Self-Consistency of the Fokker-Planck Equation

Zebang Shen ZEBANG@SEAS.UPENN.EDU
University of Pennsylvania

Zhenfu Wang ZWANG@BICMR.PKU.EDU.CN
Peking University

Satyen Kale SATYEN.KALE@GMAIL.COM
Google

Alejandro Ribeiro ARIBEIRO@SEAS.UPENN.EDU
University of Pennsylvania

Amin Karbasi AMIN.KARBASI@YALE.EDU
Yale, Google

Hamed Hassani HASSANI@SEAS.UPENN.EDU
University of Pennsylvania

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

The Fokker-Planck equation (FPE) is the partial differential equation that governs the density evolution of the Itô process and is of great importance to the literature of statistical physics and machine learning. The FPE can be regarded as a continuity equation where the change of the density is completely determined by a time varying velocity field. Importantly, this velocity field also depends on the current density function. As a result, the ground-truth velocity field can be shown to be the solution of a fixed-point equation, a property that we call *self-consistency*. In this paper, we exploit this concept to design a potential function of the hypothesis velocity fields, and prove that, if such a function diminishes to zero during the training procedure, the trajectory of the densities generated by the hypothesis velocity fields converges to the solution of the FPE in the Wasserstein-2 sense. The proposed potential function is amenable to neural-network based parameterization as the stochastic gradient with respect to the parameter can be efficiently computed. Once a parameterized model, such as Neural Ordinary Differential Equation is trained, we can generate the entire trajectory to the FPE.

Keywords: Fokker Planck equation

1. Introduction

We consider the Fokker-Planck equation (FPE) that corresponds to the Itô process with a constant diffusion coefficient, which can be written as

$$\frac{\partial}{\partial t}\alpha(t,x) + \operatorname{div}\Big(\alpha(t,x)\underbrace{(-\nabla V(t,x) - \nabla \log \alpha(t,x))}_{\text{underlying velocity field } f^*(t,x)}\Big) = 0, \tag{1}$$

subject to the initial condition

$$\alpha(0,x) = \alpha_0(x). \tag{2}$$

Here, $\alpha : [0,T] \times \mathcal{X} \to \mathbb{R}$ is a time varying density function defined on $\mathcal{X} \subseteq \mathbb{R}^d$, $V : [0,T] \times \mathcal{X} \to \mathbb{R}$ is a known potential function that determines the drifting term; div and ∇ denote the divergence and gradient operator with respect to the spatial variable x respectively. The boundary condition that we impose will be introduced in section 2.

FPE is a fundamental problem in the literature of statistical physics due to its wide applications in thermodynamic system analysis (Markowich and Villani, 2000; Lucia and Gervino, 2015; Qi and Majda, 2016) and is one of the key equations in the research of the mean field game (Cardaliaguet and Porretta, 2020; Gomes et al., 2014). Recently, it has also been used to model the dynamics of the stochastic gradient descent method on neural networks (Chizat and Bach, 2018; Sonoda and Murata, 2019; Sirignano and Spiliopoulos, 2020; Fang et al., 2021) and the dynamics of the Rényi differential privacy (Chourasia et al., 2021), and has become a fundamental tool for learning complex distributions and deep generative models due to its deep connection to the Wasserstein gradient flow (Sohl-Dickstein et al., 2015; Hashimoto et al., 2016; Liu et al., 2019; Song et al., 2020; Solin et al., 2021; Mokrov et al., 2021). There is a plethora of previous works trying to solve FPE numerically, including the classic mesh-based finite difference and finite volume methods (Carrillo et al., 2015; Bailo et al., 2018), the stochastic particle methods that are based on the discretization of the Ito SDE (Dalalyan, 2017; Li et al., 2019, 2021), the deterministic particle methods that utilize the Gaussian mollifier to approximate the dynamic (Degond and Mustieles, 1990), the variational methods that are built on the Wasserstein gradient flow interpretation of the FPE (Bernton, 2018; Liu et al., 2020; Carrillo et al., 2021; Ambrosio et al., 2005; Jordan et al., 1998), and most recently the physics-informed neural network approach that directly parameterize the solution to the FPE and cast the FPE as a root finding problem (Han et al., 2018; Long et al., 2018, 2019; Raissi et al., 2019; Blechschmidt and Ernst, 2021). We note that in all previous approaches, the entity under consideration, i.e. the function to be approximated or learned, is explicitly the solution to the PDE (1), which is a time-varying probability density function.

In this work, we take a different route: Instead of approximating the solution to the FPE, we propose to learn the *underlying velocity field* that drives the evolution of the FPE. The solution to the FPE can then be implicitly recovered by the learned velocity field. Our work is built on a concept called the *self-consistency* of the Fokker-Planck equation: A velocity field that correctly recovers the solution to the FPE should be a fixed point to a *velocity-consistency transformation* (defined in Eq. (14)) derived from the FPE. The main contribution of our work is summarized as follows.

> We establish the theoretical foundation of learning the underlying velocity field of the FPE. Specifically, we design a potential function R for the hypothesis velocity fields $\{f_n\}$ that describes the self-consistency of the Fokker-Planck equation and show that if $R(f_n) \to 0$ as $n \to \infty$, the trajectory of distributions generated by f_∞ recovers the solution to the FPE in the Wasserstein-2 sense.

Moreover, when the hypothesis velocity field is parameterized as a Neural Ordinary Differential Equation f_θ (Chen et al., 2018), we discuss how the stochastic gradient of the proposed potential function $R(f_\theta)$ with respect to the parameter θ of the neural network can be efficiently computed. Therefore, once f_θ is trained via stochastic optimization methods, our approach returns an approximate solution to the FPE, which is non-negative and has unit mass, i.e. it integrates to 1 on $\mathcal{X}$. These fundamental properties are crucial in real-world physics models and are not guaranteed in previous neural network based approaches.

2. Preliminaries

Boundary Condition We assume that the process takes place on a d-dimensional box centered around the origin, i.e. $\mathcal{X} = [-\frac{l}{2}, \frac{l}{2}]^d$. We consider the periodic boundary condition:

$$\alpha\left(t, (\cdots, -\frac{l}{2}, \cdots)\right) = \alpha\left(t, (\cdots, \frac{l}{2}, \cdots)\right) \tag{3}$$

$$\frac{\partial}{\partial x}\alpha\left(t, (\cdots, -\frac{l}{2}, \cdots)\right) = \frac{\partial}{\partial x}\alpha\left(t, (\cdots, \frac{l}{2}, \cdots)\right). \tag{4}$$

The above condition is the same as identifying the points on the corresponding boundaries which happens when the spatial domain is a *torus*. Note that on a torus, the particle that leaves the torus on the boundary will reenter the domain $\mathcal{X}$ through the boundary such that $l/2$ (resp., $-l/2$) is replaced by $-l/2$ (resp., $l/2$) in the same coordinate.

The periodic boundary condition (torus) is commonly used in the PDE analysis (e.g. see (Jabin and Wang, 2016)) with an important technical merit that the integration of a periodic function on the boundary is naturally zero and hence the analysis using integration by parts can be simplified. Moreover, it also allows us to focus on the behavior of the PDE system on compact domains without sacrificing the generality, since we can always set the diameter of the torus to be sufficiently large. We emphasize that to the ML community, this is usually the case of interest: Only in a bounded domain can we expect a neural ODE to be able to represent the underlying velocity field of the FPE, since the neural network is *not* a universal function approximator on unbounded domains.

In the following, we refer to periodic functions with a period of l as *l-periodic*.

Velocity Field and the Induced Push-forward Map A velocity field is map $f : [0, T] \times \mathcal{X} \to \mathbb{R}^d$ that determines the movement of a particle $x(t)$:

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = f(t, x(t)) \tag{5}$$

A velocity field $f(t, x)$ induces a push-forward map $X(t, x; f)$ via integrating over time

$$X(t, x_0; f) = x_0 + \int_0^t f(s, x_s)\mathrm{d}s, \tag{6}$$

where $\{x_s\}_{s=0}^t$ is the trajectory of a particle following the velocity field $f(t, x)$ with the initial position x_0. Note that the map $X(t, x; f)$ is invertible under the assumption that $f(t, x)$ is Lipschitz continuous in x for all t. Additionally $X(t, x; f) - x$ is l-periodic if we further assume that f is l-periodic: For any $i \in \{1, \ldots, d\}$

$$X(t, x_0 + le_i; f) - (x_0 + le_i) = \int_0^t f(s, x_s + le_i)\mathrm{d}s = \int_0^t f(s, x_s)\mathrm{d}s = X(t, x_0; f) - x_0. \tag{7}$$

When clear from the context, we omit the dependence of X on f and write $X(t, x)$, for simplicity.

Neural Ordinary Differential Equation The neural ordinary differential equation (NODE) is a favorable instance of the hypothesis class since neural networks are universal function approximators in a bounded domain and have achieved great recent success in machine learning (Chen et al.,

2018; Dupont et al., 2019; Choromanski et al., 2020). Let $f : \mathbb{R} \times \mathbb{R}^d \times \Theta \to \mathbb{R}^d$ be a neural network parameterized by $\theta \in \Theta$. A d-dimensional NODE in can be described as

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = f_\theta(t, x(t)). \tag{8}$$

To accommodate the periodic boundary conditions (3) and (4), we need the NODE to be l-periodic. Consider a $2d$-dimensional NODE with velocity $\tilde{f}$. We can construct a d-dimensional NODE with the following hypothesis velocity field

$$f_\theta(t, x(t)) = \tilde{f}_\theta\left(t, \begin{pmatrix} \sin \frac{2\pi}{l}x(t) \\ \cos \frac{2\pi}{l}x(t) \end{pmatrix}\right). \tag{9}$$

Here $\sin$ and $\cos$ are applied in an element-wise manner.

Notations Consider the d-dimensional index vector $a = (a_1, \ldots, a_d)$ with $a_i \in \mathbb{N}$ and $\|a\|_1 = k$ and a map $f : \mathbb{R}^d \to \mathbb{R}^d$. Denote

$$f^{(a)} = \left[\frac{\partial^k f_1}{\partial x_1^{a_1} \ldots \partial x_d^{a_d}}, \cdots, \frac{\partial^k f_d}{\partial x_1^{a_1} \ldots \partial x_d^{a_d}}\right], \tag{10}$$

where f_i denotes the ith entry of f. We define the kth order Sobolev norm of a map $f : \mathcal{X} \to \mathbb{R}^d$ with a base measure $\mu \in \mathcal{M}_+^1(\mathcal{X})$ by

$$\|f\|_{W^{k,2}(\mu)} = \left(\sum_{i=0}^{k} \int_{\mathcal{X}} \|f^{(i)}(x)\|^2 \mu(x)\mathrm{d}x\right)^{\frac{1}{2}}. \tag{11}$$

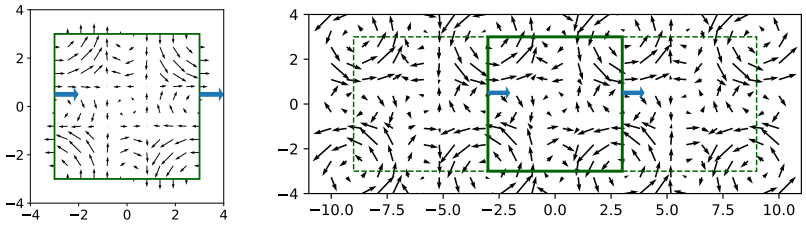
Here $f^{(k)} = \{f^{(a)}\}_{a:\|a\|_1=k}$ denotes the collection of all kth order partial derivatives of the map f and is regarded as a d^{k+1}-dimensional vector. We use $\|\cdot\|$ to denote the spectral norm for matrices and tensors and the standard ℓ_2-norm for vectors.
We use $\{e_i\}$ to denote the standard basis of $\mathbb{R}^d$ and use Δ to denote the Laplacian operator on the spatial variable. We use $\nabla^i, i \geq 2$ to denote higher order gradient.

3. Methodology

Recall that on a torus, when a particle leaves the domain on a boundary, it reappears on the other side (see Figure 1-(a)). Therefore, the velocity field of the particles are discontinuous on the boundaries, which introduces difficulties in function approximation. To avoid this issue, a useful and equivalent perspective of the periodic boundary condition is to think of the density function $\alpha(t, \cdot)$ as a l-periodic function in every coordinate, i.e.

$$\forall t, x, \quad \alpha(t, x + le_i) = \alpha(t, x), i \in [d], \tag{12}$$

which is depicted in (b) of Figure 1. While particles are allowed to leave $\mathcal{X}$, the domain of interest, due to the periodicity of the whole domain $\mathbb{R}^d$, the total mass within $\mathcal{X}$ is conserved since the influx and the outflow are balanced.



(a) (b)

Figure 1: Figure (a) depicts that when a particle leaves the torus on a boundary, it reappears on the other side. The velocity field is discontinuous on the boundary. In Figure (b), we consider the periodic extension of the density function $\alpha(t, x)$. This is equivalent to the torus since whenever a particle leaves the boundary, another particle will enter $\mathcal{X}$ from a corresponding adjoining cell. Note that in Figure (b) the velocity field is continuous on the whole domain.

3.1. Self-consistency of the Fokker-Planck Equation

Suppose that the particles are distributed initially according to the distribution α_0 defined in (2) and follow a hypothesis velocity field $f(t, x)$. From this perspective, we can write the distribution of particles on $\mathcal{X}$ at time t in a push-forward manner

$$\rho^1(t, \cdot; f) = X(t, \cdot; f)\sharp\alpha_0, \tag{13}$$

where the push-forward map X, induced by the velocity f, is defined in (6). Note that ρ^1 is well-defined on the whole domain $\mathbb{R}^d$, but we restrict our interest to $\mathcal{X}$. Based on this notation, the Fokker-Planck equation (1) induces a *velocity-consistency transformation* $\mathcal{A}$ of the velocity field in the following manner:

$$\mathcal{A}[f](t, x) = -\nabla V(t, x) - \nabla \log \rho^1(t, x; f). \tag{14}$$

Observe that, for the ground-truth velocity field f^* that drives the particle evolution of the Fokker-Planck equation, i.e. $f^*(t, x) = -\nabla V(t, x) - \nabla \log \alpha(t, x)$, we have

$$\mathcal{A}[f^*] = f^*.$$

We term this property the *self-consistency* of the Fokker-Planck equation. Similar to Eq. (13), we can define $\rho^2(t, \cdot; f) = X(t, \cdot; \mathcal{A}[f])\sharp\alpha_0$,. Indeed, the interplay between the two systems ρ^1 and ρ^2 is crucial to our analysis.

The goal of our paper is to show that if a sequence of hypothesis velocity fields $\{f_n\}$ asymptotically satisfies the above consistency property, i.e. $\|\mathcal{A}[f_n] - f_n\| \to 0$ as $n \to \infty$ for some appropriate norm $\|\cdot\|$, then the distribution $\rho^1(t, x; f_\infty)$ generated from the hypothesis velocity field f_∞ recovers $\alpha(t, x)$, the solution to the FPE (1) in the Wasserstein-2 sense.

3.2. Designing the Self-Consistency Potential Function and its Computation

Given a hypothesis velocity field f, we denote the difference between f and $\mathcal{A}[f]$ by

$$\delta(t, x; f) = f(t, x) - \mathcal{A}[f](t, x). \tag{15}$$

We propose to use the time average of the 2nd order Sobolev norm of δ with the base measure $\rho^1(t,\cdot;f)$ as the potential function of f:

$$R(f) = \int_0^T \int_{\mathcal{X}} \sum_{i=0}^{2} \|\delta^{(i)}(t,x;f)\|^2 \rho^1(t,x;f)\mathrm{d}x\mathrm{d}t = \int_0^T \|\delta(t,\cdot;f)\|^2_{W^{2,2}(\rho^1(t,\cdot;f))}\mathrm{d}t.$$

In Section 4, we show that $R(f)$ controls the Wasserstein-2 distance between $\rho^1(t,\cdot)$ and $\alpha(t,\cdot)$, i.e. for any time $t \in [0,T]$, $W_2^2(\rho^1(t,\cdot),\alpha(t,\cdot)) = O(R(f))$. This result has two direct implications: (i) Given a hypothesis velocity field f, we can use $R(f)$ to measure its quality in terms of recovering the solution to the FPE; (ii) Given a class of parameterized hypothesis velocity fields f_θ, one can find the best parameter θ by minimizing $R(f_\theta)$ with a *learning* procedure, which is discussed in details at the end of this section. By "learning", we mean to distinguish our approach from the previous numerical FPE solvers, e.g. the JKO method, which are in essence "simulating" the FPE dynamics: They iteratively update the configuration of the system using certain rules derived from the FPE. In contrast, the proposed potential function describes the self-inconsistency of a hypothesis velocity field, which can be refined through a training procedure.

The potential function $R(f)$ might seem difficult to compute at first. In the following, we present an equivalent formulation of $R(f)$ from the perspective of particle trajectory, which is critical to our analysis and to the actually computation of $R(f)$. We first introduce the following important change-of-variables formula of integrating periodic functions on $\mathcal{X}$. Recall that the standard change-of-variables formula reads as follows: for a function $g : \mathbb{R}^d \to \mathbb{R}$

$$\int_{\mathcal{X}} g\mathrm{d}X\sharp\alpha = \int_{X^{-1}(\mathcal{X})} g \circ X \mathrm{d}\alpha. \tag{16}$$

In brief, we show that for an l-periodic functions g the integration domain $X^{-1}(\mathcal{X})$ on the RHS of the above equation can be replaced by $\mathcal{X}$. The proof is deferred to Appendix A.

Lemma 1 *Consider an invertible mapping $X : \mathbb{R}^d \to \mathbb{R}^d$ such that $X(x) - x$ is l-periodic, an l-periodic function $g : \mathbb{R}^d \to \mathbb{R}$, and an l-periodic measure α. The following formula holds:*

$$\int_{\mathcal{X}} g\mathrm{d}X\sharp\alpha = \int_{\mathcal{X}} g \circ X \mathrm{d}\alpha, \tag{17}$$

where $\mathcal{X}$ is the d-dimensional box centered defined above.

Note that the push-forward map X defined in (6) is invertible and $X(t,x;f) - x$ is l-periodic (see (7)), the integrand in $R(f)$ is l-periodic, and from (12) the measure α_0 is also l-periodic. Using the above lemma, we have

$$R(f) = \int_0^T \int_{\mathcal{X}} \sum_{i=0}^{2} \|\delta^{(i)}(t,\cdot;f)\|^2 \mathrm{d}X(t,\cdot;f)\sharp\alpha_0\mathrm{d}t \tag{18}$$

$$= \int_0^T \int_{\mathcal{X}} \sum_{i=0}^{2} \|\delta^{(i)}(t,X(t,x;f);f)\|^2 \mathrm{d}\alpha_0(x)\mathrm{d}t \tag{19}$$

If we further define the trajectory-wise loss

$$R(f;x_0) = \int_0^T \sum_{i=0}^{2} \|\delta^{(i)}\left(t,X(t,x_0;f);f\right)\|^2 \mathrm{d}t, \tag{20}$$

the potential function $R(f)$ admits an equivalent formulation

$$R(f) = \int_{\mathcal{X}} R(f; x_0)\alpha_0(x_0)\mathrm{d}x_0. \tag{21}$$

Therefore, we have that $R(f; x_0)$ is an unbiased estimator of the objective $R(f)$. In the following, we elaborate on how $R(f; x_0)$ can be computed.

Computation of the trajectory-wise loss $R(f; x_0)$ We now discuss how the function $R(f; x_0)$ can be computed. We assume that we have the exact expression of f and V, and hence we can readily evaluate $f^{(i)}$ for $i \in \{0, 1, 2\}$ and $V^{(i)}$ for $i \in \{1, 2, 3\}$ (recall the notation of differentials in (10)). Use $x(t) = X(t, x_0; f)$ to denote the trajectory of a particle with the initial position x_0 and following the velocity field f. In the following, we address how $\nabla^i \log \rho^1(t, x(t); f)$ for $i \in \{1, 2, 3\}$ can be computed since these are the only unknown terms when evaluating $\mathcal{A}[f]^{(i)}(t, x(t))$ for $i \in \{0, 1, 2\}$. The proofs of the following propositions are deferred to the appendix. We first compute the first order gradient of the log-probability.

Proposition 2 *Denote $f_t(x) = f(t, x)$ and $\rho_t^1 = \rho^1(t, x; f)$ where we recall that $\rho^1(t, x; f)$ is the density function formally defined in equation (13). We have*

$$\frac{\mathrm{d}}{\mathrm{d}t}\nabla \log \rho_t^1(x(t)) = -\nabla \mathrm{div}\left(f_t(x(t))\right) - \left(\nabla f_t(x(t))\right)^\top \nabla \log \rho_t^1(x(t)),$$

The second order partial derivatives of the log-probability is computed as follows.

Proposition 3 *Denote $f_t(x) = f(t, x)$ and $\rho_t^1 = \rho^1(t, x; f)$ where we recall that $\rho^1(t, x; f)$ is formally defined in equation (13). The time evolution of the 2nd order gradient of the log probability function can be computed by*

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial^2}{\partial x_i \partial x_j}\log \rho_t^1(x(t)) = &-\frac{\partial^2}{\partial x_i \partial x_j}\mathrm{div} f_t(x_t) - \frac{\partial}{\partial x_i}\nabla \log \rho_t^1(x(t)) \cdot \frac{\partial}{\partial x_j} f_t(x(t)) \\ &-\frac{\partial}{\partial x_i} f_t(x(t)) \cdot \frac{\partial}{\partial x_j}\nabla \log \rho_t^1(x(t)) - \frac{\partial^2}{\partial x_i \partial x_j} f_t(x(t)) \cdot \nabla \log \rho_t^1(x(t)).\end{aligned}$$

The third order partial derivatives of the log-probability is computed as follows.

Proposition 4 *Denote $f_t(x) = f(t, x)$ and $\rho_t^1 = \rho^1(t, x; f)$ where we recall that $\rho^1(t, x; f)$ is formally defined in equation (13). The time evolution of the 3rd order gradient of the log probability function can be computed by*

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial^3}{\partial x_i \partial x_j \partial x_k}\log \rho_t^1(x(t)) = &-\frac{\partial^3}{\partial x_i \partial x_j \partial x_k}\mathrm{div} f_t(x(t)) - \frac{\partial^2}{\partial x_i \partial x_j}\nabla \log \rho_t^1(x(t)) \cdot \frac{\partial}{\partial x_k} f_t(x(t)) \\ &- \frac{\partial^2}{\partial x_i \partial x_k}\nabla \log \rho_t^1(x(t)) \cdot \partial_j f_t(x(t)) - \frac{\partial}{\partial x_i}\nabla \log \rho_t^1(x(t)) \cdot \partial_{j,k} f_t(x(t)) \\ &- \frac{\partial^2}{\partial x_j \partial x_j}\nabla \log \rho_t^1(x(t)) \cdot \partial_i f_t(x(t)) - \frac{\partial}{\partial x_j}\nabla \log \rho_t^1(x(t)) \cdot \partial_{i,k} f_t(x(t)) \\ &- \frac{\partial}{\partial x_k}\nabla \log \rho_t^1(x(t)) \cdot \partial_{i,j} f_t(x(t)) - \nabla \log \rho_t^1(x(t)) \cdot \frac{\partial^3}{\partial x_i \partial x_j \partial x_k} f_t(x(t)).\end{aligned}$$

The above propositions show that the evolution of the ith order differential of $\log \rho_t^1$ only depends differentials with order no more than i. This means that the differentials of $\log \rho_t^1$ can be exactly computed using only local information, even though they depend on the macroscopic distribution. Note that this is only possible along $\{x(t)\}$, the trajectory of the particle under consideration.

Parameterizing the Hypothesis Velocity Field with NODE In the following, we take the NODE as a specific parameterized instance of the hypothesis velocity field f_θ. Recall that $R(f_\theta; x_0)$ is an unbiased estimator of $R(f_\theta)$. A key step in the optimization of a neural network is to compute the stochastic gradient $\nabla_\theta R(f_\theta; x_0)$, which is elaborated as follows.

Suppose that the initial point x_0 is fixed. To compute $\nabla_\theta R(f_\theta; x_0)$, the gradient of the trajectory-wise loss with respect to the parameter θ, we write $R(f_\theta; x_0)$ in a standard ODE-constrained form:

$$R(f_\theta; x_0) = \ell(\theta) \triangleq \int_0^T g(t, s(t), \theta) \mathrm{d}t \tag{22}$$

where $\{s(t)\}_{t\in[0,T]}$ is the solution to the ODE

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t} s(t) = \psi(t, s(t); \theta) \\ s(0) = s_0(x_0). \end{cases} \tag{23}$$

Recall the definition of the differentials $f^{(i)}$ in (10). Here, the time-varying state $s(t)$ is

$$s(t) = [x(t), \zeta_1(t), \zeta_2(t), \zeta_3(t)], \tag{24}$$

where $\zeta_i(t) = (\log \rho^1)^{(i+1)}(t, x(t); f_\theta)$ for $i \in \{0, 1, 2\}$; s_0 is a function of x_0

$$s_0(x_0) = [x_0, (\log \alpha_0)^{(1)}(x_0), (\log \alpha_0)^{(2)}(x_0), (\log \alpha_0)^{(3)}(x_0)]; \tag{25}$$

Here, ψ is the velocity field that drives the evolution of the state s such that the first component of s is updated according to the ODEs in equation (8) and the last three components of s are updated according to propositions 2 to 4 respectively; and the function g is define as

$$g(t, s(t); \theta) = \sum_{i=0}^{2} \| f_\theta^{(i)}(t, x(t)) + V^{(i+1)}(x(t)) + \zeta_i \|^2,$$

so that we recover the difference function δ defined in (15). Note that by introducing the auxiliary states ζ_i, the function g depends on θ only through $f_\theta^{(i)}(t, x(t))$. With the above standard ODE-constrained form of $R(f_\theta; x_0)$, we can compute $\nabla \ell(\theta)$ in equation (22) using the classic adjoint method, which is provided in Appendix E.

Recovering an Approximate Solution to the FPE Given a hypothesis velocity field f, we return $\rho(t, \cdot; f)$ as an approximate solution to the FPE. To evaluate $\rho(t, x; f)$ for any $x \in \mathcal{X}$, let $x(s)_{s\in[0,t]}$ be the trajectory of the final value problem

$$\frac{\mathrm{d}x(s)}{\mathrm{d}s} = f(s, x(s)), x(t) = x. \tag{26}$$

We can compute that $\frac{\mathrm{d}}{\mathrm{d}t}\log\rho^1(t,x(t);f) = \frac{\partial}{\partial t}\log\rho^1(t,x(t);f) + f(t,x(t))\cdot\nabla\log\rho^1(t,x(t);f)$. Using the FPE (1), we derive $\frac{\partial}{\partial t}\log\rho^1(t,x) = -\mathrm{div}f(t,x) - \nabla\log\rho^1(t,x;f)\cdot f(t,x)$, and hence we have $\frac{\mathrm{d}}{\mathrm{d}t}\log\rho^1(t,x(t);f) = -\mathrm{div}f(t,x(t))$. Therefore, we can compute $\log\rho^1(t,x;f)$ by

$$\log\rho^1(t,x;f) = \log\alpha_0(x(0)) - \int_0^t \mathrm{div}f(t,x(s))\mathrm{d}s. \tag{27}$$

4. Analysis

In this section, we prove that the potential function $R(f)$ inspired by the self-consistency of the FPE controls the Wasserstein-2 distance between $\rho^1(t,\cdot;f)$ and $\alpha(t,\cdot)$ for all $t\in[0,T]$. We achieve this by introducing an auxiliary distribution ρ^2 induced by $\mathcal{A}[f]$ to bridge the hypothesis distribution $\rho^1(t,\cdot;f)$ induced by the velocity field and the solution to the FPE $\alpha(t,\cdot)$. This allows us to control the Wasserstein-2 distance between $\rho^1(t,\cdot;f)$ and $\rho^2(t,\cdot;f)$ and the KL-divergence between $\rho^2(t,\cdot;f)$ and $\alpha(t,\cdot)$ separately. We first present the assumptions required for our analysis.

Assumption 1 (Regularity of the initial distribution) *For any $x\in\mathcal{X}$, the Hessian of the log probability of the initial distribution $\rho_0^1=\alpha_0$ is bounded, i.e.*

$$\max\{\|\nabla\log\alpha_0(x)\|, \|\nabla^2\log\alpha_0(x)\|, \|\nabla^3\log\alpha_0(x)\|, \|\nabla^2\Delta\log\alpha\|\} \le L_0. \tag{28}$$

Assumption 2 (Regularity of the hypothesis velocity field) *The hypothesis velocity field f is l-periodic for any time t and parameter θ. Moreover, given a fixed time horizon $T>0$ of the evolution, for any space-time variables $x\in\mathcal{X}$ and $t\in[0,T]$ and any neural network parameters $\theta\in\Theta$, the hypothesis velocity field f in NODE satisfies that for all $x\in\mathcal{X}$*

$$\max\{\|\max_{i\in\{1,2,3,4\}}\nabla^i f_t(x)\|, \max_{i\in\{1,2,3,4\}}\|\nabla^i\mathrm{div}f_t(x)\|\} \le L_f. \tag{29}$$

Assumption 3 (Regularity of the drifting term) *For all t, the potential function $V(t,\cdot)$ is l-periodic and for all $x\in\mathcal{X}$ $\max\{\|\nabla^2V(t,x)\|, \|\nabla^3V(t,x)\|, \|\nabla^2\Delta V(x)\|\}\le L_v$.*

We state our main result as follows.

Theorem 5 (main result) *Suppose that the assumptions 1 to 3 hold. We have for all $t\in[0,T]$*

$$W_2^2(\rho^1(t,\cdot;f),\alpha(t,\cdot)) \le d\,l\,c\cdot R(f), \tag{30}$$

where l is length of the box $\mathcal{X}$, d is the dimension of the ambient space, and c is a constant that depends on the regularity constants L_0, L_f, L_v and the maximum evolving time T.

The following corollary states that if we can optimize over the hypothesis velocity field f such that $R(f)$ diminishes to zero, we can recover the solution to the FPE in the Wasserstein-2 sense.

Corollary 6 *Suppose that assumptions 1 to 3 hold and assume a sequence of hypothesis velocity fields f_n satisfies $R(f_n)\to 0$ as $n\to\infty$. We have $W_2^2(\rho^1(t,\cdot;f_n),\alpha(t,\cdot))\to 0$ as $n\to\infty$.*

Table 1: Summary of the notations for systems (1) and (2). Note that $Y(t,\cdot;f) = X(t,\cdot;\mathcal{A}[f])$.

	velocity field	particle map	particle trajectory	density
System (1)	$f(t,x)$	$X(t,\cdot;f)$	$\{x(t)\}$	$\rho^1(t,\cdot;f) = X(t,\cdot;f)\sharp\alpha_0$
System (2)	$\mathcal{A}[f](t,x)$	$Y(t,\cdot;f)$	$\{y(t)\}$	$\rho^2(t,\cdot;f) = Y(t,\cdot;f)\sharp\alpha_0$

Remark 7 *Assume that the class of hypothesis velocity fields is the NODE f_θ (see (8)). Also, assume that the underlying velocity field f^* is sufficiently regular such that it can be represented by f_{θ^*} for some optimal parameter θ^*. Then, we can optimize over the parameter θ and recover the solution to the FPE if $R(f_\theta)$ diminishes to zero during the training phase.*

We now present the proof of Theorem 5 which is built on the interplay between two systems: The first is described by the hypothesis velocity field f:

$$\text{System (1): } \frac{\mathrm{d}x(t)}{\mathrm{d}t} = f(t,x(t)); \tag{31}$$

and the second is driven by $\mathcal{A}[f]$ which is defined in (14):

$$\text{System (2): } \frac{\mathrm{d}y(t)}{\mathrm{d}t} = \mathcal{A}[f](t,y(t)). \tag{32}$$

Similar to the push-forward map $X(t,\cdot;f)$ defined in (6), $\mathcal{A}[f]$ also induces a map $X(t,\cdot;\mathcal{A}[f])$. To better distinguish these two systems, we denote $Y(t,x;f) = X(t,x;\mathcal{A}[f])$ and define $\rho^2(t,\cdot;f) = Y(t,\cdot;f)\sharp\alpha_0$ for system (2). These notations are summarized in Table 1.

The following lemma establishes some regularity results of the involved velocity fields.

Lemma 8 *Recall Systems (1) and (2) in Table 1. For simplicity of notations, denote their probability density functions by ρ_t^1 and ρ_t^2 respectively. Additionally, we denote $f_t(x) = f(t,x)$ and $\mathcal{A}[f]_t(x) = \mathcal{A}[f](t,x)$. We have that for all $t \in [0,T]$*

1. *Both ρ_t^1 and ρ_t^2 are l-periodic.*
2. *$\nabla \log \rho_t^1$ is bounded and Lipschitz continuous.*
3. *Both $\nabla\mathcal{A}[f]_t$ and $\nabla\mathrm{div}\mathcal{A}[f]_t$ are bounded and Lipschitz continuous.*

The following lemma shows that $R(f)$ controls the Wasserstein-2 distance between $\rho^1(t,\cdot)$ and $\rho^2(t,\cdot)$ for all $t \in [0,T]$. The full proof is provided in Appendix F.1.

Lemma 9 *Recall Systems (1) and (2) in Table 1. Denote their probability density functions by ρ_t^1 and ρ_t^2 respectively. Under Assumptions 1 to 3, there exists a constant C_1 such that*

$$\sup_{t\in[0,T]} W_2^2(\rho_t^1,\rho_t^2) \le C_1 R(f),$$

where C_1 depends on the maximum evolving time T and L_0, L_f, L_v defined in assumptions 1 to 3.

Proof [A sketch of the proof.] We first note that $P(t,\cdot;f) = Y(t,\cdot;f) \circ X(t,\cdot;f)^{-1}$ is a transport map such that $\rho^2(t,\cdot;f) = P(t,\cdot;f)\sharp\rho^1(t,\cdot;f)$. Consequently, from the definition of the Wasserstein-2 distance, we have

$$\begin{aligned} W_2^2(\rho_t^1,\rho_t^2) &\le \int_{\mathcal{X}} \|x - P(t,x;f)\|^2 \mathrm{d}\rho^1(t,x;f) = \int_{\mathcal{X}} \|X(t,x;f) - Y(t,x;f)\|^2 \mathrm{d}\alpha_0(x) \\ &= \int_{\mathcal{X}} \|x(t) - y(t)\|^2 \mathrm{d}\alpha_0(x_0), \end{aligned}$$

where we used the change-of-variables formula of the push-forward measure from Lemma 1 in the first equality and $\{x(t)\}_{t\in[0,T]}$ and $\{y(t)\}_{t\in[0,T]}$ are the trajectory of particles initialized from x_0 but driven by Systems (1) and (2) respectively. We then study the dynamic of $\frac{\mathrm{d}}{\mathrm{d}t}\|x(t) - y(t)\|^2$ and prove the lemma using the Grönwall's inequality. ■

We then show that $R(f)$ controls the distance between score functions of systems (1) and (2). The full proof is provided in Appendix F.2.

Lemma 10 *Recall Systems (1) and (2) in Table 1. For simplicity of notations, denote their probability density functions by ρ_t^1 and ρ_t^2 respectively. Denote the weighted L_2 norm by*

$$\xi_t \triangleq \|\nabla \log \rho_t^1 - \nabla \log \rho_t^2\|_{\rho_t^2}^2. \tag{33}$$

Suppose assumptions 1 to 3 hold. There exists some constant C_2 such that for any $t \in [0,T]$,

$$\int_0^t \xi_s \mathrm{d}s \le C_2 R(f), \tag{34}$$

where C_2 depends on the maximum evolving time T and L_0, L_f, L_v defined in assumptions 1 to 3.

Proof [A sketch of the proof.] With the change-of-variables lemma, we can expand

$$\begin{aligned} \xi_t &= \|\nabla \log \rho_t^1 \circ Y_t - \nabla \log \rho_t^2 \circ Y_t\|_{\alpha_0}^2 \\ &\le \|\nabla \log \rho_t^1(y(t)) - \nabla \log \rho_t^1(x(t))\|_{\alpha_0}^2 + \|\nabla \log \rho_t^1(x(t)) - \nabla \log \rho_t^2(y(t))\|_{\alpha_0}^2. \end{aligned}$$

The first term can be control by the Lipschitz continuity of $\nabla \log \rho_t^1$. We study the dynamic of the second term using Proposition 2 and use the Grönwall's inequality to establish the lemma. ■

Built on the above two lemmas, the following lemma states the most novel part of our analysis which shows that the KL-divergence between $\rho^2(t,\cdot)$ generated by system (2) and the solution to the FPE $\alpha(t,\cdot)$ is controlled by $R(f)$.

Lemma 11 *Recall Systems (1) and (2) in Table 1. For simplicity of notations, denote their probability density functions by ρ_t^1 and ρ_t^2 respectively and use α_t to denote the solution to the Fokker-Planck equation (1). Suppose assumptions 1 to 3 hold. For any $t \in [0,T]$, we have*

$$\mathrm{KL}(\rho_t^2, \alpha_t) \le \frac{C_2}{2} R(f), \tag{35}$$

where C_2 is the constant defined in Lemma 10.

Proof We study the evolution of the KL divergence between ρ_t^2 and α_t. Recall that for ρ_t^2, we have

$$\frac{\partial \rho_t^2}{\partial t} = \operatorname{div}\left(\rho_t^2 \cdot \nabla(V_t + \log \rho_t^1)\right), \tag{36}$$

and for α_t we have

$$\frac{\partial \alpha_t}{\partial t} = \operatorname{div}\left(\alpha_t \cdot \nabla(V_t + \log \alpha_t)\right). \tag{37}$$

We can compute

$$\frac{\mathrm{dKL}(\rho_t^2, \alpha_t)}{\mathrm{d}t} = \int_{\mathcal{X}} \frac{\partial \rho_t^2}{\partial t} \log \frac{\rho_t^2}{\alpha_t} + \frac{\partial \rho_t^2}{\partial t} - \rho_t^2 \frac{\partial \log \alpha_t}{\partial t} \mathrm{d}x = \int_{\mathcal{X}} \frac{\partial \rho_t^2}{\partial t} \log \frac{\rho_t^2}{\alpha_t} - \frac{\rho_t^2}{\alpha_t} \frac{\partial \alpha_t}{\partial t} \mathrm{d}x,$$

where in the second equality, we use $\int_{\mathcal{X}} \frac{\partial \rho_t^2}{\partial t} \mathrm{d}x = \frac{\mathrm{d} \int_{\mathcal{X}} \rho_t^2 \mathrm{d}x}{\mathrm{d}t} = \frac{\mathrm{d}1}{\mathrm{d}t} = 0$. Plug (36) and (37) in the above equation to derive

$$\frac{\mathrm{dKL}(\rho_t^2, \alpha_t)}{\mathrm{d}t} = \int_{\mathcal{X}} \operatorname{div}\left(\rho_t^2 \cdot \nabla(V_t + \log \rho_t^1)\right) \cdot \log \frac{\rho_t^2}{\alpha_t} - \int_{\mathcal{X}} \frac{\rho_t^2}{\alpha_t} \operatorname{div}\left(\alpha_t \cdot \nabla(V_t + \log \alpha_t)\right).$$

Using integration by part, we have that

$$\begin{aligned}
\frac{\mathrm{dKL}(\rho_t^2, \alpha_t)}{\mathrm{d}t} &= -\int_{\mathcal{X}} \rho_t^2 \cdot \nabla(V_t + \log \rho_t^1) \cdot \nabla \log \frac{\rho_t^2}{\alpha_t} + \int_{\mathcal{X}} \nabla(\frac{\rho_t^2}{\alpha_t}) \cdot \alpha_t \cdot \nabla(V_t + \log \alpha_t) \\
&= -\int_{\mathcal{X}} \rho_t^2 \cdot \nabla(V_t + \log \rho_t^1) \cdot \nabla \log \frac{\rho_t^2}{\alpha_t} + \int_{\mathcal{X}} \nabla \log \frac{\rho_t^2}{\alpha_t} \cdot \rho_t^2 \cdot \nabla(V_t + \log \alpha_t) \\
&= -\int_{\mathcal{X}} \rho_t^2 \cdot \nabla \log \frac{\rho_t^2}{\alpha_t} \cdot \nabla(\log \frac{\rho_t^2}{\alpha_t} + \log \frac{\rho_t^1}{\rho_t^2}),
\end{aligned}$$

where we note that the integrations on the boundary $\partial \mathcal{X}$ vanish due to the periodic boundary conditions (3) and (4). We hence have

$$\frac{\mathrm{dKL}(\rho_t^2, \alpha_t)}{\mathrm{d}t} = -\|\nabla \log \rho_t^2 - \nabla \log \alpha_t\|_{\rho_t^2}^2 - \int_{\mathcal{X}} \rho_t^2 \nabla \log \frac{\rho_t^1}{\rho_t^2} \cdot \nabla \log \frac{\rho_t^2}{\alpha_t}.$$

Using the Cauchy–Schwarz inequality, we have

$$-\int_{\mathcal{X}} \rho_t^2 \nabla \log \frac{\rho_t^1}{\rho_t^2} \cdot \nabla \log \frac{\rho_t^2}{\alpha_t} \mathrm{d}x \leq \frac{1}{2}\|\nabla \log \rho_t^1 - \nabla \log \rho_t^2\|_{\rho_t^2}^2 + \frac{1}{2}\|\nabla \log \rho_t^2 - \nabla \log \alpha_t\|_{\rho_t^2}^2.$$

Consequently, we obtain

$$\frac{\mathrm{dKL}(\rho_t^2, \alpha_t)}{\mathrm{d}t} \leq \frac{1}{2}\|\nabla \log \rho_t^2 - \nabla \log \rho_t^1\|_{\rho_t^2}^2 - \frac{1}{2}\|\nabla \log \rho_t^2 - \nabla \log \alpha_t\|_{\rho_t^2}^2.$$

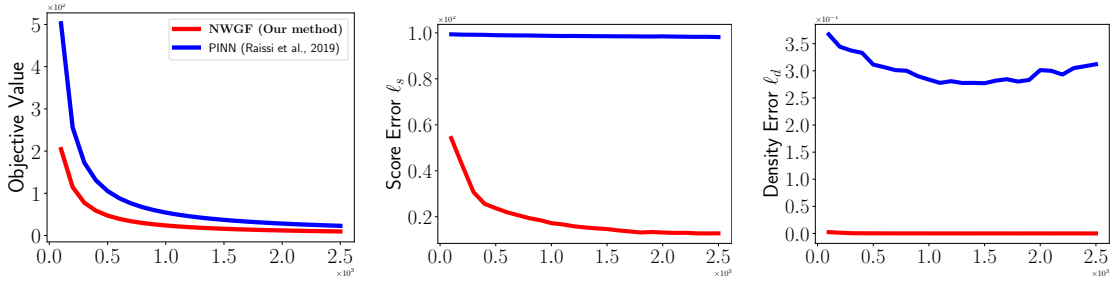
Omitting the negative term and integrating from 0 to t and using Lemma 10, we have our result. ■

We now present the proof of Theorem 5.

Proof [Proof of Theorem 5] Using Theorem 6.15 of (Villani, 2009), we have for any t

$$W_2^2(\rho^2(t,\cdot), \alpha(t,\cdot)) \leq 2ld\mathrm{TV}^2(\rho^2(t,\cdot), \alpha(t,\cdot)) \leq ld\mathrm{KL}(\rho^2(t,\cdot), \alpha(t,\cdot)), \tag{38}$$

where we use the Pinsker's inequality in the second inequality. Using the triangle inequality of the Wasserstein-2 distance, Theorem 5 is a direct consequence of Lemma 9 and Lemma 11. ■



(i) Objective Value (ii) Score Estimation Error (iii) Density Estimation Error

Figure 2: Learning the FPE with a Gaussian initial distribution α_0 and a quadratic drifting term V.

5. Experiment

Setup. In this section, we showcase the effecacy of our approach for numerically solving the FPE with the example where the initial distribution is Gaussian, i.e. $\alpha_0 = \mathcal{N}(\mu_0, \Sigma_0)$, and the drifting term is a quadratic function, i.e. $V(x) = (x - \mu_\infty)^\top \Sigma_\infty^{-1}(x - \mu_\infty)$. We use this example since we know the analytical solution of the FPE $\alpha(t, x)$ in this specific instance and hence we can explicitly calculate the difference between the learned hypothesis velocity field f_θ and the ground truth. Specifically, we know that for any time $t \geq 0$, the solution $\alpha(t, \cdot) = \mathcal{N}(\mu_t, \Gamma_t^\top \Gamma_t)$ is a Gaussian distribution where μ_t and Γ_t evolve in the following manner

$$\frac{\mathrm{d}\mu_t}{\mathrm{d}t} = \Sigma_\infty^{-1}(\mu_\infty - \mu_t), \quad \frac{\mathrm{d}\Gamma_t}{\mathrm{d}t} = -\Sigma_\infty^{-1}\Gamma_t + \Gamma_t^{-1^\top}, \Gamma_0 = \sqrt{\Sigma_0}, \tag{39}$$

if we take the the domain $\mathcal{X} = \mathbb{R}^2$ (see for example Eq. (36) and Eq. (37) in Liu et al. (2020)). In our experiment, we take $\mu_0 = (-4, -4)$, $\Sigma_0 = \mathrm{diag}(0.7, 1.3)$, and $\mu_\infty = (4, 4)$, $\Sigma_\infty = \mathrm{diag}(1.1, 0.9)$.

Performance Metrics. We grid the box $[-10, 10]^2$ with a uniform increment of 0.1 over both coordinates. This gives us $201^2 = 40401$ grid points altogether and we use β to denote the uniform distribution over these points. We then grid the time interval $[0, 3]$ with a uniform increment of 0.3. This gives us 11 distinct time stamps and we use γ to denote the uniform distribution over these time stamps. Define the score estimation error of a hypothesis velocity field f to be $\ell_s(f) = \int \|f(t, x) + \nabla \log \alpha(t, x) + \nabla V(x)\|^2 \mathrm{d}\beta(x)\mathrm{d}\gamma(t)$, where we note that $(-\nabla V - \nabla \log \alpha)$ is the ground truth velocity field. Additionally, define the density estimation error of a hypothesis density trajectory ρ as $\ell_d(\rho) = \int |\alpha(t, x) - \rho(t, x)| \mathrm{d}\beta(x)\mathrm{d}\gamma(t)$. We use the these two quantities in our experiment to measure the quality of the recovered solutions from `NWGF` (our approach) and we include a successful NN-based PDE solver `PINN` (Raissi et al., 2019) as the baseline. Note that the implementation of the continuous time `PINN` model requires a collection of spatial points $\{x_i\}$ for defining the objective loss, which are set to the grid points mentioned above.

Details. To avoid negative density values in `PINN`, instead of directly approximating $\alpha(t, x)$, we use a neural network $g_\theta(t, x)$ to approximate the ground-truth log-density trajectory $\log \alpha$ in `PINN`. For a fair comparison, the network structures of f_θ (the hypothesis velocity field used in our approach) and g_θ are identical except the last layer since g_θ outputs a scalar (log-density) while f_θ outputs a $2d$-vector (velocity). We use $\ell_s(-\nabla V - \nabla_x g_\theta)$ to measure the quality of g_θ as $(-\nabla V - \nabla g_\theta)$ is the hypothesis velocity field that corresponds to g_θ. We use the strategy discussed in Eq. (27) to

reover the density from our hypothesis velocity field f_θ. The code of our implementation can be found at https://github.com/shenzebang/self-consistency-jax.

Results. We report the results of our experiment in Figure 2 and we use NWGF (short for Neural Wasserstein Gradient Flow) to denote our approach. In plot (i), we observe that stochastic gradient descent is able to reduce the objective values of both NWGF and PINN substantially over 2500 steps. However, in plots (ii) and (iii), we observe that our method correctly learns the underlying velocity field and the density trajectories, but these two metrics of PINN barely improve after a long training procedure. This shows the advantage of our approach.

6. Conclusion

In this work, instead of directly approximating the solution to the FPE, we proposed a learning paradigm that recovers the entire velocity field, thus understanding better the evolution of the system. By introducing a velocity-consistency transformation $\mathcal{A}$ induced by the FPE, we identified a fundamental property of the system called the self-consistency of the FPE. In words, it states that the underlying velocity field of the FPE must be a fixed point of $\mathcal{A}$. Based on this novel observation, we designed a potential function $R(f)$ for any hypothesis velocity field f and proved that $R(f)$ controls the Wasserstein-2 distance between the trajectory of distributions generated by f and the exact solution to the FPE. When the hypothesis velocity field is parameterized by a time-varying neural network, we showed that the stochastic gradient of the proposed potential function with respect to the parameter of the neural network can be computed using the adjoint method.

Acknowledgments

The research of Hassani and Shen is supported by NSF Grants 1837253, 1943064, AFOSR Grant FA9550-20-1-0111, DCIST-CRA, and the AI Institute for Learning-Enabled Optimization at Scale (TILOS). Zhenfu Wang is supported by the National Key R&D Program of China, Project Number 2021YFA1002800, NSFC grant No.12171009, Young Elite Scientist Sponsorship Program by China Association for Science and Technology (CAST) No. YESS20200028 and the start-up fund from Peking University. Amin Karbasi acknowledges funding in direct support of this work from NSF (IIS-1845032), ONR (N00014-19-1-2406), NSF (2112665), and the AI Institute for Learning-Enabled Optimization at Scale (TILOS).

References

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2005.

Rafael Bailo, Jose A Carrillo, and Jingwei Hu. Fully discrete positivity-preserving and energy-dissipating schemes for aggregation-diffusion equations with a gradient flow structure. *arXiv preprint arXiv:1811.11502*, 2018.

Espen Bernton. Langevin monte carlo and jko splitting. In *Conference On Learning Theory*, pages 1777–1798. PMLR, 2018.

Jan Blechschmidt and Oliver G Ernst. Three ways to solve partial differential equations with neural networks—a review. *GAMM-Mitteilungen*, 44(2):e202100006, 2021.

Pierre Cardaliaguet and Alessio Porretta. An introduction to mean field game theory. In *Mean Field Games*, pages 1–158. Springer, 2020.

José A Carrillo, Alina Chertock, and Yanghong Huang. A finite-volume method for nonlinear nonlocal equations with a gradient flow structure. *Communications in Computational Physics*, 17(1):233–258, 2015.

Jose A Carrillo, Katy Craig, Li Wang, and Chaozhen Wei. Primal dual methods for wasserstein gradient flows. *Foundations of Computational Mathematics*, pages 1–55, 2021.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 31, 2018.

Krzysztof M Choromanski, Jared Quincy Davis, Valerii Likhosherstov, Xingyou Song, Jean-Jacques Slotine, Jacob Varley, Honglak Lee, Adrian Weller, and Vikas Sindhwani. Ode to an ode. *Advances in Neural Information Processing Systems*, 33:3338–3350, 2020.

Rishav Chourasia, Jiayuan Ye, and Reza Shokri. Differential privacy dynamics of langevin diffusion and noisy gradient descent. *Advances in Neural Information Processing Systems*, 34, 2021.

Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3): 651–676, 2017.

Pierre Degond and Francisco-José Mustieles. A deterministic approximation of diffusion equations using particles. *SIAM Journal on Scientific and Statistical Computing*, 11(2):293–310, 1990.

Emilien Dupont, Arnaud Doucet, and Yee Whye Teh. Augmented neural odes. *Advances in Neural Information Processing Systems*, 32, 2019.

Cong Fang, Jason Lee, Pengkun Yang, and Tong Zhang. Modeling from features: a mean-field framework for over-parameterized deep neural networks. In *Conference on learning theory*, pages 1887–1936. PMLR, 2021.

Diogo A Gomes et al. Mean field games models—a brief survey. *Dynamic Games and Applications*, 4(2):110–154, 2014.

Jiequn Han, Arnulf Jentzen, and E Weinan. Solving high-dimensional partial differential equations using deep learning. *Proceedings of the National Academy of Sciences*, 115(34):8505–8510, 2018.

Tatsunori Hashimoto, David Gifford, and Tommi Jaakkola. Learning population-level diffusions with generative rnns. In *International Conference on Machine Learning*, pages 2417–2426. PMLR, 2016.

Pierre-Emmanuel Jabin and Zhenfu Wang. Mean field limit and propagation of chaos for vlasov systems with bounded forces. *Journal of Functional Analysis*, 271(12):3588–3627, 2016. ISSN 0022-1236. doi: https://doi.org/10.1016/j.jfa.2016.09.014. URL https://www.sciencedirect.com/science/article/pii/S0022123616302701.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker–planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998. doi: 10.1137/S0036141096303359.

Ruilin Li, Hongyuan Zha, and Molei Tao. Sqrt (d) dimension dependence of langevin monte carlo. *arXiv preprint arXiv:2109.03839*, 2021.

Xuechen Li, Yi Wu, Lester Mackey, and Murat A Erdogdu. Stochastic runge-kutta accelerates langevin monte carlo and beyond. *Advances in neural information processing systems*, 32, 2019.

Chang Liu, Jingwei Zhuo, and Jun Zhu. Understanding mcmc dynamics as flows on the wasserstein space. In *International Conference on Machine Learning*, pages 4093–4103. PMLR, 2019.

Shu Liu, Wuchen Li, Hongyuan Zha, and Haomin Zhou. Neural parametric fokker-planck equations. *arXiv preprint arXiv:2002.11309*, 2020.

Zichao Long, Yiping Lu, Xianzhong Ma, and Bin Dong. Pde-net: Learning pdes from data. In *International Conference on Machine Learning*, pages 3208–3216. PMLR, 2018.

Zichao Long, Yiping Lu, and Bin Dong. Pde-net 2.0: Learning pdes from data with a numeric-symbolic hybrid deep network. *Journal of Computational Physics*, 399:108925, 2019.

Umberto Lucia and Gianpiero Gervino. Fokker-planck equation and thermodynamic system analysis. *Entropy*, 17(2):763–771, 2015.

Peter A Markowich and Cédric Villani. On the trend to equilibrium for the fokker-planck equation: an interplay between physics and functional analysis. *Mat. Contemp*, 19:1–29, 2000.

Petr Mokrov, Alexander Korotin, Lingxiao Li, Aude Genevay, Justin M Solomon, and Evgeny Burnaev. Large-scale wasserstein gradient flows. *Advances in Neural Information Processing Systems*, 34, 2021.

Di Qi and Andrew J Majda. Low-dimensional reduced-order models for statistical response and uncertainty quantification: Two-layer baroclinic turbulence. *Journal of the Atmospheric Sciences*, 73(12):4609–4639, 2016.

Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707, 2019.

Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A central limit theorem. *Stochastic Processes and their Applications*, 130(3):1820–1852, 2020.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

Arno Solin, Ella Tamir, and Prakhar Verma. Scalable inference in sdes by direct matching of the fokker–planck–kolmogorov equation. *Advances in Neural Information Processing Systems*, 34, 2021.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020.

Sho Sonoda and Noboru Murata. Transport analysis of infinitely deep neural network. *The Journal of Machine Learning Research*, 20(1):31–82, 2019.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

Appendix A. Proof of Lemma 1

Proof From the change-of-variables formula of the pushforward measure, we have

$$\int_{\mathcal{X}} g \mathrm{d} X \sharp \alpha = \int_{X^{-1}(\mathcal{X})} g \circ X \mathrm{d}\alpha. \tag{40}$$

Let $\Pi : \mathbb{R}^d \to \mathcal{X}$ be that modulus operator such that given any input $x \in \mathbb{R}^d$, $\Pi(x)$ is the unique element in $\mathcal{X}$ such that

$$x = \Pi(x) + \sum_{i=1}^{d} n_i l \cdot e_i, \tag{41}$$

for some $n_i \in \mathbb{Z}$, $i = 1, \ldots, d$. In the following, we show that (1) $X^{-1}(\mathcal{X})$ does not overlap with itself under the operator Π, i.e. there do not exist two points $x_1, x_2 \in X^{-1}(\mathcal{X})$ with $x_1 \neq x_2$ such that $\Pi(x_1) = \Pi(x_2)$ and (2) $\Pi(X^{-1}(\mathcal{X})) = \mathcal{X}$. Suppose that these two statements hold, we have

$$\int_{X^{-1}(\mathcal{X})} g \circ X \mathrm{d}\alpha \overset{(1)}{=} \int_{\Pi(X^{-1}(\mathcal{X}))} g \circ X \mathrm{d}\alpha \overset{(2)}{=} \int_{\mathcal{X}} g \circ X \mathrm{d}\alpha. \tag{42}$$

To prove (1), suppose that there exist $x_1, x_2 \in X^{-1}(\mathcal{X})$ with $x_1 \neq x_2$ such that $\Pi(x_1) = \Pi(x_2)$. There must exist $m_i \in \mathbb{Z}$, $i \in \{1, \ldots, d\}$ such that $x_1 = x_2 + (\cdots, m_i \times l, \cdots)$ and that m_i's cannot be all zeros. Since $X(x) - x$ is l-periodic, we have

$$X(x_1) - x_1 = X(x_2) - x_2 \Rightarrow X(x_1) = X(x_2) + (\cdots, m_i \times l, \cdots). \tag{43}$$

Since at least one of the m_i's are non-zero, it is impossible that $X(x_1)$ and $X(x_2)$ belong to $\mathcal{X}$ simultaneously, which leads to a contradiction.
To prove (2), we first observe that $\Pi(X^{-1}(\mathcal{X})) \subseteq \mathcal{X}$ holds trivially due to the definition of Π, and hence we just need to show that $\mathcal{X} \subseteq \Pi(X^{-1}(\mathcal{X}))$. We prove via contradiction. Suppose that there exists $y \in \mathcal{X}$ such that $y \notin \Pi(X^{-1}(\mathcal{X}))$. From the definition of the operator Π, we can write

$$X(y) = \Pi(X(y)) + (\cdots, m_i \times l, \cdots), \tag{44}$$

for some $m_i \in \mathbb{Z}$, $i \in \{1, \ldots, d\}$. Since X^{-1} is periodic, we have that

$$y - (\cdots, m_i \times l, \cdots) = X^{-1}(X(y) - (\cdots, m_i \times l, \cdots)) \overset{(44)}{=} X^{-1}(\Pi(X(y))) \in X^{-1}(\mathcal{X}). \tag{45}$$

However, the above statement means $y \in \Pi(X^{-1}(\mathcal{X}))$ which contradicts to the definition of y. ■

Appendix B. Proof of Proposition 2

Proof First, compute that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\nabla\log\rho_t^1(x(t)) &= \frac{\partial}{\partial t}\nabla\log\rho_t^1(x(t)) + \frac{\mathrm{d}x(t)}{\mathrm{d}t}\frac{\partial}{\partial x}\nabla\log\rho_t^1(x(t)) \\ &= \nabla\frac{\partial}{\partial t}\log\rho_t^1(x(t)) + f_t(x(t))\nabla^2\log\rho_t^1(x(t)).\end{aligned}$$

Using the Fokker Planck equation (1), we derive

$$\frac{\partial}{\partial t}\log\rho_t^1 = -\mathrm{div}f_t - \nabla\log\rho_t^1\cdot f_t, \tag{46}$$

which together with

$$\nabla(\nabla\log\rho_t^1\cdot f_t) = \nabla^2\log\rho_t^1 f_t + (\nabla f_t)^\top\nabla\log\rho_t^1$$

allows us to compute

$$\frac{\mathrm{d}}{\mathrm{d}t}\nabla\log\rho_t^1(x(t)) = -\nabla\mathrm{div}f_t(x(t)) - (\nabla f_t(x(t)))^\top\nabla\log\rho_t^1(x(t)).$$

In the above computation, we use the fact that the term $\nabla^2\log\rho_t^1(x(t))f_t(x(t))$ is canceled. ■

Appendix C. Proof of Proposition 3

Proof For compactness, we use $\partial_{i,j}$ to denote $\frac{\partial^2}{\partial x_i\partial x_j}$. First, compute that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\partial_{i,j}\log\rho_{t,\theta}^1(x(t)) &= \frac{\partial}{\partial t}\partial_{i,j}\log\rho_t^1(x(t)) + \frac{\partial}{\partial x}\partial_{i,j}\log\rho_t^1(x(t))\cdot\frac{\mathrm{d}x(t)}{\mathrm{d}t} \\ &= \partial_{i,j}\frac{\partial}{\partial t}\log\rho_t^1(x(t)) + \frac{\partial}{\partial x}\partial_{i,j}\log\rho_t^1(x(t))\cdot f_t(x(t)).\end{aligned}$$

Using the Fokker Planck equation (1), we derive

$$\frac{\partial}{\partial t}\log\rho_t^1 = -\mathrm{div}f_t - \nabla\log\rho_t\cdot f_t, \tag{47}$$

which together with

$$\begin{aligned}\partial_{i,j}(\nabla\log\rho_t^1\cdot f_t) = {}& \partial_{i,j}\nabla\log\rho_t^1\cdot f_t + \partial_i\nabla\log\rho_t^1\cdot\partial_j f_t \\ & +\partial_i f_t\cdot\partial_j\nabla\log\rho_t^1 + \partial_{i,j}f_t\cdot\nabla\log\rho_t^1\end{aligned}$$

allows us to compute

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\partial_{i,j}\log\rho_{t,\theta}^1(x(t)) = {}& -\partial_{i,j}\mathrm{div}f_t(x(t)) - \partial_i\nabla\log\rho_t(x(t))\cdot\partial_j f_t(x(t)) \\ & -\partial_i f_t(x(t))\cdot\partial_j\nabla\log\rho_t(x(t)) - \partial_{i,j}f_t(x(t))\cdot\nabla\log\rho_t(x(t)).\end{aligned}$$

In the above computation, we use the fact that the term $\frac{\partial}{\partial x}\partial_{i,j}\log\rho_t^1(x(t))\cdot f_t(x(t))$ is canceled. ■

Appendix D. Proof of Proposition 4

Proof For compactness, we use $\partial_{i,j,k}$ to denote $\frac{\partial^3}{\partial x_i \partial x_j \partial x_k}$. First, compute that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\partial_{i,j,k}\log\rho^1_{t,\theta}(x(t)) &= \frac{\partial}{\partial t}\partial_{i,j,k}\log\rho^1(t,x(t);\theta) + \frac{\partial}{\partial x}\partial_{i,j,k}\log\rho^1(t,x(t);\theta)\cdot\frac{\mathrm{d}x(t)}{\mathrm{d}t}\\ &= \partial_{i,j,k}\frac{\partial}{\partial t}\log\rho^1(t,x(t);\theta) + \frac{\partial}{\partial x}\partial_{i,j,k}\log\rho^1(t,x(t);\theta)\cdot f_t(x(t)).\end{aligned}$$

Using the Fokker Planck equation (1), we derive

$$\frac{\partial}{\partial t}\log\rho^1_t = -\mathrm{div} f_t - \nabla\log\rho_t\cdot f_t, \tag{48}$$

which together with

$$\begin{aligned}\partial_{i,j,k}(\nabla\log\rho^1_t\cdot f_t) = {}& \partial_{i,j,k}\nabla\log\rho^1_t\cdot f_t + \partial_{i,j}\nabla\log\rho^1_t\cdot\partial_k f_t\\ & \partial_{i,k}\nabla\log\rho^1_t\cdot\partial_j f_t + \partial_i\nabla\log\rho^1_t\cdot\partial_{j,k}f_t\\ & \partial_{j,k}\nabla\log\rho^1_t\cdot\partial_i f_t + \partial_j\nabla\log\rho^1_t\cdot\partial_{i,k}f_t\\ & \partial_k\nabla\log\rho^1_t\cdot\partial_{i,j}f_t + \nabla\log\rho^1_t\cdot\partial_{i,j,k}f_t\end{aligned}$$

allows us to compute

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\partial_{i,j,k}\log\rho^1_t((x(t))) = {}& -\partial_{i,j,k}\mathrm{div} f_t(x(t)) - \partial_{i,j}\nabla\log\rho^1_t(x(t))\cdot\partial_k f_t(x(t))\\ & -\partial_{i,k}\nabla\log\rho^1_t(x(t))\cdot\partial_j f_t(x(t)) - \partial_i\nabla\log\rho^1_t(x(t))\cdot\partial_{j,k}f_t(x(t))\\ & -\partial_{j,k}\nabla\log\rho^1_t(x(t))\cdot\partial_i f_t(x(t)) - \partial_j\nabla\log\rho^1_t(x(t))\cdot\partial_{i,k}f_t(x(t))\\ & -\partial_k\nabla\log\rho^1_t(x(t))\cdot\partial_{i,j}f_t(x(t)) - \nabla\log\rho^1_t(x(t))\cdot\partial_{i,j,k}f_t(x(t)).\end{aligned}$$

In the above computation, we use the fact that the term $\partial_{i,j,k}\nabla\log\rho^1_t(x(t))\cdot f_t(x(t))$ is canceled. ■

Appendix E. Gradient Computation via Adjoint Method

Consider the ODE system

$$\begin{aligned}\dot{s}(t) &= \psi(s(t),t,\theta)\\ s(0) &= s_0,\end{aligned}$$

and the objective loss

$$\ell(\theta) = \int_0^T g(s(t),t,\theta)\mathrm{d}t. \tag{49}$$

The following proposition computes the gradient of ℓ w.r.t. θ. We omit the parameters of the functions for succinctness. We note that all the functions in the integrands should be evaluated at the corresponding time stamp t, e.g. $b^\top\frac{\partial h}{\partial\theta}\mathrm{d}t$ abbreviates for $b(t)^\top\frac{\partial}{\partial\theta}h(\xi(t),x(t),t,\theta)\mathrm{d}t$.

Proposition 12

$$\frac{\mathrm{d}\ell}{\mathrm{d}\theta} = \int_0^T a^\top \frac{\partial \psi}{\partial \theta} + \frac{\partial g}{\partial \theta} \mathrm{d}t. \tag{50}$$

where $a(t)$ is solution to the following final value problems

$$\dot{a}^\top + a^\top \frac{\partial \psi}{\partial s} + \frac{\partial g}{\partial s} = 0, a(T) = 0, \tag{51}$$

Proof Let us define the Lagrange multiplier function (or the adjoint state) $a(t)$ dual to $s(t)$. Moreover, let $\mathcal{L}$ be an augmented loss function of the form

$$\mathcal{L} = \ell - \int_0^T a^\top (\dot{s} - \psi) \mathrm{d}t. \tag{52}$$

Since we have $\dot{s}(t) = \psi(s(t), t, \theta)$ by construction, the integral term in $\mathcal{L}$ is always null and a can be freely assigned while maintaining $\mathrm{d}\mathcal{L}/\mathrm{d}\theta = \mathrm{d}\ell/\mathrm{d}\theta$. Using integral by part, we have

$$\int_0^T a^\top \dot{s} \, \mathrm{d}t = a(t)^\top s(t)|_0^T - \int_0^T s^\top \dot{a} \, \mathrm{d}t. \tag{53}$$

We obtain

$$\mathcal{L} = -a(t)^\top s(t)|_0^T + \int_0^T \dot{a}^\top s + a^\top \psi + g \, \mathrm{d}t. \tag{54}$$

Now we compute the gradient of $\mathcal{L}$ w.r.t. θ as

$$\frac{\mathrm{d}\ell}{\mathrm{d}\theta} = \frac{\mathrm{d}\mathcal{L}}{\mathrm{d}\theta} = -a(T)^\top \frac{\mathrm{d}x(T)}{\mathrm{d}\theta} + \int_0^T \dot{a}^\top \frac{\mathrm{d}s}{\mathrm{d}\theta} + a^\top \left(\frac{\partial \psi}{\partial \theta} + \frac{\partial \psi}{\partial s} \frac{\mathrm{d}s}{\mathrm{d}\theta} \right) \mathrm{d}t + \int_0^T \frac{\partial g}{\partial s} \frac{\mathrm{d}s}{\mathrm{d}\theta} + \frac{\partial g}{\partial \theta} \mathrm{d}t,$$

which by rearranging terms yields to

$$\frac{\mathrm{d}\ell}{\mathrm{d}\theta} = \frac{\mathrm{d}\mathcal{L}}{\mathrm{d}\theta} = -a(T)^\top \frac{\mathrm{d}x(T)}{\mathrm{d}\theta} + \int_0^T a^\top \frac{\partial \psi}{\partial \theta} + \frac{\partial g}{\partial \theta} \mathrm{d}t + \int_0^T \left(\dot{a}^\top + a^\top \frac{\partial \psi}{\partial s} + \frac{\partial g}{\partial s} \right) \frac{\mathrm{d}s}{\mathrm{d}\theta} \mathrm{d}t.$$

Now by taking a satisfying the *final* value problems

$$\dot{a}^\top + a^\top \frac{\partial \psi}{\partial s} + \frac{\partial g}{\partial s} = 0, a(T) = 0, \tag{55}$$

we derive the result

$$\frac{\mathrm{d}\ell}{\mathrm{d}\theta} = \int_0^T a^\top \frac{\partial \psi}{\partial \theta} + \frac{\partial g}{\partial \theta} \mathrm{d}t. \tag{56}$$

■

Appendix F. Proof of Lemma 8

Proof Recall the definition of ρ_t^1 in (13). ρ_t^1 is l-periodic since it can be expressed as a push-forward measure of an l-periodic measure α_0 under an l-periodic map $X(t,\cdot)$. Consequently, $\nabla \log \rho_t^1$ is also l-periodic, which together with the l-periodicity of V shows that the map $Y(t,\cdot)$ is also l-periodic. Following a similar argument, we see that ρ_t^2 is also l-periodic.

To prove that $\|\nabla \log \rho_t^1(x)\|$ is bounded for all $x \in \mathcal{X}$, recall Proposition 2 where we show that for any $x \in \mathcal{X}$

$$\nabla \log \rho_t^1(x) = \nabla \log \alpha_0(x(0)) - \int_0^t \nabla \mathrm{div} f_s(x(s)) + \nabla f_s(x(s))^\top \nabla \log \rho_s^1(x(s)) \mathrm{d}s. \tag{57}$$

Here $x(s)_{s\in[0,t]}$ is the trajectory of the final value problem

$$\frac{\mathrm{d}x(s)}{\mathrm{d}s} = f_s(x(s)), x(t) = x. \tag{58}$$

Using Grönwall's inequality, we can bound

$$\|\nabla \log \rho_t^1(x)\| \le (L_0 + tL_f)\exp(tL_f) \le (L_0 + TL_f)\exp(TL_f). \tag{59}$$

To prove that $\|\nabla \log \rho_t^1(x)\|$ is Lipschitz continuous for all $x \in \mathcal{X}$, recall Proposition 3 where we show that for any $x \in \mathcal{X}$

$$\begin{aligned}\nabla^2 \log \rho_t^1(x) = \nabla^2 \log \alpha_0(x(0)) - \int_0^t \nabla^2 \mathrm{div} f_s(x(s)) + \left(\nabla^2 \log \rho_s^1(x(s))\right)^\top \mathcal{J}_{f_s}(x(s)) \\ + \left(\mathcal{J}_{f_s}(x(s))\right)^\top \nabla^2 \log \rho_s^1(x(s)) + \nabla^2 f_s(x(s)) \otimes_1 \nabla \log \rho_s^1(x(s)) \mathrm{d}s,\end{aligned}$$

where $x(s)$ is the trajectory defined in (58), $\mathcal{J}_f$ denotes the Jacobian matrix of a vector valued function f, and

$$\nabla^2 f_s(x(s)) \otimes_1 \nabla \log \rho_s^1(x(s)) = \begin{bmatrix} \nabla^2 (f_s)_{[1]}(x(s)) \nabla \log \rho_s^1(x(s)) \\ \cdots \\ \nabla^2 (f_s)_{[d]}(x(s)) \nabla \log \rho_s^1(x(s)) \end{bmatrix} \in \mathbb{R}^{d\times d}. \tag{60}$$

Here $f_{[i]}$ denotes the ith entry of a vector valued function f. We can bound the spectral norm $\|\nabla^2 \log \rho_t^1(x)\|_{op}$ by (note that $x(t) = x$)

$$\begin{aligned}\|\nabla^2 \log \rho_t^1(x(t))\|_{op} &\le L_0 + \int_0^t L_f + 2L_f \|\nabla^2 \log \rho_s^1(x(s))\|_{op} + L_f B_1 \mathrm{d}s \\ &= L_0 + t(L_f + L_f B_1) + \int_0^t 2L_f \|\nabla^2 \log \rho_s^1(x_s)\| \mathrm{d}s,\end{aligned}$$

where we denote $B_1 = (L_0 + TL_f)\exp(TL_f)$. Use Grönwall's inequality to derive

$$\|\nabla^2 \log \rho_t^1(x)\|_{op} \le (L_0 + t(L_f + B_1 L_f))\exp(2tL_f) \le (L_0 + T(L_f + B_1 L_f))\exp(2TL_f), \tag{61}$$

To see that $\|\nabla \mathcal{A}[f]_t\|_{op}$ is bounded over $\mathcal{X}$, observe that

$$\nabla \mathcal{A}[f]_t = -\nabla^2 V_t - \nabla^2 \log \rho_t^1, \tag{62}$$

which is bounded due to Assumption 3 and (61). To see that $\nabla\mathcal{A}[f]_t$ is Lipschitz continuous, we need to prove that the spectral norm of the following tensor is bounded

$$\nabla^2\mathcal{A}[f]_t = -\nabla^3 V_t - \nabla^3 \log \rho_t^1. \tag{63}$$

The first term is bounded due to Assumption 3. To bound the second term, use Proposition 4 to bound (note that $x(t) = x$)

$$\begin{aligned}\|\nabla^3 \log \rho_t^1(x(t))\|_{op} \leq\ & \|\nabla^3 \log \alpha_0(x(0))\|_{op} \\ &+ \int_0^t \|\nabla^3 \mathrm{div} f_s(x(s))\|_{op} + 3\|\nabla^2 f_s(x(s))\|_{op}\|\nabla^2 \log \rho_s^1(x(s))\|_{op} \\ &+ 3\|\nabla f_s(x(s))\|_{op}\|\nabla^3 \log \rho_s^1(x(s))\|_{op} + \|\nabla \log \rho_s^1(x(s))\|\|\nabla^3 f_s(x(s))\|_{op}\mathrm{d}s,\end{aligned}$$

Using Grönwall's inequality, we can bound

$$\begin{aligned}\|\nabla^3 \log \rho_t^1(x)\|_{op} &\leq (L_0 + t(L_f + B_2 L_f + B_1 L_f)) \exp(3tL_f) \\ &\leq (L_0 + T(L_f + B_2 L_f + B_1 L_f)) \exp(3TL_f),\end{aligned}$$

where we denote $B_2 = 3(L_0 + T(L_f + B_1 L_f)) \exp(2TL_f)$.

The boundedness of $\|\nabla \mathrm{div}\mathcal{A}[f]_t(x)\|$ and the Lipschitz continuity of $\nabla \mathrm{div}\mathcal{A}[f]_t$ hold following the same argument above under the assumptions 1 to 3. ■

F.1. Proof of Lemma 9

Proof In this proof, for simplicity of the notation, we use ρ_t^1 and ρ_t^2 to denote the probability density functions of systems (1) and (2) and use X_t and Y_t to denote the corresponding particle maps.

The Wasserstein-2 metric between ρ_t^1 and ρ_t^2 can be written as:

$$W_2^2(\rho_t^1, \rho_t^2) = \inf_{P:\ P_\sharp \rho_t^1 = \rho_t^2} \int_{\mathcal{X}} \|x - P(x)\|^2 \mathrm{d}\rho_t^1(x),$$

where the infimum is taken over all the pushforward maps P such that $P_\sharp \rho_t^1 = \rho_t^2$. From the Lipschitz continuity of the velocity field f in Assumption 2, the particle map X_t of System (1) is invertible. Moreover, recall that Systems (1) and (2) have the same initial distribution α_0. We have an upper bound on $W_2^2(\rho_t^1, \rho_t^2)$ by considering a special map $P_{t,\theta} = Y_t \circ X_t^{-1}$, where we use X_t and Y_t to denote the particle maps of systems (1) and (2) compactly (see Table 1). We have the feasibility of $P_{t,\theta}$ by the definitions of ρ_t^1 and ρ_t^2,

$$P_{t,\theta}\sharp\rho_t^1 = Y_t\sharp(X_t^{-1} \circ X_t)\sharp\alpha_0 = \rho_t^2. \tag{64}$$

Additionally, we have that $\|x - P_{t,\theta}(x)\|$ is l-periodic:

$$\begin{aligned}\|x + le_i - P_{t,\theta}(x + le_i)\| = \|x + le_i - Y_t \circ X_t^{-1}(x + le_i)\| &\overset{(1)}{=} \|x + le_i - Y_t(X_t^{-1}(x) + le_i)\| \\ &\overset{(2)}{=} \|x + le_i - (Y_t(X_t^{-1}(x)) + le_i)\| = \|x - P_{t,\theta}(x)\|,\end{aligned}$$

where in (1) we use $X_t^{-1}(x + le_i) = X_t^{-1}(x) + le_i$ since

$$\begin{aligned} & X_t(X_t^{-1}(x + le_i) - le_i) - (X_t^{-1}(x + le_i) - le_i) = X_t(X_t^{-1}(x + le_i)) - X_t^{-1}(x + le_i) \\ \Leftrightarrow\ & X_t(X_t^{-1}(x + le_i) - le_i) = x \qquad \Rightarrow X_t^{-1}(x + le_i) = X_t^{-1}(x) + le_i, \end{aligned}$$

and in (2) we use $Y_t(a + le_i) = Y_t(a) + le_i$ following a similar argument. Therefore, we can bound

$$\begin{aligned} W_2^2(\rho_t^1, \rho_t^2) \le \int_{\mathcal{X}} \|x - P_{t,\theta}(x)\|^2 \mathrm{d}\rho_t^1(x) &= \int_{\mathcal{X}} \|X_t(x) - Y_t(x)\|^2 \mathrm{d}\alpha_0(x) \\ &= \int_{\mathcal{X}} \|x_t - y_t\|^2 \mathrm{d}\alpha_0(x_0), \end{aligned}$$

where we used the change-of-variables formula of the push-forward measure from Lemma 1 in the first equality and $\{x_t\}_{t\in[0,T]}$ and $\{y_t\}_{t\in[0,T]}$ are the trajectory of particles initialized from x_0 but driven by Systems (1) and (2) respectively. Hence, we can bound the Wasserstein-2 distance between the trajectory of probability distributions by studying the distance between the particles driven by the two systems, which is proved to be bound by $R(f)$ in expectation ($x_0 \sim \alpha_0$) in the following.

Suppose two particles are initialized from the same position x_0, but follow System (1) and System (2) respectively. The change of their distance at time t can be computed by

$$\begin{aligned} & \frac{d}{dt}\|x_t - y_t\|^2 = 2\,(x_t - y_t)^\top \left(\frac{\mathrm{d}x_t}{\mathrm{d}t} - \frac{\mathrm{d}y_t}{\mathrm{d}t}\right) = 2\,(x_t - y_t)^\top \left(f(t, x_t) - \mathcal{A}[f](t, y_t)\right) \\ &= 2\,(x_t - y_t)^\top \left(f(t, x_t) - \mathcal{A}[f](t, x_t)\right) + 2\,(x_t - y_t)^\top \left(\mathcal{A}[f](t, x_t) - \mathcal{A}[f](t, y_t)\right) \\ &\le 2\|x_t - y_t\|^2 + \|f(t, x_t) - \mathcal{A}[f](t, x_t)\|^2 + \|\mathcal{A}[f](t, x_t) - \mathcal{A}[f](t, y_t)\|^2, \end{aligned}$$

where $\mathcal{A}[f]$ is the velocity field of System (2) and the transformation $\mathcal{A}$ is defined in equation (14). Bound the the last term on the RHS can be bounded by $L_v^2\|x_t - y_t\|^2$ using the Lipschitz continuity of $\mathcal{A}[f]$ in Lemma 8 to derive

$$\begin{aligned} & \frac{d}{dt}\|x_t - y_t\|^2 \le (2 + L_v^2)\|x_t - y_t\|^2 + \|f(t, x_t) - \mathcal{A}[f](t, x_t)\|^2 \\ \Rightarrow\ & \frac{d}{dt}\exp(-t(2 + L_v^2))\|x_t - y_t\|^2 \le \exp(-t(2 + L_v^2))\|f(t, x_t) - \mathcal{A}[f](t, x_t)\|^2 \end{aligned}$$

Integrate from $t = 0$ to τ. By noting that $x_0 = y_0$ and $\exp\left(-(2 + L_v^2)t\right) < 1$, we have

$$\exp\left(-(2 + L_v^2)\tau\right)\|x_\tau - y_\tau\|^2 \le \int_0^\tau \|f(t, x_t) - \mathcal{A}[f](t, x_t)\|^2 \mathrm{d}t.$$

Take expectation with respect to $x_0 \sim \alpha_0$. We derive that for any $\tau \in [0, T]$

$$W_2^2(\rho_t^1, \rho_t^2) \le \int_{\mathcal{X}} \|x_\tau - y_\tau\|^2 \mathrm{d}\alpha_0(x_0) \le \exp\left((2 + L_v^2)T\right) R(f). \tag{65}$$

■

F.2. Proof of Lemma 10

Proof For compactness, in this proof, we denote $f_t(x) = f(t,x)$ and $\mathcal{A}[f]_t(x) = \mathcal{A}[f](t,x)$. We use ρ_t^1 and ρ_t^2 to denote the probability density functions of systems (1) and (2) and use X_t and Y_t to denote the corresponding particle maps (see Table 1).

Since both $\nabla \log \rho_t^1$ and $\nabla \log \rho_t^2$ are l-periodic, using the change of variable formula in Lemma 1 , we have

$$\xi_t = \|\nabla \log \rho_t^1 \circ Y_t - \nabla \log \rho_t^2 \circ Y_t\|_{\alpha_0}^2 \tag{66}$$

Denote $y_t = Y_t(y_0)$ and $x_t = X_t(x_0)$ with $y_0 = x_0$. For any x_0, we have

$$\left(\nabla \log \rho_t^1 \circ Y_t\right)(x_0) = \left(\nabla \log \rho_t^1(y_t) - \nabla \log \rho_t^1(x_t)\right) + \nabla \log \rho_t^1(x_t). \tag{67}$$

Hence ξ_t can be bounded by

$$\xi_t \leq \|\nabla \log \rho_t^1(y_t) - \nabla \log \rho_t^1(x_t)\|_{\alpha_0}^2 + \|\nabla \log \rho_t^1(x_t) - \nabla \log \rho_t^2(y_t)\|_{\alpha_0}^2. \tag{68}$$

The first term is of the order $O(\|x_t - y_t\|_{\alpha_0^2}^2)$ from the Lipschitz continuity of $\nabla \log \rho_t^1$. To bound the second term, note that $\nabla \log \rho_t^1(x_t)$ can be computed from from Proposition 2,

$$\nabla \log \rho_t^1(x_t) = \nabla \log \alpha_0(x_0) - \int_0^t \nabla \mathrm{div}\left(f_\tau(x_\tau)\right) + \left[\nabla f_\tau(x_\tau)\right]^\top \nabla \log \rho_t^1(x_\tau) \mathrm{d}\tau$$

and that $\left(\nabla \log \rho_t^2 \circ Y_t\right)(y_0) = \nabla \log \rho_t^2(y_t)$ can be similarly computed as

$$\nabla \log \rho_t^2(y_t) = \nabla \log \alpha_0(y_0) - \int_0^t \nabla \mathrm{div}\left(\mathcal{A}[f]_\tau(y_\tau)\right) + \left[\nabla \mathcal{A}[f]_\tau(y_\tau)\right]^\top \nabla \log \rho_t^2(y_\tau) \mathrm{d}\tau.$$

Hence, the second term can be decomposed as follows:

$$\begin{aligned}\nabla \log \rho_t^1(x_t) - \nabla \log \rho_t^2(y_t) = &\int_0^t \underbrace{\nabla \mathrm{div}\left(\mathcal{A}[f]_\tau(y_\tau)\right) - \nabla \mathrm{div}\left(f_\tau(x_\tau)\right)}_{A_\tau} \mathrm{d}\tau \\ &+ \int_0^t \underbrace{\left[\nabla \mathcal{A}[f]_\tau(y_\tau)\right]^\top \nabla \log \rho_t^2(y_\tau) - \left[\nabla f_\tau(x_\tau) x\right]^\top \nabla \log \rho_t^1(x_\tau)}_{B_\tau} \mathrm{d}\tau.\end{aligned}$$

Recall that $\delta_\tau = f_\tau - \mathcal{A}[f]_\tau$ in (15). To bound the norm of A_τ, we have

$$A_\tau = \nabla \mathrm{div}\left(\mathcal{A}[f]_\tau(y_\tau)\right) - \nabla \mathrm{div}\left(\mathcal{A}[f]_\tau(x_\tau)\right) + \nabla \mathrm{div}(\delta_\tau(x_\tau))$$

and hence using $x_\tau \sim \rho_\tau^1 = X_\tau \sharp \alpha_0$ and the Lipschitz continuity of $\nabla \mathrm{div} \mathcal{A}[f]_t$ we have

$$\|A_\tau\|_{\alpha_0}^2 = O(\|y_\tau - x_\tau\|_{\alpha_0}^2 + \|\nabla \mathrm{div}(\delta_\tau)\|_{\rho_\tau^1}^2).$$

To bound the norm of B_τ, note that

$$\begin{aligned}B_\tau = &\nabla \mathcal{A}[f]_\tau(y_\tau)^\top \nabla \log \rho_\tau^2(y_\tau) - \nabla \mathcal{A}[f]_\tau(y_\tau)^\top \nabla \log \rho_\tau^1(y_\tau) && \text{(a)} \\ &+ \nabla \mathcal{A}[f]_\tau(y_\tau)^\top \nabla \log \rho_\tau^1(y_\tau) - \nabla \mathcal{A}[f]_\tau(x_\tau)^\top \nabla \log \rho_\tau^1(y_\tau) && \text{(b)} \\ &+ \nabla \mathcal{A}[f]_\tau(x_\tau)^\top \nabla \log \rho_\tau^1(y_\tau) - \nabla f_\tau(x_\tau)^\top \nabla \log \rho_\tau^1(y_\tau) && \text{(c)} \\ &+ \nabla f_\tau(x_\tau)^\top \nabla \log \rho_\tau^1(y_\tau) - \nabla f_\tau(x_\tau)^\top \nabla \log \rho_\tau^1(x_\tau) && \text{(d)}\end{aligned}$$

Using the boundedness of ∇f_τ and the Lipschitz continuity of $\nabla \log \rho_\tau^1$, we have $\|d\|_{\alpha_0}^2 = O(\|x_\tau - y_\tau\|_{\alpha_0}^2)$. Similarly, we have $\|b\|_{\alpha_0}^2 = O(\|x_\tau - y_\tau\|_{\alpha_0}^2)$. Note that

$$c = -\nabla \delta_\tau(x_\tau)^\top \nabla \log \rho_\tau^2(y_\tau). \tag{69}$$

Using the boundedness of $\nabla \log \rho_t^2$, we have

$$\|c\|_{\alpha_0}^2 = O(\|\nabla \delta_\tau\|_{\rho_t^1}^2). \tag{70}$$

Finally, using the boundedness of $\nabla \mathcal{A}[f]_\tau$, we have that

$$\|a\|_{\alpha_0}^2 \le L_v \|\nabla \log \rho_\tau^2 \circ Y_\tau - \nabla \log \rho_\tau^1 \circ Y_\tau\|_{\alpha_0}^2 = L_v \|\nabla \log \rho_\tau^2 - \nabla \log \rho_\tau^1\|_{\rho_\tau^{(2)}}^2 = L_v \xi_\tau. \tag{71}$$

Therefore, by noting that

$$\|\nabla \log \rho_t^1(x_t) - \nabla \log \rho_t^2(y_t)\|_{\alpha_0}^2 \le \int_0^t \|A_\tau\|_{\alpha_0}^2 + \|B_\tau\|_{\alpha_0}^2 \mathrm{d}\tau, \tag{72}$$

we bound (note that $\|\delta_\tau\|_{\rho_\tau^1} = \|\delta_\tau \circ X_\tau\|_{\alpha_0}$)

$$\begin{aligned}
\xi_t &\le \int_0^t O(\|y_\tau - x_\tau\|_{\alpha_0}^2 + \|\nabla \mathrm{div}(\delta_\tau)\|_{\rho_\tau^1}^2 + \|\nabla \delta_\tau\|_{\rho_\tau^1}^2) + L_v \xi_\tau \mathrm{d}\tau \\
&\le \int_0^t O(\|\delta_\tau\|_{\alpha_0}^2 + \|\nabla \mathrm{div}(\delta_\tau)\|_{\rho_\tau^1}^2 + \|\nabla \delta_\tau\|_{\rho_\tau^1}^2) + L_v \xi_\tau \mathrm{d}\tau \\
&\le \int_0^t O(R(f)) + L_v \xi_\tau \mathrm{d}\tau
\end{aligned}$$

where we use Lemma 9 in the second inequality. Using the Grönwall's inequality of the integral form for continuous functions, we have there exists some constant $\bar{C}(T)$ such that

$$\xi_t \le \bar{C}(T) R(f) \exp(t L_v) \le \bar{C}(T) R(f) \exp(T L_v) \tag{73}$$

Integrating τ from 0 to t, we have for any $t \in [0, T]$

$$\int_0^t \xi_\tau \mathrm{d}\tau \le \bar{C}(T) T \exp(T L_v) R(f) = C(T) R(f), \tag{74}$$

where we denote $C(T) = \bar{C}(T) T \exp(T L_v)$. ■