Annals of Mathematical Sciences and Applications
Volume 5, Number 2, 261–282, 2020

A note on parametric Bayesian inference via gradient flows

Yuan Gao* and Jian-Guo Liu

In this note, we summarize several recent developments for efficient sampling methods for parameters based on Bayesian inference. To reformulate those sampling methods, we use different formulations for gradient flows on the manifold in the parameter space, including strong form, weak form and De Giorgi type duality form. The gradient flow formulations will cover some applications in deep learning, ensemble Kalman filter for data assimilation, kinetic theory and Markov chain Monte Carlo.

Keywords and phrases: Parameter updating, KL-divergence, generalized gradient flow.

1. Introduction

We review some recent exciting developments using gradient flow in the parameter space for efficient sampling, including applications in ensemble Kalman filter for data assimilation, kinetic theory, deep learning and Markov chain Monte Carlo (MCMC). Based on Bayesian inference, all those problems can be regarded as parameter updating or reconstruction suggested by collected data. Consider a system for x, which can be determined by parameters θ. In the Bayesian inference, parameters θ are described by a probability density function (pdf) $\pi(\theta)$. We want to seek and sample the pdf $\pi(\theta)$ for parameter θ. However it is impossible to find the true $\pi(\theta)$ and the only thing we can expect is to learn more about π by gathering and analyzing data. The statistics only means we try to find an approximated pdf of θ which is the best suggested by data. More importantly, the goal is to design an efficient sampling method for this pdf $\pi(\theta)$, which is a big challenge due to the high dimensionality of the parameters and the complexity of the system.

Let x be the set of observed data with the pdf $f(x)$. Here and in the remaining of this paper, the "pdf $f(x)$" also refers to the "density $f(x)$" when $f(x)$ is absolutely continuous with respect to Lebesgue measure. Regard (x, θ) as random variables with the joint pdf $F(x, \theta)$. For given parameters

*ORCID: 0000-0002-7231-5672.

θ, assume we know the likelihood distribution $f(x|\theta)$. Bayesian formula reads

$$F(x,\theta) = f(x|\theta)\pi(\theta) = f(\theta|x)f(x). \tag{1.1}$$

The goal is to approximate $f(\theta|x)$ (as well as $\pi(\theta)$) by a sequence of $\rho_k(\theta)$ which is updated iteratively using a stream of data. In each updating step, denote $\rho_{\text{ap}}(\theta)$ as the priori density of θ. Denote $\rho_{\text{ps}}(\theta;x)$ as the posterier density of θ with given data x, which characterizes the probability density of parameter θ to be updated based on collected data x. We want to approximate

$$f(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{f(x)} \approx \frac{f(x|\theta)\rho_{\text{ap}}(\theta)}{f(x)} =: \rho_{\text{ps}}(\theta;x) \tag{1.2}$$

iteratively. Here at each step we use a shorthand notation $\rho_{\text{ps}}(\theta)$ for $\rho_{\text{ps}}(\theta;x)$ with fixed data. In the approximation (1.2),

$$f(x|\theta)\rho_{\text{ap}}(\theta) = \rho_{\text{ps}}(\theta;x)f(x) \tag{1.3}$$

can be regarded as the Bayesian formula for the random variable (x,θ) with θ-marginal given by $\rho_{\text{ap}}(\theta)$ and x-marginal given by $f(x)$. We will see later $f(x)$ does not affect the updating of parameters.

Although $\rho_{\text{ps}}(\theta) \propto f(x|\theta)\rho_{\text{ap}}(\theta)$ is known, the $\rho_{\text{ps}}(\theta)$ cannot be calculated directly by (1.3) due to high dimensionality in practice. On the other hand, we cannot directly sample $\rho_{\text{ps}}(\theta)$ due to high dimensionality and complexity of system. One way is to find a dynamic system for $\rho_t(\theta)$ to drive the initial density $\rho_0(\theta) = \rho_{\text{ap}}(\theta)$ to equilibrium $\rho_\infty(\theta) = \rho_{\text{ps}}(\theta)$. The dynamic system should has ergodicity and an invariant density, given by $\rho_{\text{ps}}(\theta)$. More precisely, we consider a manifold $\mathcal{M}$ in the parameter space and the push-forward density $\rho_t(\theta) = T_t\#\rho_{\text{ap}}(\theta) \in \mathcal{M}$, and find the best curve (under some constraints) ρ_t driving ρ_0 to ρ_∞.

We recast (1.3) as

$$\ln f(x|\theta) + \ln \rho_{\text{ap}} = \ln \rho_{\text{ps}} + \ln f(x). \tag{1.4}$$

The natural free energy describing the closeness to the equilibrium is the Kullback–Leibler divergence (KL-divergence) $KL(\rho|\rho_{\text{ps}}) := \int_\Omega \rho \ln \frac{\rho}{\rho_{\text{ps}}}\,\mathrm{d}\theta$. Consider the following minimization problem

$$\begin{aligned}\rho_{\text{ps}} =& \operatorname{argmin}_{\rho\in\mathbb{P}(\Omega)} KL(\rho|\rho_{\text{ps}}) = \operatorname{argmin} \int_\Omega \rho \ln \frac{\rho}{\rho_{\text{ps}}}\,\mathrm{d}\theta \\ =& \operatorname{argmin} \int_\Omega \rho \ln \frac{\rho}{f(x|\theta)\rho_{\text{ap}}}\,\mathrm{d}\theta.\end{aligned} \tag{1.5}$$

Due to the positivity of $f(x|\theta)\rho_{\text{ap}}$, it can always be rewritten as $\rho_{\text{ps}}(\theta) = f(x|\theta)\rho_{\text{ap}}(\theta) = e^{-U(\theta)}$ for some function $U(\theta)$. Therefore the minimization problem is recast to

(1.6)
$$\rho_{\text{ps}} = \text{argmin}_{\rho\in\mathbb{P}(\Omega)} \int_\Omega \rho \ln \frac{\rho}{f(x|\theta)\rho_{\text{ap}}} \,\mathrm{d}\theta = \text{argmin}_{\rho\in\mathbb{P}(\Omega)} \int_\Omega \rho \ln \rho + \rho U(\theta) \,\mathrm{d}\theta.$$

To obtain efficient sampling for the parameter θ such that $\rho_t(\theta)$ converges to its equilibrium $\rho_{\text{ps}}(\theta)$ for each collected data set, which is the updated/reconstructed parameter density, one need to design dynamic systems using some accelerated methods.

One efficient way to design the dynamic system on the parameter space is to construct a proper gradient flow for $\rho_t(\theta)$ with the equilibrium $\rho_\infty(\theta)$. To reformulate several recent acceleration algorithms as gradient flows on a manifold $\mathcal{M}$, we will discuss different Riemannian metrics on the tangent plane $T_{\rho_t}\mathcal{M}$, which lead to different admissible velocity set for the underlying flow map. Then the gradient flow on the flow map induce manifold $\mathcal{M}$ will give the steepest descent curve $\rho_t(\theta)$ with the steepest descent velocity $\dot\theta$ of the underlying flow map in terms of a given free energy $\mathcal{F}(\rho)$ and a Riemannian metrics g_{ρ_t}. The weak formulation, strong formulation and De Giorgi type duality formulation of gradient flows will be discussed in Section 2 and Section 4 respectively. As examples of gradient flows on manifold, we summarize in Section 3 some successful accelerated sampling methods such as Ornstein–Uhlenbeck (OU) process in kinetic theory, ensemble Kalman filter for data assimilation, Stein variation gradient descent(SVGD), parameter training in deep learning. Another efficient way to design the dynamical system on parameter space is to construct Markov chain Monte Carlo (MCMC) directly on parameter space (such as importance sampling, variance reduction, rejection method), which can also be reformulated as a generalized gradient flow using De Giorgi type duality formulations; see Section 4.2.1.

2. Gradient flows on a parameter manifold

To construct a dynamic system driving the initial density $\rho_0(\theta) = \rho_{\text{ap}}(\theta)$ to equilibrium $\rho_\infty(\theta) = \rho_{\text{ps}}(\theta)$, we regard the dynamic solution ρ_t as a curve on a parameter manifold described below. We will first illustrate a flow map (pushforward density) on a manifold and the calculations for first variation of free energy in Section 2.1. Then we will summarize several formulations for gradient flows on the manifold in Section 2.2.

2.1. Manifold in parameter space induced by pushforward density

Consider a configuration manifold $\mathcal{M} \subset \mathbb{P}(\bar{\Omega})$ and a flow map on the manifold with any given velocity field and a curve ρ_t on the manifold $\mathcal{M}$. Assume the change of Eulerian spatial variable $\theta \in \Omega \subset \mathbb{R}^d$ is described by the flow map $\theta_t(\Theta) : \Omega_0^\Theta \to \Omega_t^\theta$, satisfying

$$\begin{cases} \dot{\theta}_t(\Theta) = v_t(\theta_t(\Theta)); \\ \theta_0(\Theta) = \Theta, \end{cases} \tag{2.1}$$

where the velocity $v_t \in \mathbb{R}^d$ is to be determined. Then the curve ρ_t on the configuration manifold is defined by the pushforward density

$$\rho_t = \theta_t \# \rho_0. \tag{2.2}$$

By the definition of pushforward density, at each time $t > 0$, we have the mass conversation law

$$\partial_t \rho_t = -\nabla \cdot (\rho_t v_t), \quad \theta \in \Omega_t \tag{2.3}$$

in the distributional sense.

In general, the tangent plane at ρ_t is uniquely determined by the admissible velocity field for the flow map

$$T_{\rho_t}\mathcal{M} = \{(-\nabla \cdot (\rho_t v_t);\ v_t \in \{\text{ admissible velocity space }\}\}. \tag{2.4}$$

We will see the curve given by gradient flow on the manifold $\mathcal{M}$ can be determined by the metrics on tangent plane $T_{\rho_t}\mathcal{M}$.

2.1.1. First variation of free energy. In general, assume the free energy is

$$\mathcal{F}(\rho_t) := \int_{\Omega_t} w(\rho_t, \nabla \rho_t)\, \mathrm{d}\theta. \tag{2.5}$$

The free energy could be more general and depends also on $\nabla^k \rho_t$ for some $k \geq 1$.

In order to calculate the first variation of free energy $\mathcal{F}$ with respect to all the virtual displacements $\tilde{\theta}_s$ (another arbitrary flow map starting $s = t$ at ρ_t with virtual velocity $\tilde{v}_t$ on $T_{\rho_t}\mathcal{M}$ described below), we define the pushforward

density under the flow map of $\tilde{\theta}_s$ with flow velocity $\tilde{v}_s$

$$\tilde{\rho}_s = \tilde{\theta}_s \# \rho_t, \quad \tilde{\rho}_s|_{s=t} = \rho_t, \tag{2.6}$$

and the pushforward energy density is $\tilde{w}_s = w(\tilde{\rho}_s, \nabla\tilde{\rho}_s)$.

Assumption I: In the remaining parts of this note, we specialize ourselves to the domain $\mathbb{R}^d$. (The calculations of first variation for fixed domain Ω is same, which is equivalent to the constraint $n \cdot v_t = 0$ for $\theta \in \partial\Omega$.)

Under the Assumption I, we have

$$\frac{\mathrm{d}}{\mathrm{d}s}\big|_{s=t}\mathcal{F}(\tilde{\rho}_s) = \langle\frac{\delta\mathcal{F}}{\delta\rho_t}, \partial_t\tilde{\rho}_t\rangle \quad = \langle\frac{\delta\mathcal{F}}{\delta\rho_t}, -\nabla\cdot(\rho_t\tilde{v}_t)\rangle_\Omega = \langle\rho_t\nabla\frac{\delta\mathcal{F}}{\delta\rho_t},\ \tilde{v}_t\rangle_\Omega, \tag{2.7}$$

where $\frac{\delta\mathcal{F}}{\delta\rho_t}$ is the Fréchet derivative of $\mathcal{F}$. Choose KL-divergence free energy

$$\mathcal{F}(\rho) := \int \rho\ln(\rho/\rho_\infty)\,\mathrm{d}\theta. \tag{2.8}$$

Then $\nabla\frac{\delta\mathcal{F}}{\delta\rho_t} = \nabla(\ln\rho_t - \ln\rho_\infty)$ and

$$\rho_t\nabla\frac{\delta\mathcal{F}}{\delta\rho_t} = \rho_t\nabla(\ln\rho_t - \ln\rho_\infty) = \nabla\rho_t - \frac{\rho_t}{\rho_\infty}\nabla\rho_\infty = \nabla\rho_t + \rho_t\nabla U, \tag{2.9}$$

where $\rho_\infty = e^{-U}$.

2.2. The strong and weak formulation of gradient flow on manifold

To describe a gradient flow on manifold, there are in general three ways: strong formulation, weak formulation and De Giorgi type duality formulation. Strong formulation tells us the gradient flow holds either in tangent plane or in cotangent plane. Weak formulation is to use the virtual displacement as test function and find the optimal curve such that the free energy descents with respect to some specific metrics on tangent plane. Let us explain the strong formulation and weak formulation in detail below.

Let $A_\rho : T^*_\rho\mathcal{M} \to T_\rho\mathcal{M}$ be a symmetric nonnegative defined operator (reciprocal relation) depending on ρ from the cotangent plane to the tangent plane. Notice the Fréchet derivative (if exists) $\frac{\delta\mathcal{F}}{\delta\rho_t} \in T^*_{\rho_t}\mathcal{M}$ and $\partial_t\rho_t \in T_\rho\mathcal{M}$. We define a gradient flow in strong formulation (also known as the Onsager rate equation)

$$\partial_t\rho_t = -A_{\rho_t}\frac{\delta\mathcal{F}}{\delta\rho_t} \tag{2.10}$$

in the sense that equation holds in tangent plane $T_{\rho_t}\mathcal{M}$ for a.e. $t > 0$. Denote $A_\rho^\dagger$ as the generalized inverse of A_ρ, $A_\rho^\dagger : T_\rho\mathcal{M} \to T_\rho^*\mathcal{M}$. Then the gradient flow (2.10) can be recast as

$$A_{\rho_t}^\dagger \partial_t \rho_t = -\frac{\delta \mathcal{F}}{\delta \rho_t}. \tag{2.11}$$

To illustrate the motivation for the second and third formulations of gradient flow, define the functional $\psi(s)$ for any $s \in T_\rho\mathcal{M}$ (also known as Onsager dissipation potential [Mie16]),

$$\psi(s) = \frac{1}{2}\langle A_\rho^\dagger s, s\rangle_{\langle T^*, T\rangle}, \tag{2.12}$$

where $\langle\cdot,\cdot\rangle_{\langle T^*,T\rangle}$ means the dual pair in tangent plane $T_\rho\mathcal{M}$ and cotangent plane $T_\rho^*\mathcal{M}$. It is easy to check $\psi(s)$ is convex in s and its convex dual is the functional $\psi^*(\xi)$ for any $\xi \in T_\rho^*\mathcal{M}$ (also known as dual dissipation potential) such that

$$\psi^*(\xi) = \frac{1}{2}\langle \xi, A_\rho \xi\rangle_{\langle T^*,T\rangle}. \tag{2.13}$$

Then the strong formulation (2.10) and (2.11) imply the following identity

$$\dot{\mathcal{F}} = \langle \partial_t\rho, \frac{\delta\mathcal{F}}{\delta\rho}\rangle_{\langle T^*,T\rangle} = -2\psi(\partial_t\rho) = -2\psi^*(-\frac{\delta\mathcal{F}}{\delta\rho}) = -\psi(\partial_t\rho) - \psi^*(-\frac{\delta\mathcal{F}}{\delta\rho}). \tag{2.14}$$

When ψ is not a quadratic form defined in (2.12), the time integration of the identity

$$\dot{\mathcal{F}} = \langle \partial_t\rho, \frac{\delta\mathcal{F}}{\delta\rho}\rangle_{\langle T^*,T\rangle} = -\psi(\partial_t\rho) - \psi^*(-\frac{\delta\mathcal{F}}{\delta\rho}) \tag{2.15}$$

will lead to the definition of De Giorgi type duality formulation of gradient flow; see Section 4.1. The third formulation of gradient flow can be realized by Moreau-Yosida approximation, which inspires the construction of minimization movement and will be discussed in Section 4.1. By extending the quadratic form of ψ and its conjugate ψ^* to a general convex primitive functional, (2.15) is also called a generalized gradient flow in [RMS08]. We will discuss the generalized gradient flow formulation for MCMC in Section 4.2.1.

Next, as for the weak formulation of gradient flow, we use the virtual velocity on tangent plane as test function and define a Riemannian metrics

$g_\rho(\cdot,\cdot)$ on $T_\rho\mathcal{M}\times T_\rho\mathcal{M}$

$$g_\rho(s_1,s_2) := \langle A_\rho^\dagger s_1, s_2\rangle_{\langle T^*,T\rangle}, \quad s_1,s_2\in T_\rho\mathcal{M}. \tag{2.16}$$

The weak formulation of gradient flow with respect to g_ρ is

$$\frac{\mathrm{d}}{\mathrm{d}s}\Big|_{s=t}\mathcal{F}(\tilde\rho_s) = \langle\frac{\delta\mathcal{F}}{\delta\rho_t},\partial_t\tilde\rho_t\rangle = -g_\rho(\partial_t\rho_t,\partial_t\tilde\rho_t), \tag{2.17}$$

for any test function $\tilde\rho_s = \tilde\theta_s\#\rho_t$. Indeed, one can also define a Riemannian metrics $g_\rho^*(\cdot,\cdot)$ on $T_\rho^*\mathcal{M}\times T_\rho^*\mathcal{M}$ such that

$$g_\rho^*(\xi_1,\xi_2) := \langle\xi_1, A_\rho\xi_2\rangle_{\langle T^*,T\rangle}, \quad \xi_1,\xi_2\in T_\rho^*\mathcal{M}. \tag{2.18}$$

Although the weak formulation with respect to g_ρ^* is equivalent to (2.17), it is usually not used in this way. We will explain several gradient flow schemes using the weak formulation (2.17) in the next section.

3. Several gradient flow acceleration schemes via different metrics

3.1. Example I: JKO weighted H^{-1} metrics

The tangent plane at ρ_t is uniquely determined by the admissible velocity field $\tilde v_t$ of the flow map

$$T_{\rho_t}\mathcal{M} = \{-\nabla\cdot(\rho_t\tilde v_t);\ \text{for } \tilde v_t\in L^2(\Omega;\rho_t\,\mathrm{d}\theta)\}. \tag{3.1}$$

If we take the metric g_{ρ_t} as weighted-H^{-1} inner product

$$g_{\rho_t}(-\nabla\cdot(\rho_t v_t),-\nabla\cdot(\rho_t\tilde v_t)) := \langle\rho_t v_t,\tilde v_t\rangle_\Omega, \tag{3.2}$$

which can be regarded as the metrics induced by the operator $A_\rho^\dagger = -(\nabla\cdot(\rho\nabla))^{-1}$. Then the gradient flow of $\mathcal{F}$ with respect to the metric g_{ρ_t} is

$$\frac{\mathrm{d}}{\mathrm{d}s}\Big|_{s=t}\mathcal{F}(\tilde\rho_s) = -\langle\rho_t v_t,\tilde v_t\rangle_\Omega. \tag{3.3}$$

Therefore we obtain the steepest descent velocity for the gradient flow is

$$v_t = -\nabla\frac{\delta\mathcal{F}}{\delta\rho_t} \tag{3.4}$$

and the governing equation with (2.8) is

$$\partial_t \rho_t = \nabla \cdot (\rho_t \nabla \frac{\delta \mathcal{F}}{\delta \rho_t}) = \nabla \cdot (\nabla \rho_t + \rho_t \nabla U). \tag{3.5}$$

This is the Fokker–Planck equation of the OU process

$$\mathrm{d}\theta = -\nabla U(\theta)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t. \tag{3.6}$$

A time discretization gives a MCMC, which is an efficient way to update ρ_{ps} when the dimension of parameter space is very high. Below we use measure/distribution μ_{ap}^N and μ_{ps}^N instead of ρ_{ap} and ρ_{ps}. Indeed, (i) based on the law of large number, we can use N-samplers as the initial date of the flow map (2.1) to approximate $\mu_{\mathrm{ap}}^N \approx \frac{1}{N}\sum_{i=1}^N \delta_{\theta_i(0)}$; (ii) at each time t, the MCMC updates the random map $\theta_i(t)$ and thus μ_t can be approximated by $\mu_t^N \approx \frac{1}{N}\sum_{i=1}^N \delta_{\theta_i(t)}$ (which can be understood as the pushforward measure $\mu_t^N \approx \theta_t \# \mu_{\mathrm{ap}}^N$); (iii) by the ergodicity of OU process with a convex U, μ_t^N will converge to its equilibrium μ_{ps}^N as $t \to +\infty$.

3.2. Beyond example I: dynamics on parameter space and Kinetic theory

Assume $\rho_0 = N(\theta; \mu_0, \sigma_0^2)$, then the special solution to the Fokker-Planck equation for OU process (3.5) with harmonic trap potential $U(\theta) = \frac{1}{2}\left(\frac{\theta-\mu_\infty}{\sigma_\infty}\right)^2$ is given by $\rho_t = N(\theta; \mu_t, \sigma_t^2)$. Following the convention, here we use μ as the mean instead of measure. One can check the parameters (μ_t, σ_t) satisfy the ODE

$$\partial_t \mu(t) = \frac{\mu_\infty - \mu(t)}{\sigma_\infty^2}, \quad \partial_t \sigma(t) = \frac{\sigma_\infty^2 - \sigma^2(t)}{\sigma(t)\sigma_\infty^2}. \tag{3.7}$$

Following the notations above, we assume $\rho_t(\theta) = N(\theta; \mu(t), \sigma^2(t))$ is a curve on manifold $\mathcal{M}$, which is uniquely described by two parameters $\mu(t), \sigma^2(t)$. Recall (3.5) with KL free energy (2.8)

$$\mathcal{F}(\rho) := \int \rho(\theta) \ln(\rho(\theta)/\rho_\infty(\theta))\,\mathrm{d}\theta.$$

The velocity of the underlying flow map for $\rho_t = \theta(t)\#\rho_0$ is given by

$$\dot{\theta}(t) = -\nabla\left(\ln \frac{\rho_t}{\rho_\infty}\right) = \frac{\theta(t)-\mu(t)}{\sigma(t)^2} - \frac{\theta(t)-\mu_\infty}{\sigma_\infty^2}. \tag{3.8}$$

Combining (3.7) and (3.8), we can obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\theta(t)-\mu(t)}{\sigma(t)}=0. \tag{3.9}$$

Next we show the parameters (μ_t,σ_t), as new configurations on $\mathbb{R}^2$, possess their owe gradient flow structure with natural metrics on $\mathbb{R}^2$. Recall the equilibrium distribution satisfying $N(\theta;\mu_\infty,\sigma_\infty^2)$ is $\rho_\infty(\theta)=\frac{1}{\sqrt{2\pi\sigma_\infty^2}}e^{-\frac{(\theta-\mu_\infty)^2}{2\sigma_\infty^2}}$. Then for the special KL free energy, by the central moments formula, we have

$$\begin{aligned}\mathcal{F}(\rho)=&\mathbb{E}\left(\ln\frac{\sigma_\infty}{\sigma(t)}+\frac{(\theta-\mu_\infty)^2}{2\sigma_\infty^2}-\frac{(\theta-\mu(t))^2}{2\sigma^2(t)}\right)\\ =&\ln\frac{\sigma_\infty}{\sigma(t)}+\frac{(\mu(t)-\mu_\infty)^2}{2\sigma_\infty^2}+\frac{\sigma^2(t)}{2\sigma_\infty^2}-\frac{1}{2},\end{aligned} \tag{3.10}$$

where expectation is with respect to ρ_t. We can calculate the first variation of $\mathcal{F}$

$$\frac{\mathrm{d}}{\mathrm{d}s}\Big|_{s=t}\mathcal{F}(\tilde{\rho}_s)=-\frac{\partial_t\tilde{\sigma}(t)}{\sigma(t)}+\frac{(\mu(t)-\mu_\infty)\partial_t\tilde{\mu}(t)}{\sigma_\infty^2}+\frac{\sigma(t)\partial_t\tilde{\sigma}(t)}{\sigma_\infty^2}.$$

Define the metrics as

$$g_{\rho(t)}(\partial_t\rho_t,\partial_t\tilde{\rho}_t):=\partial_t\mu(t)\partial_t\tilde{\mu}(t)+\partial_t\sigma(t)\partial_t\tilde{\sigma}(t).$$

Then

$$\frac{\mathrm{d}}{\mathrm{d}s}\Big|_{s=t}\mathcal{F}(\tilde{\rho}_s)=-g_{\rho(t)}(\partial_t\rho_t,\partial_t\tilde{\rho}_t) \tag{3.11}$$

gives the ODEs

$$\partial_t\mu(t)=\frac{\mu_\infty-\mu(t)}{\sigma_\infty^2},\quad \partial_t\sigma(t)=\frac{\sigma_\infty^2-\sigma^2(t)}{\sigma(t)\sigma_\infty^2}. \tag{3.12}$$

Then we can solve the solution to the ODEs

$$\begin{aligned}\mu(t)&=\mu_\infty-(\mu_\infty-\mu_0)e^{-\frac{t}{\sigma_\infty^2}},\\ \sigma^2(t)&=\sigma_\infty^2-(\sigma_\infty^2-\sigma_0^2)e^{-\frac{2t}{\sigma_\infty}},\end{aligned} \tag{3.13}$$

where μ_0,σ_0^2 correspond to initial expectation and variance.

The spirit is analogue to kinetic theory for gas dynamic. Assume $f_\varepsilon(x, v, t)$ is the solution to the Boltzmann equation with mean free path ε. In the limit of $\varepsilon \to 0$, the density f_ε converges to a equilibrium density, physically known as Maxwellian distribution $\rho_t(x)N(v; u_t(x), T_t(x)I)$, where $\rho_t(x)$ is the density, $u_t(x)$ is the velocity, $T_t(x)$ is the temperature at position $x \in \mathbb{R}^3$, time t, and $N(v; u_t(x), T_t(x)I)$ is the normal distribution of v with mean $u_t(x)$ and covariance matrix $T_t(x)I$. The continuum variables $\rho_t(x), u_t(x), T_t(x)$ satisfy the compressible Euler equation for gas dynamics. The spirit that reducing equations for $f_\varepsilon(x, v, t)$ in $\mathbb{R}^7$ to equations for $(\rho_t(x), u_t(x), T_t(x))$ in $\mathbb{R}^4$ is same as reducing the Fokker-Plank equation of $\rho(\theta; \mu_t, \sigma_t)$ to an ODE of (μ_t, σ_t) presented above.

3.3. Example II: ensemble Kalman filter (EnKF) via modified JKO metrics

In this section, we summarize recent progresses for inverse problem using the continuous time ensemble Kalman filter (EnKF) studied in [SS17], the corresponding mean field limit [HV19] and the interpretation via a gradient flow with respect to a weighted Wasserstein distance [GIHLS19].

Assume a physical system for x can be described by an operator G acting on the unknown parameter θ. The operator is usually known but nonlinear and complicated, such as a PDE solver for weather prediction or optical tomography. The inverse problem we are interested in is to learn the unknown parameter θ from observation x, which is also known as data assimilation. Due to the observational noise η, which is assumed to be given by a Gaussian process, the physical system satisfies

$$x = G(\theta) + \eta. \tag{3.14}$$

Assume the covariance of the noise Γ^{-1} is known. Then given θ, the distribution of x satisfies $f(x|\theta) = N(G(\theta), \Gamma^{-1}) \propto e^{-\frac{1}{2}\langle (x-G(\theta)), \Gamma(x-G(\theta))\rangle}$ with mean $G(\theta)$ and covariance matrix Γ^{-1}. This likelihood function also means the lower cost $U(\theta) := \frac{1}{2}\langle (x - G(\theta)), \Gamma(x - G(\theta))\rangle$ has higher probability. Following the Bayesian formulation, given any priori density $\rho_{\text{ap}}(\theta)$, the posterier density of θ is

$$\rho_{\text{ps}}(\theta) \propto N(G(\theta), \Gamma^{-1}) \propto e^{-\frac{1}{2}\langle (x-G(\theta)), \Gamma(x-G(\theta))\rangle} = e^{-U(\theta;x)}. \tag{3.15}$$

Let J be the number of ensembles. Define the operator

$$M(\theta) := \frac{1}{J}\sum_{k=1}^{J}(\theta^k - \bar{\theta}) \otimes (\theta^k - \bar{\theta}),$$

which is positive semi-defined. To minimize the data-model misfit U, denote $\bar{\theta}$ as the mean of θ^j and $\bar{G}$ as the mean of $G(\theta^j)$ for $j = 1, \cdots, J$. The continuous time ensemble Kalman filter(EnKF) iteration is given by [SS17]

$$\mathrm{d}\theta^j = \frac{1}{J}\sum_{k=1}^{J}\langle G(\theta^k) - \bar{G}, \Gamma(x - G(\theta^j))\rangle(\theta^k - \bar{\theta})\,\mathrm{d}t + \sqrt{2M(\theta)}\,\mathrm{d}W^j. \tag{3.16}$$

For the linear case $G(\theta) = G\theta$, (3.16) can be recast as

$$\begin{aligned}\mathrm{d}\theta^j &= \frac{1}{J}\sum_{k=1}^{J}(\theta^k - \bar{\theta}) \otimes (\theta^k - \bar{\theta})G^T\Gamma(x - G\theta^j)\,\mathrm{d}t + \sqrt{2M(\theta)}\,\mathrm{d}W^j \\ &= -M(\theta)\nabla_\theta U(\theta^j; x)\,\mathrm{d}t + \sqrt{2M(\theta)}\,\mathrm{d}W^j\end{aligned} \tag{3.17}$$

where $\nabla_{\theta^j} U(\theta; x) = -G^T\Gamma(x - G\theta^j)$ in the linear case. Then the continuous time EnKF can be rewritten as a weighted OU process

$$\mathrm{d}\theta^j = -M(\theta)\nabla_\theta U(\theta^j; x)\,\mathrm{d}t + \sqrt{2M(\theta)}\,\mathrm{d}W^j, \tag{3.18}$$

where $M(\theta)$ is usually called acceleration matrix in Langevin dynamics. We can write down formally the corresponding Fokker Plank equation

$$\partial_t \rho_t = M(t) : \nabla^2\rho_t + \nabla\cdot(\rho_t M(t)\nabla U) = \nabla\cdot(\rho_t M(t)\nabla(\ln\rho_t + U(\theta)))\,, \tag{3.19}$$

where $M(t) := \int(\theta - \mathbb{E}(\theta)) \otimes (\theta - \mathbb{E}(\theta))\rho_t(\theta, t)\,\mathrm{d}\theta$ is a matrix depending only on time variable. Using the special KL free energy (2.8) $\mathcal{F}(\rho) = \int \rho \ln(\rho/\rho_\infty)\,\mathrm{d}\theta$. The rigorous mean field limit is studied by [HV19]. Notice (2.9) then (3.19) is recast to

$$\partial_t \rho_t = \nabla \cdot \left(\rho_t M(t)\nabla\frac{\delta\mathcal{F}}{\delta\rho_t}\right). \tag{3.20}$$

To rewrite it as a weak formulation of gradient flow, we take the Riemannian metrics g_{ρ_t} as (weighted-JKO metrics)

$$g_{\rho_t}(-\nabla\cdot(\rho_t v_t), -\nabla\cdot(\rho_t\tilde{v}_t)) := \langle M(t)^{-1}\rho_t v_t, \tilde{v}_t\rangle_\Omega, \tag{3.21}$$

then the gradient flow of $\mathcal{F}$ with respect to the metric g_{ρ_t} is

$$\frac{\mathrm{d}}{\mathrm{d}s}\Big|_{s=t}\mathcal{F}(\tilde{\rho}_s) = \langle\frac{\delta\mathcal{F}}{\delta\rho_t}, \partial_t\tilde{\rho}_t\rangle = -\langle M(t)^{-1}\rho_t v_t, \tilde{v}_t\rangle_\Omega. \tag{3.22}$$

Therefore we obtain the steepest descent velocity for the gradient flow is

$$v_t = -M(t)\nabla\frac{\delta\mathcal{F}}{\delta\rho_t} \tag{3.23}$$

and the governing equation is

$$\partial_t\rho_t = \nabla\cdot(\rho_t M(t)\nabla\frac{\delta\mathcal{F}}{\delta\rho_t}) = \nabla\cdot(\rho_t M(t)\nabla(\ln\rho_t + U(\theta)))\,. \tag{3.24}$$

The implementation of this acceleration algorithm is performed using weighted Wasserstein distance in [GIHLS19]; see more details in Section 4.1.2

3.4. Example III: Stein variational gradient descent (SVGD)

Recall the tangent plane at ρ_t is uniquely determined by the admissible velocity field $\tilde{v}_t$ of the flow map. SVGD was first introduced by [LW16, Liu17]. The idea of SVGD is to construct an efficient dynamics system by introducing a reproducing kernel K (satisfying reproducing property $\langle K(\cdot,x), K(\cdot,y)\rangle = K(x,y)$) such that the admissible velocity is chosen in reproducing kernel Hilbert space (RKHS), i.e.

$$T_{\rho_t}\mathcal{M} = \{-\nabla\cdot(\rho_t\tilde{v}_t);\ \tilde{v}_t = K*\psi := \int K(\theta,\eta)\psi(\eta)\,\mathrm{d}\eta \text{ for some } \psi\in\mathbb{R}^d\}. \tag{3.25}$$

For any $-\nabla\cdot(\rho_t u_t), -\nabla\cdot(\rho_t\tilde{v}_t) \in T_{\rho_t}\mathcal{M}$, there exist ψ^*, $\psi\in\mathbb{R}^d$ such that

$$u_t = K*\psi^* := \int K(\theta,\eta)\psi^*(\eta)\,\mathrm{d}\eta, \quad \tilde{v}_t = K*\psi := \int K(\theta,\eta)\psi(\eta)\,\mathrm{d}\eta. \tag{3.26}$$

So we define the metric g_{ρ_t} as

$$g_{\rho_t}(-\nabla\cdot(\rho_t u_t), -\nabla\cdot(\rho_t\tilde{v}_t)) := \int\int K(\theta,\eta)\psi^*(\theta)\cdot\psi(\eta)\,\mathrm{d}\eta\,\mathrm{d}\theta = \langle\tilde{v}_t,\psi^*\rangle. \tag{3.27}$$

Therefore combing (3.3) and (2.7) we can determine the velocity and equation together, i.e.

$$u_t = K*\psi^* = K*(-\rho_t\nabla\frac{\partial\mathcal{F}}{\partial\rho_t}) \tag{3.28}$$

and

$$\partial_t\rho_t = -\nabla\cdot(\rho_t u_t) = \nabla\cdot\left(\rho_t K*(\rho_t\nabla\frac{\partial\mathcal{F}}{\partial\rho_t})\right) \tag{3.29}$$

SVGD with specific $\mathcal{F}$. Using the special KL free energy (2.8) $\mathcal{F}(\rho) = \int \rho \ln(\rho/\rho_\infty)\,\mathrm{d}\theta$ and notice (2.9). Therefore we have

$$\partial_t \rho_t = -\nabla \cdot (\rho_t u_t) = \nabla \cdot (\rho_t K * (\rho_t \nabla U + \nabla \rho_t)). \tag{3.30}$$

Constructed interacting particle system. Let $\frac{1}{N}\sum_{i=1}^N \delta_{\theta_i(t)} := \mu_t^N$. Then the ODE system for (3.30) for $i = 1, \cdots, N$ is

$$\dot{\theta}_i = \frac{1}{N}\sum_{j=1}^N \nabla_y K(\theta_i, \theta_j) - \frac{1}{N}\sum_{j=1}^N K(\theta_i, \theta_j)\nabla U(\theta_j). \tag{3.31}$$

Using the same arguments as MCMC for JKO scheme, $\mu_t^N \approx \theta_t \# \mu_{\mathrm{ap}}^N$ will converge to its equilibrium μ_{ps}^N as $t \to +\infty$. We refer to [LLN19] for the convergence of the interacting particle system to its mean field limit.

3.5. Example IV: train a two-layer neural network using parametric Bayesian inference

In this section, we summarize the idea of training target function in neural network using parametric Bayesian inference [RVE19], see also [MMN18].

Given any function $g : \Omega \subset \mathbb{R}^d \to \mathbb{R}$ as the target function to be learned, consider the neural network approximation

$$g^{(n)}(x) = \frac{1}{n}\sum_{i=1}^n c_i \hat{\varphi}(x, z_i) = \frac{1}{n}\sum_{i=1}^n \varphi(x, \theta_i), \tag{3.32}$$

where $\hat{\varphi}$, usually referred as the "activation function", is given and the parameters (c_i, z_i) are to be learned. For notation simplicity, we rewrite the parameters as θ_i using function φ. Denote $\theta = \{\theta_i\}_{i=1}^n$.

Assume we have data $\{x_j\}_{j=1}^J$ and each $x_j \in \mathbb{R}^d$ with the data distribution $\nu(\mathrm{d}x)$ and the label $y_j = g(x_j)$. The goal is still using parametric Bayesian inference to update parameters θ by regarding θ as random variables with the density $\rho(\theta)$.

To measure the discrepancy between the target function g and the neural network approximation $g^{(n)}$, a natural cost function is

$$\begin{aligned} L(\theta) :=& \frac{1}{2}\mathbb{E}_\nu(|g - g^{(n)}|^2) = \frac{1}{2}\mathbb{E}_\nu(|g|^2 - 2gg^{(n)} + |g^{(n)}|^2) \\ =& \frac{1}{2}\mathbb{E}_\nu g^2 - \frac{1}{n}\sum_{i=1}^n \mathbb{E}_\nu(g(x)\varphi(x, \theta_i)) + \frac{1}{2n^2}\sum_{i,j=1}^n \mathbb{E}_\nu(\varphi(x, \theta_i)\varphi(x, \theta_j)) \end{aligned} \tag{3.33}$$

$$=:C_g - \frac{1}{n}\sum_{i=1}^{n} F(\theta_i) + \frac{1}{2n^2}\sum_{i,j=1}^{n} K(\theta_i,\theta_j),$$

where $F(\theta_i) := \mathbb{E}(g(x)\varphi(x,\theta_i))$ and $K(\theta_i,\theta_j) := \mathbb{E}_\nu(\varphi(x,\theta_i)\varphi(x,\theta_j))$.

Given data x and its label $g(x)$, $L(\theta)$ is a function of θ and we can regard $f(x|\theta) = e^{-L(\theta)}$ as the likelihood distribution. Then given any a priori density $\rho_{\mathrm{ap}}(\theta)$, following the Bayesian formulation, the posterier density of θ is given by

$$\rho_{\mathrm{ps}}(\theta) \propto f(x|\theta) = e^{-L(\theta)}, \tag{3.34}$$

which means the lower cost corresponds to higher probability. To drive $\rho_{\mathrm{ap}}(\theta)$ to this $\rho_{\mathrm{ps}}(\theta)$, one can use any gradient flow acceleration introduced above in Section 3. Alternatively, one can consider directly the SDE (interaction particle system) for θ

$$\mathrm{d}\theta = -\nabla L(\theta)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t \tag{3.35}$$

and construct proper MCMC schemes for it, which is discussed in [RVE19], see also [MMN18].

4. The De Giorgi (ψ, ψ^*)-type formulation of gradient flow

In this section, we will discuss the third formulation of gradient flow, De Giorgi type duality formulation. The (p,q)-gradient flow, as a generalization of $(2,2)$-gradient flow, can also be realized by Moreau-Yosida approximation and will be discussed in Section 4.1. The De Giorgi (ψ,ψ^*) formulation of gradient flow, as a further generalization of the (p,q)-gradient flow will be discussed in Section 4.2.

4.1. (p,q)-gradient flow via minimizing movement

Given the free energy $\mathcal{F}(u)$ and some general distance $d(u,v)$ in the ambient space X, define the Moreau-Yosida approximation

$$e(t) := \min_{u\in X}\{\mathcal{F}(u) + \frac{1}{pt^{p-1}}d^p(u,\chi)\}, \tag{4.1}$$

where χ is any given function. Assume there exists a u_t such that $u_t = \operatorname{argmin}\{\mathcal{F}(u) + \frac{1}{pt^{p-1}}d^p(u,\chi)\}$. For the case u_t is unique, which is called proximal point, the map $\chi \to u_t$ is known as proximal map. For proximal point u_t, we have

$$e(t) = \mathcal{F}(u_t) + \frac{1}{pt^{p-1}}d^p(u_t,\chi) \tag{4.2}$$

and

$$\begin{aligned} e(t) \leq & \mathcal{F}(u_s) + \frac{1}{p}\left(\frac{1}{t^{p-1}} - \frac{1}{s^{p-1}}\right) d^p(u_s, \chi) + \frac{1}{ps^{p-1}} d^p(u_s, \chi) \\ = & e(s) + \frac{s^{p-1} - t^{p-1}}{pt^{p-1}s^{p-1}} d^p(u_s, \chi). \end{aligned}$$

Exchanging s and t gives us for any $0 < s < t$,

$$\frac{d^p(u_s, \chi)}{pt^{p-1}s^{p-1}} \frac{t^{p-1} - s^{p-1}}{t - s} \leq \frac{e(s) - e(t)}{t - s} \leq \frac{d^p(u_t, \chi)}{pt^{p-1}s^{p-1}} \frac{t^{p-1} - s^{p-1}}{t - s}. \tag{4.3}$$

This implies the increase of distance $d(u_t, \chi)$ and the decay rate of $e(t)$. Indeed, taking $s \to t$, we have for any t,

$$e(t) + \frac{1}{q} \int_0^t \frac{d^p(u_s, \chi)}{s^p} \, \mathrm{d}s = e(0) \leq \mathcal{F}(\chi), \tag{4.4}$$

which is

$$\mathcal{F}(u_t) + \frac{1}{pt^{p-1}} d^p(u_t, \chi) + \frac{1}{q} \int_0^t \frac{1}{s^p} d^p(u_s, \chi) \, \mathrm{d}s \leq \mathcal{F}(\chi). \tag{4.5}$$

Now we claim the last term on the left hand side can control the local slope of the free energy

$$|\partial \mathcal{F}(u_s)| := \limsup_{v \to u_s} \frac{(\mathcal{F}(u_s) - \mathcal{F}(v))^+}{d(u_s, v)}. \tag{4.6}$$

In fact, from

$$\mathcal{F}(u_t) + \frac{1}{pt^{p-1}} d^p(u_t, \chi) \leq \mathcal{F}(v) + \frac{1}{pt^{p-1}} d^p(v, \chi),$$

we know

$$\begin{aligned} \frac{(\mathcal{F}(u_t) - \mathcal{F}(v))^+}{d(u_t, v)} \leq & \frac{1}{t^{p-1}} \left(\frac{d^p(v, \chi) - d^p(u_t, \chi)}{p(d(v, \chi) - d(u_t, \chi))^+} \frac{(d(v, \chi) - d(u_t, \chi))^+}{d(u_t, v)} \right) \\ \leq & \frac{1}{t^{p-1}} \frac{d^p(v, \chi) - d^p(u_t, \chi)}{p(d(v, \chi) - d(u_t, \chi))^+}. \end{aligned}$$

Taking limit we obtain

$$|\partial \mathcal{F}(u_s)|^q \leq \frac{d^p(u_s, \chi)}{s^p} \quad \text{for any } s < t. \tag{4.7}$$

Therefore, if u_t is the minimizer of Moreau-Yosida approximation, then (4.5) implies

$$\mathcal{F}(u_t) + \frac{1}{pt^{p-1}} d^p(u_t, \chi) + \frac{1}{q}\int_0^t |\partial\mathcal{F}(u_s)|^q \,\mathrm{d}s \le \mathcal{F}(\chi). \tag{4.8}$$

4.1.1. Minimizing movement. Based on the estimate (4.8) for the minimizer of Moreau-Yosida approximation, the backward Euler scheme is designed as

$$u^{n+1} = \mathrm{argmin}_u\{\mathcal{F}(u) + \frac{1}{p(\Delta t)^{p-1}} d^p(u, u^n)\}. \tag{4.9}$$

Denote $t^n = n\Delta t$. Then (4.8) implies

$$\mathcal{F}(u^{n+1}) + \frac{1}{p(\Delta t)^{p-1}} d^p(u^{n+1}, u^n) + \frac{1}{q}\int_{t^n}^{t^{n+1}} |\partial\mathcal{F}(u_s)|^q \,\mathrm{d}s \le \mathcal{F}(u^n). \tag{4.10}$$

By telescoping summation, we have

$$\mathcal{F}(u^{n+1}) + \frac{1}{p}\sum_{k=0}^{n}\left(\frac{d(u^{k+1}, u^k)}{\Delta t}\right)^p \Delta t + \frac{1}{q}\int_0^{t^{n+1}} |\partial\mathcal{F}(u_s)|^q \,\mathrm{d}s \le \mathcal{F}(u_0), \tag{4.11}$$

Assume $u_t(t^n)$ can be approximated by u^n, then we have

$$\mathcal{F}(u_t) + \frac{1}{p}\int_0^t |\dot{u}(s)|^p \,\mathrm{d}s + \frac{1}{q}\int_0^t |\partial\mathcal{F}(u_s)|^q \,\mathrm{d}s \le \mathcal{F}(u_0). \tag{4.12}$$

Alternatively from (4.10) we have

$$\frac{\mathcal{F}(u^{n+1}) - \mathcal{F}(u^n)}{\Delta t} + \frac{1}{p}\left(\frac{d(u^{n+1}, u^n)}{\Delta t}\right)^p + \frac{1}{q}\frac{1}{\Delta t}\int_{t^n}^{t^{n+1}} |\partial\mathcal{F}(u_s)|^q \,\mathrm{d}s \le 0. \tag{4.13}$$

This formally gives the duality (p, q)-gradient flow

$$\dot{\mathcal{F}}(u_t) + \frac{1}{p}|\dot{u}_t|^p + \frac{1}{q}|\partial\mathcal{F}(u_t)|^q \le 0. \tag{4.14}$$

If we have enough assumptions on the free energy $\mathcal{F}$, the proper interpolation of the discrete solution u^n gives a locally absolutely continuous curve u_t

starting from u_0 and the inequality (4.12) can be improved to equality, called Energy Dissipation Equality(EDE) definition of gradient flow [AGS08]

$$\mathcal{F}(u_t) + \frac{1}{p}\int_0^t |\dot{u}(s)|^p \,\mathrm{d}s + \frac{1}{q}\int_0^t |\partial\mathcal{F}(u_s)|^q \,\mathrm{d}s = \mathcal{F}(u_0). \tag{4.15}$$

From the discrete solution via backward Euler scheme, one may also obtain a strong version of gradient flow for $p = 2$, called Evolution Variation Inequality (EVI) solution, i.e. a locally absolutely continuous curve u_t starting from u_0 such that

$$\mathcal{F}(u_t) + \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}d^2(u_t, v) \le \mathcal{F}(v), \quad \forall v \in X, a.e. t > 0. \tag{4.16}$$

We refer to [AGS08] for explicit assumptions on the convexity and compatibility in free energy and metrics.

4.1.2. Application to JKO/EnKF and subdifferential with respect to Wasserstein distance. In Section 3, we have recast the JKO scheme and EnKF as gradient flow with respect to some certain Riemannian metrics. To achieve the solution ρ_t of these gradient flow via minimizing movement, an important feature is the Riemannian metrics for JKO/EnKF correspond to a Wasserstein distance. By Benamou-Brenier formula, given $\mu^0, \mu^1 \in \mathcal{P}_2(\Omega)$ and $\mu^0 \ll \mathcal{L}(\mathbb{R}^d)$ absolutely continuous with respect to the Lebesgue measure, the Wasserstein distance is

$$\begin{aligned} W_2(\mu^0,\mu^1)^2 &= \min_{\rho_0=\mu^0,\,\rho_1=\mu^1,\,\partial_t\rho_t+\nabla\cdot(\rho_t v_t)=0} \Big\{\int_0^1 \|v_t\|^2_{L^2(\rho_t)}\,\mathrm{d}t\Big\} \\ &= \min_{\rho_0=\mu^0,\,\rho_1=\mu^1,\,\partial_t\rho_t+\nabla\cdot(\rho_t v_t)=0} \big\{\langle v_t, v_t\rangle_{L^2(\rho_t)}\big\} \\ &= \min_{\tilde{T},\tilde{T}\#\nu^0=\nu^1} \int |\tilde{T}z - z|^2 \,\mathrm{d}\mu^0(z) \\ &= \int |T_0^1 z - z|^2 \,\mathrm{d}\mu^0(z), \end{aligned} \tag{4.17}$$

where T_0^1 is the optimal transport map from μ^0 to μ^1. Let $x_t(z) = [(1-t)I + tT_0^1]z$ be the optimal flow map induced by the optimal transport map T_0^1. Then the constant speed geodesic $\rho_t = x_t\#\rho_0$ induced by the optimal transport map T_0^1 is the W_2-geodesic curve. This ρ_t and the corresponding velocity field v_t of the optimal flow map give the optimal pair (ρ_t, v_t) in Benamou-Brenier formula.

Similarly, the EnKF-Wasserstein distance is introduced in [GIHLS19] as an implement of the gradient flow for EnKF

(4.18)
$$W_2(\mu^0,\mu^1)^2 = \min_{\rho_0=\mu^0,\,\rho_1=\mu^1,\,\partial_t\rho_t+\nabla\cdot(\rho_t v_t)=0}\{\int_0^1 \langle M(t)^{-1}v_t, v_t\rangle_{L^2(\rho_t)}\,\mathrm{d}t\}$$
$$= \min_{\rho_0=\mu^0,\,\rho_1=\mu^1,\,\partial_t\rho_t+\nabla\cdot(\rho_t v_t)=0}\{\langle M(t)^{-1}v_t, v_t\rangle_{L^2(\rho_t)}\}.$$

With two distance defined above in $\mathcal{P}(\Omega)$, the minimizing movement gives the gradient flow (3.5) and (3.20) respectively in EDE/EVI formulation. Especially, using the optimal map T_μ^ν from $\mu \ll \mathcal{L}(\mathbb{R}^d)$ to ν, one can define the subdifferential $\partial_w\mathcal{F}(\mu) \subset \mathrm{T}an_{\mu_t}(\mathcal{P}_2(\mathbb{R}^d))$ of F in $(\mathcal{P}_2, W_2)$ as the set of vector fields $v \in L^2(\mu,\mathbb{R}^d)$ such that

$$\mathcal{F}(\mu) + \int \langle T_\mu^\nu - I, v\rangle\,\mathrm{d}\mu \le \mathcal{F}(\nu), \quad \forall \nu \in \mathcal{P}_2. \tag{4.19}$$

We refer to [AGS08] for the detailed assumptions on $\mathcal{F}$. Then we obtain the EVI formulation of gradient flow (3.5) is equivalent to

$$v_t \in -\partial_w\mathcal{F}(\rho_t), \quad \partial_t\rho_t + \nabla\cdot(v_t\rho_t) = 0, \tag{4.20}$$

for a.e. $t > 0$.

4.2. Generalized De Giorgi (ψ, ψ^*)-type formulation of gradient flow

Recall the identity (2.15). When ψ is not a quadratic form, a generalized gradient flow formulation is introduced originally by [DGT80]. Assume the dissipation potential ψ is a convex functional defined on tangent plane $T_\rho\mathcal{M}$ and the dual dissipation potential ψ^* is the convex dual of ψ defined on cotangent plane $T_\rho^*\mathcal{M}$. If $\min_s \psi(s) = 0 = \psi(0)$, then the curve ρ_t satisfies

$$\langle \partial_t\rho, \frac{\delta\mathcal{F}}{\delta\rho}\rangle_{\langle T^*,T\rangle} + \psi(\partial_t\rho) + \psi^*(-\frac{\delta\mathcal{F}}{\delta\rho}) = 0 \tag{4.21}$$

is called the generalized De Giorgi (ψ,ψ^*)-type gradient flow. The (p,q)-gradient flow obtained via minimizing movement corresponding to the special case $\psi(s) = \frac{1}{p}|s|^p$ and $\psi^*(\xi) = \frac{1}{q}|\xi|^q$. We refer to [Mie16] for the evolutionary Γ-convergence of those generalized gradient flows. We refer to [Agu12, SR15] for gradient flows on Finsler manifolds.

4.2.1. Example V: generalized gradient flow for MCMC. We can also design dynamic systems for efficient sampling by directly constructing MCMC on parameter space. The master equation for a Markov chain with the transition probability P_{ji} (satisfies $P_{ji} \geq 0$ and $\sum_j P_{ji} = 1$ for all i) is given by

$$\frac{d}{dt}\rho_i(t) = \sum_j P_{ij}\rho_j(t) - \rho_i(t)\,, \tag{4.22}$$

where $\rho_i = \mathbb{P}(\theta = \theta_i)$ is the probability for $\theta = \theta_i$. To sample the posterior density $\{\pi_i\}$, one can use some standard MCMC schemes with transition probability satisfying detailed balance

$$P_{ij}\pi_j = P_{ji}\pi_i, \text{ for all } i, j\,; \tag{4.23}$$

for example, Metropolis-Hastings algorithm, Metropolis adjusted langevin algorithm (MALA), Gibbs sampler algorithm. Then we have the following H-theorem for the Markov chain with the transition probability P_{ji}

$$\begin{aligned}\frac{d}{dt}F(\rho) &= \frac{d}{dt}\sum_i \rho_i(t)\ln\frac{\rho_i(t)}{\pi_i} \\ &= -\frac{1}{2}\sum_{i,j} P_{ij}\pi_j\left(\frac{\rho_i}{\pi_i} - \frac{\rho_j}{\pi_j}\right)\left(\ln\frac{\rho_i}{\pi_i} - \ln\frac{\rho_j}{\pi_j}\right) \leq 0.\end{aligned}$$

Consider the free energy for KL divergence $\mathcal{F}(\rho) := KL(\rho|\pi)$ with $(\frac{\delta\mathcal{F}}{\delta\rho})_i = \log\frac{\rho_i}{\pi_i} + 1$. To reformulate (4.22) as a De Giorgi (ψ, ψ^*)-type gradient flow. Using $\sum_j P_{ji} = 1$ and detailed balance, we rewrite (4.22) as

$$\frac{d}{dt}\rho_i(t) = \sum_j P_{ij}\rho_j(t) - \rho_i(t) = \sum_j P_{ji}\pi_i\left(\frac{\rho_j}{\pi_j} - \frac{\rho_i}{\pi_i}\right).$$

[Maa11] used the logarithmic mean of a, b as $\Lambda(a, b) = \frac{a-b}{\log a - \log b} \geq 0$ to reformulate (4.22) as a generalized gradient flow. Indeed, denote $a_{ij} := P_{ji}\pi_i\Lambda(\frac{\rho_i}{\pi_i}, \frac{\rho_j}{\pi_j}) \geq 0$ and we have $a_{ij} = a_{ji}$ and

$$\frac{d}{dt}\rho_i = -\sum_j a_{ij}\left(\log\frac{\rho_i}{\pi_i} - \log\frac{\rho_j}{\pi_j}\right)$$

Define $\tilde{a}_{ij} = -a_{ij}$ for $i \neq j$ and $\tilde{a}_{ii} = \sum_{j\neq i} a_{ij}$. Then

$$\frac{d}{dt}\rho_i = -\sum_j \tilde{a}_{ij} \log \frac{\rho_j}{\pi_j} =: -A_\rho \frac{\delta F}{\delta \rho}, \tag{4.24}$$

where $(\frac{\delta F}{\delta \rho})_i = \ln \frac{\rho_i}{\pi_i} + 1$ is the Fréchet derivative. We refer to [Maa11] for the corresponding Wasserstein distance and Benamou-Brenier formula.

Following the notation in Section 2.2, denote $\psi^*(\xi) = \frac{1}{2}\langle \xi, A_\rho \xi\rangle_{\langle T^*, T\rangle}$. Since $a_{ij} = a_{ji}$,

$$\langle \xi, A_\rho \xi\rangle = \sum_{i,j;i\neq j} -\xi_i a_{ij}\xi_j + \xi_i a_{ij}\xi_i = \frac{1}{2}\sum_{i,j;i\neq j} a_{ij}(\xi_i - \xi_j)^2 \geq 0, \tag{4.25}$$

so ψ^* is convex and $\psi^*(\xi) \geq \psi^*(0) = 0$. Then its convex dual is $\psi(s) = \sup_\xi\{\langle \xi, s\rangle - \psi^*(\xi)\} = \frac{1}{2}\langle A_\rho^\dagger s, s\rangle \geq 0$, where $A_\rho^\dagger$ is the generalized inverse. We can check (4.22) is a De Giorgi (ψ, ψ^*)-type gradient flow satisfying

$$\langle \frac{\delta F}{\delta \rho}, \dot{\rho}\rangle + \psi(\partial_t \rho) + \psi^*(-\frac{\delta F}{\delta \rho}) = 0. \tag{4.26}$$

We refer to [MRP14] for more discussion on how the generalized gradient flow above is related to the L-functions in large-deviation principle for the corresponding Fokker-Planck equation.

5. Conclusion

We discussed the parametric Bayesian inference via several gradient flow formulations, including strong formulation in tangent space or cotangent space, weak formulation and De Giorgi's type dual formulation. Lots of recently developed methods in data science are covered by these gradient flow formulations. Based on those gradient flow formulations for Bayesian inference, new schemes for accelerated nonconvex optimization and effective sampling shall be studied and developed in the future.

References

[AGS08] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008. MR2401600

[Agu12] Martial Agueh. Finsler structure in the p-wasserstein space and gradient flows. *Comptes Rendus Mathematique*, 350(1–2):35–40, Jan 2012. MR2887832

[DGT80] Marino A. De Giorgi, E. and M. Tosques. Problems of evolution in metric spaces and maximal decreasing curve. *Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur.*, 68(3):180–187, 1980. MR0636814

[GIHLS19] Alfredo Garbuno-Inigo, Franca Hoffmann, Wuchen Li, and Andrew M. Stuart. Interacting langevin diffusions: Gradient structure and ensemble kalman sampler. *arXiv:1903.08866 [math]*, Oct 2019. arXiv: 1903.08866.

[HV19] Michael Herty and Giuseppe Visconti. Kinetic methods for inverse problems. *Kinetic & Related Models*, 12(5):1109–1130, 2019. MR4027079

[Liu17] Qiang Liu. Stein variational gradient descent as gradient flow. In *Advances in neural information processing systems*, pages 3115–3123, 2017.

[LLN19] Jianfeng Lu, Yulong Lu, and James Nolen. Scaling limit of the stein variational gradient descent: The mean field regime. *SIAM Journal on Mathematical Analysis*, 51(2):648–671, 2019. MR3919409

[LW16] Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. In *Advances in neural information processing systems*, pages 2378–2386, 2016.

[Maa11] Jan Maas. Gradient flows of the entropy for finite Markov chains. *Journal of Functional Analysis*, 261(8):2250–2292, Oct 2011. MR2824578

[Mie16] Alexander Mielke. On evolutionary gamma-convergence for gradient systems. In *Macroscopic and large scale phenomena: coarse graining, mean field limits and ergodicity*, pages 187–249. Springer, 2016. MR3468299

[MMN18] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018. MR3845070

[MRP14] Alexander Mielke, D. R. Michiel Renger, and Mark A. Peletier. On the relation between gradient flows and the large-deviation principle, with applications to Markov chains and diffusion. *Potential Analysis*, 41(4):1293–1327, Nov 2014. MR3269725

[RMS08] Riccarda Rossi, Alexander Mielke, and Giuseppe Savaré. A metric approach to a class of doubly nonlinear evolution equations and applications. *Annali della Scuola Normale Superiore di Pisa-Classe di Scienze*, 7(1):97–169, 2008. MR2413674

[RVE19] Grant M. Rotskoff and Eric Vanden-Eijnden. Trainability and accuracy of neural networks: An interacting particle system approach. *arXiv:1805.00915 [cond-mat, stat]*, Jul 2019. arXiv: 1805.00915.

[SR15] N. Shojaee and M. M. Rezaii. On the gradient flows on finsler manifolds. *arXiv:1502.02146 [math]*, Apr 2015. arXiv: 1502.02146. MR3611055

[SS17] Claudia Schillings and Andrew M. Stuart. Analysis of the ensemble kalman filter for inverse problems. *SIAM Journal on Numerical Analysis*, 55(3):1264–1290, Jan 2017. MR3654885

Yuan Gao
Department of Mathematics
Duke University
Durham, NC
USA
E-mail address: yuangao@math.duke.edu

Jian-Guo Liu
Department of Mathematics and Department of Physics
Duke University
Durham, NC
USA
E-mail address: jliu@math.duke.edu

Received November 6, 2019