Parametric Fokker-Planck Equation

Wuchen Li[1], Shu Liu[2](✉), Hongyuan Zha[2], and Haomin Zhou[2]

[1] University of California, Los Angeles, USA
[2] Georgia Institute of Technology, Atlanta, USA
Sliu459@gatech.edu

Abstract. We derive the Fokker-Planck equation on the parametric space. It is the Wasserstein gradient flow of relative entropy on the statistical manifold. We pull back the PDE to a finite dimensional ODE on parameter space. Some analytical examples and numerical examples are presented.

Keywords: Optimal transport · Information geometry · Statistical manifold · Fokker-Planck equation · Gradient flow

1 Introduction

Fokker-Planck equation, a linear evolution partial differential equation (PDE), plays a crucial role in stochastic calculus, statistical physics and modeling [13,16,18]. Recently, people also discover its importance in statistics and machine learning [11,15,17]. Fokker-Planck equation describes the evolution of density functions of the stochastic process driven by a stochastic differential equation (SDE).

There is another viewpoint of Fokker-Planck equation based on optimal transport theory. It treats the equation as the gradient flow of relative entropy on probability manifold equipped with Wasserstein metric [5,14]. Recently, the studies have been extended to information geometry [1–3], creating a new area known as Wasserstein information geometry [7,9,10]. Inspired by those studies, in this paper, we derive the metric tensor on parameter space by pulling back the Wasserstein metric via the parameterized pushforward map. Then we compute the Wasserstein gradient flow (an ODE system) of relative entropy defined on parameter space. This leads to a statistical manifold version of Fokker Planck equation, which can be viewed as an approximation of the original PDE.

Our work is motivated by two purposes, (1) reducing the evolution PDE to a finite dimensional ODE system on parameter space; (2) applying parameterized pushforward map to obtain an efficient sampling method to generate samples from SDE. This is different from Markov Chain Monte Carlo (MCMC) methods [12] or momentum methods [16]. In this brief presentation, we sketch the theoretical framework with illustrations on several examples. The complete results will be reported in an extended version [8].

F. Nielsen and F. Barbaresco (Eds.): GSI 2019, LNCS 11712, pp. 715–724, 2019.
https://doi.org/10.1007/978-3-030-26980-7_74

2 Parametric Fokker-Planck Equation

In this section, we briefly review the fact that Fokker-Planck equation is a Wasserstein gradient flow of relative entropy. We then introduce a Wasserstein statistical manifold generated by parameterized mapping function. Based on it, we derive the parametric Fokker-Planck equation as the gradient flow of parameterized relative entropy.

2.1 Fokker-Planck Equation

Consider the Fokker-Planck equation:

$$\frac{\partial \rho(t,x)}{\partial t} = \nabla \cdot (\rho(t,x)\nabla V(x)) + \beta \Delta \rho(t,x), \quad \rho(0,x) = \rho_0(x). \tag{1}$$

Here $\nabla \cdot$, ∇ is the divergence and gradient operator in $\mathbb{R}^d$, ∇V is the drift function and $\beta > 0$ is a diffusion constant. There are several understandings for the Eq. (1).

On the one hand, consider the stochastic differential equation:

$$d\boldsymbol{X}_t = -\nabla V(\boldsymbol{X}_t) + \sqrt{2\beta} d\boldsymbol{B}_t, \quad X_0 \sim \rho_0. \tag{2}$$

Here $\{\boldsymbol{B}_t\}_{t\geq 0}$ is the standard Brownian motion. It is well known that the density function $\rho(t,x)$ of stochastic process $\boldsymbol{X}_t$, i.e. $\boldsymbol{X}_t \sim \rho(t,x)$, satisfies the Fokker-Planck equation (1).

On the other hand, Eq. (1) is the Wasserstein gradient flow of relative entropy. Denote the probability space supported on $\mathbb{R}^d$:

$$\mathcal{P} = \left\{ \rho \colon \int \rho(x) dx = 1, \ \rho(x) \geq 0, \ \int |x|^2 \rho(x) \ dx < \infty \right\}$$

Equipped with the Wasserstein metric [6,14], $\mathcal{P}$ is an infinite dimensional Riemmanian manifold. Denote

$$T_\rho \mathcal{P} = \left\{ \dot\rho \colon \int \dot\rho(x) dx = 0 \right\}.$$

Consider a specific $\rho \in \mathcal{P}$ and $\dot\rho_i \in T_\rho\mathcal{P}$, $i = 1,2$. The Wasserstein metric tensor g^W is defined as:

$$g^W(\rho)(\dot\rho_1, \dot\rho_2) = \int \nabla\psi_1(x) \cdot \nabla\psi_2(x)\rho(x) \ dx,$$

where $\dot\rho_i = -\nabla \cdot (\rho_i \nabla\psi_i)$ for $i = 1,2$. Here g^W is a metric tensor, which is a positive definite bilinear form defined on tangent bundle $T\mathcal{P} = \{(\rho, \dot\rho) \colon \rho \in \mathcal{P}, \ \dot\rho \in T_\rho\mathcal{P}\}$.

The Riemannian gradient in $(\mathcal{P}, g^W)$ is given as follows. Consider a smooth functional $\mathcal{F}\colon \mathcal{P} \to \mathbb{R}$, then

$$\begin{aligned} \mathrm{grad}_W \mathcal{F}(\rho) &= g^W(\rho)^{-1}\left(\frac{\delta \mathcal{F}}{\delta \rho}\right)(x) \\ &= -\nabla \cdot (\rho(x) \nabla \frac{\delta}{\delta \rho(x)} \mathcal{F}(\rho)), \end{aligned} \tag{3}$$

where $\frac{\delta}{\delta\rho(x)}$ is the L^2 first variation at variable $x \in \mathbb{R}^d$. In particular, consider the relative entropy

$$\mathcal{F}(\rho) = \beta \int \rho(x) \log \frac{\rho(x)}{\frac{1}{Z} e^{-\frac{V(x)}{\beta}}}\ dx = \int V(x)\rho(x)dx + \beta \int \rho(x) \log \rho(x) dx + \beta \log(Z). \tag{4}$$

Here $Z = \int e^{\frac{V(x)}{\beta}}\ dx$ is the normalizing constant for $e^{\frac{V(x)}{\beta}}$.
Then $\nabla\left(\frac{\delta\mathcal{F}}{\delta\rho}\right) = \nabla V + \beta \nabla \log \rho$, and (3) forms

$$\frac{\partial \rho}{\partial t} = -\mathrm{grad}_W \mathcal{F}(\rho) = \nabla \cdot (\rho \nabla V) + \beta \nabla \cdot (\rho \nabla \log \rho)).$$

Notice $\nabla \log \rho = \frac{\nabla \rho}{\rho}$, then $\nabla \cdot (\rho \nabla \log \rho) = \nabla \cdot (\nabla \rho) = \Delta \rho$. The above equation is exactly Fokker-Planck equation (1).

From now on, we apply the above geometric gradient flow formulation and derive the Fokker-Planck equation (1) on parameter space.

2.2 Parameter Space Equipped with Wasserstein Metric

We consider a parameter space $\varTheta$ as an open set in $\mathbb{R}^m$. Denote the sample space $M = \mathbb{R}^d$. Suppose T_θ is a pushforward map from M to M, which is parametrized by θ. For example, we can set $T_\theta(x) = Ux + b$, with $\theta = (U, b), U \in GL_d(\mathbb{R}),\ b \in \mathbb{R}^d$; we can also let T_θ be a neural network with parameter θ. We further assume that T_θ is invertible and smooth with respect to parameter θ and variable x.

Denote $p \in \mathcal{P}$ as a reference probability measure with positive density defined on M. For example, we can choose p as the standard Gaussian. We denote ρ_θ as the density of $T_{\theta\#}p$.[1] We further require: $\int |T_\theta(x)|^2\ dp(x) < \infty$ holds for all $\theta \in \varTheta$. Then $\rho_\theta \in \mathcal{P}$ for each $\theta \in \varTheta$. Denote $\mathcal{P}_\varTheta = \{\rho_\theta = \rho(\theta, x) | \theta \in \varTheta\}$, then $\mathcal{P}_\varTheta \subset \mathcal{P}$.

Now the connection between $\mathcal{P}$ and $\varTheta$ is the pushforward operation $T_\# : \varTheta \to \mathcal{P}_\varTheta \subset \mathcal{P}, \theta \mapsto \rho_\theta$. In order to introduce the Wasserstein metric to parameter space $\varTheta$, we assume that $T_\#$ is an isometric immersion from $\varTheta$ to $\mathcal{P}$. Under this assumption, the pullback $(T_\#)^* g^W$ of the Wasserstein metric g^W by $T_\#$ is the metric tensor on $\varTheta$. Let us denote $G = (T_\#)^* g^W$. Then for each θ, $G(\theta)$ is a

[1] Let X, Y be two measurable spaces, λ is a probability measure defined on X; let $T : X \to Y$ be a measurable map, then $T_\#\lambda$ is defined as: $T_\#\lambda(E) = \lambda(T^{-1}(E))$ for all measurable $E \subset Y$. We call $T_\# p$ the pushforward of measure p by map T.

bilinear form on $T_\theta \Theta \simeq \mathbb{R}^m$, thus $G(\theta)$ can be treated as an $m \times m$ matrix. Computation of $G(\theta)$ is illustrated in the following theorem:

Theorem 1. *Suppose $T_\# : \Theta \to \mathcal{P}$ is isometric immersion from Θ to $\mathcal{P}$. Then the metric tensor $G(\theta)$ at $\theta \in \Theta$ is $m \times m$ non-negative definite symmetric matrix and can be computed as:*

$$G(\theta) = \int \nabla \boldsymbol{\Psi}(T_\theta(x)) \nabla \boldsymbol{\Psi}(T_\theta(x))^T \; dp(x), \tag{5}$$

Or in entry-wised form:

$$G_{ij}(\theta) = \int \nabla \psi_i(T_\theta(x)) \cdot \nabla \psi_j(T_\theta(x)) \; dp(x), \quad 1 \leq i, j \leq m.$$

Here $\boldsymbol{\Psi} = (\psi_1, ... \psi_m)^T$ and $\nabla \boldsymbol{\Psi}$ is $m \times d$ Jacobian matrix of $\boldsymbol{\Psi}$. For each $k = 1, 2, ..., m$, ψ_k solves the following equation:

$$\nabla \cdot (\rho_\theta \nabla \psi_k(x)) = \nabla \cdot (\rho_\theta \; \partial_{\theta_k} T_\theta(T_\theta^{-1}(x))). \tag{6}$$

Proof. Suppose $\xi \in T\Theta$ is a vector field on Θ, for a fixed $\theta \in \Theta$, we first compute the pushforward $(T_\#|_\theta)_* \xi(\theta)$ of ξ at point θ: We choose any differentiable curve $\{\theta_t\}_{t \geq 0}$ on Θ with $\theta_0 = \theta$ and $\dot{\theta}_0 = \xi(\theta)$. If we denote $\rho_{\theta_t} = T_{\theta_t \#} p$, then we have $(T_\#)_* \xi(\theta) = \frac{\partial \rho_{\theta_t}}{\partial t}\Big|_{t=0}$. To compute $\frac{\partial \rho_{\theta_t}}{\partial t}\Big|_{t=0}$, we consider for any $\phi \in C_0^\infty(M)$:

$$\begin{aligned}\int \phi(y) \frac{\partial \rho_{\theta_t}}{\partial t}(y) dy = \frac{\partial}{\partial t}\left(\int \phi(T_{\theta_t}(x)) dp\right) &= \int \dot{\theta}_t^{\,T} \partial_\theta T_{\theta_t}(x) \nabla \phi(T_{\theta_t}(x)) dp \\ &= \int \dot{\theta}_t^T \partial_\theta T_{\theta_t}(T_{\theta_t}^{-1}(x)) \nabla \phi(x) \; \rho_{\theta_t}(x) \; dx \\ &= \int \phi(x) \left(-\nabla \cdot (\rho_{\theta_t} \partial_\theta T_{\theta_t}(T_{\theta_t}^{-1}(x))^T \; \dot{\theta}_t)\right) \; dx\end{aligned}$$

This weak formulation reveals that

$$(T_\#|_\theta)_* \xi(\theta) = \frac{\partial \rho_{\theta_t}}{\partial t}\Big|_{t=0} = -\nabla \cdot (\rho_\theta \; \partial_\theta T_\theta(T_\theta^{-1}(x))^T \; \xi(\theta)) \tag{7}$$

Now let us compute the metric tensor G. Since $T_\#$ is isometric immersion from Θ to $\mathcal{P}$, the pullback of g^W by $T_\#$ gives G, i.e. $(T_\#)^* g^W = G$. By definition of pullback map, for any $\xi \in T\Theta$ and for any $\theta \in \Theta$, we have:

$$G(\theta)(\xi(\theta), \xi(\theta)) = g^W(\rho_\theta)((T_\#|_\theta)_* \xi(\theta), (T_\#|_\theta)_* \xi(\theta)) \tag{8}$$

To compute the right hand side of (8), recall (3), we need to solve for φ from:

$$\frac{\partial \rho_{\theta_t}}{\partial t}\Big|_{t=0} = -\nabla \cdot (\rho_\theta \nabla \varphi(x)) \tag{9}$$

By (7), (9) is:

$$\nabla \cdot (\rho_\theta \nabla \varphi(x)) = \nabla \cdot (\rho_\theta \partial_\theta T_\theta(T_\theta^{-1}(\cdot))^T \; \xi(\theta)) \tag{10}$$

We can straightforwardly check that $\varphi(x) = \boldsymbol{\Psi}^T(x)\xi(\theta)$ is the solution of (10). Then $G(\theta)$ is computed as:

$$\begin{aligned} G(\theta)(\xi,\xi) &= \int |\nabla\varphi(y)|^2 \ \rho_\theta(y) \ dy = \int |\nabla\varphi(T_\theta(x))|^2 \ dp(x) \\ &= \int |\nabla\boldsymbol{\Psi}(T_\theta(x))^T\xi|^2 dp(x) = \xi^T\left(\int \nabla\boldsymbol{\Psi}(T_\theta(x))\nabla\boldsymbol{\Psi}(T_\theta(x))^T dp(x)\right)\xi \end{aligned}$$

Thus we can verify that:

$$G(\theta) = \int \nabla\boldsymbol{\Psi}(T_\theta(x))\nabla\boldsymbol{\Psi}(T_\theta(x))^T \ dp(x)$$

Generally speaking, the metric tensor G doesn't have an explicit form when $d \geq 2$; but for $d = 1$, G has an explicit form and can be computed directly.

Corollary 1. *When dimension d of M equals 1. And we further assume that: $\rho_\theta > 0$ on M and $\lim_{x\to\pm\infty}\rho_\theta(x) = 0$. Then $G(\theta)$ has an explicit form:*

$$G(\theta) = \int \partial_\theta T_\theta(x)^T \partial_\theta T_\theta(x) \ dp(x). \tag{11}$$

The following theorem ensures the positive definiteness of the metric tensor G:

Theorem 2. *We follow the notations and conditions in Sects. 2.2 and 2.3. Then G is Riemmanian metric on $T\Theta$ iff For each $\theta \in \Theta$, for any $\xi \in T_\theta\Theta$ $(\xi \neq 0)$, we can find $x \in M$ such that $\nabla \cdot (\rho_\theta \ \partial_\theta T_\theta(T_\theta^{-1}(x)\xi) \neq 0$.*

To keep our discussion concise, in the following sections, we will always assume that G is positive definite on $T\Theta$.
From now on, following [9,10], we call (Θ, G) Wasserstein statistical manifold.

2.3 Fokker-Planck Equation on Statistical Manifold

Recall the relative entropy functional $\mathcal{F}$ defined in (4), we consider $F = \mathcal{F} \circ T_\# : \Theta \to \mathbb{R}$. Then:

$$F(\theta) = \mathcal{F}(\rho_\theta) = \int V(x)\rho_\theta(x) \ dx + \beta \int \rho_\theta(x)\log\rho_\theta(x) \ dx. \tag{12}$$

As in [1], the gradient flow of F on Wasserstein statistical manifold (Θ, G) satisfies

$$\dot{\theta} = -G(\theta)^{-1}\nabla_\theta F(\theta). \tag{13}$$

We call (13) *parametric Fokker-Planck equation*. The ODE (13) as the Wasserstein gradient flow on parameter space (Θ, G) is closely related to Fokker-Planck equation on probability submanifold $\mathcal{P}_\Theta$. We have the following theorem, which is a natural result derived from submanifold geometry:

Theorem 3. *Suppose $\{\theta_t\}_{t\geq 0}$ solves (13). Then $\{\rho_{\theta_t}\}$ is the gradient flow of $\mathcal{F}$ on probability submanifold $\mathcal{P}_\Theta$.*

3 Example on Fokker-Planck Equations with Quadratic Potential

The solution of Fokker-Planck equation on statistical manifold (13) can serve as an approximation to the solution of the original Eq. (1). However, in some special cases, ρ_{θ_t} exactly solves (1). In this section, we demonstrate such examples.

Let us consider Fokker-Planck equations with quadratic potentials whose initial conditions are Gaussian, i.e.

$$V(x) = \frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu) \quad \text{and} \quad \rho_0 \sim \mathcal{N}(\mu_0, \Sigma_0). \tag{14}$$

Consider parameter space $\Theta = (\Gamma, b) \subset \mathbb{R}^m$ ($m = d(d+1)$), where Γ is a $d \times d$ invertible matrix with $\det(\Gamma) > 0$ and $b \in \mathbb{R}^d$. We define the parametric map as $T_\theta(x) = \Gamma x + b$. We choose the reference measure $p = \mathcal{N}(0, I)$. Here is the lemma we have to use:

Lemma 1. *Let $\mathcal{F}$ be the relative entropy defined in (4) and F defined in (12). For $\theta \in \Theta$, If the vector function $\nabla\left(\frac{\delta\mathcal{F}}{\delta\rho}\right) \circ T_\theta$ can be written as the linear combination of $\{\frac{\partial T_\theta}{\partial\theta_1}, ..., \frac{\partial T_\theta}{\partial\theta_m}\}$, i.e. there exists $\zeta \in \mathbb{R}^m$, such that $\nabla\left(\frac{\delta\mathcal{F}}{\delta\rho}\right) \circ T_\theta(x) = \partial_\theta T_\theta(x)\zeta$. Then:*

(1) $\zeta = G(\theta)^{-1}\nabla_\theta F(\theta)$, which is the Wasserstein gradient of F at θ.
(2) If we denote the gradient of $\mathcal{F}$ on $\mathcal{P}$ as $\mathrm{grad}\mathcal{F}(\rho_\theta)$ and the gradient of $\mathcal{F}$ on the submanifold $\mathcal{P}_\Theta$ as $\mathrm{grad}\mathcal{F}(\rho_\theta)|_{\mathcal{P}_\Theta}$, then $\mathrm{grad}\mathcal{F}(\rho_\theta)|_{\mathcal{P}_\Theta} = \mathrm{grad}\mathcal{F}(\rho_\theta)$.

Proof. The detailed proof is provided in [8]. Here is an intuitive explanation: $\nabla\left(\frac{\delta\mathcal{F}}{\delta\rho}\right) = \nabla V + \beta\nabla\log\rho_\theta$ is the real vector field that moves the particles in Fokker-Planck equation; and $\partial_\theta T_\theta(T_\theta^{-1}(\cdot))\dot{\theta}$ is the approximate vector field induced by the pushforward map T_θ. If such approximate is perfect with zero error, i.e. exits ζ such that $\nabla\left(\frac{\delta\mathcal{F}}{\delta\rho}\right) \circ T_\theta(x) = \partial_\theta T_\theta(x)\zeta$, then $\zeta = \dot{\theta} = G(\theta)^{-1}\nabla_\theta F(\theta)$ and the submanifold gradient agrees with entire manifold gradient.

Now, let us come back to our example, we can compute

$$\rho_\theta(x) = T_{\theta\#}p(x) = \frac{f(T_\theta^{-1}(x))}{|\det(\Gamma)|} = \frac{f(\Gamma^{-1}(x-b))}{|\det(\Gamma)|}, \ f(x) = \frac{\exp(-\frac{1}{2}|x|^2)}{(2p)^{\frac{d}{2}}}.$$

Then we have:

$$\nabla\left(\frac{\delta\mathcal{F}(\rho_\theta)}{\delta\rho}\right) \circ T_\theta(x) = \nabla(V + \beta\log\rho_\theta) \circ T_\theta(x) = \Sigma^{-1}(\Gamma x + b - \mu) - \beta\Gamma^{-T}x$$

is affine w.r.t. x.

Notice that $\partial_{\Gamma_{ij}} T_\theta(x) = (..0..\underset{i-\text{th}}{x_j}..0..)^T$ and $\partial_{b_i} T_\theta = (..0..\underset{i-\text{th}}{1}..0..)^T$. We can verify that $\zeta = (\Sigma^{-1}\Gamma - \beta\Gamma^{-T}, \Sigma^{-1}(b-\mu))$ solves $\nabla\left(\frac{\delta\mathcal{F}(\rho_\theta)}{\delta\rho}\right)\circ T_\theta(x) = \partial_\theta T_\theta(x)\zeta$. By (1) of Lemma 1, $\zeta = G(\theta)^{-1}\nabla_\theta F(\theta)$. Thus ODE (13) for our example is:

$$\dot{\Gamma} = -\Sigma^{-1}\Gamma + \beta\Gamma^{-T} \quad \Gamma_0 = \sqrt{\Sigma_0} \tag{15}$$

$$\dot{b} = \Sigma^{-1}(\mu - b) \quad b_0 = \mu_0 \tag{16}$$

By (2) of Lemma 1, we know $\text{grad}\mathcal{F}(\rho_\theta)|_{\mathcal{P}_\Theta} = \text{grad}\mathcal{F}(\rho_\theta)$ for all $\theta \in \Theta$. This indicates that there is no local error for our approximation, one can verify that the solution to the parametric Fokker-Planck equation also solves the original equation.

In addition to previous results, we have the following corollary:

Corollary 2. *The solution of Fokker-Planck equation (1) with condition (14) is Gaussian distribution for all* $t > 0$.

Proof. If we denote $\{\Gamma_t, b_t\}$ as the solutions to (15), (16), set $\theta_t = (\Gamma_t, b_t)$, then $\rho_t = T_{\theta_t\#}p$ solves the Fokker Planck Equation (1) with conditions (14). Since the pushforward of Gaussian distribution p by an affine transform T_θ is still a Gaussian, we conclude that for any $t > 0$, the solution $\rho_t = T_{\theta_t\#}p$ is always Gaussian distribution. This is already a well known result about Fokker-Planck equation. We reprove it under our framework.

4 Numerical Examples for 1D Fokker-Planck Equation

Since the Wasserstein metric tensor G has an explicit solution when dimension $d = 1$, it is convenient to numerically compute ODE (13).

For example, we can choose a series of basis functions $\{\varphi_k\}_{k=1}^n$. Each φ_k can be chosen as a sinusoidal function or a piece-wise linear function defined on a certain interval $[-l, l]$. It is also beneficial to choose orthogonal or near-orthogonal basis functions because they will keep the metric tensor G far away from ill-posedness. We set $T_\theta(x) = \sum_{k=1}^m \theta_k\varphi_k(x)$[2]. Then according to (11), we can compute G as

$$G_{ij}(\theta) = \mathbb{E}_{\mathbf{X}\sim p}\Big[\varphi_i(\mathbf{X})\varphi_j(\mathbf{X})\Big] \quad 1 \le i, j \le m$$

Recall that $F(\theta) = \int V(x)\rho_\theta(x)dx + \beta\int \rho_\theta(x)\log\rho_\theta(x)dx$. The second part of F is the entropy of ρ_θ, which can be computed by solving the following optimization problem [4]:

$$\int \rho_\theta(x)\log\rho_\theta(x)\;dx = \sup_h\Big\{\int h(x)\rho_\theta(x)\;dx - \int e^{h(x)}dx\Big\} + 1 \tag{17}$$

[2] In application, carefully choosing T_θ which is not necessarily invertibile or smooth can still provide valid results.

We can solve (17) by parametrizing h. Suppose the optimal solution is h^*. Then by envelope theorem, we know $\nabla_\theta F(\theta)$ can be computed as

$$\begin{aligned}\nabla_\theta F(\theta) &= \partial_\theta \left(\int V(x)\rho_\theta(x)\ dx + \beta \int h^*(x)\rho_\theta(x)\ dx \right) \\ &= \mathbb{E}_{\mathbf{X}\sim p}\left[\partial_\theta T_\theta(\mathbf{X})^T \nabla_y (V(y) + \beta h^*(y))|_{y=T_\theta(\mathbf{X})} \right] \end{aligned} \tag{18}$$

Notice that both the metric tensor G and $\nabla_\theta F(\theta)$ are written in forms of expectations, thus we can compute them by Monte Carlo simulations. And finally, (13) can be computed by forward Euler method.

Our numerical results are always demonstrated by sample points: For each time node t, we sample points $\{\mathbf{X}_1, ..., \mathbf{X}_N\}$ from p, then $\{T_{\theta_t}(\mathbf{X}_1), ..., T_{\theta_t}(\mathbf{X}_N)\}$ are our numerical samples from distribution ρ_t which solves the Fokker-Planck equation.

Here are several numerical results based on our method. We exhibit them in the form of histograms. Consider the potential $V(x) = (x+1)^2(x-1)^2$. Suppose the initial distribution is $\rho_0 = \mathcal{N}(0, I)$. Figure 1 contains histograms of ρ_t which solves $\frac{\partial \rho}{\partial t} = \nabla \cdot (\rho \nabla V)$ at different time nodes; we know ρ_t converges to $\frac{\delta_{-1}+\delta_{+1}}{2}$

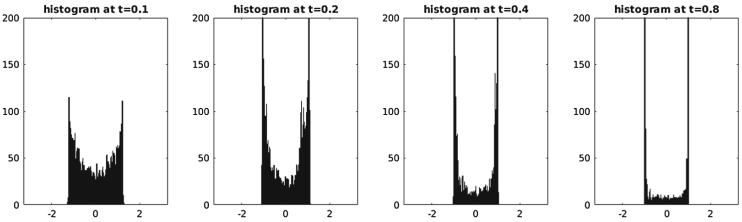

Fig. 1. Histograms of ρ_t solving $\frac{\partial \rho}{\partial t} = \nabla \cdot (\rho \nabla V)$

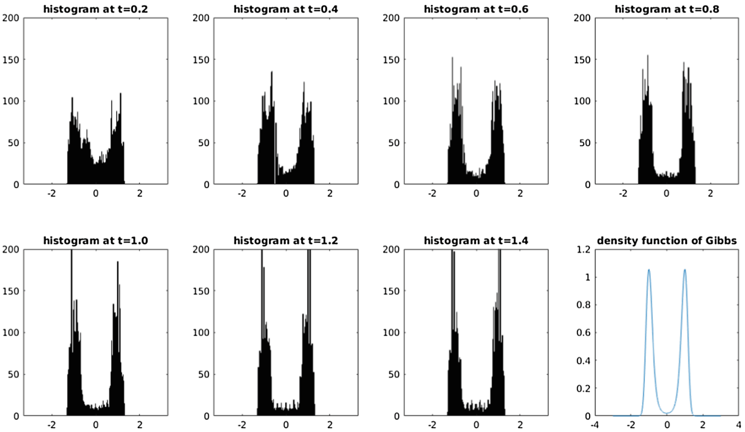

Fig. 2. Histograms of ρ_t solving $\frac{\partial \rho}{\partial t} = \nabla \cdot (\rho \nabla V) + \frac{1}{4}\Delta \rho$

as $t \to \infty$. Here δ_a is the Dirac distribution concentrated on point a. Figure 2 contains histograms of ρ_t which solves $\frac{\partial \rho}{\partial t} = \nabla \cdot (\rho \nabla V) + \frac{1}{4}\Delta \rho$ at different time nodes, we know ρ_t will converge to Gibbs distribution $\rho_* = \frac{1}{Z}\exp(-4(x+1)^2(x-1)^2)$, with Z being a normalizing constant, as $t \to \infty$. The density function of ρ_* is exhibited in Fig. 2.

5 Discussion

We presented a new approach for approximating Fokker-Planck equations by parameterized push-forward mapping functions. Compared to the classical moment method and MCMC method, we propose a systematic way for obtaining a finite dimensional ODE on parameter space. The ODE represents the evolution of statistical information conveyed in the original Fokker-Planck equation. In the future, we will study its geometric and statistical properties, and derive practical numerical methods for applications in scientific computing and machine learning. To be specific, in scientific computing, our techniques can be used to provide numerical solutions (samples) to those evolution PDEs that can be treated as Wasserstein gradient flows of certain functions defined on probability manifold; in area of machine learning, we wish to create efficient sampling methods based on our computational tools designed for Wasserstein gradients.

Acknowledgement. This project has received funding from AFOSR MURI FA9550-18-1-0502 and NSF Awards DMS–1419027, DMS-1620345, and ONR Award N000141310408.

References

1. Amari, S.: Natural gradient works efficiently in learning. Neural Comput. **10**(2), 251–276 (1998)
2. Amari, S.: Information Geometry and Its Applications. AMS, vol. 194. Springer, Tokyo (2016). https://doi.org/10.1007/978-4-431-55978-8
3. Ay, N., Jost, J., Lê, H.V., Schwachhöfer, L.: Information Geometry. EMG-FASMSM, vol. 64. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-56478-4
4. Essid, M., Laefer, D., Tabak, E.G.: Adaptive Optimal Transport. arXiv:1807.00393 [math] (2018)
5. Jordan, R., Kinderlehrer, D., Otto, F.: The variational formulation of the fokker-planck equation. SIAM J. Math. Anal. **29**(1), 1–17 (1998)
6. Lafferty, J.D.: The density manifold and configuration space quantization. Trans. Am. Math. Soc. **305**(2), 699–741 (1988)
7. Li, W.: Geometry of probability simplex via optimal transport. arXiv:1803.06360 [math] (2018)
8. Li, W., Liu, S., Zha, H., Zhou, H.: Scientific computing via parametric fokker-planck equations. In preparation (2019)
9. Li, W., Montufar, G.: Natural gradient via optimal transport. arXiv:1803.07033 [cs, math] (2018)

10. Li, W., Montufar, G.: Ricci curvature for parametric statistics via optimal transport (2018)
11. Liu, Q., Wang, D.: Stein variational gradient descent: a general purpose bayesian inference algorithm. arXiv:1608.04471 [cs, stat] (2016)
12. Liu, Q., Wang, D.: Stein variational gradient descent as moment matching. arXiv:1810.11693 [cs, stat] (2018)
13. Nelson, E.: Quantum Fluctuations. Princeton Series in Physics. Princeton University Press, Princeton (1985)
14. Otto, F.: The geometry of dissipative evolution equations: the porous medium equation. Commun. Partial Differ. Eqn. **26**(1–2), 101–174 (2001)
15. Pavon, M., Tabak, E.G., Trigilal, G.: The data-driven Schroedinger bridge. arXiv:1806.01364 [math] (2018)
16. Qi, D., Majda, A.J.: Low-dimensional reduced-order models for statistical response and uncertainty quantification: barotropic turbulence with topography. Phys. D **343**, 7–27 (2017)
17. Rezende, D.J., Mohamed, S.: Variational inference with normalizing flows. arXiv:1505.05770 [cs, stat] (2015)
18. Risken, H.: The Fokker-Planck Equation. Springer Series in Synergetics, vol. 18. Springer, Heidelberg (1989)