

# A Geometric Variational Approach to Bayesian Inference

Abhijoy Saha<sup>1</sup>, Karthik Bharath<sup>2</sup>, Sebastian Kurtek<sup>1</sup>

<sup>1</sup>Department of Statistics, The Ohio State University

<sup>2</sup>School of Mathematical Sciences, University of Nottingham

## 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 density representation, the manifold can be identified with the positive orthant of the unit hypersphere  $S^\infty$  in  $\mathbb{L}^2$ , and the Fisher–Rao metric reduces to the standard  $\mathbb{L}^2$  metric. Exploiting such a Riemannian structure, we formulate the task of approximating the posterior distribution as a variational problem on the hypersphere based on the  $\alpha$ -divergence. This provides a tighter lower bound on the marginal distribution when compared to, and a corresponding upper bound unavailable with, approaches based on the Kullback–Leibler divergence. We propose a novel gradient-based algorithm for the variational problem based on Fréchet derivative operators motivated by the geometry of  $S^\infty$ , and examine its properties. Through simulations and real data applications, we demonstrate the utility of the proposed geometric framework and algorithm on several Bayesian models.

*Keywords:* Infinite-dimensional Riemannian optimization; Gradient ascent algorithm; Square-root density; Bayesian density estimation; Bayesian logistic regression.

# 1 Introduction

Various algorithms based on optimization techniques, such as variational inference (VI) (Ghahramani and Beal, 1999), variational Bayes (VB) (Attias, 2000; Jaakkola and Jordan, 2000), Black Box- $\alpha$  (BB- $\alpha$ ) (Hernández-Lobato et al., 2016) and expectation propagation (EP) (Minka, 2001), have been successfully used to approximate the posterior distribution in the Bayesian setting. Recent advancements have made variational methods very useful for complex high-dimensional Bayesian models in view of their applicability in large scale data analysis (Hoffman et al., 2013; Broderick et al., 2013). In particular, VB methods have proved to be popular (Li and Turner, 2016) since they provide a lower bound on (the logarithm of) the marginal density or model evidence, thus offering a natural model selection criterion (Ueda and Ghahramani, 2002; McGrory and Titterton, 2007).

In essence, VB and Markov Chain Monte Carlo (MCMC) sampling techniques are distinct approaches to resolve the same problem of approximating the posterior distribution in a Bayesian model. In certain problems, VB methods are preferred to standard MCMC for two main reasons: MCMC suffers from high computational complexity when scaling to high dimensions, and assessing convergence of an MCMC algorithm (Carlin and Louis, 2008; Cowles and Carlin, 1996) is problematic. For a recent comparative account of the main issues with MCMC- and VB-based approaches, and for guidelines on preferring one over the other, see Blei et al. (2017).

While geometric information of the statistical model has been previously considered for improving MCMC techniques (Girolami and Calderhead, 2011), there is a striking paucity of the same in variational approaches to Bayesian inference; one exception to this is the work by Chen et al. (2015). The aim of our work is to demonstrate the utility in the explicit use of the intrinsic geometry of the space of probability density functions (PDFs) in variational approaches to Bayesian inference. We achieve this in two complementary ways: (1) we show how the Fisher–Rao Riemannian geometry of the space of nonparametric PDFs can

be used profitably to design a parameterization-invariant variational framework; and (2) we combine the geometric framework with the use of the  $\alpha$ -divergence in obtaining lower and upper bounds on the marginal density for a large class of Bayesian models, noted recently as an important extension within the  $\alpha$ -divergence framework of Li and Turner (2016).

## 1.1 Background

The inference problem is the following. For a given dataset  $x$ , the variational problem is to find a density  $q \in \mathcal{Q}$  over the unknown, hidden parameters (or latent variables)  $\boldsymbol{\theta}$  that best approximates the true posterior density  $p(\boldsymbol{\theta}|x)$ , by solving  $\arg \min_{q \in \mathcal{Q}} \mathcal{L}(p, q)$  for a suitable distance or divergence function  $\mathcal{L}$ . Traditional VB methods are typified by the use of the Kullback–Leibler divergence (KLD) for  $\mathcal{L}$  under the *mean-field* approximation that the class  $\mathcal{Q}$  consists of densities with independent marginals:  $\mathcal{Q} := \{q(\boldsymbol{\theta}) = \prod_i q_i(\theta_i)\}$ . In conditionally conjugate models, the  $q_i$ s belong to the same exponential family as the complete conditional distribution  $q_i(\theta_i|\boldsymbol{\theta}_{-i}, x)$ , where  $\boldsymbol{\theta}_{-i}$  denotes all of  $\boldsymbol{\theta}$  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, which often simplifies computation. The (approximate) solution to the variational problem is usually obtained by a gradient ascent (or descent) approach along the individual coordinates of  $\boldsymbol{\theta}$ , where the updates are simple and available as members of the same family (Beal, 2003; Bishop, 2006). Wang and Blei (2013) extended the VB approach to nonconjugate models and proposed two generic methods which use Gaussian approximations: Laplace variational inference and delta method variational inference. More recently, there have been a few approaches in literature to relax the mean-field approximation in variational Bayes (Rezende and Mohamed, 2015; Hoffman and Blei, 2015; Kingma et al., 2016; Kucukelbir et al., 2017). The assumption of a specific parametric form for the class of approximating densities for the posterior, e.g., Gaussian, is a common restriction for some of the aforementioned techniques including Hoffman and Blei (2015) and Kucukelbir et al. (2017).

The approximating class  $\mathcal{Q}$  should be large enough to include densities close to  $p(\boldsymbol{\theta}|x)$ . The restriction to a parametric family of distributions, e.g., exponential family, imposes restrictions on the statistical model under consideration. Moreover, the geometry of  $\mathcal{Q}$  plays an important role in the performance of a gradient-based or line-search algorithm. The lack of geometric considerations of  $\mathcal{Q}$  in the KLD-based VB framework was noted by Hoffman et al. (2013) wherein the (approximate) natural gradient, proposed by Amari (1998), capturing the curvature of the space through the Fisher information matrix, was used for updates in the gradient descent algorithm.

## 1.2 Motivation and Contributions

The proposed framework is mainly motivated by nonconjugate Bayesian models, but is equally applicable to conjugate ones, as demonstrated in the simulation examples in the sequel. Utility of VB procedures for nonconjugate models are influenced by three main inter-dependent factors: (1) choice of the variational family  $\mathcal{Q}$ , (2) choice of the loss function  $\mathcal{L}$ , and (3) computation of the gradient, and efficiency of exploration of  $\mathcal{Q}$  in gradient-based algorithms. The interplay between these three factors, and their impact on the quality of posterior approximations, can be captured and quantified under a geometric framework:  $\mathcal{Q}$  can be chosen in order to make it compatible with a Riemannian structure, with the resulting distance governing the choice of  $\mathcal{L}$ , under which local moves in the parameter space can be carried out using (Fréchet) directional derivatives.

To this end, for a continuous parameter set of  $d$  dimensions, we choose as  $\mathcal{Q}$  the nonparametric manifold of *all* probability densities in  $d$  dimensions that factorize. We equip  $\mathcal{Q}$  with the nonparametric Fisher–Rao (FR) Riemannian metric (simply referred to as the FR metric hereafter). The distinguishing feature of our approach lies in the fact that the variational problem is not defined directly on  $\mathcal{Q}$ , but instead on the space of all square-root PDFs. The square-root map transforms  $\mathcal{Q}$  onto the positive orthant of the infinite-dimensional unit sphere in  $\mathbb{L}^2$ . This simplifies computations through explicit expressions

for useful geometric quantities and operations (e.g., geodesic path and distance, exponential and inverse-exponential maps, parallel transport). Under such a setup, it is possible to obtain a ‘local linear’ representation of the  $d$ -dimensional density as a vector in the tangent space, a subspace of a suitable Hilbert space. This allows for a representation of the density with a basis set containing an infinite number of orthonormal functions spanning the tangent space. In practice, one is required to choose a finite number of basis functions resulting in a finite-dimensional representation of the density; the theoretical framework is unencumbered by such a restriction. An upside of the truncation to an  $N$ -basis representation is that the choice of  $N$  acts as a ‘bandwidth parameter’ when approximating the posterior density, and can hence be tuned to improve the quality of the approximation.

Leveraging the metric structure of  $\mathcal{Q}$ , we use the Rényi  $\alpha$ -divergence (Rényi, 1961) as the loss function  $\mathcal{L}$  in the variational formulation. The  $\alpha$ -divergence subsumes a large family of divergences (including the KLD). Our choice of  $\mathcal{L}$  is motivated by the fact that the FR metric is closely related to the  $\alpha$ -divergence for  $\alpha = 1/2$ , and also by the possibility of obtaining lower and upper bounds on the marginal density by suitably varying  $\alpha$  (Section 3.1 and Proposition 2), currently unavailable in existing literature. However, we note that the Riemannian framework under the FR metric can be employed using *any* divergence function as a choice for  $\mathcal{L}$  with appropriate adjustments. Armed with a versatile loss function  $\mathcal{L}$  defined through a Riemannian metric, the gradient direction in an ascent/descent algorithm is now defined as a Fréchet directional derivative along directions given by the orthonormal basis elements in the tangent space of the current iterate. This results in efficient explorations of  $\mathcal{Q}$ , as evidenced in the simulation and data analysis examples. We additionally prove the existence of an optimal step size for the gradient (Proposition 4). As with any VB procedure, computing the gradient direction requires us to approximate  $d$ -dimensional integrals. We present a novel approximation of the gradient based on a general first-order Taylor approximation of high-dimensional integrals developed by Olson and Weissfeld (1991). Such an approximation works quite well, even in fairly high dimensions.

The generality of the approximation makes it possible, in principle, to extend our framework to the non mean-field setting. We comment on this extension in Section 6 and leave it for future work.

To summarize, the main contributions of this paper are:

- We propose a Riemannian-geometric framework for variational inference for continuous  $d$ -dimensional densities based on the intrinsic geometry of the manifold of all PDFs equipped with the nonparametric FR metric. The approximating family  $\mathcal{Q}$  contains all  $d$ -dimensional densities on the parameter space with independent marginals.
- We show, theoretically and numerically, that the proposed approach using the  $\alpha$ -divergence loss function results in a tighter lower bound on the marginal density than the KLD-based VB approach. Our approach is also able to provide an upper bound on the marginal, which cannot be obtained with the standard KLD-based VB.
- We utilize the geometry of the space of PDFs to define a gradient ascent algorithm based on Fréchet derivatives to solve the variational problem. We also specify a technique to approximate the gradient function efficiently based on a novel first-order Taylor approximation argument.

The rest of the paper is organized as follows. Section 2 introduces the FR Riemannian geometric framework and describes the tools relevant to our analysis. In Section 3, we review the  $\alpha$ -divergence, and provide a detailed formulation of the variational problem within the FR framework. Further, we derive bounds for the marginal density based on an appropriate energy function closely related to the  $\alpha$ -divergence. In Section 4, we present a gradient ascent algorithm for approximating the posterior distribution, and examine its properties. In Section 5, we present a simulation study along with a few applications of the proposed method using various models including linear regression, density estimation and logistic regression. Section 6 includes a discussion of future work directions including possible ways to extend the proposed methodology to non-mean-field variational families.

## 2 Fisher–Rao Riemannian Geometry of PDFs

In this section, we introduce a representation space of PDFs, and associated geometric tools, which are useful in formulating the proposed variational method; most of these concepts have been previously summarized in Kurtek and Bharath (2015); Kurtek (2017).

For simplicity, we restrict our attention to the case of univariate densities on  $[0, 1]$ . We note however, that the framework is equally valid for all finite-dimensional distributions. Denote by  $\mathcal{P}$ , the Banach manifold of PDFs defined as  $\mathcal{P} = \{p : [0, 1] \rightarrow \mathbb{R}_{>0} \mid \int_0^1 p(x)dx = 1\}$ . Next, for a point  $p \in \mathcal{P}$ , consider a vector space that contains the set of tangent vectors at this point. This is defined as the tangent space at the point  $p$ ,  $T_p(\mathcal{P}) = \{\delta p : [0, 1] \rightarrow \mathbb{R} \mid \int_0^1 \delta p(x)p(x)dx = 0\}$ . Intuitively, the tangent space  $T_p(\mathcal{P})$  at any point  $p$  contains all possible perturbations of the PDF  $p$ . This tangent space can be used to define a suitable metric on the manifold  $\mathcal{P}$  as follows. For any  $p \in \mathcal{P}$  and any two tangent vectors  $\delta p_1, \delta p_2 \in T_p(\mathcal{P})$ , the nonparametric FR metric is given by  $\langle\langle \delta p_1, \delta p_2 \rangle\rangle_p = \int_0^1 \delta p_1(x)\delta p_2(x)\frac{1}{p(x)}dx$  (Rao, 1945; Kass and Vos, 2011). This metric is closely related to the Fisher information matrix, rendering it attractive to use in various statistical methods. An important property of this metric is that it is invariant to reparameterization (Cencov, 2000), i.e., smooth transformations of the domain of PDFs. However, since the FR metric changes from point to point on  $\mathcal{P}$ , it leads to cumbersome computations, which makes it difficult to use in practice. Thus, instead of working on  $\mathcal{P}$  directly under the FR metric, we use a suitable transformation that simplifies the Riemannian geometry of this space.

The square-root representation (Bhattacharyya, 1943) provides an elegant simplification. We define a mapping  $\phi : \mathcal{P} \rightarrow \Psi$ , where  $\phi(p) = \psi = \sqrt{p}$  is the square-root density (SRD) of a PDF  $p$ ; the inverse mapping is simply given by  $\phi^{-1}(\psi) = p = \psi^2$  (Kurtek and Bharath, 2015). The space of all SRDs is  $\Psi = \{\psi : [0, 1] \rightarrow \mathbb{R}_{>0} \mid \int_0^1 \psi^2(x)dx = 1\}$ , i.e., the positive orthant of the unit Hilbert sphere (Lang, 2012). Since the differential geometry of the sphere is well-known, one can define standard geometric tools on this space for

analyzing PDFs analytically. Let  $T_\psi(\Psi) = \{\delta\psi \mid \langle \delta\psi, \psi \rangle = 0\}$  denote the tangent space at  $\psi \in \Psi$ . Under the SRD representation, it is straightforward to show that, for any two vectors  $\delta\psi_1, \delta\psi_2 \in T_\psi(\Psi)$ , the FR metric reduces to the standard  $\mathbb{L}^2$  Riemannian metric:  $\langle \delta\psi_1, \delta\psi_2 \rangle = \int_0^1 \delta\psi_1(t)\delta\psi_2(t)dt$ . The corresponding geodesic distance between two PDFs  $p_1, p_2 \in \mathcal{P}$ , now represented by the SRDs  $\psi_1, \psi_2 \in \Psi$ , is now simply defined as the length of the shortest arc connecting them on  $\Psi$ :  $d_{FR}(p_1, p_2) = \cos^{-1}(\langle \psi_1, \psi_2 \rangle) = v$ .

We will use additional geometric tools to solve the variational inference problem in subsequent sections. 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  $\psi$ ,  $\exp_\psi : T_\psi(\Psi) \rightarrow \Psi$  is defined as  $\exp_\psi(\delta\psi) = \cos(\|\delta\psi\|)\psi + \sin(\|\delta\psi\|)\frac{\delta\psi}{\|\delta\psi\|}$ , where  $\|\cdot\|$  is the  $\mathbb{L}^2$  norm. Similarly for  $\psi_1, \psi_2 \in \Psi$ , the inverse-exponential map denoted by  $\exp_\psi^{-1} : \Psi \rightarrow T_\psi(\Psi)$  is  $\exp_{\psi_1}^{-1}(\psi_2) = \frac{v}{\sin(v)}(\psi_2 - \cos(v)\psi_1)$ ,  $v = d_{FR}(p_1, p_2)$ . With the help of these two tools from differential geometry, we can travel between  $\Psi$ , the representation space of SRDs, and  $T_\psi(\Psi)$ . Finally, we define parallel transport, which is used to map tangent vectors from one tangent space to another. We use the parallel transport along geodesic paths (great circles) in  $\Psi$ . For  $\psi_1, \psi_2 \in \Psi$ , and a vector  $\delta\psi \in T_{\psi_1}(\Psi)$ , the parallel transport of  $\delta\psi$  from  $\psi_1$  to  $\psi_2$  along the geodesic path is defined as  $\delta\psi^{\parallel} = \delta\psi_{\psi_1 \rightarrow \psi_2} = \delta\psi - \frac{2\langle \delta\psi, \psi_2 \rangle}{\|\psi_1 + \psi_2\|}(\psi_1 + \psi_2)$ , where  $\delta\psi^{\parallel} \in T_{\psi_2}(\Psi)$ . This defines a mapping  $\kappa : T_{\psi_1}(\Psi) \rightarrow T_{\psi_2}(\Psi)$  such that  $\delta\psi^{\parallel} = \kappa(\delta\psi)$ . An important property of parallel transport is that the mapping  $\kappa$  is an isometry between two tangent spaces, i.e., for  $\delta\psi_1, \delta\psi_2 \in T_{\psi_1}(\Psi)$ ,  $\langle \delta\psi_1, \delta\psi_2 \rangle = \langle \kappa(\delta\psi_1), \kappa(\delta\psi_2) \rangle$ .

### 3 Variational Inference Based on the $\alpha$ -Divergence

Our objective is to synthesize the benefits of using a divergence measure that leads to lower and upper bounds on the marginal density in a Bayesian model with the Riemannian geometric structure of the space of PDFs induced by the FR metric. To this end, starting with a review of Rényi's  $\alpha$ -divergence in Section 3.1, we outline the variational problem of



interest in Section 3.2 and formulate the corresponding optimization problem. In Section 3.3, we show how the use of the  $\alpha$ -divergence provides a tighter lower bound on the marginal density compared to the standard KLD-based VB setup, and in addition, an upper bound.

### 3.1 Rényi $\alpha$ -Divergence

Let us consider two probability distributions  $p$  and  $q$  on an  $d$ -dimensional set  $\Theta \subset \mathbb{R}^d$ . Then, the  $\alpha$ -divergence  $D_\alpha$  (Rényi, 1961) defined for  $\{\alpha : \alpha > 0, \alpha \neq 1\}$  is given by  $D_\alpha[p||q] = \frac{1}{\alpha - 1} \ln \int_{\Theta} p(\boldsymbol{\theta})^\alpha q(\boldsymbol{\theta})^{1-\alpha} d\boldsymbol{\theta}$ . The full 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  $q$ . Although  $D_\alpha$  can be defined for any  $\alpha > 0$ , certain special cases are noteworthy. In particular,  $D_\alpha$  is connected to KLD in two ways: (1)  $\lim_{\alpha \rightarrow 0} D_\alpha[p||q] = KL(q||p)$ , and (2)  $\lim_{\alpha \rightarrow 1} D_\alpha[p||q] = KL(p||q)$ . These limiting cases are defined using continuity of  $D_\alpha$  (Van Erven and Harremos, 2014). With specific regard to variational inference, VB attempts to minimize  $KL(q||p)$  globally, whereas EP attempts to minimize  $KL(p||q)$  locally. Another special case of  $D_\alpha$  is that for  $\alpha = 1/2$ , which is very closely related to the aforementioned FR metric. In fact, this is the only choice of  $\alpha$ , which results in a proper distance between PDFs.

### 3.2 Problem Formulation

Let  $x \in \mathcal{X}$  denote the observed data and  $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_d) \in \Theta$  denote the unknown  $d$ -dimensional parameter, where  $\{\Theta = (\Theta_1, \Theta_2, \dots, \Theta_d) : \theta_i \in \Theta_i\}$ . Let  $f(\boldsymbol{\theta}, x) = f(x|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$  denote the joint density of  $x$  and  $\boldsymbol{\theta}$  where  $f(x|\boldsymbol{\theta})$  is the likelihood function and  $\pi(\boldsymbol{\theta})$  is the prior distribution on  $\boldsymbol{\theta}$ . The posterior distribution is then given by  $p(\boldsymbol{\theta}|x) = \frac{f(x, \boldsymbol{\theta})}{m(x)}$  where  $m(x) = \int_{\Theta} f(x, \boldsymbol{\theta}) d\boldsymbol{\theta}$  denotes the marginal density of  $x$ , sometimes also called the model evidence. In practice, calculating the posterior is difficult because evaluating  $m(x)$  is hard in general, especially when analytical solutions are not available. In such scenarios, we have to resort to approximate Bayesian inference methods as discussed in Section 1. To

this effect, we consider a variational framework based on  $D_\alpha$ , where we wish to find a PDF to approximate the true posterior among the class of *all* joint PDFs that factorize.

Based on the mean-field approximation, let  $\mathcal{Q} = \{q \mid q = \prod_{i=1}^d q_i\}$  denote the class of strictly positive probability densities with support  $\Theta$  that contain independent marginals. Note that  $\mathcal{Q}$  is an infinite-dimensional set of PDFs on  $\Theta$ , and not a parametric class. Then, the  $\alpha$ -divergence between the posterior and an element of  $\mathcal{Q}$  is

$$D_\alpha[p \parallel \prod_{i=1}^d q_i] = \frac{1}{\alpha - 1} \ln \int_{\Theta} p(\boldsymbol{\theta} \mid x)^\alpha \left( \prod_{i=1}^d q_i(\theta_i) \right)^{1-\alpha} d\boldsymbol{\theta}, \quad \alpha > 0.$$

Note that for the limiting case of  $\alpha \rightarrow 1$ ,  $D_\alpha$  converges to the KLD between  $p$  and  $q$ , i.e.,  $\int_{\Theta} \ln \left( \frac{p(\boldsymbol{\theta} \mid x)}{q(\boldsymbol{\theta})} \right) p(\boldsymbol{\theta} \mid x) d\boldsymbol{\theta}$ . Since the integral in this case is with respect to the computationally intractable posterior density  $p$ , the optimization problem becomes difficult to handle. Thus, we do not consider this limiting case in our setup.

Minimizing  $D_\alpha$  over  $\mathcal{Q}$  is not straightforward for two reasons: (1) the nonlinear manifold structure of  $\mathcal{Q}$ ; (2) unavailability of analytical expressions for corresponding geometric quantities. In order to exploit the FR geometry of the space of probability densities for the task of minimizing  $D_\alpha$ , we use the SRD representation defined in Section 2. Accordingly, the set  $\mathcal{Q}_\psi = \{\psi_q \mid \psi_q = \prod_{i=1}^d \psi_{q_i}\}$  consists of elements of the  $d$ -fold product space  $\Psi_d = \Psi \times \Psi \times \dots \times \Psi$  of SRDs. Suppose the SRDs of the joint, marginal and the posterior are denoted by  $\psi_f$ ,  $\psi_m$  and  $\psi_p$ , respectively, and observe the following equivalence relationships:

$$\begin{aligned} (q_1^*, q_2^*, \dots, q_d^*) &= \arg \min_{\mathcal{Q}} D_\alpha[p \parallel \prod_{i=1}^d q_i] = \arg \min_{\Psi_d} \frac{1}{\alpha - 1} \int_{\Theta} \psi_p(\boldsymbol{\theta} \mid x)^{2\alpha} \left( \prod_{i=1}^d \psi_{q_i}(\theta_i) \right)^{2-2\alpha} d\boldsymbol{\theta} \\ &= \arg \min_{\Psi_d} \frac{1}{\alpha - 1} \int_{\Theta} \psi_f(x, \boldsymbol{\theta})^{2\alpha} \left( \prod_{i=1}^d \psi_{q_i}(\theta_i) \right)^{2-2\alpha} d\boldsymbol{\theta}. \end{aligned}$$

The last equality follows from the fact that  $\psi_m(x)$ , the SRD of the marginal  $m(x)$ , is constant in  $\boldsymbol{\theta}$ . Furthermore, when  $\alpha < 1$ , the factor  $(\alpha - 1)^{-1} < 0$ , and thus the minimization problem can be written as one of maximization. We can hence transfer the variational

problem defined on the manifold  $\mathcal{Q}$  of PDFs on  $\Theta$  to the  $d$ -fold product space  $\Psi_d$  of SRDs whose geometry is well-understood. Consequently, we define the energy functional  $\mathcal{E}_\alpha(\psi_q; \boldsymbol{\theta}) : \Psi_d \rightarrow \mathbb{R}_{>0}$  for a given element  $\psi_q = \prod_{i=1}^d \psi_{q_i}$  of  $\Psi_d$  as

$$\mathcal{E}_\alpha(\psi_q; \boldsymbol{\theta}) := \int_{\Theta} \psi_f(x, \boldsymbol{\theta})^{2\alpha} \left( \prod_{i=1}^d \psi_{q_i}(\theta_i) \right)^{2-2\alpha} d\boldsymbol{\theta}.$$

The case  $\alpha = 1/2$ , as mentioned earlier, links to the intrinsic FR Riemannian metric on the space of probability densities, and is therefore coordinate-invariant. A convenient byproduct of this is that the energy functional enjoys a certain invariance.

**Proposition 1** *Consider injective, differentiable coordinate reparametrizations  $\phi_i : \Theta_i \rightarrow \Theta_i$  such that  $\eta_i = \phi_i(\theta_i)$  for  $i = 1, \dots, d$  and  $\boldsymbol{\eta} = (\eta_1, \dots, \eta_d)$ . The energy functional  $\mathcal{E}_{1/2}(\psi_q; \cdot)$  satisfies the invariance property*

$$\mathcal{E}_{1/2}(\psi_q; \boldsymbol{\eta}) = \mathcal{E}_{1/2}(\psi_q; \boldsymbol{\theta}).$$

**Remark 1** *For simplicity, the coordinate reparameterizations  $\phi$  were defined as self-maps of  $\Theta_i$ . Indeed, the  $\phi_i$  can map  $\Theta_i$  to another space altogether, but the result of Proposition 1 would still hold as long as  $\phi_i$  is injective and differentiable for each  $i = 1, \dots, d$ . Importantly, it is easy to see that Proposition 1 holds only for  $\alpha = 1/2$  when integrating with respect to Lebesgue measure; the result does not hold for general reparameterizations  $\eta_i = \phi_i(\theta_1, \dots, \theta_d)$ ,  $i = 1, \dots, d$  since the Jacobian matrix is no longer diagonal and the corresponding determinant of the Jacobian cannot be expressed as a product of differentials.*

For a general  $\alpha > 0$ , we define the variational problem for approximating the posterior as

$$\arg \max_{\Psi_d} \mathcal{E}_\alpha(\psi_q; \cdot) \text{ if } \alpha \in (0, 1) \quad \text{or} \quad \arg \min_{\Psi_d} \mathcal{E}_\alpha(\psi_q; \cdot) \text{ if } \alpha \in (1, \infty).$$

The definition of the energy functional  $\mathcal{E}_\alpha$  distinguishes our approach to alternative varia-

tional formulations on the space  $\mathcal{Q}$  under a class of distance or divergence measures: in our setup, the variational problem is defined on  $\Psi_d$ , and we explicitly incorporate and utilize the underlying geometry of  $\Psi_d$  in minimizing  $\mathcal{E}_\alpha$ .

### 3.3 Bounds on the Marginal Density

The two important reasons for using  $D_\alpha$  (and not necessarily  $D_{1/2}$  or KLD) are:

1. It leads to a tighter lower bound on the marginal density than KLD.
2. It leads to an upper bound on the marginal density, which is not possible under KLD.

Recall that, under the traditional KLD-based VB setup, one minimizes the KLD between a member of the approximating class  $q$  and the true posterior  $p$ :

$$\begin{aligned} (q_1^{*KL}, q_2^{*KL}, \dots, q_d^{*KL}) &= \arg \min_{q \in \mathcal{Q}} \int_{\Theta} \ln \left( \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|x)} \right) q(\boldsymbol{\theta}) d\boldsymbol{\theta} = \arg \max_{q \in \mathcal{Q}} \int_{\Theta} \ln \left( \frac{p(\boldsymbol{\theta}|x)}{q(\boldsymbol{\theta})} \right) q(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= \arg \max_{q \in \mathcal{Q}} \int_{\Theta} \ln \left( \frac{f(x, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} \right) q(\boldsymbol{\theta}) d\boldsymbol{\theta} =: \arg \max_{q \in \mathcal{Q}} \mathcal{H}(f, q), \end{aligned}$$

where the third equality again stems from the fact that the marginal does not depend on  $\boldsymbol{\theta}$ . Thus, instead of minimizing  $KL(q||p)$ , one can choose to maximize  $\mathcal{H}(f, q)$  to obtain an equivalent solution to the original optimization problem.

For a general variational family (not necessarily one which factorizes), we formally state the two results given earlier on the logarithmic scale for ease of comparison with the KLD-based bound on the marginal density.

**Proposition 2** *The following inequalities hold for the marginal  $m(x)$ :*

- (i) For  $0 < \alpha < 1$  :  $\mathcal{H}(f, q) \leq \frac{1}{\alpha} \ln \mathcal{E}_\alpha(\psi_q; \cdot) \leq \ln m(x)$ , i.e.,  $D_\alpha$  provides a tighter lower bound on the marginal than KLD.
- (ii) For  $\alpha > 1$  :  $\ln m(x) \leq \frac{1}{\alpha} \ln \mathcal{E}_\alpha(\psi_q; \cdot)$ , i.e.,  $D_\alpha$  provides an upper bound on the marginal.

This proposition motivates the study of the properties of variational inference based on  $D_\alpha$ . In addition, the ability to compute a tighter lower bound and an upper bound on the marginal provides a novel approach to approximate Bayesian statistical inference. For example, we are able to bound the Bayes factor (ratio of two marginal densities under two models) above and below, providing better evidence for model choice.

## 4 Optimization via Gradient Ascent

The definition of the energy functional  $\mathcal{E}_\alpha : \Psi_d \rightarrow \mathbb{R}_{>0}$  does not require the geometric tools or the novel representation space of PDFs defined in Section 2. Indeed, the minimum of  $\mathcal{E}_\alpha$  on  $\Psi_d$  is independent of the Riemannian metric and the corresponding geometric tools. However, its determination through a line-search algorithm based on gradients of  $\mathcal{E}_\alpha$  is inextricably linked to the geometry of  $\Psi_d$  through the Fréchet or directional derivatives. Without restricting the class of approximating densities to parametric families, we will utilize Riemannian optimization tools under the FR framework and propose a gradient-based algorithm. Throughout this section, the subscript  $i = 1, \dots, d$  indexes quantities related to the parameter  $\theta_i$ .

The tangent space  $T_{\psi_{q_i}}(\Psi) = \{\delta\psi_{q_i} : \Theta_i \rightarrow \mathbb{R} \mid \int_{\Theta_i} \delta\psi_{q_i}(\theta_i)\psi_{q_i}(\theta_i)d\theta_i = 0\}$  at  $\psi_{q_i} \in \Psi$  is the vector subspace of square-integrable functions from  $\Theta_i$  to  $\mathbb{R}$ . This space is spanned by the set  $\mathcal{B}_i = \{b_i^k, k = 1, 2, \dots\}$  of orthonormal basis functions such that  $\int_{\Theta_i} b_i^k(\theta_i)\psi_{q_i}(\theta_i)d\theta_i = 0 \quad \forall k$ . The mean-field approximation on the class  $\Psi_d$  ensures that the gradient of  $\mathcal{E}_\alpha$  can be computed for its restriction  $\mathcal{E}_{\alpha|\Psi} : \Psi \rightarrow \mathbb{R}_{>0}$  to  $\Psi$  for each  $i = 1, \dots, d$ . The Hilbert space structure of the tangent space plays a crucial role in this computation.

**Proposition 3** *For each  $i = 1, \dots, d$ , the gradient  $\nabla\mathcal{E}_\alpha^i$  along direction  $b_i^k$  is given by:*

$$\nabla\mathcal{E}_\alpha^i = \sum_{k=1}^{\infty} D^i\mathcal{E}_\alpha(b_i^k)b_i^k = 2(1-\alpha) \sum_{k=1}^{\infty} \left[ \int_{\Theta} \psi_f(x, \boldsymbol{\theta})^{2\alpha} \prod_{j \neq i} \psi_{q_j}(\theta_j)^{2-2\alpha} \psi_{q_i}(\theta_i)^{1-2\alpha} b_i^k(\theta_i) d\boldsymbol{\theta} \right] b_i^k.$$

**Remark 2** *The gradient  $\nabla\mathcal{E}_\alpha^i$  represents an ascent or a descent direction depending on whether  $\alpha$  is lesser or greater than one, respectively. To unify the two cases, we use  $|\nabla\mathcal{E}_\alpha^i|$  to denote the value of the map  $\Theta_i \ni \theta_i \mapsto |\nabla\mathcal{E}_\alpha^i(b_i^k(\theta_i))| \in \mathbb{R}_{>0}$  at a fixed  $\theta_i$ . This ensures that the gradient always represents an ascent direction regardless of the value of  $\alpha$ .*

We use the geometry of the space  $\Psi$  to define an appropriate basis set  $\mathcal{B}_i$ ,  $i = 1, \dots, d$ . We explain the construction of this basis for  $\theta_i \in [0, 1]$  and note that it is easily extended to a general compact support. For this purpose, we use the tangent space at the SRD of the uniform distribution  $u_i$  on  $[0, 1]$  defined as  $T_{\psi_{u_i}}(\Psi) = \{\delta\psi_{u_i} : [0, 1] \rightarrow \mathbb{R} \mid \int_0^1 \delta\psi_{u_i}(\theta_i)d\theta_i = 0\}$ . We define the basis set  $\tilde{\mathcal{B}}_i = \{\sin(2\pi n\theta_i), \cos(2\pi n\theta_i), 1 - \theta_i \mid n \in \mathbb{Z}_+\}$ . It is easy to verify that all elements of this set are orthogonal to  $\psi_{u_i}$ . This basis is then orthonormalized using the Gram-Schmidt procedure under the  $\mathbb{L}^2$  metric to result in  $\mathcal{B}_i$ .

The above construction leads to an orthonormal basis only for  $T_{\psi_{u_i}}(\Psi)$ ; it can be extended to every point of  $\Psi$  using parallel transport (Section 2). The explicit expressions for parallel transport ensure that this can be done exactly, and that the resulting basis elements in the tangent space of the new point are orthonormal and remain orthogonal to the representation space. For implementing the algorithm practically, we need to choose a finite basis set. We let  $N$  denote the number of basis functions. This leads to the following gradient ascent algorithm for optimizing  $\mathcal{E}_\alpha$  on  $\Psi_d$  (Algorithm 1).

A key aspect of the algorithm is the availability of an explicit expression for the exponential map, which ensures that we remain in the space of SRDs. Our approach is then to separately update each  $\psi_{q_i}$  at every iteration until convergence. As this is a gradient-based approach, we are not guaranteed to arrive at the global solution. There are many approaches to initialize the algorithm. However, through simulation, we found that initialization does not play a crucial role with respect to convergence. In related work, Minka (2005) defined optimization algorithms for  $D_\alpha$ , but under the assumption that the approximating class is an exponential family. The proposed geometric approach is more general.

**Algorithm 1:** Gradient-ascent algorithm on  $\Psi_d$ .

Initialize:  $\psi_{q_1}^0 = \sqrt{q_1^0}, \dots, \psi_{q_d}^0 = \sqrt{q_d^0}$ ,  $l = 0$ ,  $\delta > 0$ ,  $l_{max}$  and  $N$ ;  
 For each  $i = 1, \dots, d$ , select orthonormal bases  $\mathcal{B}_i = \{b_i^k : \Theta_i \rightarrow \mathbb{R}, k = 1, \dots, N\}$   
 for tangent spaces  $T_{\psi_{q_i}^0}(\Psi)$ ;  
 Choose step size  $\epsilon > 0$ ;  
**while**  $\min\{|\nabla \mathcal{E}_\alpha^i|, i = 1 \dots, d\} > \delta$  **and**  $l < l_{max}$  **do**  
     Compute ascent direction  $|\nabla \mathcal{E}_\alpha^i|$ ;  
     Update:  $\psi_{q_i}^{l+1} = \exp_{\psi_{q_i}^l}(\epsilon |\nabla \mathcal{E}_\alpha^i|)$ ;  
     Parallel transport basis:  $b_i^k = b_{\psi_{q_i}^0 \rightarrow \psi_{q_i}^{l+1}}^k$  for  $i = 1, \dots, d$  and  $k = 1, \dots, N$ ;  
      $l = l + 1$ ;  
**end**  
 Return  $\psi_{q_i}^l, i = 1, \dots, d$ .

#### 4.1 Choice of Step Size and Approximation of the Gradient

The performance of the algorithm on  $\Psi_d = \Psi \times \dots \times \Psi$  is governed by its performance on the individual  $\Psi$ . The computation of the gradient  $\nabla \mathcal{E}_\alpha^i$  and the choice of the step size  $\epsilon$  are crucial in order for the algorithm to efficiently explore  $\Psi$ . For finite-dimensional optimization problems, the existence of an optimal  $\epsilon$  that guides its selection is given by the so-called Wolfe-conditions.

The proposed ascent algorithm is defined on an infinite-dimensional manifold; the corresponding Wolfe-conditions can be defined in terms of the functional  $\tilde{\mathcal{E}}_\alpha^i : T_{\psi_{q_i}}(\Psi) \rightarrow \mathbb{R}_{>0}$  with  $\tilde{\mathcal{E}}_\alpha^i(v_i) = \mathcal{E}_\alpha^i \circ \exp(v_i)$  for a tangent vector  $v_i$ . Note that  $\tilde{\mathcal{E}}_\alpha^i$  is now an element of the dual space of  $T_{\psi_{q_i}}(\Psi)$ , which is a linear subspace of  $\mathbb{L}^2(\Theta_i)$ . For a given ascent direction  $v_i \in T_{\psi_{q_i}}(\Psi)$ , the corresponding (weak) Wolfe-conditions that specify guidelines for the choice of the step size  $\epsilon$  are given by (Ring and Wirth, 2012):

$$\tilde{\mathcal{E}}_\alpha^i(\epsilon v_i) \geq \mathcal{E}_\alpha^i(\psi_{q_i}) + c_1 \epsilon D^i \mathcal{E}_\alpha^i(v_i) \quad \text{and} \quad D^i \tilde{\mathcal{E}}_\alpha^i(\epsilon v_i) D^i \exp(\epsilon v_i) v_i \leq c_2 D^i \mathcal{E}_\alpha^i(v_i), \quad (1)$$

where  $D^i \exp(\epsilon v_i)$  is the derivative of the exponential map at  $v_i \in T_{\psi_{q_i}}(\Psi)$ ,  $D^i \tilde{\mathcal{E}}_\alpha^i(\epsilon v_i)$  is the

directional derivative of  $\tilde{\mathcal{E}}_{\alpha|\Psi}$ , the restriction of  $\tilde{\mathcal{E}}_{\alpha}$  to  $\Psi$ , and  $0 < c_1 < c_2 < 1$ . It does not follow directly that, for a given algorithm on the infinite-dimensional manifold, an  $\epsilon$  satisfying Equation 1 exists. The following result clarifies this for the proposed approach.

**Proposition 4** *For an ascent direction  $v_i \in T_{\psi_{q_i}}(\Psi)$ , an  $\epsilon$  satisfying the Wolfe conditions in Equation 1 exists.*

One significant issue encountered when computing the gradient is the evaluation of an integral over the  $d$ -dimensional  $\Theta$ . While the mean-field approximation on  $\psi_q$  helps, the presence of the (square-root) joint density  $\psi_f(x, \boldsymbol{\theta})$  in the integrand complicates matters. We use a nested univariate first-order Taylor approximation of the multivariate integral proposed in Olson and Weissfeld (1991), which reduces a multivariate integral to functions of univariate ones. Briefly, the basis of the approximation method is 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  $g(y) = g(\mu) + g'(\mu)(y - \mu) + O_p(y - \mu)^2$ . Taking expectations on both sides, we obtain  $E(g(y)) = g(\mu) + 0 + O(V(y))$ . Thus,  $E(g(y))$  is approximated with  $g(\mu)$ .

For  $d$ -dimensional  $\boldsymbol{\theta}$ , consider the approximation of  $E(g(\boldsymbol{\theta})) = \int_{\Theta} g(\boldsymbol{\theta})f(\boldsymbol{\theta})d\boldsymbol{\theta}$ . Using the above argument,  $E(g(\boldsymbol{\theta}))$  can be expressed as:

$$\begin{aligned} E(g(\boldsymbol{\theta})) &= \int_{\theta_d} \left[ \int_{\theta_1 \times \theta_2 \times \dots \times \theta_{d-1}} g(\theta_1, \theta_2, \dots, \theta_d) f(\theta_1, \theta_2, \dots, \theta_{d-1} | \theta_d) d\theta_1 d\theta_2 \dots d\theta_{d-1} \right] f(\theta_d) d\theta_d \\ &= E_{\theta_d} \left[ \int_{\theta_1 \times \theta_2 \times \dots \times \theta_{d-1}} g(\theta_1, \theta_2, \dots, \theta_d) f(\theta_1, \theta_2, \dots, \theta_{d-1} | \theta_d) d\theta_1 d\theta_2 \dots d\theta_{d-1} \right], \end{aligned}$$

where  $f(\theta_1, \theta_2, \dots, \theta_{d-1} | \theta_d)$  is the density of  $(\theta_1, \theta_2, \dots, \theta_{d-1})$  conditional on  $\theta_d$ , and  $E_{\theta_d}$  denotes expectation with respect to  $\theta_d$ . Let  $\mu_d = E_{\theta_d}(\theta_d) = \int_{\theta_d} \theta_d f(\theta_d) d\theta_d$ . We use a first-order Taylor expansion to approximate the conditional expectation above about  $\mu_d$ :  $E(g(\boldsymbol{\theta})) \approx \int_{\theta_1 \times \theta_2 \times \dots \times \theta_{d-1}} g(\theta_1, \theta_2, \dots, \theta_{d-1}, \mu_d) f(\theta_1, \theta_2, \dots, \theta_{d-1} | \mu_d) d\theta_1 d\theta_2 \dots d\theta_{d-1}$ . We can keep on repeating the above approximation technique until we obtain the univariate integral



$E(g(\boldsymbol{\theta})) \approx \int_{\theta_1} g(\theta_1, \mu_{2|3,\dots,d}, \mu_{3|4,\dots,d}, \dots, \mu_d) f(\theta_1 | \mu_{2|3,\dots,d}, \mu_{3|4,\dots,d}, \dots, \mu_d) d\theta_1$ , where  $\mu_{j|j+1,\dots,d}$  is the conditional expectation of  $\theta_j | \theta_{j+1}, \dots, \theta_d$ ,  $j = 2, \dots, d - 1$ .

Consider the expression for the gradient  $\nabla \mathcal{E}_\alpha^i$ . Bearing in mind that in our setting the joint density is  $q = \prod_j q_j$ , applying the above approximation we can rewrite the integral in the expression for the gradient as

$$\int_{\Theta_d} \dots \int_{\Theta_2} \int_{\Theta_1} \frac{\psi_f(x, \theta_1, \theta_2, \dots, \theta_d)^{2\alpha}}{\prod_{j \neq i} \psi_{q_j}(\theta_j)^{2\alpha}} \prod_{j \neq i} q_j(\theta_j) \psi_{q_i}(\theta_i)^{1-2\alpha} b_i^k(\theta_i) d\theta_1 d\theta_2 \dots d\theta_d, \quad (2)$$

since  $q_j(\theta_j) = \psi_{q_j}^2(\theta_j)$ . We first compute the expectations  $\mu_j = \int_{\Theta_j} \theta_j q_j(\theta_j) d\theta_j$ ,  $\forall j \neq i$ . We then use these expected values to redefine the high dimensional integral as a one dimensional integral given by

$$\int_{\Theta_i} \frac{\psi_f(x, \mu_{-i}, \theta_i)^{2\alpha}}{\prod_{j \neq i} \psi_{q_j}(\mu_j)^{2\alpha}} \psi_{q_i}(\theta_i)^{1-2\alpha} b_i^k(\theta_i) d\theta_i,$$

where  $\mu_{-i}$  denotes all of  $\boldsymbol{\mu}$  except  $\mu_i$ . We apply the same first-order Taylor expansion technique to approximate the bounds on the marginal density as defined in Proposition 2.

## 5 Simulations and Real Data Examples

In this section, we present several examples that validate the proposed framework. In the first example, we consider a simulation study from a normal-gamma conjugate model where the posterior distribution is bivariate. Since the true value of the marginal density is known in this case, we can compare the marginal for a given dataset  $x$  to the bounds computed under our setup and to the lower bound obtained using KLD. Next, we assess the performance of our method in the context of Bayesian multiple linear regression and Bayesian density estimation using logistic Gaussian process priors. The last model we consider is logistic regression. In this case, we compare classification performance of our method to various other techniques. Finally, we consider a real signature verification experiment using novel shape-based signature descriptors.

## 5.1 Low-Dimensional Simulation Study

We consider the following hierarchical model:

$$x|\mu, \tau \stackrel{iid}{\sim} N(\mu, \tau^{-1}), \quad \mu|\tau \sim N(0, \tau^{-1}), \quad \tau \sim Ga(0.01, 0.01).$$

Because the posterior in this case is bivariate, we can evaluate the proposed method 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:  $q(\mu, \tau) = q(\mu)q(\tau)$ . It is easy to show that under the KLD-based VB, 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  $\mu_0^* = \frac{n\bar{x}}{1+n}$ ,  $\lambda_0^* = (1+n) \int_{\mathbb{R}_{>0}} \tau q(\tau) d\tau$ ,  $a^* = 0.01 + \frac{n+1}{2}$  and  $b^* = 0.01 + \frac{1}{2} \int_{\mathbb{R}} (2\mu^2 + (\sum_{i=1}^n x_i)^2 - 2\mu \sum_{i=1}^n x_i) q(\mu) d\mu$ , where  $n$  is the sample size and  $\bar{x}$  is the sample mean of the data. In the proposed algorithm, we use only 99 basis elements to show the efficiency of our method. Multiple simulation studies reveal that increasing the number of basis elements can lead to better approximations of the posterior.

We compare three different approaches: KLD-based VB (KLD), the proposed method with the gradient evaluated using a numerical integral (PM), and the proposed method with the gradient evaluated using the approximation described in Section 4.1 (PMA). Using Proposition 2, the lower ( $LB_{PM}$ ,  $LB_{KLD}$ ) and upper ( $UB_{PM}$ ) bounds on the marginal can be computed exactly in this scenario, since it only involves a two-dimensional integral. To show the efficiency of the proposed first-order integral approximation technique in this low-dimensional study, we calculate the lower ( $LB_{PMA}$ ) and upper ( $UB_{PMA}$ ) bounds for our method using the approximation described in Section 4.1 as well. 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.

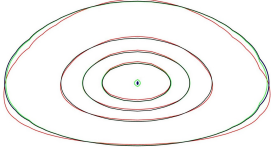
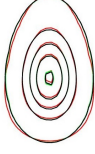
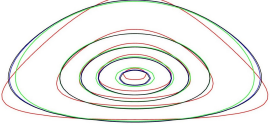
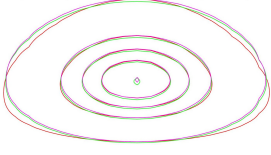

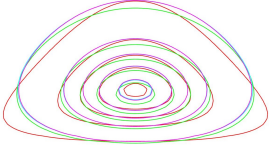
(a) $x_i \stackrel{iid}{\sim} N(0, 1), i = 1, \dots, 100$	(b) $x_i \stackrel{iid}{\sim} U(0, 2), i = 1, \dots, 100$	(c) $x_i \stackrel{iid}{\sim} t_2, i = 1, \dots, 20$
		
$LB_{PM} = 0.9995$ $UB_{PM} = 1.0005$ $LB_{KLD} = 0.9950$	$LB_{PM} = 0.9995$ $UB_{PM} = 1.0005$ $LB_{KLD} = 0.9950$	$LB_{PM} = 0.9975$ $UB_{PM} = 1.0025$ $LB_{KLD} = 0.9757$
		
$LB_{PMA} = 0.9992$ $UB_{PMA} = 1.0008$ $LB_{KLD} = 0.9950$	$LB_{PMA} = 0.9992$ $UB_{PMA} = 1.0008$ $LB_{KLD} = 0.9950$	$LB_{PMA} = 0.9960$ $UB_{PMA} = 1.0042$ $LB_{KLD} = 0.9757$

Figure 1: Contour plots of the approximated posteriors and the true posterior for three different simulated datasets. LB = lower bound, UB = upper bound, PM = proposed method, KLD = Kullback-Leibler divergence and PMA = proposed method with approximated integral. All of the values are to be compared to the optimal value of 1.

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 same results computed 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 the bound 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 and PMA is always tighter than the KLD one. Furthermore, the main advantage of PM/PMA is that it can also

Posterior mean of $\mu$		Posterior mean of $\tau$		Bayes factor	
$LB_{PM}$	$UB_{PM}$	$LB_{PM}$	$UB_{PM}$	$LB_{PM}$	$UB_{PM}$
-0.0480	-0.0480	1.0195	1.0205	7.9505	7.9664
$LB_{PMA}$	$UB_{PMA}$	$LB_{PMA}$	$UB_{PMA}$	$LB_{PMA}$	$UB_{PMA}$
-0.0481	-0.0480	1.0192	1.0208	7.9472	7.9697

Table 1: Lower (LB) and upper bounds (UB) on the Bayes factor and posterior means of  $\mu$  and  $\tau$ .

compute an UB on the marginal. Panel (c) shows that the proposed method is better at estimating the tails of the posterior than KLD. Table 1 shows the utility of the proposed method 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  $N(0, 1)$ ). These results suggest that the proposed approach has promise when extended to higher-dimensional and more complex Bayesian models.

## 5.2 Bayesian Linear Regression

In this section, we apply the proposed method to a Bayesian linear regression model. Let  $y = (y_1, y_2, \dots, y_n)$  be an  $n$ -dimensional vector denoting the continuous response variable, where  $n$  is the number of observations. Let  $X$  be an  $n \times d$  matrix, where  $d$  is the number of covariates and let  $\beta$  be a  $d$ -dimensional coefficient vector of regression parameters. Using matrix notation, the linear regression model can be written as  $y = X\beta + e$ , where  $e \sim N(0, \sigma^2 I_n)$ . For Bayesian inference, we assume a vague independent Gaussian prior distribution over all of the unknown regression parameters,  $\beta \sim N(0, s_0^2 I_d)$ . The true

posterior distribution can be easily determined, and is given by:

$$\boldsymbol{\beta}|y \sim N \left( \frac{1}{\sigma^2} \left( \frac{1}{\sigma^2} X'X + \frac{1}{s_0^2} I_d \right)^{-1} X'Y, \left( \frac{1}{\sigma^2} X'X + \frac{1}{s_0^2} I_d \right)^{-1} \right).$$

To assess the performance of our method, we use simulation studies with a varying number of covariates and estimate the  $q_i$ s for various choices of  $\alpha$ . For each value of  $d$ , we generate the design matrix  $X$  and the regression coefficients  $\boldsymbol{\beta}$  from a continuous uniform distribution,  $U(-1, 1)$ . We then proceed to estimate the unknown regression coefficients using different techniques. Under the proposed  $D_\alpha$ -based approach, the estimated posterior  $q(\boldsymbol{\beta})$  is a product of all of the  $q_i$ s,  $q(\boldsymbol{\beta}) = \prod_{i=1}^d q_i(\beta_i)$ , and the estimated individual regression coefficients are evaluated using the posterior means corresponding to each  $q_i$ . We compare our approach to a simple Gibbs sampling algorithm for each value of  $d$ , and estimate the coefficients using the posterior sample mean after suitable burn-in. To account for the variation in the randomly generated datasets and regression coefficients, we replicate each study  $rep$  times. Since the true posterior is known, we calculate the mean squared error (MSE) between the estimator  $\hat{\boldsymbol{\beta}}$  and the true value  $\beta_{\boldsymbol{\mu}}$ :  $MSE = \frac{1}{rep} \frac{1}{d} \sum_{k=1}^{rep} \sum_{i=1}^d \left( \hat{\beta}_{rep}^i - \beta_{\boldsymbol{\mu}}^i \right)^2$ .

Table 2 reports the results. For each choice of  $d$ , the estimated regression parameters obtained using the proposed method result in a very small MSE. Although the number of iterations and burn-in is quite large for the Gibbs sampler, it still results in a higher MSE than the  $D_\alpha$ -based approach. Further, to evaluate the efficiency of our method in a high-dimensional setting, we simulated a single dataset with  $d = 500$  covariates and  $n = 1000$  observations. The MSE obtained using the proposed method with  $\alpha = 0.5$  was  $4.9336 \times 10^{-7}$ , which shows its utility in the high-dimensional setting.

The true marginal is also available in closed form for the Bayesian linear regression setup:  $y \sim N(0, \sigma^2 I_n + s_0^2 X'X)$ . We use Proposition 2 to compute bounds on the logarithm of the marginal. For evaluating the high-dimensional integrals in  $\mathcal{H}(f, q)$  and  $\mathcal{E}_\alpha(\psi_q; \cdot)$  for KLD-based VB and  $D_\alpha$ -based VB, respectively, we use the proposed first-order Taylor

	$d = 25$	$d = 50$	$d = 100$	$d = 200$
	$n = 100$	$n = 100$	$n = 500$	$n = 500$
	$rep = 100$	$rep = 100$	$rep = 50$	$rep = 25$
Gibbs sampling ( <i>iter/burn-in</i> )	7.4368e-07 (50000/20000)	2.7044e-06 (50000/20000)	1.0359e-07 (60000/25000)	2.7866e-07 (60000/25000)
$\alpha = 0.5$	2.8065e-11	3.9600e-10	9.4248e-12	4.3790e-09
$\alpha = 0.9$	9.1023e-11	9.5112e-10	3.9182e-11	1.8072e-08
$\alpha = 1.1$	1.6681e-10	1.8017e-09	9.9131e-11	4.5736e-08

Table 2: MSE for Gibbs sampler and  $D_\alpha$ -based VB for  $\alpha = 0.5, 0.9, 1.1$ .  $d$ : number of unknown regression parameters,  $n$ : sample size,  $rep$ : number of simulated datasets for each choice of  $d$  and  $n$ ,  $\sigma^2 = 1$ ,  $s_0^2 = 100^2$ .

$d$	$n$	$LB_{KLDA}$	$LB_{PMA}$ $\alpha = 0.9$	$UB_{PMA}$ $\alpha = 1.1$	True log marginal
3	10	-27.5491	-27.5481	-27.2727	-27.5285
5	20	-54.2927	-54.2899	-53.6364	-54.0821
20	100	-273.5986	-273.5864	-272.7273	-273.0204
20	200	-425.8311	-425.7470	-425.4545	-425.8824
50	250	-695.2683	-695.2685	-694.5455	-694.8856

Table 3: Lower (LB) and upper bounds (UB) on the logarithm of the marginal using KLD- and  $D_\alpha$ -based VB.

approximation technique. Let  $LB_{KLDA}$  denote the lower bound obtained using the KLD-based VB framework, and  $LB_{PMA}$  and  $UB_{PMA}$  denote the lower and upper bounds obtained using the proposed methodology with  $\alpha = 0.9$  and  $\alpha = 1.1$ , respectively (these bounds are again computed using the method discussed in Section 4.1). Table 3 reports the results for different choices of  $d$  and  $n$ . In all cases, the PMA lower bound is tighter than the KLDA lower bound, with highest differences seen when  $d$  is large. The upper bound provided by PMA is also close to the true value of the log-marginal. One could potentially use the average of the lower and upper bounds as an estimate of the true value.

Finally, in Figure 2, we report 95% equal-tailed posterior credible intervals based on the estimated  $q_i$ s using the proposed method with  $\alpha = 0.5$ . The  $x$ -axis represents the regression parameter number and the  $y$ -axis represents the value of the parameter. In all cases, the intervals do a good job of capturing the value of the true regression coefficient,

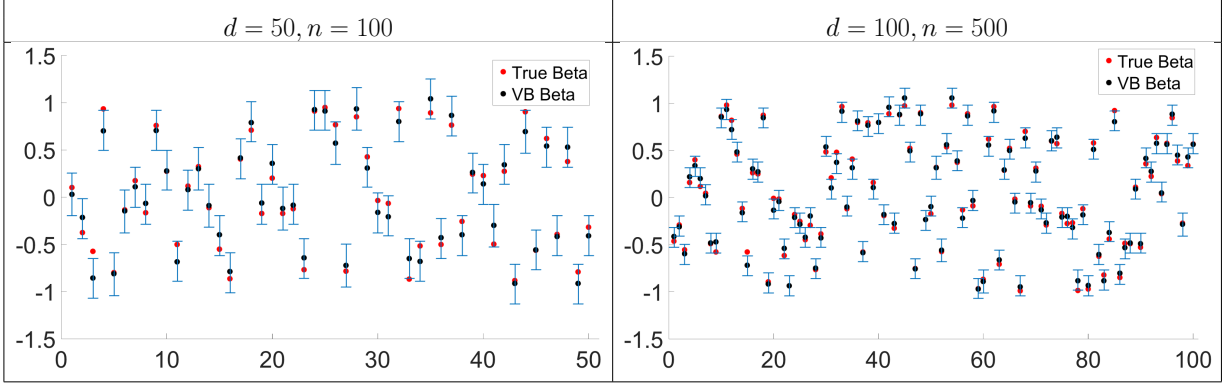


Figure 2: Equal-tailed 95% posterior credible intervals for  $\alpha = 0.5$ .

including the second example ( $d = 100, n = 500$ ), where the intervals get smaller due to a larger sample size.

### 5.3 Bayesian Density Estimation

Logistic Gaussian process (LGP) priors (Leonard, 1978) have been efficiently used as a flexible tool for Bayesian nonparametric density estimation. Theoretical properties of this model have been studied extensively (Tokdar and Ghosh, 2007; van der Vaart and van Zanten, 2009). Further, a quick approximation using Laplace’s method for LGP density estimation and regression was proposed in Riihimäki and Vehtari (2014). The resulting posterior distribution obtained using the LGP prior is analytically intractable because of the integral term, which appears in the likelihood function. Before proceeding to show how the proposed method can be used in this setting, we briefly review the LGP model.

Let  $x_1, x_2, \dots, x_n$  denote a random sample of size  $n$  drawn from an unknown univariate density function,  $f$ . Let  $\mathcal{X}$  denote the support of the distribution. To estimate  $f$ , we use the logistic density transform (Leonard, 1978)  $f(x) = e^{g(x)} / \int_{\mathcal{X}} e^{g(x)} dx$ , where  $g$  is an unconstrained function. Thus, the problem of estimating the unknown density function  $f$  reduces to estimating the function  $g$ . This transformation is useful as it introduces two necessary constraints for  $f$  to be a valid pdf:  $f(x) > 0$  and  $\int_{\mathcal{X}} f(x) dx = 1$ . To estimate the function  $g$ , we use a basis expansion model, i.e.,  $g(x) = \sum_{i=1}^d c_i b_i(x)$ , where

$c_i$ s are the basis coefficients,  $b_i$ s are the basis functions, and  $d$  denotes the number of basis functions used to estimate  $g$ . We place a noninformative Gaussian prior  $\pi_i$  on the unknown coefficients:  $c_i \sim N(0, s_0^2)$ ,  $\forall i = 1, \dots, d$ . Let  $\mathbf{x} = (x_1, \dots, x_n)$  and  $\mathbf{c} = (c_1, \dots, c_d)$ . The joint density function can then be written as  $f(\mathbf{x}, \mathbf{c}) = \prod_{j=1}^n f(x_j|\mathbf{c}) \prod_{i=1}^d \pi_i(c_i)$ , where  $f(x_j|\mathbf{c}) = \frac{\exp\{\sum_{i=1}^d c_i b_i(x_j)\}}{\int_{\mathcal{X}} \exp\{\sum_{i=1}^d c_i b_i(x)\} dx}$ .

We then use the proposed method to approximate the posterior  $p(\mathbf{c}|\mathbf{x})$  using  $q(\mathbf{c})$ , where  $q(\mathbf{c}) = \prod_{i=1}^d q_i(c_i)$ . Once the approximation to the posterior distribution for each coefficient  $c_i$  has been obtained, we calculate the posterior mean,  $\hat{c}_i$ ,  $\forall i = 1, \dots, d$ . The expression for the estimated density function is finally given by:

$$\hat{f}(x) = \frac{\exp\{\sum_{i=1}^d \hat{c}_i b_i(x)\}}{\int_{\mathcal{X}} \exp\{\sum_{i=1}^d \hat{c}_i b_i(x)\} dx}.$$

To validate the efficiency of our method for estimating density functions, we performed several simulation studies. A random sample was generated from the true underlying distribution in each case, and a histogram corresponding to the sample was used to represent the data. In all of the figures in this section, we plot the true density function in red and the estimated density function in blue. Based on the random sample, we also plot the kernel density estimate in black to provide a visual comparison between the two estimation techniques. The value of  $\alpha$  for the proposed method in Figures 3 and 4 was chosen to be 0.5. Among the multiple choices for basis functions that can be used to estimate the function  $g$ , we used B-splines of order four for all of the simulation studies; we also found that Fourier basis provided comparable results. A set of MATLAB code files, supplemental to the book by Ramsay et al. (2009) is available for download, and was used to generate the basis functions for all of the examples.

First, we generated datasets from various distributions which exhibit different features as shown in Figure 3. The third row of Figure 3 shows two plots generated from a Gamma distribution and Beta distribution in the left and right panels, respectively. While imple-



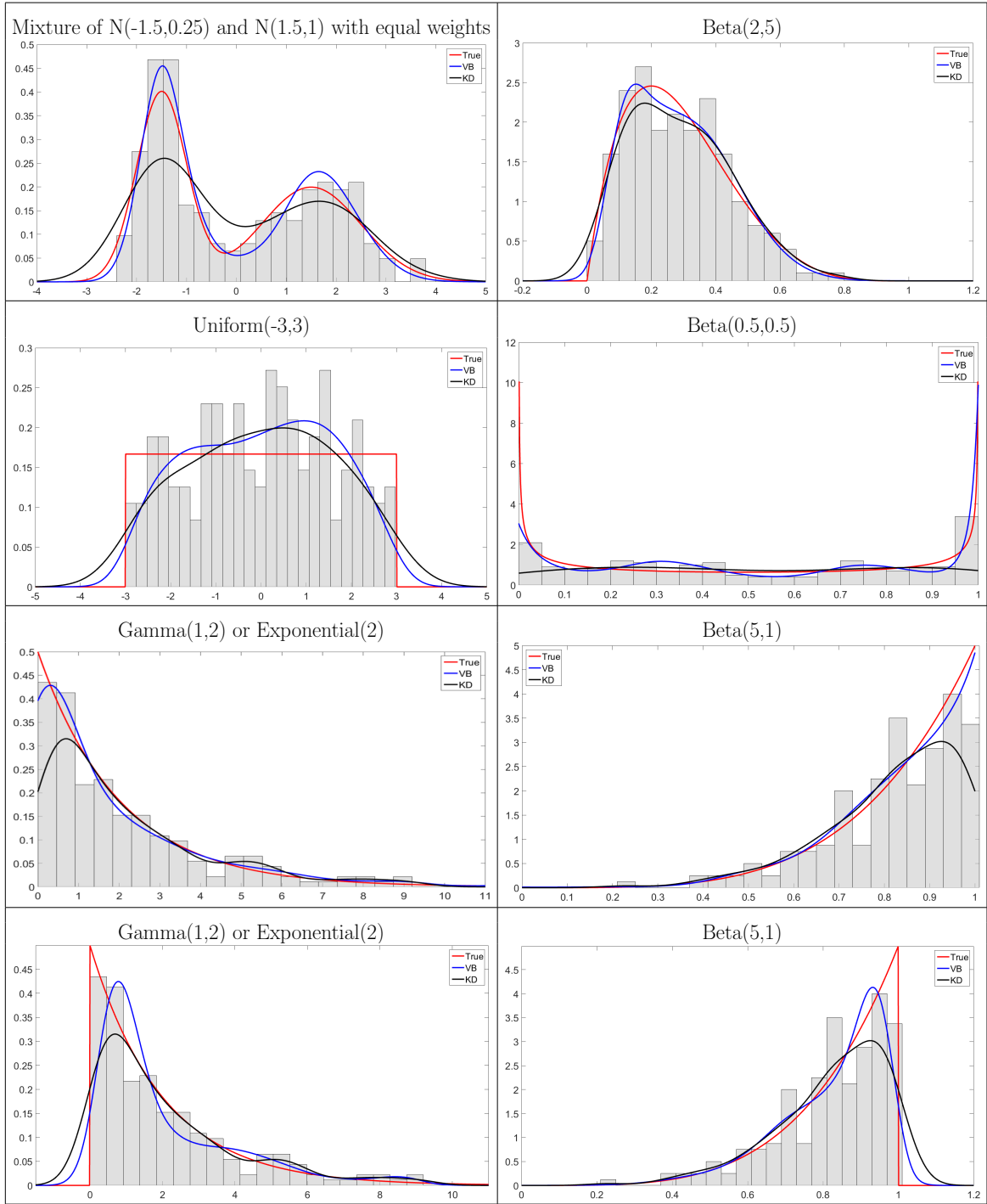


Figure 3: Examples of Bayesian density estimation for various density functions. The simulated data is displayed as a histogram with a plot of the true density (red), density estimate using the proposed method with  $\alpha = 0.5$  (blue), and a kernel density estimate (black).

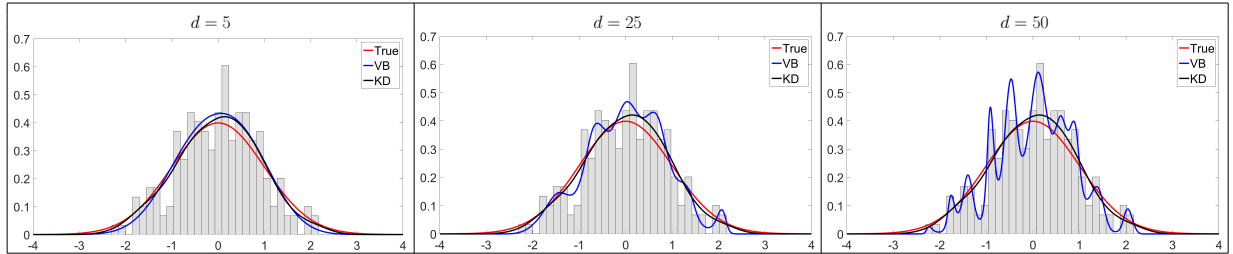


Figure 4: Effect of increasing the number of basis elements on density estimation. Data was generated from a  $N(0,1)$ .

menting our algorithm for density estimation in this example, we set the lower bound for density estimation as 0, since the support of the Gamma distribution is  $(0, \infty)$ . Similarly, owing to the support of the Beta distribution, i.e.,  $[0, 1]$ , we set the lower and upper bounds as 0 and 1, respectively. In practice, the support of the distribution may be unknown; thus, in the last panel we show the same results but without the use of information about the support of the true density. In all cases, the proposed method performs very well compared to standard kernel density estimation.

Figure 4 shows the effect of increasing the number of B-spline basis functions used to model  $g$ . The number of basis functions,  $d$ , used for estimating the density has a large impact on the final estimate, and behaves similarly to the bandwidth parameter in the kernel density estimator. As we increase the value of  $d$ , the smoothness of the resulting estimate decreases, and we tend to overfit the data.

#### 5.4 Bayesian Logistic Regression for Real Data Applications

We examine the performance on the proposed methodology on binary classification problems using Bayesian logistic regression models. Our choice is motivated by the fact that this is a nonconjugate model that does not fit into the VB setup with conjugate updates. Jaakkola and Jordan (1997) considered variational methods for such models and extended them to binary belief networks. We illustrate that the performance of the proposed geometry-based method is comparable to other approximations, and even better in

certain scenarios.

First, we give a brief description of the problem and our classification scheme based on the  $D_\alpha$  framework. Let  $X$  be a  $d \times n$  matrix, where  $d$  is the number of covariates (features) and  $n$  is the number of observations (cases). Also, let  $\boldsymbol{\theta}$  be a  $d$ -dimensional coefficient vector and  $y$  be an  $n$ -dimensional vector of class labels corresponding to the observations. The class labels take binary values in  $\{-1, 1\}$ . Under this setup, the logistic regression model is given by  $P(y|X, \boldsymbol{\theta}) = g(\boldsymbol{\theta}^T X)$  for class label  $y = 1$ , and  $P(y|X, \boldsymbol{\theta}) = g(-\boldsymbol{\theta}^T X)$  for class label  $y = -1$ , where  $g(r) = \frac{\exp(r)}{1 + \exp(r)}$ . Our final goal is to estimate  $\boldsymbol{\theta}$ , the vector of unknown coefficients. We again assume vague independent Gaussian priors over all of the unknown parameters in this setting, in the same manner as in Section 5.2. Since the posterior under this setup does not have a closed form expression, we approximate it using  $q(\boldsymbol{\theta}) = \prod_{i=1}^d q_i(\theta_i)$  via the proposed variational approach. Finally, for classification purposes, we need to compute the probability  $P(y|X, \boldsymbol{\theta})$ . There exist various choices based on different features of the posterior that can be used in this scenario; we calculate the following summaries: maximum a posteriori (MAP), posterior mean (PMEA), posterior median (PMED) and posterior predictive (PPRED). If the optimality criterion is chosen to be KLD instead of  $D_\alpha$ , we can still use the proposed gradient-based algorithm to approximate the posterior. Thus, all of the aforementioned summaries (KLMAP, KLPMEA, KLPMED and KLPPRED) can be obtained using the proposed algorithm for a standard KLD VB framework as well. We use this approach for comparison to  $D_\alpha$  and present classification results in terms of accuracy (in %) for each of the methods.

For both of the examples that follow, we use a 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 a test dataset, and evaluate the classification accuracy. We select a threshold for the binary partition, which minimizes the training error rate based on the posterior predictive in the training set. If the predicted probability is greater than the cutoff, we set  $y = 1$ , and  $y = -1$  otherwise. Further, we also calculate

	MAP	PMEA	PMED	PPRED	KLMAP	KLPMEA	KLPMED	KLPPRED
Accuracy (in %)	96.0	96.0	96.0	96.0	94.04	94.70	94.70	94.70
ALPL	-0.1980	-0.1879	-0.1979	-0.1883	-0.2217	-0.1886	-0.2042	-0.1895

Table 4: Classification results for the ionosphere dataset.

the average log predictive likelihood (ALPL) based on the test set. Given an observation from the test set, we calculate  $\log P(y|X, \hat{\theta})$  based on the value of the binary class label  $y$ , where  $\hat{\theta}$  is one of the posterior summaries considered above. A high value of the likelihood signifies better fit of the model.

In the first example, we use a standard benchmark dataset to compare the classification results obtained using the proposed methodology to many other approaches. Further, we compute bounds on the marginal density of the data using the proposed method. In the second example, we apply our approach to the problem of signature verification. We first define a novel set of shape-based descriptors, and then use them as features in a binary genuine vs. forgery classification problem.

#### 5.4.1 Ionosphere Data

The ionosphere dataset (Sigillito et al., 1989) is a standard binary classification benchmark, which we obtained from the UCI Machine Learning Repository (Bache and Lichman, 2013). This data contains 34 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. Good radar returns were defined as those showing some type of structure in the ionosphere. 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 use the same split as reported at <http://www.is.umk.pl/~duch/projects/projects/datasets.html#Ionosphere>. This split is very unbalanced: in the training set, the sizes of the two classes are 101 (50.5%) and 99 (49.5%), whereas in the test set, the sizes are

124 (82%) and 27 (18%), respectively. This website also provides classification results on the same training-testing split for various classification methods. We used the four summaries listed above, both for  $D_\alpha$  (with  $\alpha = 0.9$ ) and KLD-based VB, to compute the classification rate. In both cases, we used 499 basis elements to approximate the energy gradient. Table 4 presents the results. The proposed method clearly outperforms the KLD-based VB approach, both in terms of classification accuracy and ALPL. We can also compare our results to those listed on the previously mentioned webpage. With six misclassifications, the proposed method ranks fifth best in a list of 23 total methods.

The marginal distribution for the Bayesian logistic regression setup is unavailable in closed form. However, using the same technique as discussed in Section 5.2, we can find bounds on the logarithm of the marginal. For calculating bounds using  $D_\alpha$ -based VB, we choose  $\alpha = 0.9$  and  $\alpha = 1.1$  for lower and upper bounds, respectively. The lower bound obtained using KLD-based VB is  $-459.5$ . Using the proposed method, the lower bound is  $-456.7$ , and the upper bound is  $-448.2$ .

#### 5.4.2 Application to Signature Verification

In this section, we consider the problem of signature verification. The data used here are a subset of the SVC 2004 signature dataset (Yeung et al., 2004), which 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 variational Bayes framework for this binary classification problem. Figure 5 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. (2011), which provides tools for registering, comparing and averaging







(a)	(b)	(a)	(b)	(a)	(b)
					

Figure 5: Three examples of (a) genuine and (b) forged signatures.

shapes of curves. Let  $\beta : [0, 1] \rightarrow \mathbb{R}^2$  denote a planar, open, parameterized signature curve. In order to analyze its shape,  $\beta$  is represented by a special function, called the square-root velocity function (SRVF)  $q : [0, 1] \rightarrow \mathbb{R}^2$ , defined as  $q(t) = \dot{\beta}(t)|\dot{\beta}(t)|^{-1/2}$ , where  $\dot{\beta} = \frac{d}{dt}\beta$  and  $|\cdot|$  is the standard Euclidean norm in  $\mathbb{R}^2$ . 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 reparameterization, 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 reparameterizations (orientation preserving diffeomorphisms of  $[0, 1]$ ). For a curve  $\beta$ , a rotation  $O \in SO(2)$  and a reparameterization  $\gamma \in \Gamma$ , the transformed curve is given by  $O(\beta \circ \gamma)$ . The SRVF of the transformed curve is given by  $O(q \circ \gamma)\sqrt{\dot{\gamma}}$ . Using this, one can define equivalence classes of the type  $[q] = \{O(q \circ \gamma)\sqrt{\dot{\gamma}} \mid O \in SO(2), \gamma \in \Gamma\}$ . 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 reparameterization of  $q_2$  by solving

$$(O^*, \gamma^*) = \arg \min_{O \in SO(2), \gamma \in \Gamma} \|q_1 - O(q_2 \circ \gamma)\sqrt{\dot{\gamma}}\|^2.$$

This procedure optimally registers these two shapes. Minimization over the rotation group is performed using Procrustes analysis. Optimization over the reparameterization group requires the dynamic programming algorithm. One can also compute an average shape in

		MAP	PMEA	PMED	PPRED
Accuracy (in %)	(a)	100	91	96.5	83.3
	(b)	100	99.8	99.6	99.8
ALPL		-2.8848e-07	-0.0075	-0.0115	-0.0178

Table 5: Classification results for the signature dataset averaged over 40 different signature types for two methods. (a) Minimum training error cutoff and (b) empirical 0.5 cutoff.

this framework using the Karcher mean (minimizer of the sum of squared distances).

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 shape using Equation 5.4.2. 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 variational approach. We use 99 basis elements to approximate the energy gradient with  $\alpha = 0.9$ . As before, we use summaries of the approximate posterior to compute the classification performance. The results averaged over all test signatures (total of 800) are given in Table 5. Note that the proposed shape-based signature descriptors perform extremely well on this signature verification task, both in terms of accuracy and ALPL. Since the split of the training and test set in this case is very balanced, we also present classification results obtained using an empirical cutoff of 0.5. Interestingly, this choice of cutoff performs better than the results obtained using the minimum training error cutoff based on the posterior predictive. Overall, the proposed method is very successful in this application.

## 6 Discussion

The use of Fisher–Rao Riemannian geometry for the analysis of PDFs has been demonstrated in various settings, including diffeomorphic density matching Bauer et al. (2015), random sampling via optimal information transport Bauer et al. (2017), sensitivity analysis in Bayesian models Kurtek and Bharath (2015), and computer vision Srivastava et al. (2007). The unified metric structure and availability of high-speed computing resources provide a natural habitat for the formulation of variational versions of several tasks involving high-dimensional data. Theoretical study of resulting estimates and their comparison with ones currently used within the statistical literature, with a view towards inference, will be highly beneficial.

By moving to the space of nonparametric densities, the availability of explicit expressions for the exponential and inverse-exponential maps under the SRD representation plays a crucial role in the scalability of the proposed gradient ascent algorithm. Our approximation of the gradient direction is based on nested univariate first-order Taylor expansions of a high-dimensional integral; while this worked well in our investigations, better approximation schemes can be explored. There are multiple direction for future work including (1) examination of choice of appropriate basis functions in the tangent space to better capture modalities of the posterior, (2) building upon Proposition 4 to obtain theoretical guarantees for the proposed algorithm (encouragingly, the SRD representation space is a convex subset of the Hilbert sphere, and this will assist us in studying convergence properties), (3) development of efficient initialization schemes for different problems of interest, and (4) extending the proposed framework to a variety of other Bayesian models including generalized linear models, graphical models, spatial models.



## 6.1 Extension to Non-mean-field Setting

We comment now on how the proposed approach can be extended to the setting where we do not assume that joint densities  $q$  on  $\Theta$  factorize. The definition of the variational family and the square-root map remain unchanged. The definition of the loss function  $\mathcal{L}$  can be easily modified to reflect the new variational family. The significant changes lie in the implementation of the proposed algorithm. Recent work by Tan (2018) considered a model-dependent reparameterization trick that can capture posterior dependencies between parameters. The invertible affine transformation that they propose is similar in nature to the reparameterization considered in Proposition 1, and can thus be used in our setting. However, the reparameterization invariance only applies when  $\alpha = 1/2$ , limiting the applicability of this approach.

The key ingredients of the algorithm are the orthonormal bases, the exponential map, the gradient direction and the parallel transport. The exponential map and the parallel transport can appropriately be modified to reflect the  $d$ -dimensional nature of the density space. Given the  $d$ -dimensional basis set of orthonormal basis functions, the expression for the gradient can be written down explicitly. The computation of the gradient however is not straightforward. The key observation here lies in our approximation method based on nested approximations. Denote by  $f_d(\theta_1) := f(\theta_1 | \mu_{2|3,\dots,d}, \mu_{3|4,\dots,d}, \dots, \mu_d)$  the density  $f$  of  $\theta_1$  conditioned on the conditional expectations (e.g.,  $\mu_{2|3,\dots,d}$  denotes the conditional expectation of  $\theta_2$  given  $\theta_3, \dots, \theta_d$ ), and  $\tilde{f}_d(\theta_1) := f(\theta_1, \mu_{2|3,\dots,d}, \mu_{3|4,\dots,d}, \dots, \mu_d)$ . The resulting modification of Equation 2 for  $i = 1$  is  $\int_{\Theta_1} \left[ \frac{\psi_f(x, \theta_1, \theta_2, \dots, \theta_d)}{\psi_{q_d}(\theta_1)} \right]^{2\alpha} \tilde{b}_d^k(\theta_1) q_d(\theta_1) d\theta_1$ , where  $\tilde{b}_d^k$  is the  $k$ th element of the  $d$ -dimensional orthonormal basis function set. This requires us to compute only one-dimensional conditional expectations. One approach to this is to start with a parametric family for the approximating density and embed it into the nonparametric space of all  $d$ -dimensional densities. If  $d$  is too large, we can consider a more generalized block structure similar to structured mean-field approximation (Saul and Jordan, 1996;

Barber and Wierginck, 1999). Instead of assuming that all the parameters are mutually independent and controlled by their individual marginals, we can exploit the presence of a substructure in the collection of parameters, assume partial factorization, and continue along the lines mentioned above. The key point however is that the proposed framework can, in principle, be extended to the non-mean-field setting. Much remains to be done in this direction, and is currently work in progress.

**Supplementary Material:** The supplementary material includes proofs of all propositions as well as additional results for Bayesian linear regression, Bayesian density estimation and Bayesian logistic regression.

**Acknowledgements:** The authors would like to thank Prof. Steven MacEachern for valuable discussions and suggestions. They are also grateful for the comments provided by two anonymous reviewers that improved the contents of this manuscript. This research was partially supported by NSF DMS 1613054 and NIH R01 CA214955-01A1 (to KB and SK), and NSF CCF 1740761 (to SK).

## References

- Amari, S. (1998). Natural gradient works efficiently in learning. *Neural Computation* 10(2), 251–276.
- Attias, H. (2000). A variational Bayesian framework for graphical models. In *Neural Information Processing Systems*, Volume 12, pp. 209–215.
- Bache, K. and M. Lichman (2013). UCI machine learning repository.
- Barber, D. and W. Wierginck (1999). Tractable variational structures for approximating graphical models. In *Neural Information Processing Systems*, pp. 183–189.
- Bauer, M., S. Joshi, and K. Modin (2015). Diffeomorphic density matching by optimal information transport. *SIAM Journal on Imaging Sciences* 8(3), 1718–1751.

- Bauer, M., S. Joshi, and K. Modin (2017). Diffeomorphic random sampling using optimal information transport. In *Geometric Science of Information*, pp. 135–142.
- Beal, M. J. (2003). *Variational algorithms for approximate Bayesian inference*. PhD thesis, University College London.
- Bhattacharyya, A. (1943). On a measure of divergence between two statistical population defined by their population distributions. *Bulletin of the Calcutta Mathematical Society* 35, 99–109.
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York.
- Blei, D. M., A. Kucukelbir, and J. D. McAuliffe (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association* 112(518), 859–877.
- Broderick, T., N. Boyd, A. Wibisono, A. C. Wilson, and M. I. Jordan (2013). Streaming variational Bayes. In *Neural Information Processing Systems*, pp. 1727–1735.
- Carlin, B. P. and T. A. Louis (2008). *Bayesian Methods for Data Analysis*. CRC Press.
- Cencov, N. N. (2000). *Statistical Decision Rules and Optimal Inference*. Number 53. American Mathematical Society.
- Chen, T., J. Streets, and B. Shahbaba (2015). A geometric view of posterior approximation. *arXiv:1510.00861*.
- Cowles, M. K. and B. P. Carlin (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association* 91(434), 883–904.
- Ghahramani, Z. and M. J. Beal (1999). Variational inference for Bayesian mixtures of factor analysers. In *Neural Information Processing Systems*, Volume 12, pp. 449–455.
- Girolami, M. and B. Calderhead (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society, Series B* 73(2), 123–214.

- Hernández-Lobato, J., Y. Li, M. Rowland, T. Bui, D. Hernández-Lobato, and R. Turner (2016). Black-box  $\alpha$ -divergence minimization. In *International Conference on Machine Learning*, pp. 1511–1520.
- Hoffman, M. and D. Blei (2015). Stochastic structured variational inference. In *Artificial Intelligence and Statistics*, pp. 361–369.
- Hoffman, M. D., D. M. Blei, C. Wang, and J. W. Paisley (2013). Stochastic variational inference. *Journal of Machine Learning Research* 14(1), 1303–1347.
- Jaakkola, T. and M. I. Jordan (1997). A variational approach to Bayesian logistic regression models and their extensions. In *International Workshop on Artificial Intelligence and Statistics*, Volume 82.
- Jaakkola, T. S. and M. I. Jordan (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing* 10(1), 25–37.
- Kass, R. E. and P. W. Vos (2011). *Geometrical Foundations of Asymptotic Inference*, Volume 908. John Wiley & Sons.
- Kingma, D. P., T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling (2016). Improved variational inference with inverse autoregressive flow. In *Advances in neural information processing systems*, pp. 4743–4751.
- Kucukelbir, A., D. Tran, R. Ranganath, A. Gelman, and D. M. Blei (2017). Automatic differentiation variational inference. *Journal of Machine Learning Research* 18(1), 430–474.
- Kurtek, S. (2017). A geometric approach to pairwise Bayesian alignment of functional data using importance sampling. *Electronic Journal of Statistics* 11(1), 502–531.
- Kurtek, S. and K. Bharath (2015). Bayesian sensitivity analysis with the Fisher-Rao metric. *Biometrika* 102(3), 601–616.

- Lang, S. (2012). *Fundamentals of Differential Geometry*, Volume 191. Springer Science & Business Media.
- Leonard, T. (1978). Density estimation, stochastic processes and prior information. *Journal of the Royal Statistical Society, Series B*, 113–146.
- Li, Y. and R. E. Turner (2016). Rényi divergence variational inference. In *Neural Information Processing Systems*, pp. 1073–1081.
- McGrory, C. A. and D. Titterton (2007). Variational approximations in Bayesian model selection for finite mixture distributions. *Computational Statistics & Data Analysis* 51(11), 5352–5367.
- Minka, T. P. (2001). Expectation propagation for approximate Bayesian inference. In *Seventeenth Conference on Uncertainty in Artificial Intelligence*, pp. 362–369.
- Minka, T. P. (2005). Divergence measures and message passing. Technical report.
- Olson, J. M. and L. A. Weissfeld (1991). Approximation of certain multivariate integrals. *Statistics & Probability Letters* 11(4), 309–317.
- Ramsay, J. O., G. Hooker, and S. Graves (2009). *Functional Data Analysis with R and MATLAB*. Springer Science & Business Media.
- Rao, C. R. (1945). Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society* 37, 81–91.
- Rényi, A. (1961). On measures of entropy and information. In *Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1, pp. 547–561.
- Rezende, D. and S. Mohamed (2015). Variational inference with normalizing flows. In *International Conference on Machine Learning*, pp. 1530–1538.

- Riihimäki, J. and A. Vehtari (2014). Laplace approximation for logistic Gaussian process density estimation and regression. *Bayesian Analysis* 9(2), 425–448.
- Ring, W. and B. Wirth (2012). Optimization methods on Riemannian manifolds and their applications to shape space. *SIAM Journal of Optimization* 22(2), 596–627.
- Saul, L. K. and M. I. Jordan (1996). Exploiting tractable substructures in intractable networks. In *Neural Information Processing Systems*, pp. 486–492.
- Sigillito, V. G., S. P. Wing, L. V. Hutton, and K. B. Baker (1989). Classification of radar returns from the ionosphere using neural networks. *Johns Hopkins APL Technical Digest* 10(3), 262–266.
- Srivastava, A., I. H. Jermyn, and S. H. Joshi (2007). Riemannian analysis of probability density functions with applications in vision. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8.
- Srivastava, A., E. Klassen, S. H. Joshi, and I. H. Jermyn (2011). Shape analysis of elastic curves in Euclidean spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33(7), 1415–1428.
- Tan, L. S. (2018). Model reparametrization for improving variational inference. *arXiv preprint arXiv:1805.07267*.
- Tokdar, S. T. and J. K. Ghosh (2007). Posterior consistency of logistic Gaussian process priors in density estimation. *Journal of Statistical Planning and Inference* 137(1), 34–42.
- Ueda, N. and Z. Ghahramani (2002). Bayesian model search for mixture models based on optimizing variational bounds. *Neural Networks* 15(10), 1223–1241.
- van der Vaart, A. W. and J. H. van Zanten (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse Gamma bandwidth. *Annals of Statistics* 37(5B), 2655–2675.

- Van Erven, T. and P. Harremoës (2014). Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory* 60(7), 3797–3820.
- Wang, C. and D. M. Blei (2013). Variational inference in nonconjugate models. *Journal of Machine Learning Research* 14, 1005–1031.
- Yeung, D., H. Chang, Y. Xiong, S. George, R. Kashi, T. Matsumoto, and G. Rigoll (2004). SVC2004: First international signature verification competition. In *Biometric Authentication*, pp. 16–22.