Bayesian models including generalized linear models, graphical models, spatial models, etc.
References


