THE BAYESIAN UPDATE: VARIATIONAL FORMULATIONS AND GRADIENT FLOWS

NICOLÁS GARCÍA TRILLOS * AND DANIEL SANZ-ALONSO *

Abstract. The Bayesian update can be viewed as a variational problem by characterizing the posterior as the minimizer of a functional. The variational viewpoint is far from new and is at the heart of popular methods for posterior approximation. However, some of its consequences seem largely unexplored. We focus on the following one: defining the posterior as the minimizer of a functional gives a natural path towards the posterior by moving in the direction of steepest descent of the functional. This idea is made precise through the theory of gradient flows, allowing to bring new tools to the study of Bayesian models and algorithms. Since the posterior may be characterized as the minimizer of different functionals, several variational formulations may be considered. We study three of them and their three associated gradient flows. We show that, in all cases, the rate of convergence of the flows to the posterior can be bounded by the geodesic convexity of the functional to be minimized. Each gradient flow naturally suggests a nonlinear diffusion with the posterior as invariant distribution. These diffusions may be discretized to build proposals for Markov chain Monte Carlo (MCMC) algorithms. By construction, the diffusions are guaranteed to satisfy a certain optimality condition, and rates of convergence are given by the convexity of the functionals. We use this observation to propose a criterion for the choice of metric in Riemannian MCMC methods.

Key words. Gradient flows, Wasserstein Space, Geodesic Convexity, Classification, Image Denoising, Riemannian MCMC

1. Introduction In this paper we revisit the old idea of viewing the posterior as the minimizer of an energy functional. The use of variational formulations of Bayes rule seems to have been largely focused on one of its methodological benefits: restricting the minimization to a subclass of measures is the backbone of variational Bayes methods for posterior approximation. Our aim is to bring attention to two other theoretical and methodological benefits, and to study in some detail one of these: namely, that each variational formulation suggests a natural path, defined by a gradient flow, towards the posterior. We use this observation to propose a criterion for the choice of metric in Riemannian MCMC methods.

Let us recall informally a variational formulation of Bayes rule. Given a prior $p(u)$ on an unknown parameter u and a likelihood function $L(y|u)$, the posterior $p(u|y) \propto L(y|u)p(u)$ can be characterized [Zellner, 1988] as the distribution $q^*(u)$ that minimizes

$$J_{\mathrm{KL}}\big(q(u)\big) = D_{\mathrm{KL}}\big(q(u)\|p(u)\big) - \int \log L(y|u) q(u) du. \tag{1.1}$$

Indeed, minimizing $J_{\mathrm{KL}}\big(q(u)\big)$ is equivalent to minimizing the Kullback-Leibler divergence $D_{\mathrm{KL}}(q(u)\|p(u|y))$, and so clearly the minimizer is $q^*(u) = p(u|y)$. Other variational formulations may be considered by minimizing, for instance, other divergence rather than Kullback-Leibler. In this paper we consider three variational formulations of the Bayesian update. The first two characterize the posterior *measure* as the minimizer of functionals J_{KL} or J_{χ^2} constructed by penalizing deviations from the prior measure in Kullback-Leibler or χ^2 divergence —definitions of these functionals and divergences are given in equations (3.2), (3.5), (2.2), (2.3). The third one characterizes the posterior *density* as the minimizer of the Dirichlet energy D^μ —see (2.4).

Why is it useful to view the posterior $p(u|y)$ as the minimizer of an energy? We list below three advantages of this viewpoint, the third of which will be the focus of our paper.

*Division of Applied Mathematics, Brown University.

1. The variational formulation provides a natural way to approximate the posterior by restricting the minimization problem to distributions $q(u)$ satisfying some computationally desirable property. For instance, variational Bayes methods often restrict the minimization to $q(u)$ with product structure [Attias, 1999], [Wainwright and Jordan, 2008], [Fox and Roberts, 2012]. A similar idea is studied in [Pinski et al., 2015], where $q(u)$ is restricted to a class of Gaussian distributions. An iterative variational procedure that progressively improves the posterior approximation by enriching the family of distributions was introduced in [Guo et al., 2016].
2. If the prior $p_{\varepsilon_n}(u)$ or the likelihood $L_{\varepsilon_n}(y|u)$ depend on a parameter ϵ_n, then the variational formulation allows to show large n convergence of posteriors $p_{\varepsilon_n}(u|y)$ by establishing the Γ-convergence of the associated energies. This method of proof has been employed by the authors in [Garcia Trillos and Sanz-Alonso, 2017], [Garcia Trillos et al., 2017b] to analyze the large-data consistency of graph-based Bayesian semi-supervised learning.
3. Each variational formulation gives a natural path, defined by a gradient flow, towards the posterior. These flows can be thought of as time-parameterized curves in the space of probability measures, converging in the large-time limit towards the posterior.

In this paper we study three gradient flows associated with the variational formulations defined by minimization of the functionals J_{KL}, J_{χ^2}, and D^μ. For intuition, we recall that a gradient flow in Euclidean space defines a curve whose tangent always points in the direction of steepest descent of a given function –see equation (2.5). In the same fashion, a gradient flow in a more general metric space can be thought of as a curve on said space that always points in the direction of steepest descent of a given functional [Ambrosio et al., 2008]. In Euclidean space the direction of steepest descent is naturally defined as that in which an *Euclidean* infinitesimal increment leads to the largest decrease on the value of the function. In a general metric space the direction of steepest descent is the one in which an infinitesimal increment *defined in terms of the distance* leads to the largest decrease on the value of the functional. In this paper we study:

(i) The gradient flows defined by J_{KL} and J_{χ^2} in the space of probability measures with finite second moments endowed with the Wasserstein distance—definitions are given in (2.1). By construction, these flows give curves of probability measures that evolve following the direction of steepest descent of J_{KL} and J_{χ^2} in Wasserstein distance, converging to the posterior measure in the large-time limit.

(ii) The gradient flow defined by the Dirichlet energy D^μ in the space of square integrable densities endowed with the L^2 distance. By construction, this flow gives a curve of densities in L^2 that evolves following the direction of steepest descent of D^μ in L^2 distance, converging to the posterior density in the large-time limit. Interestingly, the curve of measures associated with these densities is the exact same as the curve defined by the J_{KL} flow on Wasserstein space [Jordan et al., 1998].

A question arises: what is the rate of convergence of these flows to the posterior? The answer is, to a large extent, provided by the theory of optimal transport and gradient flows [Ambrosio et al., 2008], [Villani, 2003], [Villani, 2008], [Santambrogio, 2015]. We will review and provide a unified account of these results in the main body of the paper, section 3. In the remainder of this introduction we discuss how rates of conver-

gence may be studied in terms of convexity of functionals, and how these rates may be used as a guide for the choice of proposals for MCMC methods.

Rates of convergence of the flows hinge on the convexity level of each of the functionals J_{KL}, J_{χ^2}, and D^μ. Recalling the Euclidean case may be helpful: gradient descent on a highly convex function will lead to fast convergence to the minimizer. What is, however, a sensible notion of convexity for functionals defined over measures or densities? Our presentation highlights that the notion of geodesic (or displacement) convexity [McCann, 1997] nicely unifies the theory: it guarantees the existence and uniqueness of the three gradient flows and it also provides a bound on their rate of convergence to the posterior. In the L^2 setting one can show that positive geodesic convexity is equivalent to the posterior satisfying a Poincaré inequality, and also to the existence of a spectral gap —see subsection 3.2.2. On the other hand, the geodesic convexity of J_{KL} and J_{χ^2} in Wasserstein space is determined by the Ricci curvature of the manifold, as well as by the likelihood function and prior density —see (3.2), (3.3), (3.5), (3.6). Typically the three functionals J_{KL}, J_{χ^2}, and D^μ will have different levels of geodesic convexity, and establishing a sharp bound on each of them may not be equally tractable.

The theory of gradient flows and optimal transport gives, for each of the flows, an associated Fokker-Planck partial differential equation (PDE) that governs the evolution of densities [Ohta and Takatsu, 2011], [Santambrogio, 2015]. Such PDEs are typically costly to discretize if the parameter space is of moderate or high dimension, but they may be used in small-dimensional problems as a way to define tempering schemes. Here we do not explore this idea any further. Instead, we focus on the (nonlinear) diffusion processes associated with the PDEs. These diffusions are Langevin-type stochastic differential equations, whose evolving densities satisfy the Fokker-Planck equations. By construction, the invariant distribution of each of these diffusions is the sought posterior, and a bound on the rate of convergence of the diffusions to the posterior is given by the geodesic convexity of the corresponding functional. The gradient flow perspective automatically gives a sense in which the diffusions are optimal: the associated densities move locally (in Wasserstein or L^2 sense) in the direction of steepest descent of the functional. From this it immediately follows, for instance, that the law of a standard Langevin diffusion in Euclidean space evolves locally in Wasserstein space in the direction that minimizes Kullback-Leibler, and that it also evolves locally in L^2 in the direction that minimizes the Dirichlet energy.

The MCMC methodology allows to use a proposal based on a discretization of a diffusion —combined with an accept-reject mechanism to remove the discretization bias— to produce, in the large-time asymptotic, correlated posterior samples. Heuristically, the rate of convergence of the un-discretized diffusion may guide the choice of proposal. Proposals based on Langevin diffusions were first suggested in [Besag, 1994], and the exponential ergodicity of the resulting algorithms was analysed in [Roberts and Tweedie, 1996]. The paper [Girolami and Calderhead, 2011] considered changing the metric on the parameter space in order to accelerate MCMC algorithms by taking into account the geometric structure that the posterior defines in the parameter space. This led to a new family of Riemannian MCMC algorithms. Our paper is concerned with the study of un-discretized diffusions; the effect of the accept-reject mechanism on rates and ergodicity of MCMC methods will be studied elsewhere. We suggest that a way to guide the choice of metric of Riemannian MCMC methods is to choose the one that leads to a faster rate of convergence of the diffusion under certain constraints. We emphasize that despite working with un-discretized diffusions, our guidance for choice of proposals accounts for the fact that discretization will eventually

be needed. Our criterion weeds out choices of metric that lead to diffusions that achieve fast rate of convergence by merely speeding-up the drift. This is crucial, since a larger drift typically leads to a larger discretization error, and therefore to more rejections in the MCMC accept-reject mechanism in order to remove the bias in the discrete chain. This important constraint on the size of the drift seems to have been overlooked in existing continuous-time analyses of MCMC methods.

In summary, the following points highlight the key elements and common structure of the variational formulations of the Bayesian update and of the study of the associated gradient flows:

- The posterior can be characterized as the minimizer of different functionals on probability measures or densities.
- One can then study the gradient flows of these functionals with respect to a metric on the space of probability measures or densities; the resulting curve is a curve of maximal slope and its endpoint is the posterior.
- The gradient flows are characterized by a Fokker-Planck PDE that governs the evolution of the density of an associated diffusion process.
- By studying the convexity of the functionals (with respect to a given metric) one can obtain rates of convergence of the gradient flows towards the posterior. In particular, the level of convexity determines the speed of convergence of the densities of the associated diffusion process towards the posterior, and hence can be used as a criterion to guide the choice of proposals for MCMC methods; here we emphasize that care must be taken when comparing different diffusions if a higher speed of convergence is at the cost of a more expensive discretization.

The ideas in this paper immediately extend beyond the Bayesian interpretation stressed here to any application (e.g. the study of conditioned diffusions) where a measure of interest is defined in terms of a reference measure and a change of measure. Also, we consider only Kullback-Leibler and χ^2 prior penalizations to define the functionals J_{KL} and J_{χ^2}, but it would be possible to extend the analysis to the family of m-divergences introduced in [Ohta and Takatsu, 2011]. Kullback-Leibler and χ^2 prior penalization correspond to $m \to 1$ and $m = 2$ within this family. In what follows we point out some of the features of the different functionals and gradient flows that we consider in this paper.

1.1. Comparison of Functionals and Flows We now provide a comparison of the three choices of functionals that we consider.

1. The two gradient flows in Wasserstein space (arising from the functionals J_{KL} and J_{χ^2}) are fundamentally connected with the variational formulation: these variational formulations can be used to *define* posterior-type measures via a penalization of deviations from the prior and deviations from the data in situations where establishing the existence of conditional distributions by disintegration of measures is technically demanding. On the other hand, the variational formulation for the Dirichlet energy is less natural and requires previous knowledge of the posterior.
2. The precise level of geodesic convexity of the functionals J_{KL} (and J_{χ^2}) can be computed from point evaluation of the Ricci tensor (of the parameter space) and derivatives of the densities. In particular, knowledge of the underlying metric suffices to compute these quantities. In contrast, establishing a sharp Poincaré inequality —the level of geodesic convexity of the Dirichlet energy in $L^2(\mathcal{M}, \mu)$— is in practice unfeasible, as it effectively requires solving an infinite dimensional optimization problem. It is for this reason —and because of the

explicit dependence of the convexity in Wasserstein space with the geometry induced by the manifold metric tensor— that our analysis of the choice of metric in Riemannian MCMC methods is based on the J_{KL} functional (see section 4, and in particular Theorem 4.1).

3. On the flip side of point 2, a Poincaré inequality for the posterior with a not necessarily optimal constant can be established using only tail information. In particular, even when the functional J_{KL} is not geodesically convex in Wasserstein space, one may still be able to obtain a Poincaré inequality (see subsection 5.2 for an example).
4. In contrast to the diffusions arising from the J_{KL} or Dirichlet flows, the stochastic processes arising from the J_{χ^2} formulation are inhomogeneous, and hence simulation seems more challenging unless further structure is assumed on the prior measure and likelihood function. Also, the evolution of densities of the gradient flow of J_{χ^2} in Wasserstein space is given by a porous medium PDE.

1.2. Outline The rest of the paper is organized as follows. Section 2 contains some background material on the Wasserstein space, geodesic convexity of functionals, and gradient flows in metric spaces. The core of the paper is section 3, where we study the geodesic convexity, PDEs, and diffusions associated with each of the three functionals J_{KL}, J_{χ^2}, and D^{μ}. In section 4 we consider an application of the theory to the choice of metric in Riemannian MCMC methods [Girolami and Calderhead, 2011], and in section 5 we illustrate the main concepts and ideas through examples arising in Bayesian formulations of semi-supervised learning [Garcia Trillos and Sanz-Alonso, 2017], [Garcia Trillos et al., 2017b], [Bertozzi et al., 2017]. We close in section 6 by summarizing our main contributions and pointing to open directions.

1.3. Set-up and Notation $(\mathcal{M},g)$ will denote a smooth connected m-dimensional Riemannian manifold with metric tensor g representing the parameter space. We will denote by d the associated Riemannian distance, and assume that $(\mathcal{M},d)$ is a complete metric space. By the Hopf-Rinow theorem it follows that $\mathcal{M}$ is a *geodesic space* —we refer to subsection (2.1) for a discussion on geodesic spaces and their relevance here. We denote by vol_g the associated volume form. To emphasize the dependence of differential operators on the metric with which $\mathcal{M}$ is endowed, we write ∇_g, div_g, Hess_g and Δ_g for the gradient, divergence, Hessian, and Laplace Beltrami operators on $(\mathcal{M},g)$. The reader not versed in Riemannian geometry may focus on the case $\mathcal{M}=\mathbb{R}^m$ with the usual metric tensor, in which case d is the Euclidean distance and $dvol_g = dx$ is the Lebesgue measure. However, in section 4 where we discuss applications to Riemannian MCMC, we endow $\mathbb{R}^m$ with a general metric tensor g and hence familiarity with some notions from differential geometry is desirable.

We denote by $\mathcal{P}(\mathcal{M})$ the space of probability measures on $\mathcal{M}$ (endowed with the Borel σ-algebra). We will be concerned with the update of a *prior* probability measure $\pi \in \mathcal{P}(\mathcal{M})$ —that represents various degrees of belief on the value of a quantity or parameter of interest— into a *posterior* probability measure $\mu \in \mathcal{P}(\mathcal{M})$, based on observed data y. We will assume that the prior is defined as a change of measure from vol_g, and that the posterior is defined as a change of measure from π as follows:

$$\pi = e^{-\Psi} vol_g, \qquad \mu \propto e^{-\phi}\pi. \tag{1.2}$$

The data is incorporated in the Bayesian update through the negative log-likelihood function $\phi(\cdot) = \phi(\cdot; y)$.

2. Preliminaries

In this section we provide some background material. The Wasserstein space, and the notion of λ-geodesic convexity of functionals are reviewed in subsection 2.1. Gradient flows in metric spaces are reviewed in subsection 2.2.

2.1. Geodesic Spaces and Geodesic Convexity of Functionals A geodesic space (X,d_X) is a metric space with a notion of length of curves that is compatible with the metric, and where every two points in the space can be connected by a curve whose length achieves the distance between the points (see [Burago et al., 2001] for more details). Geodesic spaces constitute a large family of metric spaces with a rich theory of gradient flows. Here we consider three geodesic spaces. First, the *base space* $(\mathcal{M},d)$, i.e. the manifold $\mathcal{M}$ equipped with its Riemannian distance. Second, the space $\mathcal{P}_2(\mathcal{M})$ of square integrable Borel probability measures defined on $\mathcal{M}$, endowed with the Wasserstein distance $\mathcal{W}_2$. Third, the space of functions $f\in L^2(\mathcal{M},\mu)$, with $\int_{\mathcal{M}} f d\mu = 1$, equipped with the $L^2(\mathcal{M},\mu)$ norm.

We spell out the definitions of $\mathcal{P}_2(\mathcal{M})$ and $\mathcal{W}_2$:

$$\begin{aligned}\mathcal{P}_2(\mathcal{M}) &:= \Big\{\nu\in\mathcal{P}(\mathcal{M}) : \int_{\mathcal{M}} d^2(x,x_0)d\nu(x) < \infty, \text{ for some } x_0\in\mathcal{M}\Big\},\\ \mathcal{W}_2^2(\nu_1,\nu_2) &:= \inf_{\alpha}\int_{\mathcal{M}\times\mathcal{M}} d(x,y)^2 d\alpha(x,y), \quad \nu_1,\nu_2\in\mathcal{P}_2(\mathcal{M}).\end{aligned} \tag{2.1}$$

The infimum in the previous display is taken over all *transportation plans* between ν_1 and ν_2, i.e. over $\alpha\in\mathcal{P}(\mathcal{M}\times\mathcal{M})$ with marginals ν_1 and ν_2 on the first and second factors. The space $(\mathcal{P}_2(\mathcal{M}),\mathcal{W}_2)$ is indeed a geodesic space: geodesics in $(\mathcal{P}_2(\mathcal{M}),\mathcal{W}_2)$ are induced by those in $(\mathcal{M},d)$. All it takes to construct a geodesic connecting $\nu_0\in\mathcal{P}_2(\mathcal{M})$ and $\nu_1\in\mathcal{P}_2(\mathcal{M})$ is to find an optimal transport plan between ν_0 and ν_1 to determine source locations and target locations, and then transport the mass along geodesics in $\mathcal{M}$ (see [Villani, 2003] and [Santambrogio, 2015]).

The space of functions $f\in L^2(\mathcal{M},\mu)$, with $\int_{\mathcal{M}} f d\mu = 1$, equipped with the $L^2(\mathcal{M},\mu)$ norm is also a geodesic space, where a constant speed geodesic connecting f_0 and f_1 is given by linear interpolation: $t\in[0,1]\mapsto (1-t)f_0+tf_1$.

We will consider several functionals $E: X\to\mathbb{R}\cup\{\infty\}$ throughout the paper. They will all be defined in one of our three geodesic spaces —that is, $X=\mathcal{M}$, $X=\mathcal{P}_2(\mathcal{M})$ or $X=L^2(\mathcal{M},\mu)$. Important examples will be, respectively:

1. Functions $\Psi:\mathcal{M}\to\mathbb{R}\cup\{\infty\}$.
2. The Kullback-Leibler and χ^2 divergences $D_{\mathrm{KL}}(\cdot\|\pi)$, $D_{\chi^2}(\cdot\|\pi):\mathcal{P}(\mathcal{M})\to[0,\infty]$, where π is a given (prior) measure and, for $\nu_1,\nu_2\in\mathcal{P}(\mathcal{M})$,

$$D_{\mathrm{KL}}(\nu_1\|\nu_2) := \begin{cases}\int_{\mathcal{M}} \frac{d\nu_1}{d\nu_2}(u)\log\Big(\frac{d\nu_1}{d\nu_2}(u)\Big)d\nu_2(u), & \nu_1\ll\nu_2,\\ \infty, & \text{otherwise,}\end{cases} \tag{2.2}$$

$$D_{\chi^2}(\nu_1\|\nu_2) := \begin{cases}\int_{\mathcal{M}} \Big(\frac{d\nu_1}{d\nu_2}(u)-1\Big)^2 d\nu_2(u), & \nu_1\ll\nu_2,\\ \infty, & \text{otherwise;}\end{cases} \tag{2.3}$$

and the potential-type functional $J:\mathcal{P}(\mathcal{M})\to\mathbb{R}\cup\{\infty\}$ given by

$$J(\nu) := \int_{\mathcal{M}} h(u)d\nu(u),$$

where h is a given potential function.

3. The *Dirichlet* energy $D^\mu : L^2(\mathcal{M},\mu) \to [0,\infty]$ defined by

$$D^\mu(f) = \begin{cases} \int_{\mathcal{M}} \|\nabla_g f(u)\|^2 d\mu(u), & f \in L^2(\mathcal{M},\mu) \cap H^1(\mathcal{M}), \\ +\infty, & \text{otherwise.} \end{cases}, \tag{2.4}$$

Recall that here and throughout, ∇_g denotes the gradient in $(\mathcal{M},g)$ and $\|\cdot\|$ is the norm on each tangent space $\mathcal{T}_x\mathcal{M}$.

A crucial unifying concept will be that of λ-geodesic convexity of functionals. We recall it here:

Definition 2.1. *Let (X,d_X) be a geodesic space and let $\lambda \in \mathbb{R}$. A functional $E : X \to \mathbb{R} \cup \{\infty\}$ is called λ-geodesically convex provided that for any $x_0, x_1 \in X$ there exists a constant speed geodesic $t \in [0,1] \mapsto \gamma(t) \in X$ such that $\gamma(0) = x_0$, $\gamma(1) = x_1$, and*

$$E\big(\gamma(t)\big) \leq (1-t)E(x_0) + tE(x_1) - \lambda \frac{t(1-t)}{2} d_X^2(x_0,x_1), \quad \forall t \in [0,1].$$

The following remark characterizes the λ-convexity of functionals when $X = \mathcal{M}$.

Remark 2.2. *Let $\Psi \in C^2(\mathcal{M})$ so that we can define its Hessian at all points in $\mathcal{M}$ (see the proof of Theorem 4.1 in the appendix for the definition). Then the following conditions are equivalent:*

(i) Ψ is λ-geodesically convex.

(ii) $\mathrm{Hess}_g \Psi_x(v,v) \geq \lambda$ for all $x \in \mathcal{M}$ and all unit vectors $v \in T_x\mathcal{M}$.

If $(\mathcal{M},d)$ is the Euclidean space, (i) and (ii) are also equivalent to:

(iii) $\Psi - \frac{\lambda}{2}|\cdot|^2$ is a convex function.

This latter condition is known in the optimization literature as strong convexity.

2.2. Gradient Flows in Metric Spaces In this subsection we review the basic concepts needed to define gradient flows in a metric space (X,d_X). We follow Chapter 8 of [Santambrogio, 2015]; a standard technical reference is [Ambrosio et al., 2008].

To guide the reader, we first recall the formulation of gradient flows in Euclidean space, where $X = \mathbb{R}^d$ and d_X is the Euclidean metric. Let $E : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function, and consider the equation

$$\begin{cases} \dot{x}(t) &= -\nabla E\big(x(t)\big), \quad t \geq 0, \\ x(0) &= x_0. \end{cases} \tag{2.5}$$

Then, the solution x to (2.5) is the *gradient flow* of E in Euclidean space with initial condition x_0; it is a curve whose tangent vector at every point in time is the negative of the gradient of the function E at that time. In order to generalize the notion of a gradient flow to functionals defined on more general metric spaces, and in particular when the metric space has no differential structure, we reformulate (2.5) in integral form by using that $\frac{d}{dt} E\big(x(t)\big) = \langle \nabla E\big(x(t)\big), \dot{x}(t) \rangle = -\frac{1}{2}|\dot{x}(t)|^2 - \frac{1}{2}|\nabla E\big(x(t)\big)|^2$ as follows:

$$E(x_0) = E\big(x(t)\big) + \frac{1}{2}\int_0^t |\dot{x}(r)|^2 dr + \frac{1}{2}\int_0^t \big|\nabla E\big(x(r)\big)\big|^2 dr, \quad t > 0. \tag{2.6}$$

This identity, known as energy dissipation equality, is equivalent to (2.5) —see Chapter 8 of [Santambrogio, 2015] for further details and other possible formulations. Crucially (2.6) involves notions that can be defined in an arbitrary metric space (X,d_X): the metric derivative of a curve $t \mapsto x(t) \in X$ is given by

$$|\dot{x}(t)| := \lim_{s \to t} \frac{d_X\big(x(t),x(s)\big)}{|s-t|},$$

and the *slope* of a functional $E: X \to \mathbb{R} \cup \{\infty\}$ is defined as the map $|\nabla E|: \{x \in X : E(x) < \infty\} \to \mathbb{R} \cup \{\infty\}$ given by

$$|\nabla E|(x) := \limsup_{y \to x} \frac{\big(E(x) - E(y)\big)^+}{d_X(x,y)}.$$

The identity (2.6) is the standard way to introduce gradient flows in arbitrary metric spaces. In this paper we consider gradient flows in L^2 and Wasserstein spaces, where the notion of tangent vector is available. L^2 has Hilbert space structure, whereas the Wasserstein space can be seen as an infinite dimensional manifold (see [Ambrosio et al., 2008], [Santambrogio, 2015]).

3. Variational Characterizations of the Posterior and Gradient Flows

In this section we lay out the main elements of the theory of variational formulations and gradient flows in regards to the Bayesian update. Subsection 3.1 details three variational formulations defined in terms of the functionals J_{KL}, J_{χ^2} and the Dirichlet energy D^μ. Subsection 3.2 studies the geodesic convexity of J_{KL} and J_{χ^2} in Wasserstein space and of D^μ in L^2. Finally, subsection 3.3 collects the PDEs that characterize the gradient flows, as well as the corresponding diffusion processes.

3.1. Variational Formulation of the Bayesian Update The variational formulation of the posterior as the minimizer of J_{KL} and J_{χ^2} share the same structure, and is outlined in subsection 3.1.1. The variational formulation in terms of the Dirichlet energy is given in subsection 3.1.2.

3.1.1. The Functionals J_{KL} and J_{χ^2} In mathematical analysis [Jordan and Kinderlehrer, 1996] and probability theory [Dupuis and Ellis, 2011] it is often useful to note that a probability measure μ defined by

$$\mu(du) = \frac{1}{Z} \exp\big(-\phi(u)\big)\pi(du) \tag{3.1}$$

is the minimizer of the functional

$$J_{\mathrm{KL}}(\nu) := D_{\mathrm{KL}}(\nu \| \pi) + F_{\mathrm{KL}}(\nu; \phi), \quad \nu \in \mathcal{P}(\mathcal{M}), \tag{3.2}$$

where

$$F_{\mathrm{KL}}(\nu; \phi) := \int_{\mathcal{M}} \phi(u) d\nu(u), \tag{3.3}$$

and the integral is interpreted as $+\infty$ if ϕ is not integrable with respect to ν. In physical terms, the Kullback-Leibler divergence represents an internal energy, F_{KL} represents a potential energy, and the constant Z is known as the partition function. Here we are concerned with a statistical interpretation of equation (3.1), and view it as defining a posterior measure as a change of measure from a prior measure. In this context, the Kullback-Leibler term $D_{\mathrm{KL}}(\cdot \| \pi)$ in (3.2) represents a penalization of deviations from prior beliefs, the term $F_{\mathrm{KL}}(\nu; \phi)$ penalizes deviations from the data, and the normalizing constant Z represents the marginal likelihood. For brevity, we will henceforth suppress the data y from the negative log-likelihood function ϕ, writing $\phi(u)$ instead of $\phi(u; y)$.

We remark that the fact that μ minimizes J_{KL} follows immediately from the identity

$$D_{\mathrm{KL}}(\cdot \| \mu) = J_{\mathrm{KL}}(\cdot) + \log Z. \tag{3.4}$$

Minimizing $J_{\mathrm{KL}}(\cdot)$ or $D_{\mathrm{KL}}(\cdot\|\mu)$ is thus equivalent, but the functional J_{KL} makes apparent the roles of the prior and the likelihood.

The posterior μ also minimizes the functional

$$J_{\chi^2}(\nu) := D_{\chi^2}(\nu\|\pi) + F_{\chi^2}(\nu;\phi), \quad \nu \in \mathcal{P}(\mathcal{M}), \tag{3.5}$$

where

$$F_{\chi^2}(\nu;\phi) := \int_{\mathcal{M}} \tilde{\phi}(u) d\nu(u), \quad \tilde{\phi} = g\big(\exp(\phi(u))\big), \quad g(t) := t-1, \quad t>0. \tag{3.6}$$

We refer to [Ohta and Takatsu, 2011] for details. Note that both J_{KL} and J_{χ^2} are defined in terms of the two starting points of the Bayesian update: the prior π and the negative log-likelihood ϕ. The associated variational formulations suggest a way to *define* posterior-type measures based on these two ingredients in scenarios where establishing the existence of conditional distributions via desintegration of measures is technically demanding. This appealing feature of the two variational formulations above is not shared by the one described in the next subsection.

3.1.2. The Dirichlet Energy D^μ Let now the posterior μ be given, and consider the space $L^2(\mathcal{M},\mu)$ of functions defined on $\mathcal{M}$ which are square integrable with respect to μ. Recall the Dirichlet energy

$$D^\mu(f) = \int_{\mathcal{M}} \|\nabla_g f(u)\|^2 d\mu(u),$$

introduced in equation (2.4). Now, since the measure μ can be characterized as the probability measure with density $\rho_\mu \equiv 1$ a.s. with respect to μ, it follows that the posterior density $\rho_\mu \equiv 1$ is the minimizer of the Dirichlet energy D^μ over probability densities $\rho \in L^2(\mathcal{M},\mu)$ with $\int_{\mathcal{M}} \rho d\mu = 1$.

3.2. Geodesic Convexity and Functional Inequalities In this section we study the geodesic convexity of the functionals J_{KL}, J_{χ^2}, and D^μ. The geodesic convexity of J_{KL} and J_{χ^2} in Wasserstein space is considered in subsection 3.2.1, and the geodesic convexity of D^μ in L^2 in subsection 3.2.2, where we show its equivalence to the posterior satisfying a Poincaré inequality.

3.2.1. Geodesic Convexity of J_{KL} and J_{χ^2} The next proposition can be found in [von Renesse and Sturm, 2005] and [Sturm, 2006]. It shows that the convexity of J_{KL} can be determined by the so-called curvature-dimension condition —a condition that involves the curvature of the manifold and the Hessian of the combined change of measure $\Psi + \phi$. We recall the notation $\pi = e^{-\Psi} vol_g$ and $\mu \propto e^{-\phi}\pi$.

Proposition 3.1. *Suppose that $\Psi, \phi \in C^2(\mathcal{M})$. Then J_{KL} (or $D_{\mathrm{KL}}(\cdot\|\mu)$) is λ-geodesically convex if, and only if,*

$$\mathrm{Ric}_g(v,v) + \mathrm{Hess}_g \Psi(v,v) + \mathrm{Hess}_g \phi(v,v) \geq \lambda, \quad \forall x \in \mathcal{M}, \quad \forall v \in T_x\mathcal{M} \text{ with } g(v,v) = 1,$$

where Ric_g denotes the Ricci curvature tensor.

We recall that the Ricci curvature provides a way to quantify the disagreement between the geometry of a Riemannian manifold and that of ordinary Euclidean space. The Ricci tensor is defined as the trace of a map involving the Riemannian curvature (see [do Carmo Valero, 1992]).

The following example illustrates the geodesic convexity of $D_{\mathrm{KL}}(\cdot\|\mu)$ for Gaussian μ.

Example 1. *Let $\mu = N(\theta, \Sigma)$ be a Gaussian measure in $\mathbb{R}^m$ (endowed with the Euclidean metric), with Σ positive definite. Then $D_{\mathrm{KL}}(\cdot\|\mu)$ is $1/\Lambda_{\max}(\Sigma)$-geodesically convex, where $\Lambda_{\max}(\Sigma)$ is the largest eigenvalue of Σ. This follows immediately from the above, since here $\Psi(x) = \frac{1}{2}\langle x-\theta, \Sigma^{-1}(x-\theta)\rangle$, and the Euclidean space is flat (its Ricci curvature is identically equal to zero). Note that the level of convexity of the functional depends only on the largest eigenvalue of the covariance, but not on the dimension m of the underlying space.*

The λ-convexity of J_{KL} guarantees the existence of the gradient flow of J_{KL} in Wasserstein space. Moreover, it determines the rate of convergence towards the posterior μ. Precisely, if μ_0 is absolutely continuous with respect to μ, and if $\lambda > 0$, then the gradient flow $t \in [0,\infty) \mapsto \mu_t$ of J_{KL} with respect to the Wasserstein metric starting at μ_0 is well defined and we have:

$$\begin{aligned} &D_{\mathrm{KL}}(\mu_t\|\mu) \leq e^{-\lambda t} D_{\mathrm{KL}}(\mu_0\|\mu), \quad t \geq 0, \\ &\mathcal{W}_2(\mu_t,\mu)^2 \leq \lambda e^{-\lambda t} D_{\mathrm{KL}}(\mu_0\|\mu), \quad t \geq 0. \end{aligned} \tag{3.7}$$

The second inequality, known as Talagrand inequality [Villani, 2003], establishes a comparison between Wasserstein geometry and information geometry. It can be established directly combining the λ-geodesic convexity of J_{KL} (for positive λ) with the first inequality. From (3.7) we see that a higher level of convexity of J_{KL} allows to guarantee a faster rate of convergence towards the posterior distribution μ.

We now turn to the geodesic convexity properties of J_{χ^2}. We recall that m denotes the dimension of the manifold $\mathcal{M}$. The following proposition can be found in [Ohta and Takatsu, 2011, Theorem 4.1].

Proposition 3.2. *J_{χ^2} is λ-geodesically convex if and only if both of the following two properties are satisfied:*

1. $\mathrm{Ric}_g(v,v) + \mathrm{Hess}_g\Psi(v,v) + \frac{1}{m+1}\langle \nabla_g \Psi, v\rangle^2 \geq 0, \quad \forall x \in \mathcal{M}, \quad \forall v \in T_x\mathcal{M}.$
2. *ϕ is λ-geodesically convex as a real valued function defined on $\mathcal{M}$.*

There are two main conclusions we can extract from the previous proposition. First, that condition 1) is only related to the prior distribution π whereas condition 2) is only related to the likelihood; in particular, the convexity properties of J_{χ^2} can indeed be studied by studying separately the prior and the likelihood (notice that the proposition gives an equivalence). Secondly, notice that condition 1) is a qualitative property and if it is not met there is no hope that the functional J_{χ^2} has any level of global convexity even when the likelihood function is a highly convex function. In addition, if 1) is satisfied, the convexity of ϕ determines completely the level of convexity of J_{χ^2}. These features are markedly different from the ones observed in the Kullback-Leibler case.

As for the functional J_{KL}, one can establish the following functional inequalities, under the assumption of λ-geodesic convexity of J_{χ^2} for $\lambda > 0$:

$$\begin{aligned} J_{\chi^2}(\mu_t) - J_{\chi^2}(\mu) &\leq e^{-\lambda t}\big(J_{\chi^2}(\mu_0) - J_{\chi^2}(\mu)\big), \quad t \geq 0, \\ \mathcal{W}_2(\mu_t,\mu)^2 &\leq \lambda e^{-\lambda t}\big(J_{\chi^2}(\mu_0) - J_{\chi^2}(\mu)\big), \quad t \geq 0. \end{aligned} \tag{3.8}$$

The above inequalities exhibit the fact that a higher level of convexity of J_{χ^2} guarantees a faster convergence towards the posterior distribution μ.

3.2.2. Geodesic Convexity of Dirichlet Energy We now study the geodesic convexity of the Dirichlet energy functional defined in equation (2.4). In what follows we denote by $\|\cdot\|$ the L^2 norm with respect to μ. Let us start recalling Poincaré inequality.

Definition 3.3. *We say that a Borel probability measure μ on $\mathcal{M}$ has a Poincaré inequality with constant λ if for every $f \in L^2(\mathcal{M},\mu)$ satisfying $\int_{\mathcal{M}} f d\mu = 0$ we have*

$$\|f\|_\mu^2 \leq \frac{1}{\lambda} D^\mu(f).$$

We now show that Poincaré inequalities are directly related to the geodesic convexity of the functional D_μ in the $L^2(\mathcal{M},\mu)$ space.

Proposition 3.4. *Let λ be a positive real number and let μ be a Borel probability measure on $\mathcal{M}$. Then, the measure μ has a Poincaré inequality with constant λ if and only if the functional D^μ is 2λ-geodesically convex in the space of functions $f \in L^2(\mathcal{M},\mu)$ satisfying $\int f d\mu = 1$.*

Proof. First of all we claim that

$$D^\mu(tf_0 + (1-t)f_1) + t(1-t)D^\mu(f_0 - f_1) = tD^\mu(f_0) + (1-t)D^\mu(f_1), \tag{3.9}$$

for all $f_0, f_1 \in L^2(\mathcal{M},\mu)$ and every $t \in [0,1]$. To see this, it is enough to assume that both $D^\mu(f_0)$ and $D^\mu(f_1)$ are finite and then notice that equality (3.9) follows from the easily verifiable fact that for an arbitrary Hilbert space V with induced norm $|\cdot|$ one has

$$|tv_0 + (1-t)v_1|^2 + t(1-t)|v_0 - v_1|^2 = t|v_0|^2 + (1-t)|v_1|^2, \quad \forall v_0, v_1 \in V, \quad \forall t \in [0,1].$$

Now, suppose that μ has a Poincaré inequality with constant λ and consider two functions $f_0, f_1 \in L^2(\mathcal{M},\mu)$ satisfying $\int_{\mathcal{M}} f_0 d\mu = \int_{\mathcal{M}} f_1 d\mu = 1$. Then, (3.9) combined with Poincaré inequality (taking $f := f_0 - f_1$) gives:

$$D^\mu\big(tf_0 + (1-t)f_1\big) + \lambda t(1-t)\|f_0 - f_1\|_\mu^2 \leq tD^\mu(f_0) + (1-t)D^\mu(f_1), \tag{3.10}$$

which is precisely the 2λ-geodesic convexity condition for D^μ.

Conversely, suppose that D^μ is 2λ-geodesic convex in the space of $L^2(\mathcal{M},\mu)$ functions that integrate to one. Let $f \in L^2(\mathcal{M},\mu)$ be such that $\int_{\mathcal{M}} f d\mu = 0$ and without the loss of generality assume that $D^\mu(f) < \infty$ and that $\|f\|_\mu \neq 0$. Under these conditions, the positive and negative parts of f, f^+ and f^-, satisfy $D^\mu(f^+), D^\mu(f^-) < \infty$ and $\int_{\mathcal{M}} f^+ d\mu = r = \int_{\mathcal{M}} f^- d\mu$ where $r > 0$. The inequality

$$\|f\|_\mu^2 \leq \frac{1}{\lambda} D^\mu(f)$$

is obtained directly from (3.9) and (3.10) applied to

$$f_0 := \frac{1}{r} f^-, \quad f_1 := \frac{1}{r} f^+, \quad t = 1/2.$$

□

Remark 3.5. *It is well known that the best Poincaré constant for a measure μ is equal to the smallest non-trivial eigenvalue of the operator $-\Delta_g^\mu$ defined formally as*

$$-\Delta_g^\mu f := -\frac{1}{Z} \mathrm{div}_g\left(e^{-\phi-\psi} \nabla_g f\right),$$

where div_g and ∇_g are the divergence and gradient operators in $(\mathcal{M},g)$. This eigenvalue can be written variationally as

$$\lambda_2 := \min_{f \in L^2(\mathcal{M},\mu)} \frac{D_\mu(f)}{\|f - f_\mu\|_\mu^2},$$

where

$$f_\mu := \int_{\mathcal{M}} f d\mu.$$

Remark 3.6. *Spectral gaps are used in the theory of MCMC as a means to bound the asymptotic variance of empirical expectations [Kipnis and Varadhan, 1986].*

Let us now consider $t \in (0,\infty) \mapsto \mu_t$ the flow of D^μ in $L^2(\mathcal{M},\mu)$ with some initial condition $\frac{d\mu_0}{d\mu} = \rho_0$. It is well known that this flow coincides with that of the functional J_{KL} in Wasserstein space. However, taking the Dirichlet-L^2 point of view, one can use Poincaré inequality (i.e. the geodesic convexity of D^μ) to deduce the exponential convergence of μ_t towards μ in the χ^2-sense. Indeed, let

$$\rho_t := \frac{d\mu_t}{d\mu}, \quad t \in (0,\infty).$$

A standard computation then shows that,

$$\frac{1}{2}\frac{d}{dt}\|\rho_t - 1\|_\mu^2 = \int_{\mathcal{M}} (\rho_t - 1)\frac{\partial \rho}{\partial t} d\mu = -\int_{\mathcal{M}} |\nabla_g(\rho_t(u) - 1)|^2 d\mu(u) \leq -\lambda \|\rho_t - 1\|_\mu^2.$$

In the second equality we have used that $\frac{\partial \rho}{\partial t} = \Delta_g^\mu \rho$, as discussed in subsection 3.3 below. Hence by Gronwall's inequality

$$\|\rho_t - 1\|_\mu \leq \exp(-\lambda t)\|\rho_0 - 1\|_\mu, \quad t > 0.$$

3.3. PDEs and Diffusions Here we describe the PDEs that govern the evolution of densities of the three gradient flows, and the stochastic processes associated with these PDEs. We consider first the flows defined with the functionals J_{KL} and D^μ in subsection 3.3.1, and then the flow defined by the functional J_{χ^2} in subsection 3.3.2.

3.3.1. J_{KL}-Wasserstein and D^μ-$L^2(\mathcal{M},\mu)$ It was shown in [Jordan et al., 1998] —in the Euclidean setting and in the unweighted case $\pi = dx$— that the gradient flow of the Kullback-Leibler functional $D_{\mathrm{KL}}(\cdot\|\pi)$ in Wasserestein space produces a solution to the Fokker-Planck equation. More generally, under the convexity conditions guaranteeing the existence of the gradient flow $t \in (0,\infty) \mapsto \mu_t$ of $D_{\mathrm{KL}}(\cdot\|\mu)$ (equivalently of J_{KL}) starting from $\mu_0 \in \mathcal{P}(\mathcal{M})$, the densities

$$\rho_t := \frac{d\mu_t}{d\mu}, \quad \theta_t := \frac{d\mu_t}{dvol_g}, \quad t \in (0,\infty)$$

satisfy (formally) the following Fokker-Planck equations

$$\frac{\partial \rho}{\partial t} = \Delta_g^\mu \rho. \tag{3.11}$$

$$\frac{\partial \theta}{\partial t} = \Delta_g \theta + \mathrm{div}_g\big(\theta(\nabla_g \phi + \nabla_g \Psi)\big). \tag{3.12}$$

Equation (3.12) can be identified as the evolution of the densities (w.r.t. $dvol_g$) of the diffusion

$$dX_t = -\nabla_g\Big(\Psi(X_t) + \phi(X_t)\Big)dt + \sqrt{2}dB_t^g, \tag{3.13}$$

where B^g denotes a Brownian motion defined on $(\mathcal{M},g)$ and ∇_g is the gradient on $(\mathcal{M},g)$. Naturally, the D^μ flow in L^2 has the same associated Fokker-Planck equation (3.11) and diffusion process (3.13).

3.3.2. J_{χ^2}-Wasserstein The PDE satisfied (formally) by the densities

$$\tilde{\rho}_t := \frac{d\mu_t}{d\pi},$$

of the J_{χ^2}-Wasserstein flow $t \in (0,\infty) \mapsto \mu_t$ is the (weighted) porous medium equation:

$$\frac{\partial \tilde{\rho}}{\partial t} = \Delta_g^\pi \tilde{\rho}^2 + \mathrm{div}_g^\pi(\tilde{\rho}\nabla_g \phi), \tag{3.14}$$

where the weighted Laplacian and divergence are defined formally as

$$\begin{aligned} \Delta_g^\pi f &:= \Delta_g f - \langle \nabla_g f, \nabla_g \Psi \rangle, \\ \mathrm{div}_g^\pi F &:= \mathrm{div}_g F - \langle F, \nabla_g \Psi \rangle. \end{aligned} \tag{3.15}$$

Consider now the stochastic process $\{u_t\}_{t\geq 0}$ formally defined as the solution to the nonlinear diffusion

$$dX_t = -\big(\tilde{\rho}(t,X_t)\nabla_g \Psi(X_t) + \nabla_g \phi(X_t)\big)dt + \sqrt{2\tilde{\rho}(t,X_t)}dB_t^g, \quad u_0 \sim \rho_0, \tag{3.16}$$

where $\tilde{\rho}$ is the solution to (3.14). Let θ_t be the evolution of the densities (with respect to $dvol_g$) of the above diffusion. Then a formal computation shows that θ satisfies the Fokker-Plank equation:

$$\frac{\partial \theta}{\partial t} = -\mathrm{div}_g\Big(\theta\big(-\tilde{\rho}\nabla_g \Psi - \nabla_g \phi\big)\Big) + \Delta_g(\tilde{\rho}\theta).$$

If we let $\beta = \frac{1}{Z}\exp(-\Psi)\theta$ we see, using (3.15), that

$$\begin{aligned} \Delta_g^\pi \beta^2 &= e^{\Psi}\Big(\Delta_g(\beta^2 e^{-\Psi}) + \mathrm{div}_g\big((\beta^2 e^{-\Psi})\nabla_g \Psi\big)\Big), \\ \mathrm{div}_g^\pi(\beta \nabla_g \phi) &= e^{\Psi}\mathrm{div}_g\big((\beta e^{-\Psi})\nabla_g \Psi\big), \end{aligned}$$

implying that the distributions of the stochastic process (3.16) are those generated by the gradient flow of J_{χ^2} in Wasserstein space.

Remark 3.7. *In contrast with the Langevin diffusion* (3.13)*, the process* (3.16) *is defined in terms of the solution of the equation satisfied by its densities. In particular, if one wanted to simulate* (3.16) *one would need to know the solution of* (3.14) *before hand.*

4. Application: Sampling and Riemannian MCMC So far we have treated the Riemannian manifold $(\mathcal{M}, g)$ as fixed. In this section we take a different perspective and treat the metric g as a free parameter. Precisely, we will now consider a family of gradient flows of the functional J_{KL} with respect to Wasserstein distances induced by different metrics g on the parameter space. We do this motivated by the so called Riemannian MCMC methods for sampling, where a change of metric in the base space is introduced in order to produce Langevin-type proposals that are adapted to the geometric features of the target, thereby exploring regions of interest and accelerating the convergence of the chain to the posterior. There are different heuristics regarding the choice of metric (see [Girolami and Calderhead, 2011]), but no principled way to compare different metrics and rank their performance for sampling purposes. With the developments presented in this paper we propose one such principled criterion as we describe below. We restrict our attention to the case $\mathcal{M} = \mathbb{R}^m$.

Let g be a Riemannian metric tensor on $\mathbb{R}^m$ defined via

$$g_x(u,v) := \langle G(x)u, v\rangle, \quad u,v \in \mathcal{T}_x\mathcal{M},$$

where for every $x \in \mathbb{R}^m$, $G(x)$ is a $m \times m$ positive definite matrix. In what follows we identify g with G and refer to both as 'the metric' and we use terms such as g-geodesic, g-Wassertein distance, etc. to emphasize that the notions considered are being constructed using the metric g. Let d_g be the distance induced by the metric tensor g and let vol_g be the associated volume form. Notice that in terms of the Lebesgue measure and the metric G, we can write

$$dvol_g(x) = \sqrt{\det\big(G(x)\big)}\, dx.$$

We use the canonical basis for $\mathbb{R}^m$ as global chart for $\mathbb{R}^m$ and consider the canonical vector fields $\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_m}$. The *Christoffel symbols* associated to the Levi-Civita connection of the Riemannian manifold $(\mathbb{R}^m, g)$ can be written in terms of derivatives of the metric as

$$\Gamma_{ij}^l = \frac{1}{2}\left(\frac{\partial}{\partial x_j}G_{ki} + \frac{\partial}{\partial x_i}G_{kj} - \frac{\partial}{\partial x_k}G_{ij}\right) G^{-1}_{lk}, \tag{4.1}$$

where in the right hand-side —and in what follows— we use Einstein's summation convention. The proof of the following result is in the Appendix.

Theorem 4.1. *Let $F \in C^2(\mathbb{R}^m)$ and*

$$\mu(du) \propto \exp\big(-F(u)\big)du.$$

The sharp constant λ for which J_{KL} (or $D_{\mathrm{KL}}(\cdot\|\mu)$) is λ-geodesically convex in the g-Wasserstein distance is equal to

$$\lambda_G := \inf_{x\in\mathbb{R}^m} \Lambda_{\min}\Big(G^{-1/2}\big(B + \operatorname{Hess} F - C\big)G^{-1/2}\Big),$$

where $\operatorname{Hess} F$ *is the usual (Euclidean) Hessian matrix of F, B is the matrix with coordinates*

$$B_{ij} := \frac{\partial \Gamma_{ij}^l}{\partial x_l} - \Gamma_{il}^k \Gamma_{jk}^l, \tag{4.2}$$

and C is the matrix with coordinates

$$C_{ij} := \Gamma_{ij}^l \frac{\partial F}{\partial x_l}. \tag{4.3}$$

Moreover, for any $a > 0$,

$$\lambda_{aG} = \frac{1}{a}\lambda_G. \tag{4.4}$$

Note that λ_G is a key quantity in evaluating the quality of a metric G in building geometry-informed Langevin diffusions for sampling purposes, as it gives the exponential rate at which the evolution of probabilities built using the metric G converges towards the posterior: larger λ_G corresponds to faster convergence. However, in order to establish a fair performance comparison, the metrics need to be scaled appropriately.

Indeed a faster rate can be obtained by scaling down the metric (which can be thought of as time-rescaling), as it is clearly seen by the scaling property (4.4) of the functional λ_G. It is important to note that scaling down the metric leads to a faster diffusion, but also makes its discretization more expensive. Indeed the error of Euler discretizations is largely influenced by the Lipschitz constant of the drift. This motivates that a fair criterion for choosing the metric could be to maximize λ_G with the constraint

$$\mathrm{Lip}(\nabla_g F) = \mathrm{Lip}(G^{-1}\nabla F) \leq 1, \tag{4.5}$$

since $\nabla_g F = G^{-1}\nabla F$ (where ∇ denotes the standard Euclidean gradient) is the drift of the diffusion (3.13). Note that the constraint (4.5) ensures that the metric cannot be scaled down arbitrarily while also guarantees that the discretizations do not become increasingly expensive. We remark that other constraints involving higher regularity requirements may be useful if higher order discretizations are desired.

Remark 4.2. *The functional λ_G can be used to determine the optimal metric among a certain subclass of metrics of interest satisfying the condition* (4.5). *For instance, it may be of interest to find the optimal constant metric G (see Proposition 4.3 below), or to find the best metric within a finite family of metrics. On the other hand the constraint* (4.5) *forces feasible metrics to induce diffusions that are not expensive to discretize.*

To illustrate the previous remark we show that for a Gaussian target measure the optimal preconditioner is, unsurprisingly, given by the Fisher information. More precisely we have the following proposition:

Proposition 4.3. *Let $\mu = N(0,\Sigma)$. Then*

$$G^* := \Sigma^{-1} \tag{4.6}$$

maximizes λ_G over the class of constant metrics G satisfying $\|G^{-1}\Sigma^{-1}\| \leq 1$, as in (4.5). *Moreover, the maximum value is*

$$\lambda_{G^*} = 1.$$

Proof. Suppose for the sake of contradiction that there exists a constant metric G that satisfies condition (4.5), which in this case reads $\|G^{-1}\Sigma^{-1}\| \leq 1$ and is such that $\lambda_G > \lambda_{G^*}$.

Let u be a unit norm eigenvector of G with eigenvalue $\lambda > 0$. Notice that by definition of λ_G we must have

$$\langle G^{-1/2}\Sigma^{-1}G^{-1/2}u, u\rangle \geq \lambda_G > \lambda_{G^*} = 1. \tag{4.7}$$

The left hand side of the above display can be rewritten as

$$\langle G^{-1/2}\Sigma^{-1}G^{-1/2}u, u\rangle = \langle G^{-1}\Sigma^{-1}G^{-1/2}u, G^{1/2}u\rangle$$

and by Cauchy-Schwartz inequality we see that

$$\langle G^{-1}\Sigma^{-1}G^{-1/2}u, G^{1/2}u\rangle \leq \|G^{-1}\Sigma^{-1}G^{-1/2}u\| \cdot \|G^{1/2}u\| \leq \|G^{-1/2}u\| \cdot \|G^{1/2}u\|$$

Since u is an eigenvector of G with eigenvalue λ, it follows that u is also an eigenvector of $G^{1/2}$ with eigenvalue $\sqrt{\lambda}$ and of $G^{-1/2}$ with eigenvalue $\frac{1}{\sqrt{\lambda}}$. Therefore the right hand side of the above display is equal to one. This however contradicts (4.7). From this we deduce the optimality of G^* among feasible metrics.

□

Example 2. *Suppose that* $F(u)=\frac{1}{2}\langle\Sigma^{-1}u,u\rangle$, *with*

$$\Sigma=\begin{bmatrix}1 & 0\\ 0 & \epsilon\end{bmatrix}.$$

Consider the optimal metric

$$G^*=\begin{bmatrix}1 & 0\\ 0 & \epsilon^{-1}\end{bmatrix}$$

given by the previous proposition and the rescaled Euclidean metric G_e

$$G_e=\begin{bmatrix}\epsilon^{-1} & 0\\ 0 & \epsilon^{-1}\end{bmatrix},$$

where the scalings have been chosen so that

$$\mathrm{Lip}(G^{*-1}\nabla F)=\mathrm{Lip}(G_e^{-1}\nabla F)=1.$$

A calculation then shows that $\lambda_{G^*}=1$ *while* $\lambda_{G_e}=\epsilon$. *Note that if the Euclidean metric is not rescaled by* ϵ^{-1} *—violating the constraint* (4.5)*— then the same unit rate of convergence as with the metric* G^* *is achieved. However, the drift of the associated diffusion*

$$dX_t=-\Sigma^{-1}X_t\,dt+\sqrt{2}\,dB_t$$

is of order ϵ^{-1}, *making the discretization increasingly expensive in the small* ϵ *limit. On the other hand, since both* G^{*-1} *and* G_e^{-1} *are of order* ϵ, *the drifts for both associated diffusions are order* 1. *This motivates our choice of constraint in equation* (4.5).

5. Example: Semi-Supervised Learning

In this section we study the geodesic convexity of functionals arising in the Bayesian formulation of semi-supervised classification. Our purpose is to illustrate the concepts in a tangible setting, and to show that establishing sharp levels of geodesic convexity may be more tractable for some functionals than others.

In semi-supervised classification one is interested in the following task: given a data cloud $X=\{x_1,\dots,x_n\}$ together with (noisy) labels $y_i\in\{-1,1\}$ for some of the data points x_i, $i\in\mathcal{Z}\subset\{1,\dots,n\}$, classify the unlabeled data points by assigning labels to them. Here we assume to have access to a weight matrix W quantifying the level of similarity between the points in X. Thus, we focus on the graph-based approach to semi-supervised classification, which boils down to propagating the known labels to the whole cloud, using the geometry of the weighted graph (X,W). We will investigate the existence and convergence of gradient flows for several Bayesian graph-based classification models proposed in [Bertozzi et al., 2017]. In the Bayesian approach, the geometric structure that the weighted graph imposes on the data cloud is used to build a prior on a latent space, and the noisy given labels are used to build the likelihood. The Bayesian solution to the classification problem is a measure on the latent space, that is then push-forwarded into a measure on the label space $\{-1,1\}^n$. This latter measure contains information on the most likely labels, and also provides a principled way to quantify the remaining uncertainty on the classification process.

Let (X,W) then be a weighted graph, where $X=\{x_1,\dots,x_n\}$ is the set of nodes of the graph and W is the weight matrix between the points in X. All the entries of W

are non-negative real numbers and we assume that W is symmetric. Let L be the graph Laplacian matrix defined by

$$L := D - W,$$

where D is the degree matrix of the weighted graph, i.e., the diagonal matrix with diagonal entries $D_{ii} = \sum_{j=1}^{n} W_{ij}$. The above corresponds to the *unnormalized* graph Laplacian, but different normalizations are possible [Von Luxburg, 2007]. The graph-Laplacian will be used in all the models below to favor prior draws of the latent variables that are consistent with the geometry of the data cloud.

Remark 5.1. *A special case of a weighted graph (X,W) frequently found in the literature is that in which the points in X are i.i.d. points sampled from some distribution on a manifold $\mathcal{M}$ embedded in $\mathbb{R}^d$, and the similarity matrix W is obtained as*

$$W_{ij} = K\left(\frac{|x_i - x_j|}{r}\right).$$

In the above, K is a compactly supported kernel function, $|x_i - x_j|$ is the Euclidean distance between the points x_i and x_j, and $r>0$ is a parameter controlling data density. It can be shown (see [Burago et al., 2013] and [Garcia Trillos et al., 2017a]) that the smallest non-trivial eigenvalue of a rescaled version of the resulting graph Laplacian is close to the smallest non-trivial eigenvalue of a weighted Laplacian on the manifold, provided that r is scaled with n appropriately.

We will now study the probit and logistic models in subsection 5.1, and then the Ginzburg-Landau model in 5.2.

5.1. Probit and Logistic Models Traditionally, the *probit* approach to semi-supervised learning is to classify the unlabeled data points by first optimizing the functional $G:\mathbb{R}^n \to \mathbb{R}$ given by

$$G(u) := \frac{1}{2}\langle L^\alpha u, u\rangle - \sum_{j\in\mathcal{Z}} \log\big(H(y_j u_j);\gamma\big), \quad H(w;\gamma) := \int_{-\infty}^{w} \exp(-t^2/2\gamma^2)\,dt \tag{5.1}$$

over all $u\in\mathbb{R}^n$ satisfying $\sum_{i=1}^n u_i = 0$, and then thresholding the optimizer with the sign function; the parameter $\alpha>0$ is used to regularize the functions u. The minimizer of the functional G can be interpreted as the MAP (maximum a posteriori estimator) in the Bayesian formulation of probit semi-supervised learning (see [Bertozzi et al., 2017]) that we now recall:

Prior: Consider the subspace $U := \{u\in\mathbb{R}^n : \sum_{i=1}^n u_i = 0\}$ and let π be the Gaussian measure on U defined by

$$\frac{d\pi}{du}(u) \propto \exp\left(-\frac{1}{2}\langle L^\alpha u, u\rangle\right) =: \exp\big(-\Psi(u)\big). \tag{5.2}$$

The measure π is interpreted as a prior distribution on the space of real valued functions on the point cloud X with average zero. Larger values of $\alpha>0$ force more regularization of the functions u.

Likelihood function: For a fixed $u\in U$ and for $j\in\mathcal{Z}$ define

$$y_j = S(u_j + \eta_j),$$

where the η_j are i.i.d. $N(0,\gamma^2)$, and S is the sign function. This specifies the distribution of observed labels given the underlying latent variable u. We then define, for given data

y, the negative log-density function

$$\phi(u;y) := -\sum_{j\in\mathcal{Z}} \log\big(H(y_j u_j;\gamma)\big), \quad u\in U, \tag{5.3}$$

where H is given by (5.1).

Posterior distribution: As shown in [Bertozzi et al., 2017], a simple application of Bayes' rule gives the posterior distribution of u given y (denoted by μ^y):

$$\frac{d\mu^y}{d\pi}(v) = \exp\big(-\phi(v;y)\big); \quad \frac{d\mu^y}{dv}(v) \propto \exp\big(-\Psi(v) - \phi(v;y)\big),$$

where Ψ is given by (5.2), and ϕ is given by (5.3).

From what has been discussed in the previous sections, the posterior μ^y can be characterized as the unique minimizer of the energy

$$J_{\mathrm{KL}}(\nu) := D_{\mathrm{KL}}(\nu\|\pi) + \int_{\mathbb{R}^n} \phi(u;y)d\nu(u), \quad \nu\in\mathcal{P}(\mathbb{R}^n). \tag{5.4}$$

Let us first consider the gradient flow of J_{KL} with respect to the usual Wassertsein space (i.e. the one induced by the Euclidean distance).

We can study the geodesic convexity of this functional by studying independently the convexity properties of $D_{\mathrm{KL}}(\nu\|\pi)$ and of $\phi(\cdot;y)$. Precisely:

i) Since π is a Gaussian measure with covariance $L^{-\alpha}$, Example 1 shows that $D_{\mathrm{KL}}(\nu\|\pi)$ is $(\Lambda_{\min}(L))^\alpha$-geodesically convex in Wasserstein space, where $\Lambda_{\min}(L)$ is the smallest non-trivial eigenvalue of L.
ii) The function $\phi(\cdot;y)$ is convex —see the appendix of [Bertozzi et al., 2017]. Hence, the functional $F_{\mathrm{KL}}(\nu) = \int_{\mathbb{R}^n} \phi(u;y)d\nu(u)$ is 0-geodesicaly convex in Wasserstein space.

It then follows from Proposition 3.1 that J_{KL} is $(\Lambda_{\min}(L))^\alpha$-geodesically convex in Wasserstein space. As a consequence, if we consider $t\in[0,\infty)\mapsto\mu_t$, the gradient flow of J_{KL} with respect to the Wasserstein distance starting at μ_0 (an absolutely continuous measure with respect to μ), geometric inequalities can be immediately obtained from (3.7); such inequalities will not deteriorate with n —see Remark 5.1.

However, the diffusion associated to this flow is given by

$$dX_t = (-L^\alpha X_t - \nabla\phi(X;y))\,dt + \sqrt{2}dB_t, \tag{5.5}$$

and in particular its drift (more precisely the term $L^\alpha X_t$) deteriorates as n gets larger. Notice that if we wanted to control the cost of discretization by rescaling the Euclidean metric (as exhibited in Example 2), the geodesic convexity of the resulting flow would vanish as n gets larger.

The previous discussion shows that the flow of J_{KL} in the usual Wasserstein sense does not produce a flow with good convergence properties that at the same time is cheap to discretize (robustly in n). This motivates considering the gradient flow of J_{KL} with respect to the Wasserstein distance induced by a certain constant metric g. Indeed, inspired by Proposition 4.3, let us consider the constant metric tensor

$$G := L^\alpha.$$

Since the metric tensor is constant, in particular its induced volume form vol_g is proportional to the Lebesgue measure and hence we can write

$$d\mu^y(u) \propto \exp\left(-\phi(u;y) - \frac{1}{2}\langle L^\alpha u, u\rangle\right) dvol_g(u), \quad u\in U.$$

On the other hand, from the discussion in Section 3.3.1 we know that the densities of the stochastic process

$$dX_t = -\nabla_g(\phi(X_t;y) + \frac{1}{2}\langle L^\alpha X_t, X_t\rangle)dt + \sqrt{2}dB_t^g$$

correspond to to the gradient flow of the energy J_{KL} with respect to the Wasserstein distance induced by the metric g, where B^g is a Brownian motion on $(\mathbb{R}^m, g)$. This diffusion can be rewritten in terms of the standard Euclidean gradient ∇ and Brownian motion B as

$$dX_t = -(X_t + L^{-\alpha}\nabla\phi(X_t;y))dt + \sqrt{2L^{-\alpha}}dB_t, \tag{5.6}$$

after noticing that

$$\nabla_g = G^{-1}\nabla, \quad B^g = \sqrt{G^{-1}}B,$$

where for the second identity we have used the fact that G is constant. How convex is the energy J_{KL} with respect to the Wasserstein distance induced by g? Since the metric tensor G is constant it follows that

$$\lambda_G := \inf_{x\in\mathbb{R}^m} \Lambda_{\min}\Big(G^{-1/2}\big(\text{Hess}\, F\big)G^{-1/2}\Big),$$

where $F(u) := \phi(u;y) + \frac{1}{2}\langle L^\alpha u, u\rangle$. Finally, due to the convexity of $\phi(u;y)$ we deduce that

$$\lambda_G \geq \Lambda_{\min}\Big(G^{-1/2}L^\alpha G^{-1/2}\Big) = 1.$$

We notice that in (5.6) L appears as $L^{-\alpha}$. This is a fundamental difference from (5.5) (where L appears as L^α) with computational advantages, given that the eigenvalues of L grow towards infinity.

Remark 5.2. *A carefully designed discretization of* (5.6) *induces the so called Langevin pCN proposal for MCMC computing (see [Cotter et al., 2013]).*

Remark 5.3. *In the above we have considered a probit model for the likelihood function. The ideas generalize straightforwardly to other settings, notably the logistic model*

$$\phi(u;y) := -\sum_{j\in\mathcal{Z}} \log\big(\sigma(y_j u_j;\gamma)\big), \quad u\in U, \tag{5.7}$$

where

$$\sigma(t;\gamma) := \frac{1}{1+e^{-t/\gamma}}.$$

The convexity of ϕ for the logistic model (5.7) *can be established by direct computation of the second derivative of σ.*

5.2. Ginzburg-Landau Model We now present the Ginzburg-Landau model for semi-supervised learning. This model will provide us with an example of a functional J_{KL} whose geodesic convexity with respect to Wasserstein distance is not positive (and hence one can not deduce geometric inequalities describing the rate of convergence towards the posterior), but for which one can obtain a positive spectral gap giving the rate of convergence of the flow of Dirichlet energy in the L^2 sense.

Let

$$W_\epsilon(t) := \frac{1}{4\epsilon}(t^2-1)^2, \quad t \in \mathbb{R}.$$

We consider the following Bayesian model.

Prior:

$$\frac{d\pi}{dv}(u) \propto \exp\Big(-\frac{1}{2}\langle u, L^\alpha u\rangle - \sum_{j\in\mathcal{Z}} W_\epsilon\big(u_j\big)\Big) =: \exp\big(-\Psi(u)\big), \quad u \in U. \tag{5.8}$$

Likelihood function: For $j \in \mathcal{Z}$,

$$y_j = u_j + \eta_j, \qquad \eta_j \sim N(0,\gamma^2).$$

This leads to the following negative log-density function:

$$\phi(u;y) = \frac{1}{2\gamma^2}\sum_{j\in\mathcal{Z}} |y_j - u_j|^2. \tag{5.9}$$

Posterior distribution: Combining the prior and the likelihood via Bayes' formula gives the posterior distribution

$$\frac{d\mu^y}{d\pi}(u) = \exp\big(-\phi(u;y)\big); \quad \frac{d\mu^y}{du}(u) \propto \exp\big(-\Psi(u) - \phi(u;y)\big), \quad u \in U,$$

where Ψ is given by (5.8), and ϕ is given by (5.9).

For this model, the negative prior log-density Ψ is not convex, and Wasserstein λ-geodesic convexity of the functional $D_{\mathrm{KL}}(\cdot\|\pi)$ only holds for negative λ. In particular, it is not possible to deduce exponential decay taking the Wasserstein flow point of view. However, in the L^2/Dirichlet energy setting we can still show exponential convergence towards the posterior μ^y. Indeed, because the negative log-likelihood of μ^y satisfies:

$$\frac{|\nabla\Psi(u) + \nabla\phi(u;y)|^2}{2} - \Delta\Psi(u) - \Delta\phi(u;y) \to \infty, \quad \text{as } |u| \to \infty,$$

there exists some $\lambda > 0$ for which μ^y has a Poincaré inequality with constant λ (see Chapter 4.5 in [Pavliotis, 2014]). In this example we can say more, and in particular we are able to find a Poincaré constant that depends explicitly on ε, the smallest non-trivial eigenvalue of L, and $k := |\mathcal{Z}|$.

Let $\psi(u) := \sum_{j\in\mathcal{Z}} W_\varepsilon(u_j)$ and let ψ_c be its convex envelope, i.e. let ψ_c be the largest convex function that is below ψ. It is straightforward to show that $\psi_c(0) = 0$ and that

$$\inf_{u\in\mathbb{R}^n}\{\exp(\psi_c(u) - \psi(u))\} = \exp\left(-\frac{k}{4\varepsilon}\right).$$

Consider now the probability measure μ_c with Lebesgue density

$$d\mu_c(u) = \frac{1}{Z_c}\exp\left(-\frac{1}{2}\langle L^\alpha u, u\rangle - \phi(u;y) - \psi_c(u)\right)du,$$

and define λ_2 and $\lambda_{2,c}$ as in Remark 3.5 using μ^y and μ_c instead of μ. For any given $f \in L^2(\mu)$ we then have

$$\frac{\int|\nabla f|^2 d\mu}{\int|f - f_\mu|^2 d\mu} \geq \frac{\int|\nabla f|^2 d\mu}{\int|f - f_{\mu_c}|^2 d\mu} \geq \exp\left(-\frac{k}{\varepsilon}\right)\frac{\int|\nabla f|^2 d\mu_c}{\int|f - f_{\mu_c}|^2 d\mu_c},$$

where the first inequality follows from the fact that $f_\mu = \text{argmin}_{a\in\mathbb{R}} \int |f-a|^2 d\mu$ and the second inequality follows directly from the fact that $0 \geq \psi_c - \psi \geq -\frac{k}{\varepsilon}$. It follows that

$$\lambda_2 \geq \exp\left(-\frac{k}{\varepsilon}\right)\lambda_{2,c} \geq \exp\left(-\frac{k}{\varepsilon}\right)(\Lambda_{\min}(L))^\alpha.$$

where the last inequality follows from the fact that the negative log-likelihood of μ^y satisfies the Bakry-Emery condition with constant $\Lambda_{\min}(L)$ (see Chapter 4.5 in [Pavliotis, 2014]). Clearly, the Poincaré constant above is very large for small ϵ or for large k (number of labeled data points). We also notice that the cost of discretization of the diffusion associated to this flow increases with n (as in Section 5.1).

Remark 5.4. *A similar analysis can be carried out now using the constant metric*

$$G := L^\alpha.$$

More precisely, consider the flow of the Dirichlet energy

$$D_g^\mu(f) := \int |\nabla_g f|_g^2 d\mu(u)$$

with respect to $L^2(\mu)$. How convex is this functional? For every $f \in U$ we have

$$\frac{\int |\nabla_g f|_g^2 d\mu}{\int |f-f_\mu|^2 d\mu} \geq \frac{\int |\nabla_g f|_g^2 d\mu}{\int |f-f_{\mu_c}|^2 d\mu} \geq \exp\left(-\frac{k}{\varepsilon}\right)\frac{\int |\nabla_g f|_g^2 d\mu_c}{\int |f-f_{\mu_c}|^2 d\mu_c},$$

from where it follows that

$$\lambda_2^g \geq \exp\left(-\frac{k}{\varepsilon}\right).$$

A similar remark to the one at the end of section 5.1 regarding the dependence in L of the resulting diffusion applies here as well.

6. Conclusions and Future Work The main contribution of this paper is to explore three variational formulations of the Bayesian update and their associated gradient flows. We have shown that, for each of the three variational formulations, the geodesic convexity of the objective functionals gives a bound on the rate of convergence of the flows to the posterior. As an application of the theory, we have suggested a criterion for the optimal choice of metric in Riemannian MCMC schemes. We summarize below some additional outcomes and directions for further work.

- We bring attention to different variational formulations of the Bayesian update. These formulations have the potential of extending the theory of Bayesian inverse problems in function spaces, in particular in cases with infinite dimensional, non-additive, and non-Gaussian observation noise. Moreover, they suggest numerical approximations to the posterior by restricting the space of allowed measures in the minimization, by discretization of the associated gradient flows, or by sampling via simulation of the associated diffusion.
- The variational framework considered in this paper provides a natural setting for the study of robustness of Bayesian models, and for the analysis of convergence of discrete to continuum Bayesian models. Indeed, the authors [Garcia Trillos and Sanz-Alonso, 2017], [Garcia Trillos et al., 2017b] have recently established the consistency of Bayesian semi-supervised learning in the regime with fixed number of labeled data points and growing number of unlabeled data. The analysis relies on the variational formulation based on Kullback-Leibler prior penalization in equation (5.4).

- The results in the paper give new understanding of the ubiquity of Kullback-Leibler penalizations in sampling methodology. In practice Kullback-Leibler is often used for computational and analytical tractability. The results presented in section 3.3 show that Kullback-Leibler prior penalization leads to a heat-type flow and, therefore, to an easily discretized diffusion process. On the other hand, χ^2 prior penalization leads to a nonlinear diffusion process.

Acknowledgments. We are thankful to Matías Delgadino for pointing to us the reference [Ohta and Takatsu, 2011] while participating in the CNA Ki-net workshop "Dynamics and Geometry from High Dimensional Data" that took place at Carnegie Mellon University in March 2017. We are also thankful to Sayan Mukherjee for the reference [Zellner, 1988]. Finally, we thank the anonymous editor and referees for their immense help in improving the readability of our manuscript.

REFERENCES

[Ambrosio et al., 2008] Ambrosio, L., Gigli, N., and Savaré, G. (2008). *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media.

[Attias, 1999] Attias, H. (1999). A Variational Bayesian Framework for Graphical Models. In *NIPS*, volume 12.

[Bertozzi et al., 2017] Bertozzi, A. L., Luo, X., Stuart, A. M., and Zygalakis, K. C. (2017). Uncertainty quantification in the classification of high dimensional data.

[Besag, 1994] Besag, J. E. (1994). Comments on "representations of knowledge in complex systems" by u. grenander and m. i. miller. *J. Roy. Statist. Soc. Ser. B*, 56:591–592.

[Burago et al., 2001] Burago, D., Burago, Y., and Ivanov, S. (2001). *A course in metric geometry*, volume 33 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI.

[Burago et al., 2013] Burago, D., Ivanov, S., and Kurylev, Y. (2013). A graph discretization of the Laplace-Beltrami operator. *arXiv preprint arXiv:1301.2222*.

[Cotter et al., 2013] Cotter, S. L., Roberts, G. O., Stuart, A. M., and White, D. (2013). MCMC methods for functions: modifying old algorithms to make them faster. *Statistical Science*, 28(3):424–446.

[do Carmo Valero, 1992] do Carmo Valero, M. P. (1992). *Riemannian Geometry*.

[Dupuis and Ellis, 2011] Dupuis, P. and Ellis, R. S. (2011). *A weak convergence approach to the theory of large deviations*, volume 902. John Wiley & Sons.

[Fox and Roberts, 2012] Fox, C. W. and Roberts, S. J. (2012). A tutorial on variational Bayesian inference. *Artificial intelligence review*, 38(2):85–95.

[Garcia Trillos et al., 2017a] Garcia Trillos, N. ., Gerlach, M. ., Hein, M., and Slepcev, D. (2017a). Spectral convergence of empirical graph Laplacians. *In preparation*.

[Garcia Trillos et al., 2017b] Garcia Trillos, N., Kaplan, Z., Samakhoana, T., and Sanz-Alonso, D. (2017b). On the consistency of graph-based bayesian learning and the scalability of sampling algorithms. *arXiv preprint arXiv:1710.07702*.

[Garcia Trillos and Sanz-Alonso, 2017] Garcia Trillos, N. and Sanz-Alonso, D. (2017). Continuum limit of posteriors in graph Bayesian inverse problems. *arXiv preprint arXiv:1706.07193*.

[Girolami and Calderhead, 2011] Girolami, M. and Calderhead, B. (2011). Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214.

[Guo et al., 2016] Guo, F., Wang, X., Fan, K., Broderick, T., and Dunson, D. B. (2016). Boosting variational inference. *arXiv preprint arXiv:1611.05559*.

[Jordan and Kinderlehrer, 1996] Jordan, R. and Kinderlehrer, D. (1996). 18, an extended variational principle. *Partial differential equations and applications: collected papers in honor of Carlo Pucci*, page 187.

[Jordan et al., 1998] Jordan, R., Kinderlehrer, D., and Otto, F. (1998). The variational formulation of the fokker–planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17.

[Kipnis and Varadhan, 1986] Kipnis, C. and Varadhan, S. R. S. (1986). Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions. *Communications in Mathematical Physics*, 104(1):1–19.

[McCann, 1997] McCann, R. J. (1997). A convexity principle for interacting gases. *Advances in mathematics*, 128(1):153–179.

[Ohta and Takatsu, 2011] Ohta, S. and Takatsu, A. (2011). Displacement convexity of generalized relative entropies. *Advances in Mathematics*, 228(3):1742–1787.

[Pavliotis, 2014] Pavliotis, G. A. (2014). Stochastic processes and applications. *Texts in Applied Mathematics. Springer, Berlin.*

[Pinski et al., 2015] Pinski, F. J., Simpson, G., Stuart, A. M., and Weber, H. (2015). Kullback–Leibler approximation for probability measures on infinite dimensional spaces. *SIAM Journal on Mathematical Analysis*, 47(6):4091–4122.

[Roberts and Tweedie, 1996] Roberts, G. O. and Tweedie, R. L. (1996). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, pages 341–363.

[Santambrogio, 2015] Santambrogio, F. (2015). Optimal transport for applied mathematicians. *Birkäuser, NY.*

[Sturm, 2006] Sturm, K. T. (2006). On the geometry of metric measure spaces. *Acta mathematica*, 196(1):65–131.

[Villani, 2003] Villani, C. (2003). *Topics in optimal transportation.* Number 58. American Mathematical Soc.

[Villani, 2008] Villani, C. (2008). *Optimal transport: old and new*, volume 338. Springer Science & Business Media.

[Von Luxburg, 2007] Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416.

[von Renesse and Sturm, 2005] von Renesse, M. K. and Sturm, K. T. (2005). Transport inequalities, gradient estimates, entropy and Ricci curvature. *Communications on pure and applied mathematics*, 58(7):923–940.

[Wainwright and Jordan, 2008] Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305.

[Zellner, 1988] Zellner, A. (1988). Optimal information processing and Bayes's theorem. *The American Statistician*, 42(4):278–280.

Appendix A. Proof of Theorem 4.1.

Notice that the measure μ can be rewritten as

$$d\mu(x) \propto \exp\left(-F(x) - \log\left(\sqrt{\det(G(x))}\right)\right) dvol_g(x) = \exp(-F_g(x)) dvol_g(x),$$

where

$$F_g(x) := F(x) + \log\left(\sqrt{\det(G)}\right), \quad x \in \mathbb{R}^m.$$

From Proposition 3.1 the sharp constant λ for which $D_{\mathrm{KL}}(\cdot\|\mu)$ is λ-geodesically convex in the g-Wasserstein space is given by

$$\inf_{x\in\mathbb{R}^m} \min_{v: g(v,v)=1} \mathrm{Ric}_g(v,v) + \mathrm{Hess}_g F_g(v,v),$$

where Ric_g and Hess_g stand for the Ricci curvature and Hessian in the g-metric. To establish the proposition, it suffices to show that for any given $x \in \mathbb{R}^m$, $\min_{v:g(v,v)=1} \mathrm{Ric}_g(v,v) + \mathrm{Hess}_g F_g$ is equal to the smallest eigenvalue of the matrix $B + \mathrm{Hess}\, F - C$.

Let us start by recalling that the g-Hessian of a C^2 function I, denoted by $\mathrm{Hess}_g I$, is the symmetric $(2,0)$-tensor satisfying

$$\mathrm{Hess}_g I(v,v) = \frac{d^2}{dt^2} I\big(\gamma(t)\big)\Big|_{t=0},$$

for every $v \in \mathbb{R}^m$ and every constant speed g-geodesic curve with $\gamma(0) = x$ and $\dot\gamma(0) = v$. It is convenient to rewrite $\mathrm{Hess}_g I(v,v)$ in terms of the Christoffel symbols of the metric g, the Euclidean inner product, and the regular (Euclidean) gradient and Hessian of the

function I. Let $\gamma:(-\varepsilon,\varepsilon)\to\mathbb{R}^m$ be a constant speed g-geodesic with $\gamma(0)=x$, $\dot\gamma(0)=v$. We can then write:

$$\begin{aligned}\frac{d^2}{dt^2}I\big(\gamma(t)\big)&=\frac{d}{dt}\Big(\big\langle\nabla I(\gamma(t)),\dot\gamma(t)\big\rangle\Big)\\&=\Big\langle\mathrm{Hess}\,I\big(\gamma(t)\big)\dot\gamma(t),\dot\gamma(t)\Big\rangle+\Big\langle\nabla I(\gamma(t)),\ddot\gamma(t)\Big\rangle,\end{aligned}\tag{1.1}$$

where ∇I and $\mathrm{Hess}\,I$ are the usual gradient and Hessian matrix of I, respectively. The acceleration of the curve γ can be written in terms of the Christoffel symbols. Namely, if we write γ in coordinates as

$$\gamma(t)=\big(\gamma_1(t),\dots,\gamma_m(t)\big),\quad t\in(-\varepsilon,\varepsilon),$$

the following system of second order ODEs holds:

$$\ddot\gamma_l(t)=-\sum_{ij}\Gamma^l_{ij}\dot\gamma_i(t)\dot\gamma_j(t),\quad l=1,\dots,m.$$

Plugging the geodesic equations back into (1.1), and setting $t=0$, it follows that

$$\frac{d^2}{dt^2}I\big(\gamma(t)\big)|_{t=0}=\Big\langle\big(\mathrm{Hess}\,I-A(I)\big)v,v\Big\rangle,$$

where $A(I)$ is the matrix with coordinates

$$A(I)_{ij}=\Gamma^l_{ij}\frac{\partial I}{\partial x_l}.$$

Hence,

$$\mathrm{Hess}_g\,I(v,v)=\Big\langle\big(\mathrm{Hess}I-A(I)\big)v,v\Big\rangle,\quad\forall v\in\mathbb{R}^m.\tag{1.2}$$

Taking $I:=F_g$ we can write the coordinate ij of the matrix $\mathrm{Hess}\,I-A(I)$ as

$$\frac{\partial^2F}{\partial x_ix_j}+\frac{\partial^2}{\partial x_ix_j}\log\Big(\sqrt{\det(G)}\Big)-\Gamma^l_{ij}\frac{\partial F}{\partial x_l}-\Gamma^l_{ij}\frac{\partial}{\partial x_l}\log\Big(\sqrt{\det(G)}\Big).$$

Using now the fact that the g-divergence of the vector field $\frac{\partial}{\partial x_l}$ can be written in terms of the Christoffel symbols as

$$\frac{\partial}{\partial x_l}\log\Big(\sqrt{\det(G)}\Big)=\Gamma^k_{lk},$$

we deduce that

$$\big(\mathrm{Hess}\,I-A(I)\big)_{ij}=\frac{\partial^2F}{\partial x_i\partial x_j}+\frac{\partial\Gamma^l_{il}}{\partial x_j}-\Gamma^l_{ij}\frac{\partial F}{\partial x_l}-\Gamma^l_{ij}\Gamma^k_{lk}.$$

On the other hand, the Ricci curvature $\mathrm{Ric}_g(v)$ can be written in terms of the Christoffel symbols and its derivatives (alternatively in terms of the metric and its first and second order derivatives) as

$$\mathrm{Ric}_g(v,v)=\langle Rv,v\rangle,\quad v\in\mathbb{R}^m,\tag{1.3}$$

where R is the (symmetric) matrix with entries

$$R_{ij} = \frac{\partial \Gamma^l_{ij}}{\partial x_l} - \frac{\partial \Gamma^l_{il}}{\partial x_j} + \Gamma^l_{ij}\Gamma^k_{kl} - \Gamma^k_{il}\Gamma^l_{jk},$$

see [do Carmo Valero, 1992] for details. After some cancellations using the symmetry of the symbols, we obtain that

$$\big(R + \mathrm{Hess}\, I - A(I)\big)_{ij} = \frac{\partial \Gamma^l_{ij}}{\partial x_l} - \Gamma^k_{il}\Gamma^l_{jk} + \frac{\partial^2 F}{\partial x_i \partial x_j} - \Gamma^l_{ij}\frac{\partial F}{\partial x_l},$$

and so

$$R + \mathrm{Hess}\, I - A(I) = B + \mathrm{Hess}\, F - C.$$

Using (1.2) and (1.3) we deduce that

$$\mathrm{Ric}_g(v,v) + \mathrm{Hess}_g F_g(v,v) = \big\langle (B + \mathrm{Hess}\, F - C)v, v\big\rangle, \quad \forall v \in \mathbb{R}^m.$$

Therefore the variational problem (for every fixed $x \in \mathbb{R}^m$)

$$\min_{v: g(v,v)=1} \mathrm{Ric}_g(v,v) + \mathrm{Hess}_g F(v,v)$$

can be rewritten, applying the change of variables $w := G^{1/2}v$, as

$$\min_{v:\langle w,w\rangle = 1} \Big\langle G^{-1/2}\big(B + \mathrm{Hess}\, F - C\big)G^{-1/2}w, w\Big\rangle.$$

In turn this coincides with

$$\Lambda_{\min}\Big(G^{-1/2}\big(B + \mathrm{Hess}\, F - C\big)G^{-1/2}\Big),$$

i.e., the smallest eigenvalue of the matrix

$$G^{-1/2}\big(B + \mathrm{Hess}\, F - C\big)G^{-1/2}.$$

This concludes the proof of the first part of the theorem.

Now we show the scaling property (4.4). Let $\widetilde{G} = aG$. By definition

$$\lambda_{\widetilde{G}} = \inf_{x \in \mathbb{R}^m} \Lambda_{\min}\Big(\widetilde{G}^{-1/2}\big(\widetilde{B} + \mathrm{Hess}\, F - \tilde{C}\big)\widetilde{G}^{-1/2}\Big),$$

where $\widetilde{B}$ and $\widetilde{C}$ are defined as in (4.2) and (4.3) but in terms of the metric $\tilde{G}$. From the expression (4.1) for the Christoffel symbols, it follows that they are invariant under rescaling of the metric and, since $B, \widetilde{B}$ and $C, \widetilde{C}$ depend on the metric only through the symbols, we deduce that $\widetilde{B} = B$, $\widetilde{C} = C$. Therefore,

$$\begin{aligned}\lambda_{\widetilde{G}} &= \frac{1}{a}\inf_{x \in \mathbb{R}^m} \Lambda_{\min}\Big(G^{-1/2}\big(B + \mathrm{Hess}\,(F) - C\big)G^{-1/2}\Big) \\ &= \frac{1}{a}\lambda_G.\end{aligned}$$

□