SIAM J. NUMER. ANAL.
Vol. 60, No. 3, pp. 1385–1449

NEURAL PARAMETRIC FOKKER–PLANCK EQUATION*

SHU LIU†, WUCHEN LI‡, HONGYUAN ZHA§, AND HAOMIN ZHOU†

Abstract. In this paper, we develop and analyze numerical methods for high-dimensional Fokker–Planck equations by leveraging generative models from deep learning. Our starting point is a formulation of the Fokker–Planck equation as a system of ordinary differential equations (ODEs) on finite-dimensional parameter space with the parameters inherited from generative models such as normalizing flows. We call such ODEs *neural parametric Fokker–Planck equations*. The fact that the Fokker–Planck equation can be viewed as the L^2-Wasserstein gradient flow of Kullback–Leibler (KL) divergence allows us to derive the ODEs as the constrained L^2-Wasserstein gradient flow of KL divergence on the set of probability densities generated by neural networks. For numerical computation, we design a variational semi-implicit scheme for the time discretization of the proposed ODE. Such an algorithm is sampling-based, which can readily handle the Fokker–Planck equations in higher dimensional spaces. Moreover, we also establish bounds for the asymptotic convergence analysis of the neural parametric Fokker–Planck equation as well as the error analysis for both the continuous and discrete versions. Several numerical examples are provided to illustrate the performance of the proposed algorithms and analysis.

Key words. optimal transport, transport information geometry, deep learning, neural parametric Fokker–Planck equation, implicit Euler scheme, numerical analysis

AMS subject classifications. 65M15, 68T07, 49Q22

DOI. 10.1137/20M1344986

1. Introduction.

The Fokker–Planck equation is a parabolic partial differential equation (PDE) that plays a crucial role in stochastic calculus, statistical physics, biology, and many other disciplines [44, 55, 59]. Recently, it has seen many applications in machine learning as well [39, 52, 64]. The Fokker–Planck equation describes the evolution of probability density of a stochastic differential equation (SDE). In this paper, we mainly focus on the following linear Fokker–Planck equation:

$$\frac{\partial \rho(t,x)}{\partial t} = \nabla \cdot (\rho(t,x)\nabla V(x)) + D\Delta \rho(t,x), \quad \rho(0,x) = p(x), \tag{1.1}$$

where $x \in \mathbb{R}^d$, $V\colon \mathbb{R}^d \to \mathbb{R}$ is a given potential function, $D > 0$ is a diffusion coefficient, and $p(x)$ is the initial (or reference) density function. In numerical algorithms, there exist several classical methods [54] such as finite difference [14] or finite element [29] for solving the Fokker–Planck equation. Most of the existing methods are grid based, which may be able to approximate the solution accurately if the grid sizes become small. However, they find limited usage in high-dimensional problems, especially for

*Received by the editors June 16, 2020; accepted for publication (in revised form) January 24, 2022; published electronically June 14, 2022.
https://doi.org/10.1137/20M1344986
Funding: This work was partially supported by National Science Foundation grants DMS-1620345 and DMS-1830225 and by ONR grant N000141310408. The work of the second author was supported by a start-up fund from the University of South Carolina and NSF grant RTG:2038080. The work of the third author was partially supported by a grant from Shenzhen Research Institute of Big Data.

†School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332 USA (sliu459@gatech.edu, hmzhou@math.gatech.edu).
‡Department of Mathematics, University of South Carolina, Columbia, SC 29208 USA (wuchen@mailbox.sc.edu).
§School of Data Science, Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen, China, 518172 (zhahy@cuhk.edu.cn).

$d > 3$, because the number of unknowns grows exponentially fast as the dimension increases. This is known as the curse of dimensionality. The main goal of this paper is providing an alternative strategy, with provable error estimates, to solve high-dimensional Fokker–Planck equations.

1.1. Neural parametric Fokker–Planck equation. To overcome the challenges imposed by high dimensionality, we leverage the generative models in machine learning [58] and a new interpretation of the Fokker–Planck equation in the theory of optimal transport [68]. We first introduce the Kullback–Leibler (KL) divergence, also known as relative entropy, defined by

$$\mathcal{D}_{\mathrm{KL}}(\rho||\rho_*) = \int_{\mathbb{R}^d} \rho(x) \log\left(\frac{\rho(x)}{\rho_*(x)}\right) dx, \quad \rho_*(x) = \frac{1}{Z_D} e^{-\frac{V(x)}{D}}, \text{ with } Z_D = \int_{\mathbb{R}^d} e^{-\frac{V(x)}{D}}\, dx.$$

Here $\rho_*(x)$ is the Gibbs distribution. A well-known fact is that the Fokker–Planck equation (1.1) can be viewed as the gradient flow of the functional $D\ \mathcal{D}_{\mathrm{KL}}(\rho||\rho_*)$ on the probability space $\mathcal{P}$ equipped with Wasserstein metric g^W [23, 46]. Recently, this line of research was extended to parameter space in the field of information geometry [2, 3, 6], leading to an emergent area called transport information geometry [33, 38, 36, 37].

Inspired by the aforementioned works, we study the Fokker–Planck equation defined on parameter manifold (space) $\Theta \subset \mathbb{R}^m$ equipped with metric tensor G, which is obtained by pulling back the Wasserstein metric g^W to Θ. Here the metric tensor G can be viewed as an $m \times m$ matrix that contains all the metric information on Θ. In this paper, we focus on the parameter space from generative models using neural networks. Our train of thought can be summarized as following. We start with a given reference distribution p and consider a suitable family of parametric maps $\{T_\theta\}_{\theta\in\Theta}$. Such $T_\theta : \mathbb{R}^d \to \mathbb{R}^d$ is also called a parametric pushforward map since it generates a family of parametric distributions $\{T_{\theta\sharp}p\}$ by pushing p forward using T_θ (see Definition 3.2). Then we consider the map $T_{(\cdot)\sharp} : \Theta \to \mathcal{P}, \theta \mapsto T_{\theta\sharp}p$, which can be treated as an immersion from the parameter manifold Θ to the probability manifold $\mathcal{P}$. We derive the metric tensor $G(\theta)$ by pulling back the Wasserstein metric via $T_{(\cdot)\sharp}$. Once establishing (Θ, G), we can compute the G-gradient flow of function $H(\theta) = D\ \mathcal{D}_{\mathrm{KL}}(T_{\theta\sharp}p \,||\, \rho_*)$ defined on the parameter manifold. This leads to an ODE system that can be viewed as a parametric version of the Fokker–Planck equation:

$$\dot{\theta}_t = -G(\theta_t)^{-1}\nabla_\theta H(\theta_t). \tag{1.2}$$

Here (and for the rest of the paper) the dot symbol $\dot{\theta}$ stands for time derivative $\frac{d\theta_t}{dt}$. Using the pushforward $\rho_\theta = T_{\theta\sharp}p$, in which θ is the solution of (1.2), we can approximate the solution ρ_t in (1.1).

There are many potential applications for the parametric Fokker–Planck equation. For example, the solution of (1.2) can be immediately used for sampling, which is a crucial task in statistics and machine learning. To be more precise, if the goal is drawing a large number of samples from ρ_t at N different time instances $\{t_1, t_2, \ldots, t_N\}$ along the solution of (1.1), we can acquire N sets of parameters $\theta_{t_1}, \ldots, \theta_{t_N}$ from the solution of (1.2), which provide N pushforward maps $T_{\theta_{t_1}}, \ldots, T_{\theta_{t_N}}$. Thus the desired samples at time t_k are $\{T_{\theta_{t_k}}(\boldsymbol{Z}_1), \ldots, T_{\theta_{t_k}}(\boldsymbol{Z}_M)\}$, in which $\{\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_M\}$ are samples drawn from the reference distribution p. If needed, the pushforward maps can be conveniently reused to generate more samples with negligible additional cost.

1.2. Computational method. For the computation of (1.2), we want to point out that metric tensor $G(\theta)$ doesn't have an explicit form, and thus the direct com-

putation of $G(\theta)^{-1}\nabla_\theta H(\theta)$ is not tractable. To deal with this issue, we design a numerical algorithm based on the semi-implicit Euler scheme of (1.2) with time step size h. To be more precise, at each time step, the algorithm seeks to solve the following double-minimization problem:

(1.3)
$$\min_{\theta}\left\{\left(\int\left(2\ \nabla\phi(x)\cdot((T_\theta - T_{\theta_k})\circ T_{\theta_k}^{-1}(x)) - |\nabla\phi(x)|^2\right)\rho_{\theta_k}(x)\ dx\right) + 2hH(\theta)\right\},$$
$$\text{with the constraint: } \phi \text{ solves } \min_{\phi}\left\{\int|\nabla\phi(x) - ((T_\theta - T_{\theta_k})\circ T_{\theta_k}^{-1}(x))|^2\rho_{\theta_k}(x)\ dx\right\}.$$

Here ρ_{θ_k} is the density of the pushed forward distribution $T_{\theta_k\sharp}p$ (cf. Definition 3.2), and $\phi\colon \mathbb{R}^d \to \mathbb{R}$ is the Kantorovich dual potential variable for constrained probability models in optimal transport theory. Hence (1.3) is derived following the semi-implicit Euler scheme in the dual variable. The advantage of using this formulation is that it allows us to design an efficient implementation, purely based on sampling techniques which are computationally friendly in high-dimensional problems, to compute the solution of the parametric Fokker–Planck equation (1.2). In our implementation, we endow the pushforward map T_θ with certain kinds of deep neural networks known as normalizing flows [58], because they are friendly to our scheme evaluations. The dual variable ϕ in the inner maximization is parametrized by the deep rectified linear unit (ReLU) networks [53]. Once the network structures for T_θ and ϕ are chosen, the optimizations are carried out by stochastic gradient descent method [62], in which all terms involved can be computed using samples from the reference distribution p. We stress that this is critical in scaling up the computation in high dimensions. It is worth mentioning that we use neural network as a computational tool without any actual data. Such "data-poor" computation is in significant contrast to the mainstream of deep learning research.

1.3. Major innovations of the proposed method. There are two main innovative points regarding our proposed method:

- (Dimension reduction.) Reducing the high-dimensional evolution PDE to a finite-dimensional ODE system on parameter space. Equivalently, we use the dynamics in a finite dimension to approximate the density evolution of particles that follow the Vlasov-type SDE
$$\dot{\boldsymbol{X}}_t = -\nabla V(\boldsymbol{X}_t) - D\nabla\log\rho_t(\boldsymbol{X}_t),$$
$$\rho_t \text{ is the density function of distribution of } \boldsymbol{X}_t.$$
Here D is the diffusion coefficient as mentioned in (1.1). The density function ρ_t corresponds to the Fokker–Planck equation (1.1).
- (Sampling-friendly.) We distill the information of ρ_t into parameters $\{\theta_t\}$ by solving the parametric Fokker–Planck equation (1.2). By doing so, we are able to obtain an efficient sampling technique to generate samples from ρ_t for any time step t. To be more precise, once we have applied our algorithm to solve (1.2) for the time-dependent parameters $\{\theta_t\}$, we can then generate samples from ρ_t by pushing forward the samples drawn from a reference distribution p using the pushforward map T_{θ_t} with very little computational cost. Such "implementing once for free future uses" mechanism is one of the significant advantages of our proposed algorithm. It is worth mentioning that in the view of both theoretical derivation and numerical implementation, our method is

very different from Langevin Monte Carlo (LMC, MALA) methods [19, 60], which aim at targeting the stationary distribution of the SDE associated to (1.1), or moment methods [55], which focus on keeping track of certain statistical information of the density ρ_t.

1.4. Sketch of numerical analysis. In addition to the method proposed for solving (1.1), we also conducted a mathematical analysis on (1.2) and our algorithm. We established asymptotic convergence and error estimates for the parametric Fokker–Planck equation (1.2), which are summarized in the following two theorems.

THEOREM 5.1 (asymptotic convergence). *Consider the Fokker–Planck equation* (1.1) *with potential* V *and diffusion coefficient* D. *Suppose* V *can be decomposed as* $V = U + \phi$ *with* $U \in \mathcal{C}^2(\mathbb{R}^d)$, $\nabla^2 U \succeq KI$[1] *with* $K > 0$ *and* $\phi \in L^\infty(\mathbb{R}^d)$, *and* $\{\theta_t\}$ *solves* (1.2). *Then the following inequality holds:*

$$\mathcal{D}_{KL}(\rho_{\theta_t}\|\rho_*) \leq \frac{\delta_0}{\tilde{\lambda}_D D^2}(1 - e^{-D\tilde{\lambda}_D t}) + \mathcal{D}_{KL}(\rho_{\theta_0}\|\rho_*)e^{-D\tilde{\lambda}_D t},$$

where ρ_* *is the Gibbs distribution,* $\tilde{\lambda}_D > 0$ *is a constant related to the potential function* V *and* D, *and* δ_0 *is a constant depending on the approximation power of pushforward map* T_θ.

THEOREM 5.11 (approximation error). *Consider the Fokker–Planck equation* (1.1) *with potential* V, *diffusion coefficient* D, *and initial density* ρ_0. *Assume that* λ *is a lower bound of Hessian of potential* V, *i.e.,* $\nabla^2 V \succeq \lambda I$, δ_0 *is defined in Theorem* 5.1, $E_0 = W_2(\rho_{\theta_0}, \rho_0)$, *and* $\delta_0, E_0 \ll 1$; *then the following uniform bounds for the* L^2*-Wasserstein error* $W_2(\rho_{\theta_t}, \rho_t)$ *hold:*

- *When* $\lambda > 0$, $W_2(\rho_{\theta_t}, \rho_t) \leq \max\{\sqrt{\delta_0}/\lambda, E_0\} \sim O(\sqrt{\delta_0} + E_0)$.
- *When* $\lambda = 0$, $W_2(\rho_{\theta_t}, \rho_t) \leq \frac{\sqrt{\delta_0}}{\mu_D} \log \frac{B}{\sqrt{\delta_0}+E_0} + E_0 \sim O(\sqrt{\delta_0} \log \frac{1}{\sqrt{\delta_0}+E_0} + E_0)$.
- *When* $\lambda < 0$, $W_2(\rho_{\theta_t}, \rho_t) \leq A\sqrt{\delta_0} + C\left(E_0 + \sqrt{\delta_0}/|\lambda|\right)^\alpha \sim O((E_0 + \sqrt{\delta_0})^\alpha)$.

Here δ_0 *is a constant depending on the approximation power of pushforward map* T_θ. $\mu_D, A, B, C > 0$ *are constants depending only on* V, D, ρ_0, θ_0. $\alpha = \frac{\mu_D}{|\lambda|+\mu_D}$ *is a certain exponent between* 0 *and* 1.

This result reveals that the difference between the solutions of the parametric Fokker–Planck equation (1.2) and the original equation (1.1), measured by their Wasserstein distance $W_2(\rho_{\theta_t}, \rho_t)$, has a *uniformly* small upper bound if both the initial error E_0 and δ_0 are small enough. Most of the techniques used in our analysis for establishing such a result rely on the theory of optimal transport and Wasserstein manifold, which are still not commonly used for numerical analysis in the relevant literature.

Besides error analysis for the continuous version of (1.2), we are able to provide the order of W_2-error for the numerical scheme when (1.2) is computed at discrete time by numerical schemes. To be more precise, if we apply the forward Euler scheme to (1.2) and compute $\{\theta_k\}$ at different time nodes $\{t_k\}$, we can show that the error at t_k: $W_2(\rho_{\theta_k}, \rho_{t_k})$ is of order $O(\sqrt{\delta_0}) + O(Ch) + O(E_0)$ for finite time t. This is summarized in the following theorem.

THEOREM 5.14 (error for discrete scheme). *Assume that* $\{\rho_t\}_{t\geq 0}$ *is the solution of* (1.1) *with potential satisfying* $\lambda I \preceq \nabla^2 V \preceq \Lambda I$, *and* $\{\theta_k\}_{k=0}^N$ *is the numerical solution of* (1.2) *at time nodes* $t_k = kh$ *for* $k = 0, 1, \ldots, N$ *computed by the forward*

[1] The matrix $\nabla^2 U(x) - KI_{d\times d}$ is nonnegative definite for any $x \in \mathbb{R}^d$.

Euler scheme with time step h. Recalling δ_0 as mentioned in Theorem 5.1*, we denote $E_0 = W_2(\rho_{\theta_0}, \rho_0)$, giving us*

$$W_2(\rho_{\theta_k}, \rho_{t_k}) \le (\sqrt{\delta_0}h + Ch^2)\frac{(1 - e^{-\lambda t_k})}{1 - e^{-\lambda h}} + e^{-\lambda t_k}E_0 \sim O(\sqrt{\delta_0}) + O(Ch) + O(E_0), \quad 0 \le k \le N,$$

where C is a constant depending on N and h.

This indicates that the W_2-error is dominated by three different terms: $O(\sqrt{\delta_0})$ is the intrinsic error originated from the approximation mechanism of the parametric Fokker–Planck equation; the $O(Ch)$ term is induced by the time discretization; and the $O(E_0)$ term is the initial error. We further prove that the difference between the forward Euler scheme and our semi-implicit Euler scheme is of order $O(h^2)$, which implies that the proposed semi-implicit Euler scheme can achieve error bounds similar to the one presented in Theorem 5.14.

It is worth mentioning that we establish Theorem 5.14 based on totally different techniques than those used for Theorem 5.11. Since the ODE (1.2) contains the term $G(\theta)^{-1}$, which is hard to handle using traditional strategies, we interpret it as a particle system governed by stochastic differential equations (SDEs) of Vlasov type and obtain the analysis results shown in Theorem 5.14.

1.5. Literature review. Numerous works exist for solving the Fokker–Planck equations. A finite difference scheme is proposed in [14] that preserves the equilibrium of the original equation. A more general class of equations possessing Wasserstein gradient flow structures is solved in [12] in which the method is based on a space discretization of a proximal-type scheme (also known as the JKO method [23]). Besides direct solutions, particle simulation techniques also serve as an efficient way of solving the equation. The "blob" method proposed in [11] solves the equations by evolving a certain interacting particle system. A related swarming system is also studied in [32, 13, 27, 21, 10]. In [41], the authors propose another type of interacting system in order to approximate $\nabla \log \rho$, which plays the role of the diffusion term in the Fokker–Planck equation, with higher accuracy and less fluctuation. In [50, 57], the authors mainly focus on exploiting the gradient flow structure, i.e., a particle discretization of the Fokker–Planck equation, to deal with Bayesian inference problems.

In addition to the literature focusing on solving the Fokker–Planck equations, there are existing works on applying neural networks to solve PDEs of various types in high-dimensional spaces [70, 56, 24, 25, 73, 45]. Among the listed works, algorithms for general types of high-dimensional PDEs are provided in [56, 24]; a sampling-friendly method is proposed in [45] to deal with the general optimal control problem of diffusion processes. This is equivalent to solving an associated Hamilton–Jacobi–Bellman equation, and such technique can also be applied to importance sampling and rare event simulation. Moreover, numerical methods for high-dimensional parabolic PDEs, to which the Fokker–Planck equation belongs, are studied in [70] and [25]. Our approach differs from these existing works in many aspects, including motivations, strategies, and the associated numerical analysis.

For example, in [70], the authors propose to use the nonlinear Feynman–Kac formula to rewrite certain parabolic PDEs such as the backward stochastic differential equation (BSDE), which is then reformulated as a stochastic control problem (also known as reinforcement learning in the machine learning community). By applying a deep neural network as the control function and optimizing over network parameters,

the solution at any given space-time location can be evaluated. Another example is [25], which mainly focuses on computing the committor function that solves a steady-state (time-independent) Fokker–Planck equation with specific boundary conditions. This committor function can be treated as the solution to a variational problem associated with an energy functional. A neural network is used to replace the solution in the variational problem. When optimizing over network parameters, the neural network can be used to approximate the committor function.

In this paper, we focus on designing a sampling-friendly method for the time-dependent Fokker–Planck equation. There are two main reasons that motivate us for this investigation. One, as mentioned before, is to design a sample-based algorithm to solve PDEs in high dimensions. The other is to provide an alternative sampling strategy that can be potentially faster than Langevin Monte Carlo. Our approaches are different in terms of how deep networks are leveraged to approximate the solution of the PDE. We use pushforward of a given reference measure by neural networks to create a generative model. This is to approximate the stream of probability distributions, which can be used to generate samples not only at the terminal time, but also any time in between. More importantly, we prove results, obtained by using newly developed techniques based on a Wasserstein metric on probability manifold, on the asymptotic convergence and error control of our numerical schemes. To the best of our knowledge, similar results are still lacking in existing studies.

1.6. Organization of this paper. We organize the paper as follows. In section 2, we briefly introduce some background knowledge of the Fokker–Planck equation, including its relation with SDEs and its Wasserstein gradient flow structure. In section 3, we introduce the Wasserstein statistical manifold (Θ, G) and derive our parametric Fokker–Planck equation as the manifold gradient flow of relative entropy on Θ. We study the geometric property of this equation, including an insightful particle motion–based interpretation of the parametric Fokker–Planck equation. In section 4, we design a numerical scheme that is tractable for computing our parametric Fokker–Planck equation using a deep learning framework. Some important details of implementation will be discussed. We present asymptotic convergence and error estimates for the parametric Fokker–Planck equation in section 5 and provide some numerical examples in section 6.

2. Background on the Fokker–Planck equation. In this section, we present two different perspectives regarding the Fokker–Planck equations. More discussion can be found in [35].

2.1. As the density evolution of stochastic differential equations. The general form of the Fokker–Planck equation is [51, 31]

$$\begin{aligned}\frac{\partial \rho(x,t)}{\partial t} &= -\nabla \cdot (\rho(x,t)\boldsymbol{\mu}(x,t)) + \frac{1}{2}\nabla^2 : (\boldsymbol{D}(x,t)\rho(x,t)) \\ &= -\sum_{i=1}^{d} \frac{\partial}{\partial x_i}(\rho(x,t)\mu_i(x,t)) + \frac{1}{2}\sum_{i,j=1}^{d} \frac{\partial^2}{\partial x_i \partial x_j}(D_{ij}(x,t)\rho(x,t)), \quad \rho(x,0) = \rho_0(x).\end{aligned}$$

Here $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)^{\mathrm{T}}$ is the drift function and $\boldsymbol{D} = \{D_{ij}\}$ is the $d \times d$ diffusion tensor. Furthermore, $\boldsymbol{D}$ can be written as $\boldsymbol{D} = \boldsymbol{\sigma}\boldsymbol{\sigma}^{\mathrm{T}}$, where $\boldsymbol{\sigma}(x,t)$ is a $d \times \tilde{d}$ matrix. One derivation of the Fokker–Planck equation originates from the following SDE [51, 31]:

$$d\boldsymbol{X}_t = \boldsymbol{\mu}(\boldsymbol{X}_t, t)\ dt + \boldsymbol{\sigma}(\boldsymbol{X}_t, t)\ d\boldsymbol{B}_t, \quad \boldsymbol{X}_0 \sim \rho_0,$$

where $\{\boldsymbol{B}_t\}_{t\geq 0}$ is the standard Brownian motion in $\mathbb{R}^{\tilde{d}}$, and ρ_0 is the distribution of the initial state. It is well known that the evolution of the density $\rho(x,t)$ of the stochastic process $\{\boldsymbol{X}_t\}_{t\geq 0}$ is described by the above Fokker–Planck equation.

In this paper, we consider a more specific type of (2.1) by setting $\boldsymbol{\mu}(x,t) = -\nabla V(x)$, $\boldsymbol{\sigma}(x,t) = \sqrt{2D}\ I_{d\times d}$ $(D>0)$, where $I_{d\times d}$ is the $d\times d$ identity matrix, and so $\boldsymbol{D} = 2D\ I_{d\times d}$. Then (2.1) is

$$d\boldsymbol{X}_t = -\nabla V(\boldsymbol{X}_t)\ dt + \sqrt{2D}\ d\boldsymbol{B}_t, \quad \boldsymbol{X}_0 \sim \rho_0. \tag{2.1}$$

This equation is also called overdamped Langevin dynamics, which has broad applications in computational physics, computational biology, and Bayesian statistics [19, 63, 71]. The corresponding Fokker–Planck equation is simplified to

$$\frac{\partial \rho(x,t)}{\partial t} = \nabla\cdot(\rho(x,t)\nabla V(x)) + D\Delta\rho(x,t), \quad \rho(x,0) = \rho_0(x). \tag{2.2}$$

In addition, we would like to mention that there is a Vlasov-type SDE corresponding to the Fokker–Planck equation (2.2):

$$\frac{d\boldsymbol{X}_t}{dt} = -\nabla V(\boldsymbol{X}_t) - D\ \nabla \log \rho(\boldsymbol{X}_t,t), \quad \boldsymbol{X}_0 \sim \rho_0, \tag{2.3}$$

in which $\rho(\cdot,t)$ is the density of $\boldsymbol{X}_t$. This Vlasov-type SDE (2.3) will be very useful in our proofs for the error estimates of our proposed numerical algorithms.

2.2. As the Wasserstein gradient flow of relative entropy. Another useful viewpoint states that (2.2) is the Wasserstein gradient flow of relative entropy. We briefly present some of the notation and basic results in this regard. We only provide in sections 2.2.1 and 2.2.2 an informal discussion on the Wasserstein manifold and Wasserstein gradient flow. More rigorous treatments on the topics can be found in [4].

2.2.1. Wasserstein manifold. Denote the probability space supported on $\mathbb{R}^d$ with densities having finite second order moments as

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

Here the integral is computed over the sample space $\mathbb{R}^d$. In the following discussion, if not specified, we always write $\int_{\mathbb{R}^d}$ as $\int$ for simplicity.

The so-called Wasserstein distance (also known as L^2-Wasserstein distance) on $\mathcal{P}$ is defined as [68]

$$W_2(\rho_1,\rho_2) = \left(\inf_{\pi\in\Pi(\rho_1,\rho_2)} \iint |x-y|^2\ d\pi(x,y)\right)^{1/2}, \tag{2.4}$$

where $\Pi(\rho_1,\rho_2)$ is the set of joint distributions defined on $\mathbb{R}^d\times\mathbb{R}^d$ with fixed marginal distributions whose densities are ρ_1,ρ_2. If we treat $\mathcal{P}$ as an infinite-dimensional manifold, the Wasserstein distance W_2 can induce a metric g^W on the tangent bundle $\mathcal{TP}$, with which $\mathcal{P}$ becomes a Riemannian manifold. For simplicity, here we directly give the definition of g^W. One can identify the tangent space at ρ as

$$\mathcal{T}_\rho\mathcal{P} = \left\{f\colon \int f(x)dx = 0\right\}.$$

For a specific $\rho \in \mathcal{P}$ and $f_i \in \mathcal{T}_\rho\mathcal{P}$, $i = 1, 2$, we define the Wasserstein metric tensor g^W as [30, 46]

$$g^W(\rho)(f_1, f_2) = \int \nabla\psi_1(x) \cdot \nabla\psi_2(x)\rho(x)\ dx, \tag{2.5}$$

where ψ_1, ψ_2 satisfy

$$f_i = -\nabla \cdot (\rho\nabla\psi_i), \quad i = 1, 2, \tag{2.6}$$

with boundary conditions

$$\lim_{x\to\infty} \rho(x)\nabla\psi_i(x) = 0, \quad i = 1, 2.$$

Using the above definition, we can also write

$$g^W(\rho)(f_1, f_2) = \int \psi_1(-\nabla \cdot (\rho\nabla\psi_2))\ dx = \int (-\nabla \cdot (\rho\nabla))^{-1}(f_1) \cdot f_2\ dx.$$

Thus, we can identify $g^W(\rho)$ as $(-\nabla \cdot (\rho\nabla))^{-1}$. When $\mathrm{supp}(\rho) = \mathbb{R}^d$, $g^W(\rho)$ is a positive definite bilinear form defined on tangent bundle $\mathcal{TP} = \{(\rho, f) \colon \rho \in \mathcal{P},\ f \in \mathcal{T}_\rho\mathcal{P}\}$. Hence we can treat $\mathcal{P}$ as a Riemannian manifold, which we call *Wasserstein manifold*, denoted by $(\mathcal{P}, g^W)$ [46]. In order to keep our notation concise, in what follows, we denote $g^W(\rho)$ as g^W if no confusion is caused.

2.2.2. Wasserstein gradient. We denote the Wasserstein gradient grad_W as the manifold gradient on $(\mathcal{P}, g^W)$. In Riemannian geometry, the manifold gradient must be compatible with the metric, implying that for any smooth functional $\mathcal{F}$ defined on $\mathcal{P}$ and any $\rho \in \mathcal{P}$, considering an arbitrary differentiable curve $\{\rho_t\}_{t\in(-\delta,\delta)}$ with $\rho_0 = \rho$, we have

$$\frac{d}{dt}\mathcal{F}(\rho_t)\Big|_{t=0} = g^W(\rho)(\mathrm{grad}_W\mathcal{F}(\rho),\ \dot\rho_0).$$

Since we can write

$$\frac{d}{dt}\mathcal{F}(\rho_t)\Big|_{t=0} = \int \frac{\delta\mathcal{F}(\rho)}{\delta\rho(x)}(x) \cdot \dot\rho_0(x)\ dx = \left\langle \frac{\delta\mathcal{F}(\rho)}{\delta\rho}, \dot\rho_0 \right\rangle_{L^2},$$

where $\frac{\delta\mathcal{F}(\rho)}{\delta\rho(x)}(x)$ is the L^2 variation of $\mathcal{F}$ at point $x \in \mathbb{R}^d$, we then have

$$\left\langle \frac{\delta\mathcal{F}(\rho)}{\delta\rho}, \dot\rho_0 \right\rangle_{L^2} = g^W(\rho)(\mathrm{grad}_W\mathcal{F}(\rho),\ \dot\rho_0) \quad \forall\ \dot\rho_0 \in \mathcal{T}_\rho\mathcal{P}.$$

This leads to the following useful formula for computing the Wasserstein gradient of functional $\mathcal{F}$:

$$\mathrm{grad}_W\mathcal{F}(\rho) = g^W(\rho)^{-1}\left(\frac{\delta\mathcal{F}}{\delta\rho}\right)(x) = -\nabla \cdot \left(\rho(x)\nabla\ \frac{\delta\mathcal{F}(\rho)}{\delta\rho(x)}(x)\right). \tag{2.7}$$

In particular, if $\mathcal{F}$ is taken as the relative entropy functional given by

$$\mathcal{H}(\rho) = D\ \mathcal{D}_{\mathrm{KL}}\left(\rho\ \middle|\middle|\ \rho_*\right) = \left(\int V(x)\rho(x) + D\rho(x)\log\rho(x)\ dx\right) + D\ \log Z_D, \tag{2.8}$$

we have $\nabla \frac{\delta \mathcal{H}(\rho)}{\delta \rho} = \nabla V + D\nabla \log \rho$. Using (2.7), and noticing $\nabla \log \rho = \frac{\nabla \rho}{\rho}$, then $\nabla \cdot (\rho \nabla \log \rho) = \nabla \cdot (\nabla \rho) = \Delta \rho$, and the Wasserstein gradient flow of $\mathcal{H}$ can be written as

$$\frac{\partial \rho}{\partial t} = -\mathrm{grad}_W \mathcal{H}(\rho) = \nabla \cdot (\rho \nabla V) + D\nabla \cdot (\rho \nabla \log \rho),$$

which is exactly the Fokker–Planck equation (2.2).

3. Parametric Fokker–Planck equation.

In this section, we provide a detailed derivation for our parametric Fokker–Planck equation.

3.1. Wasserstein statistical manifold.

Consider a parameter space Θ as an open, convex set in $\mathbb{R}^m$, and assume the sample space is $\mathbb{R}^d$. Let T_θ be a map from $\mathbb{R}^d$ to $\mathbb{R}^d$ parametrized by θ. In our discussion, we always assume the invertibility of $T_\theta(x)$, and it is second order differentiable with respect to x and θ, i.e., $T_\theta(x) \in C^2(\Theta \times \mathbb{R}^d)$.

Remark 3.1. There are many different choices for T_θ:

- We can set $T_\theta(x) = Ux + b$, with $\theta = (U, b)$, where U is a $d \times d$ invertible matrix and $b \in \mathbb{R}^d$.
- We may also choose T_θ as the linear combination of basis functions $T_\theta(x) = \sum_{k=1}^m \theta_k \vec{\Phi}_k(x)$, where $\{\vec{\Phi}_k\}_{k=1}^m$ are the basis functions and the parameter θ will be the coefficients $\theta = (\theta_1, \ldots, \theta_m)$.
- We can also treat T_θ as a neural network. Its general structure can be written as the composition of l affine and nonlinear activation functions: $T_\theta(x) = \sigma_l(W_l(\sigma_{l-1}(\ldots \sigma_1(W_1 x + b_1)\ldots)) + b_l)$. In this case, the parameter θ will be the weight matrices and bias vectors of the neural network, i.e., $\theta = (W_1, b_1, \ldots, W_l, b_l)$.

Definition 3.2. *Suppose X, Y are two measurable spaces, and λ is a probability measure defined on X; let $f : X \to Y$ be a measurable map. We define $f_\sharp \lambda$ as $f_\sharp \lambda(E) = \lambda(f^{-1}(E))$ for all measurable $E \subset Y$. We call $f_\sharp \lambda$ the pushforward of measure λ by map f.*

Let $p \in \mathcal{P}$ be a reference probability measure with positive density defined on $\mathbb{R}^d$, such as the standard Gaussian. We denote by ρ_θ the density of $T_{\theta\sharp} p$. Such a mechanism of producing parametric probability distributions is also known as a *generative model*, which has broad applications in deep learning research [18, 5, 8]. We further assume our T_θ satisfy the following two conditions:

$$\text{Condition 1:} \quad \int |z|^2 \rho_\theta(z)\, dz = \int |T_\theta(x)|^2\, dp(x) < \infty \quad \forall\, \theta \in \Theta. \tag{3.1}$$

This ensures that $\rho_\theta \in \mathcal{P}$ for each $\theta \in \Theta$. In order to introduce Wasserstein metric to the parameter space Θ, we also assume that the Frobenius norm of the operator $\partial_\theta T_\theta(x) : \mathbb{R}^d \to \mathbb{R}^{d \times m}$ is locally bounded in the following sense: for any fixed $\theta_* \in \Theta$, there exist $r(\theta_*) > 0$ and two functions $L_1(\cdot|\ \theta_*), L_2(\cdot|\ \theta_*)$ satisfying the following condition:

$$\begin{aligned} \text{Condition 2:} \quad & \|\partial_\theta T_\theta(x)\|_F \le L_1(x|\ \theta_*), \quad \|\partial_\theta T_\theta(x)\|_F^2 \le L_2(x|\ \theta_*), \\ & \forall\, \theta, |\theta - \theta_*| < r(\theta_*) \text{ and } x \in \mathbb{R}^d, \text{ and} \\ & \int L_1(x|\ \theta_*)\, dx < \infty, \quad \int L_2(x|\ \theta_*)\, dx < \infty. \end{aligned} \tag{3.2}$$

We define the parametric submanifold $\mathcal{P}_\Theta \subset \mathcal{P}$ as

$$\mathcal{P}_\Theta = \{\rho_\theta \text{ is density function of } \ T_{\theta\sharp}p \mid \theta \in \Theta\}.$$

Clearly, the connection between $\mathcal{P}$ and Θ is through the pushforward operation $T_{\theta\sharp} : \Theta \to \mathcal{P}_\Theta, \theta \mapsto \rho_\theta$. Hence it is natural to define the Wasserstein metric $G(\theta)$ on parameter space Θ as the pullback of g^W by $T_{\theta\sharp}$. To be specific, we define $G(\theta) = (T_{\theta\sharp})^* g^W$. Using this definition, $T_{\theta\sharp}$ becomes an isometric immersion from Θ to $\mathcal{P}$. For each θ, $G(\theta)$ is a bilinear form defined on $\mathcal{T}_\theta\Theta \simeq \mathbb{R}^m$, which can be identified as an $m \times m$ matrix.

Before computing $G(\theta)$, we introduce a lemma which can help us to better understand $G(\theta)$.

LEMMA 3.3. *Suppose $\vec{u}, \vec{v}$ are two vector fields defined on $\mathbb{R}^d$, and suppose φ, ψ solves $-\nabla\cdot(\rho\nabla\varphi) = -\nabla\cdot(\rho\vec{u})$ and $-\nabla\cdot(\rho\nabla\psi) = -\nabla\cdot(\rho\vec{v})$, or equivalently $Proj_\rho[\vec{u}] = \nabla\varphi$ and $Proj_\rho[\vec{v}] = \nabla\psi$ (cf. Definition 4.2). Then*

$$\int \vec{u}(x)\cdot\nabla\psi(x)\rho(x)\ dx = \int \nabla\varphi(x)\cdot\nabla\psi(x)\rho(x)\ dx, \tag{3.3}$$

$$\int |\nabla\psi(x)|^2\rho(x)\ dx \le \int |\vec{v}(x)|^2\rho(x)\ dx. \tag{3.4}$$

We prove Lemma 3.3 in Appendix A. The metric tensor $G(\theta)$ is computed in the following theorem.

THEOREM 3.4. *Assume Θ satisfies (3.1), (3.2). T_θ is invertible and $T_\theta(x) \in C^2(\Theta \times \mathbb{R}^d)$. Then Θ can be equipped with the metric tensor $G = (T_{\theta\sharp})^* g^W$, which is an $m \times m$ nonnegative definite symmetric matrix of the form*

$$G(\theta) = \int \nabla\mathbf{\Psi}(T_\theta(x))\nabla\mathbf{\Psi}(T_\theta(x))^T\ dp(x) \tag{3.5}$$

at every $\theta \in \Theta$. More precisely, in entrywise form,

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

in which $\mathbf{\Psi} = (\psi_1, \ldots, \psi_m)^T$ and $\nabla\mathbf{\Psi}$ is an $m \times d$ Jacobian matrix of $\mathbf{\Psi}$. For each $j = 1, 2, \ldots, m$, ψ_j solves the equation

$$\nabla\cdot(\rho_\theta\nabla\psi_j(x)) = \nabla\cdot\left(\rho_\theta\ \frac{\partial T_\theta}{\partial\theta_j}(T_\theta^{-1}(x))\right), \tag{3.6}$$

with boundary conditions

$$\lim_{x\to\infty} \rho_\theta(x)\nabla\psi_j(x) = 0.$$

Proof. Supposing $\xi \in \mathcal{T}\Theta$ is a vector field on Θ, for a fixed $\theta \in \Theta$, we first compute the pushforward $(T_{\theta\sharp})_*\xi(\theta)$ of ξ at point θ. We choose any smooth curve $\{\theta_t\}_{t\ge0}$ on Θ with $\theta_0 = \theta$ and $\dot{\theta}_0 = \xi(\theta)$. If we denote $\rho_{\theta_t} = T_{\theta_t\sharp}p$, we have $(T_{\theta\sharp})_*\xi(\theta) = \frac{\partial\rho_{\theta_t}}{\partial t}\big|_{t=0}$.

To compute $\frac{\partial\rho_{\theta_t}}{\partial t}\big|_{t=0}$, we consider an arbitrary $\phi \in C_0^\infty(M)$.

On the one hand, $\frac{\rho_{\theta_{\Delta t}}(y) - \rho_{\theta_0}(y)}{\Delta t} = \frac{\partial}{\partial t}\rho(\theta_{\tilde{t}_1}, y)$, where $\tilde{t}_1$ is some point between $0, \Delta t$; since $\phi \in C_0^\infty$ and $\rho(\theta_t, x)$ is at least C^1 with respect to t, y, we can show that

the function $\varphi(x) = \sup_{s\in[0,\Delta t]} |\phi(x)\frac{\partial}{\partial t}\rho(\theta_s, y)|$ is continuous on a compact set and thus integrable on $\mathbb{R}^d$. Using the dominated convergence theorem, we obtain

$$\frac{\partial}{\partial t}\left(\int \phi(y)\rho_{\theta_t}(y)\ dy\right)\Big|_{t=0} = \int \phi(y)\frac{\partial \rho_{\theta_t}(y)}{\partial t}\Big|_{t=0}\ dy. \tag{3.7}$$

On the other hand, we have

$$\frac{\phi(T_{\theta_{\Delta t}}(y)) - \phi(T_{\theta_0}(y))}{\Delta t} = \dot\theta_{\tilde t_2}^{\mathrm{T}}\ \partial_\theta T_{\theta_{\tilde t_2}}(x)^{\mathrm{T}}\ \nabla\phi(T_{\theta_{\tilde t_2}}(y)), \tag{3.8}$$

in which $\tilde t_2$ is also between $0, \Delta t$. For any Δt small enough and $\tilde t \in [0, \Delta t]$, we can easily find upper bounds for $\|\dot\theta_{\tilde t}\| \le A$ and $\|\nabla\phi(\cdot)\|_\infty \le B$. Recall the condition (3.2); when Δt is small enough, we have $|\theta_{\Delta t} - \theta_0| < r(\theta_0)$, and thus we obtain the following upper bound for (3.8):

$$|\dot\theta_{\tilde t}^{\mathrm{T}}\ \partial_\theta T_{\theta_{\tilde t}}(x)^{\mathrm{T}}\ \nabla\phi(T_{\theta_{\tilde t}}(y))| \le AB\|\partial_\theta T_{\theta_{\tilde t}}(x)\|_F \le ABL_1(x|\theta_0).$$

By (3.2), we know $L_1(\cdot|\theta_0) \in L^1(p)$, and we can apply the dominated convergence theorem to obtain

$$\frac{\partial}{\partial t}\left(\int \phi(T_{\theta_t}(x))dp\right)\Big|_{t=0} = \int \dot\theta_t^{\ \mathrm{T}}\partial_\theta T_{\theta_t}(x)^{\mathrm{T}}\nabla\phi(T_{\theta_t}(x))|_{t=0}dp. \tag{3.9}$$

Since $\frac{\partial}{\partial t}\int \phi(y)\rho_{\theta_t}(y)\ dy = \frac{\partial}{\partial t}\int \phi(T_{\theta_t}(x))\ dp(x)$, we use (3.7) and (3.9) to get

$$\begin{aligned}
\int \phi(y)\frac{\partial \rho_{\theta_t}}{\partial t}(y)\Big|_{t=0}\ dy &= \int \dot\theta_t^{\ \mathrm{T}}\partial_\theta T_{\theta_t}(x)^{\mathrm{T}}\nabla\phi(T_{\theta_t}(x))|_{t=0}\ dp(x)\\
&= \int \dot\theta_t^{\mathrm{T}}\left(\frac{\partial T_{\theta_t}}{\partial\theta}(T_{\theta_t}^{-1}(x))\right)^{\mathrm{T}}\nabla\phi(x)\ \rho_{\theta_t}(x)|_{t=0}\ dx\\
&= \int \phi(x)\left(-\nabla\cdot\left(\rho_{\theta_t}(x)\frac{\partial T_{\theta_t}}{\partial\theta}(T_{\theta_t}^{-1}(x))\ \dot\theta_t\right)\right)|_{t=0}\ dx.
\end{aligned}$$

Because $\phi(x)$ is arbitrary, this weak formulation reveals that

$$(T_{\theta\sharp})_*\xi(\theta) = \frac{\partial \rho_{\theta_t}}{\partial t}\Big|_{t=0} = -\nabla\cdot\left(\rho_\theta(x)\ \frac{\partial T_\theta}{\partial\theta}(T_\theta^{-1}(x))\xi(\theta)\right). \tag{3.10}$$

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

$$G(\theta)(\xi(\theta), \xi(\theta)) = g^W(\rho_\theta)((T_{\theta\sharp})_*\xi(\theta), (T_{\theta\sharp})_*\xi(\theta)). \tag{3.11}$$

To compute the right-hand side of (3.11), recalling (2.5), we need to solve for φ from

$$\frac{\partial \rho_{\theta_t}}{\partial t}\Big|_{t=0} = -\nabla\cdot(\rho_\theta(x)\nabla\varphi(x)). \tag{3.12}$$

By (3.10), (3.12) is

$$\nabla\cdot(\rho_\theta(x)\nabla\varphi(x)) = \nabla\cdot\left(\rho_\theta(x)\frac{\partial_\theta T_\theta}{\partial\theta}(T_\theta^{-1}(\cdot))\xi(\theta)\right). \tag{3.13}$$

We can straightforwardly check that $\varphi(x) = \mathbf{\Psi}^{\mathrm{T}}(x)\xi(\theta)$ is the solution of (3.13). Now by definition of g^W as mentioned in subsection 2.2.1, we write the right-hand side of (3.11) as

(3.14)
$$\begin{aligned} g^W(\rho_\theta)((T_{\theta\sharp})_*\xi(\theta),(T_{\theta\sharp})_*\xi(\theta)) &= \int |\nabla\varphi(y)|^2\rho_\theta(y)\ dy \\ &= \xi(\theta)^{\mathrm{T}}\left(\int \nabla\mathbf{\Psi}(y)\nabla\mathbf{\Psi}(y)^{\mathrm{T}}\rho_\theta(y)\ dy\right)\xi(\theta) \\ &= \sum_{i,j=1}^m \left(\int \nabla\psi_i(y)\cdot\nabla\psi_j(y)\rho_\theta(y)\ dy\right)\xi_i(\theta)\xi_j(\theta). \end{aligned}$$

Here we assume components of $\xi(\theta)$ as $(\xi_1(\theta),\dots,\xi_m(\theta))^{\mathrm{T}}$. Before we compute $G(\theta)$, we first verify that the inner product in (3.14) is finite for any $\xi \in \mathcal{T}\Theta$. To show this, by the Cauchy–Schwarz inequality we obtain

$$\int \nabla\psi_i(y)\cdot\nabla\psi_j(y)\rho_\theta(y)\ dy \le \left(\int |\nabla\psi_i(y)|^2\rho_\theta(y)\ dy\right)^{\frac{1}{2}}\left(\int |\nabla\psi_j(y)|^2\rho_\theta(y)\ dy\right)^{\frac{1}{2}}.$$

Recall ψ_j as defined in (3.6); then applying (3.4) of Lemma 3.3 yields

$$\begin{aligned} \int |\nabla\psi_j(y)|^2\rho_\theta(y)\ dy &\le \int \left|\frac{\partial T_\theta}{\partial\theta_j}(T_\theta^{-1}(y))\right|^2\rho_\theta(y)\ dy \\ &= \int \left|\frac{\partial T_\theta}{\partial\theta_j}(x)\right|^2 dp(x) \le \int L_2(y|\theta)p(y)\ dy < \infty. \end{aligned}$$

The last two inequalities are due to condition (3.2). As a result, we proved the finiteness of (3.14).

Finally, let us compute

$$\begin{aligned} G(\theta)(\xi(\theta),\xi(\theta)) &= g^W(\rho_\theta)((T_{\theta\sharp})_*\xi(\theta),(T_{\theta\sharp})_*\xi(\theta)) \\ &= \xi(\theta)^{\mathrm{T}}\left(\int \nabla\mathbf{\Psi}(T_\theta(x))\nabla\mathbf{\Psi}(T_\theta(x))^{\mathrm{T}}dp(x)\right)\xi(\theta). \end{aligned}$$

Thus we can verify that

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

which completes the proof. ∎

Generally speaking, the metric tensor G does not have an explicit form when $d \ge 2$. It is worth mentioning that G has an explicit form and can be computed directly when $d = 1$ [35].

Remark 3.5 (well-posedness of (3.6)). It is worth commenting on the existence and the regularity question of equations like (3.6). Determining what properties or conditions that T_θ should have to guarantee the well-posedness of (3.6) is an interesting and important problem on its own. In references such as [48] and [69], there are sufficient conditions that guarantee the well-posedness of elliptic PDEs defined on $\mathbb{R}^d$. Most of the existing results require a uniform lower bound on ρ_θ, i.e., $\rho_\theta(x) > \epsilon > 0$ for all $x \in \mathbb{R}^d$. Such a coercive condition is not applicable in our case since $\int \rho_\theta(x)dx = 1$ is finite.

It is worth pointing out that under certain situations discussed in section 3.4, (3.6) does have classical solutions. For example, if we select T_θ as an affine transform and consider the Fokker–Planck equation (2.2) with quadratic potential V and Gaussian initial ρ_0, we can prove that (3.6) is well-posed along the trajectory of the ODE (3.18), i.e., the elliptic equation

$$-\nabla\cdot(\rho_{\theta_t}\nabla\psi) = -\nabla\cdot\left(\rho_{\theta_t}\frac{\partial_\theta T_{\theta_t}}{\partial\theta}(T_{\theta_t}^{-1}(x))\dot{\theta}_t\right), \quad \text{where } \{\theta_t\} \text{ solves (3.18)},$$

always admits a classical solution $\psi(x) = V(x) + D\log\rho_\theta(x) + \text{Const}$.

In general, the conditions imposed on T_θ to guarantee well-posedness of (3.6) are a fundamental and interesting subject for further investigation. A good reference related to the topic can be found in [4].

The following theorem provides several criteria for examining whether G is a Riemannian metric, i.e., whether $G(\theta)$ is positive definite.

THEOREM 3.6. *For $\theta\in\Theta$, $\{\psi_k\}_{k=1}^m$ satisfies* (3.6)*, and the following four statements are equivalent:*

1. *$G(\theta)$ is positive definite.*
2. *For any $\xi\in\mathcal{T}_\theta\Theta$ ($\xi\neq 0$), there exists $z\in M$ such that $\nabla\cdot(\rho_\theta(z)\frac{\partial T_\theta}{\partial\theta}(T_\theta^{-1}(z))\xi)\neq 0$.*
3. *$\{\nabla\psi_k\}_{k=1}^m$, as m functions in the space $L^2(\mathbb{R}^d;\mathbb{R}^d,\rho_{\theta_k})$, are linearly independent.*
4. *$\frac{d}{dt}(T_{\theta+t\xi\sharp}p)|_{t=0}\neq 0$ for any $\xi\in\mathbb{R}^m$.*

Proof. We first verify that 1 and 2 are equivalent. We need the following identity, used in Theorem 3.4: For any θ,ξ,x, we have

$$\nabla\cdot(\rho_\theta(x)\nabla(\xi^{\mathrm{T}}\mathbf{\Psi}(x))) = \nabla\cdot\left(\rho_\theta(x)\frac{\partial T_\theta}{\partial\theta}(T_\theta^{-1}(x))\xi\right). \tag{3.15}$$

($\Leftarrow$): Suppose for any $\theta\in\Theta$ and $\xi\in\mathcal{T}_\theta\Theta$, at certain $z\in\mathbb{R}^d$, that $\nabla\cdot(\rho_\theta(z)\frac{\partial T_\theta}{\partial\theta}(T_\theta^{-1}(z)\xi)\neq 0$; then $\nabla\cdot(\rho_\theta(z)\nabla(\xi^{\mathrm{T}}\mathbf{\Psi}(z)))\neq 0$, and thus $\rho_\theta\nabla(\xi^{\mathrm{T}}\mathbf{\Psi})$ is not identically $\mathbf{0}$. Using continuity of $\rho_\theta\nabla(\xi^{\mathrm{T}}\mathbf{\Psi})$, we know that $|\nabla(\xi^{\mathrm{T}}\mathbf{\Psi}(x))|^2\rho_\theta(x)>0$ in some small neighborhood of z. Thus we have that

$$\xi^{\mathrm{T}}G(\theta)\xi = \int|\nabla\mathbf{\Psi}(x)^{\mathrm{T}}\xi|^2\rho_\theta(x)\,dx > 0 \tag{3.16}$$

holds for any θ and ξ, which leads to the positive definiteness of G.

($\Rightarrow$): Now suppose (3.16) holds for all θ,ξ; then we have

$$\int -\nabla\cdot(\rho_\theta(x)\nabla(\xi^{\mathrm{T}}\mathbf{\Psi}(x)))\cdot\xi^{\mathrm{T}}\mathbf{\Psi}(x)\;dx > 0.$$

This leads to the existence of a $z\in\mathbb{R}^d$ such that $-\nabla\cdot(\rho_\theta(z)\nabla(\xi^{\mathrm{T}}\mathbf{\Psi}(z)))\neq 0$. Combining (3.15), we have verified the equivalence between 1 and 2.

Recalling (3.10), we then have $\frac{d}{dt}(T_{\theta+t\xi\sharp}p)|_{t=0} = (T_{\theta\sharp})_*\xi = -\nabla\cdot(\rho_\theta(x)\frac{\partial T_\theta}{\partial\theta}(T_\theta^{-1}(x))\xi)$, which verifies the equivalence between 2 and 3.

Finally, as stated before, we can verify $\xi^{\mathrm{T}}G(\theta)\xi = \|\sum_{k=1}\xi_k\nabla\psi_k\|^2_{L^2(\rho_\theta)}$; this formula will directly lead to the equivalence between 1 and 4, and we have proved the equivalence among statements 1, 2, 3, and 4. ∎

To keep our discussion concise in the following sections, we will always assume $G(\theta)$ is positive definite for every $\theta\in\Theta$.

3.2. Parametric Fokker–Planck equation. We consider the pushforward $T_{(\cdot)\sharp}$ induced relative entropy functional $H = \mathcal{H} \circ T_{(\cdot)\sharp} : \Theta \to \mathbb{R}$:

$$\begin{aligned} H(\theta) = \mathcal{H}(\rho_\theta) &= \left(\int V(x)\rho_\theta(x) + D\rho_\theta(x) \log \rho_\theta(x) \, dx \right) + D \log Z_D \\ &= \left(\int V(T_\theta(x)) + D \log \rho_\theta(T_\theta(x)) \, dp(x) \right) + D \log Z_D. \end{aligned} \tag{3.17}$$

Following the theory in [2], the gradient flow of H on the Wasserstein parameter manifold (Θ, G) satisfies

$$\dot{\theta} = -G(\theta)^{-1} \nabla_\theta H(\theta). \tag{3.18}$$

We call (3.18) the *parametric Fokker–Planck equation*. The ODE (3.18) as the Wasserstein gradient flow on parameter space (Θ, G) is closely related to the Fokker–Planck equation on probability submanifold $\mathcal{P}_\Theta$. We have the following theorem, which is a natural result derived from submanifold geometry.

THEOREM 3.7. *Suppose $\{\theta_t\}_{t\geq 0}$ solves (3.18). Then $\{\rho_{\theta_t}\}$ is the gradient flow of $\mathcal{H}$ on probability submanifold $\mathcal{P}_\Theta$. Furthermore, at any time t, $\dot{\rho}_{\theta_t} = \frac{d}{dt}\rho_{\theta_t} \in \mathcal{T}_{\rho_{\theta_t}}\mathcal{P}_\Theta$ is the orthogonal projection of $-\mathrm{grad}_W \mathcal{H}(\rho_{\theta_t}) \in \mathcal{T}_{\rho_{\theta_t}}\mathcal{P}$ onto the subspace $\mathcal{T}_{\rho_{\theta_t}}\mathcal{P}_\Theta$ with respect to the Wasserstein metric g^W.*

We prove this theorem in Appendix B.

The following theorem is an important new statement closely related to Theorem 3.7.

THEOREM 3.8 (Wasserstein gradient as solution to a least squares problem). *For a fixed $\theta \in \Theta$, $\mathbf{\Psi} \subset \mathbb{R}^m$ as defined in Theorem 3.4,*

$$G(\theta)^{-1}\nabla_\theta H(\theta) = \underset{\eta \in \mathcal{T}_\theta\Theta \cong \mathbb{R}^m}{\arg\min} \left\{ \int |(\nabla \mathbf{\Psi}(T_\theta(x)))^T \eta - \nabla \left(V + D \log \rho_\theta \right) \circ T_\theta(x)|^2 dp(x) \right\}. \tag{3.19}$$

Proof. Direct computation shows that minimizing the function in (3.19) is equivalent to minimizing

$$\begin{aligned} &\eta^{\mathrm{T}} \left(\int \nabla \mathbf{\Psi}(T_\theta(x)) \nabla \mathbf{\Psi}(T_\theta(x))^{\mathrm{T}} \, dp(x) \right) \eta \\ -2 \, &\eta^{\mathrm{T}} \left(\int \nabla \mathbf{\Psi}(y) \nabla (V(y) + D \log \rho\theta(y)) \rho_\theta(y) \, dy \right). \end{aligned}$$

For each entry in the second term, we have

$$\begin{aligned} &\int \nabla \psi_k(y) \cdot \nabla (V(y) + D \log \rho_\theta(y)) \rho_\theta(y) \, dy \\ &= \int -\nabla \cdot (\rho_\theta(y) \nabla \psi_k(y)) \cdot (V(y) + D \log \rho_\theta(y)) \, dy \\ &= \int -\nabla \cdot (\rho_\theta(y) \partial_{\theta_k} T_\theta(T_\theta^{-1}(y))) \cdot (V(y) + D \log \rho_\theta(y)) \, dy \\ &= \int (\nabla V(T_\theta(x)) + D \nabla \log \rho_\theta(T_\theta(x))) \cdot \partial_{\theta_k} T_\theta(x) \, dp(x) \\ &= \int \nabla V(T_\theta(x)) \cdot \partial_{\theta_k} T_\theta(x) + \partial_{\theta_k} [D \log \rho_\theta(T_\theta(x))] \, dp(x) - \underbrace{\int D \, \partial_{\theta_k} \log \rho_\theta(T_\theta(x)) \, dp(x)}_{= D \int \nabla_\theta \rho_\theta(y) dy = 0} \\ &= \partial_{\theta_k} \left(\int (V(T_\theta(x)) + D \log \rho_\theta(T_\theta(x))) \, dp(x) \right) = \partial_{\theta_k} H(\theta). \end{aligned}$$

Recall the definition (3.5) of $G(\theta)$; the target function to be minimized is $\eta^{\mathrm{T}}G(\theta)\eta - 2\eta^{\mathrm{T}}\nabla_\theta H(\theta)$, and the minimizer is clearly $G(\theta)^{-1}\nabla_\theta H(\theta)$. ∎

In addition to the direct proof, the result in Theorem 3.8 can also be understood in a different way. Let us denote $\xi = G(\theta)^{-1}\nabla_\theta H(\theta)$, where $\{\theta_t\}$ solves (3.18) with initial value $\theta_0 = \theta$. By Theorem 3.7, $\frac{d}{dt}\rho_{\theta_t}\big|_{t=0} = (T_{\theta\sharp})_*\xi \in \mathcal{T}_{\rho_\theta}\mathcal{P}_\Theta$ is the orthogonal projection of $\mathrm{grad}_W\mathcal{H}(\rho_\theta)$ onto $\mathcal{T}_{\rho_\theta}\mathcal{P}_\Theta$ with respect to the metric g^W. This is equivalent to saying that η solves the following least squares problem:

$$\min_\eta g^W(\mathrm{grad}_W\mathcal{H}(\rho_\theta) - (T_{\theta\sharp})_*\eta,\ \mathrm{grad}_W\mathcal{H}(\rho_\theta) - (T_{\theta\sharp})_*\eta). \tag{3.20}$$

Recalling the definition of g^W in section 2.2.1, by (2.7) we have $\mathrm{grad}_W\mathcal{H}(\rho_\theta) = -\nabla\cdot(\rho_\theta\nabla(V + D\log\rho_\theta))$. Because of (3.10), $(T_{\theta\sharp})_*\eta = -\nabla\cdot(\rho_\theta\partial_\theta T_\theta(T_\theta^{-1}(\cdot))\eta)$, and solving $-\nabla\cdot(\rho_\theta\nabla\varphi) = \mathrm{grad}_W\mathcal{H}(\rho_\theta) - (T_{\theta\sharp})_*\eta$ gives

$$\varphi = (V + D\log\rho_\theta) - \boldsymbol{\Psi}^{\mathrm{T}}\eta,$$

and thus the least squares problem (3.20) can be written as

$$\min_\eta \left\{ \int |\nabla\boldsymbol{\Psi}(x)^{\mathrm{T}}\eta - \nabla(V(x) + D\log\rho_\theta(x))|^2\rho_\theta(x)\ dx \right\},$$

which is exactly (3.19).

3.3. A particle viewpoint of the parametric Fokker–Planck equation. The motion of parameter θ_t solving (3.18) naturally induces a stochastic dynamics on $\mathbb{R}^d$ whose density evolution is exactly $\{\rho_{\theta_t}\}$. To see this, notice that $\{\theta_t\}$ directly leads to a time-dependent map $\{T_{\theta_t}\}$. Let us denote a random variable $\boldsymbol{Z} \sim p$, i.e., $\boldsymbol{Z}$ is distributed according to the reference distribution p. We set $\boldsymbol{Y}_0 = T_{\theta_0}(\boldsymbol{Z}) \sim \rho_{\theta_0}$. At any time t, the map T_{θ_t} sends $\boldsymbol{Y}_0$ to $\boldsymbol{Y}_t = T_{\theta_t}(T_{\theta_0}^{-1}(\boldsymbol{Y}_0)) \sim \rho_{\theta_t}$. Thus, we construct a sequence of random variables $\{\boldsymbol{Y}_t\}$ whose density evolution is exactly $\{\rho_{\theta_t}\}$. We can characterize the dynamical system satisfied by $\{\boldsymbol{Y}_t\}$ by taking the time derivative: $\dot{\boldsymbol{Y}}_t = \partial_\theta T_{\theta_t}(\boldsymbol{Z})\dot\theta_t = \partial_\theta T_{\theta_t}(T_{\theta_t}^{-1}(\boldsymbol{Y}_t))\dot\theta_t$. It is actually more insightful to consider the following dynamic:

$$\dot{\boldsymbol{X}}_t = \nabla\boldsymbol{\Psi}_t(\boldsymbol{X}_t)^{\mathrm{T}}\ \dot\theta_t, \quad \boldsymbol{X}_0 = T_{\theta_0}(\boldsymbol{Z}) \sim \rho_{\theta_0}. \tag{3.21}$$

Here $\boldsymbol{\Psi}_t$ is obtained from (3.6) with parameter θ_t. It is not hard to show that for any time t, $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$ have the same distribution. Thus $\boldsymbol{X}_t \sim \rho_{\theta_t}$ for all $t \ge 0$. Recalling $\dot\theta_t = -G(\theta_t)^{-1}\nabla_\theta H(\theta_t)$, we are able to rewrite (3.21) as

$$\begin{aligned}\dot{\boldsymbol{X}}_t = \nabla\boldsymbol{\Psi}_t(\boldsymbol{X}_t)^{\mathrm{T}}&\underbrace{\left(\int \nabla\boldsymbol{\Psi}_t(x)\nabla\boldsymbol{\Psi}_t(x)^{\mathrm{T}}\ \rho_{\theta_t}(x)\ dx\right)}_{G(\theta_t)}{}^{-1}\\ &\times\underbrace{\left(\int \nabla\boldsymbol{\Psi}_t(\eta)(-\nabla V(\eta) - D\nabla\log\rho_{\theta_t}(\eta))\ \rho_{\theta_t}(\eta)\ d\eta\right)}_{-\nabla_\theta H(\theta_t)}.\end{aligned} \tag{3.22}$$

We define the kernel function $K_\theta : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{d\times d}$ as

$$K_\theta(x,\eta) = \nabla\boldsymbol{\Psi}^{\mathrm{T}}(x)\left(\int \nabla\boldsymbol{\Psi}(x)\nabla\boldsymbol{\Psi}(x)^{\mathrm{T}}\ \rho_\theta(x)\ dx\right)^{-1}\nabla\boldsymbol{\Psi}(\eta).$$

This K_θ induces a linear operator $\mathcal{K}_\theta : L^2(\mathbb{R}^d;\mathbb{R}^d,\rho_\theta) \to L^2(\mathbb{R}^d;\mathbb{R}^d,\rho_\theta)$ by

$$\mathcal{K}_\theta[\vec{v}] = (\mathcal{K}_\theta * \vec{v})(\cdot) = \int K_\theta(\cdot,\eta)\ \vec{v}(\eta)\ \rho_\theta(\eta)\ d\eta.$$

It can be verified that $\mathcal{K}_\theta$ is an orthogonal projection defined on the Hilbert space $L^2(\mathbb{R}^d;\mathbb{R}^d,\rho_\theta)$. The range of such a projection is the subspace span $\{\nabla\psi_1,\dots,\nabla\psi_m\} \subset L^2(\mathbb{R}^d;\mathbb{R}^d,\rho_\theta)$. Here $\psi_1,\dots,\psi_m$ are the m components of $\boldsymbol{\Psi}$ solved from (3.6). Using the linear operator, we can rewrite (3.22) as

(3.23)
$$\dot{\boldsymbol{X}}_t = -\mathcal{K}_{\theta_t}[\nabla V + D\nabla\log\rho_{\theta_t}](\boldsymbol{X}_t), \quad \text{where } \rho_{\theta_t} \text{ is the probability density of } \boldsymbol{X}_t, \ \boldsymbol{X}_0 \sim \rho_{\theta_0}.$$

We can compare (3.23) with the following dynamic without projection:
(3.24)
$$\dot{\tilde{\boldsymbol{X}}}_t = -(\nabla V + D\nabla\log\rho_t)(\tilde{\boldsymbol{X}}_t), \quad \text{where } \rho_t \text{ is the probability density of } \tilde{\boldsymbol{X}}_t,\ \tilde{\boldsymbol{X}}_0 \sim \rho_0.$$

As discussed in section 2.1, (3.24) is the Vlasov-type SDE that involves the density of a random particle. If assuming (3.24) admits a regular solution, we have $\rho(x,t) = \rho_t(x)$, which solves the original Fokker–Planck equation (2.2). From an orthogonal projection viewpoint, the parametric approximation ρ_{θ_t} of ρ_t originates from the projection of the vector field that drives the SDE (3.24).

We would like to mention that the expectation of the ℓ^2 discrepancy between $\nabla V + D\nabla\log\rho$ and its $\mathcal{K}_\theta$ projection is

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{X}\sim\rho_\theta}|\mathcal{K}_\theta[\nabla V + D\nabla\log\rho_\theta](\boldsymbol{X}) - (\nabla V + D\nabla\log\rho_\theta)(\boldsymbol{X})|^2\\ &\quad = \int |\nabla\boldsymbol{\Psi}(x)^{\mathrm{T}}\xi - (-\nabla V - D\nabla\log\rho_\theta)(x)|^2\rho_\theta(x)\ dx,\end{aligned} \tag{3.25}$$

in which $\xi = -G(\theta)^{-1}\nabla_\theta H(\theta)$. This is an essential term appearing in our error analysis part.

Remark 3.9. We should mention the relationship between our kernel K_{θ_t} and the neural tangent kernel (NTK) introduced in [22]. Using our notation, the NTK can be written as $K_\theta^{NTK} = \partial_\theta T_\theta(x)\partial_\theta T_\theta(\xi)^{\mathrm{T}}$. If we consider the flat gradient flow $\dot\theta = -\nabla_\theta H(\theta)$ of relative entropy on Θ, its corresponding particle dynamic is

$$\dot{\boldsymbol{X}}_t = \int K_{\theta_t}^{NTK}(T_{\theta_t}^{-1}(\boldsymbol{X}_t), T_{\theta_t}^{-1}(\eta))(-\nabla V(\eta) - D\nabla\log\rho_{\theta_t}(\eta))\rho_{\theta_t}(\eta)\ d\eta.$$

Different from our K_θ, which introduces an orthogonal projection, the NTK introduces a nonnegative definite transform to the vector field $-\nabla V - D\nabla\log\rho_{\theta_t}$.

Remark 3.10. Figure 1 illustrates the relation among (2.2), (3.18), (3.24), and (3.23). It is worth mentioning that the probability manifold point of view discussed in Theorem 3.7 is useful for our analysis of the continuous dynamics (3.18), while the particle viewpoint helps us in establishing the numerical analysis for the time discrete scheme (i.e., forward Euler) of (3.18).

3.4. An example of the parametric Fokker–Planck equation with quadratic potential. The solution of the parametric Fokker–Planck equation (3.18)

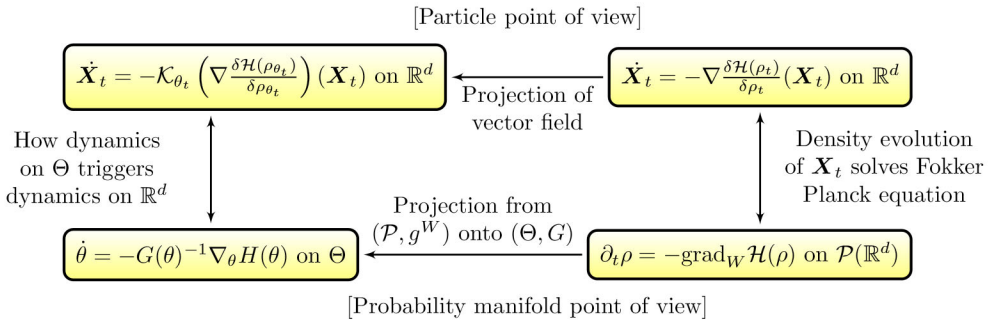


FIG. 1. *Illustrative diagram.*

can serve as an approximation to the solution of the original equation (2.2). In some special cases, ρ_{θ_t} exactly solves (2.2). In this section, we provide such examples.

Let us consider the Fokker–Planck equations with quadratic potentials whose initial conditions are Gaussian:

$$V(x) = \frac{1}{2}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu) \quad \text{and} \quad \rho_0 \sim \mathcal{N}(\mu_0, \Sigma_0). \tag{3.26}$$

Here $\mathcal{N}(\mu, \Sigma)$ denotes Gaussian distribution with mean μ and covariance Σ. We consider parameter space $\Theta = (\Gamma, b) \subset \mathbb{R}^m$ ($m = \frac{1}{2}d(d+1)+d$), where Γ is a $d \times d$ symmetric positive definite matrix and $b \in \mathbb{R}^d$. We define the parametric map as $T_\theta(x) = \Gamma x + b$, and choose the reference measure $p = \mathcal{N}(0, I)$.

LEMMA 3.11. *Let $\mathcal{H}$ be the relative entropy defined in* (2.8)*, and let H be defined as in* (3.17)*. For $\theta \in \Theta$, if the vector function $\nabla\left(\frac{\delta \mathcal{H}}{\delta \rho}\right) \circ T_\theta$ can be written as the linear combination of $\{\frac{\partial T_\theta}{\partial \theta_1}, \ldots, \frac{\partial T_\theta}{\partial \theta_m}\}$, i.e., there exists $\zeta \in \mathbb{R}^m$, such that $\nabla\left(\frac{\delta \mathcal{H}}{\delta \rho}\right) \circ T_\theta(x) = \partial_\theta T_\theta(x)\zeta$, then*

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

Proof. Suppose that $\zeta \in \mathbb{R}^m$ satisfies $\nabla\left(\frac{\delta \mathcal{H}}{\delta \rho}\right) \circ T_\theta(x) = \partial_\theta T_\theta(x)\zeta$; then we have

$$\int \left| \partial_\theta T_\theta(x)\zeta - \nabla\left(\frac{\delta \mathcal{H}}{\delta \rho}\right) \circ T_\theta(x) \right|^2 dp(x) = 0.$$

By definition of $\boldsymbol{\Psi}$ in Theorem 3.4, one can verify

$$-\nabla \cdot \left(\rho_\theta \left((\nabla \boldsymbol{\Psi})^{\mathrm{T}}\zeta - \nabla\left(\frac{\delta \mathcal{H}}{\delta \rho}\right)\right)\right) = -\nabla \cdot \left(\rho_\theta \left(\partial_\theta T_\theta \circ T_\theta^{-1}\zeta - \nabla\left(\frac{\delta \mathcal{H}}{\delta \rho}\right)\right)\right).$$

Now we apply (3.3) of Lemma 3.3 to obtain

$$\int \left| (\nabla \boldsymbol{\Psi}(T_\theta(x)))^{\mathrm{T}}\zeta - \nabla\left(\frac{\delta \mathcal{H}}{\delta \rho}\right) \circ T_\theta(x) \right|^2 dp(x) \le 0.$$

This implies

$$\inf_{\eta} \int \left| (\nabla \mathbf{\Psi}(T_\theta(x)))^{\mathrm{T}} \eta - \nabla \left(\frac{\delta \mathcal{H}}{\delta \rho} \right) \circ T_\theta(x) \right|^2 dp(x)$$
$$= \int \left| (\nabla \mathbf{\Psi}(T_\theta(x)))^{\mathrm{T}} \zeta - \nabla \left(\frac{\delta \mathcal{H}}{\delta \rho} \right) \circ T_\theta(x) \right|^2 dp(x) = 0.$$

By Theorem 3.8, we get $\zeta = G(\theta)^{-1} \nabla_\theta H(\theta)$ and $\|(T_{\theta\sharp})_* \zeta - \mathrm{grad}_W \mathcal{H}(\rho_\theta)\|_{g^W(\rho_\theta)} = 0$. The latter leads to $(T_{\theta\sharp})_* \zeta = \mathrm{grad}_W \mathcal{H}(\rho_\theta)$. According to Theorem 3.7, $(T_{\theta\sharp})_* \zeta = \mathrm{grad}_W \mathcal{H}(\rho_\theta)|_{\mathcal{P}_\Theta}$. As a result, we have $\mathrm{grad}_W \mathcal{H}(\rho_\theta)|_{\mathcal{P}_\Theta} = \mathrm{grad}_W \mathcal{H}(\rho_\theta)$. ∎

Back to our example with quadratic potential (3.26) and $T_\theta(x) = \Gamma x + b$, we can compute

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

Then we have

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

which is affine with respect to x.

Noticing that $\partial_{\Gamma_{ij}} T_\theta(x) = (\dots, 0, \dots, \underset{i\text{th}}{x_j}, \dots, 0, \dots)^{\mathrm{T}}$ and $\partial_{b_i} T_\theta = (\dots, 0, \dots, \underset{i\text{th}}{1}, \dots, 0, \dots)^{\mathrm{T}}$, we can verify that $\zeta = (\Sigma^{-1}\Gamma - D\Gamma^{-T}, \Sigma^{-1}(b-\mu))$ solves $\nabla\big(\frac{\delta \mathcal{H}(\rho_\theta)}{\delta \rho}\big) \circ T_\theta(x) = \partial_\theta T_\theta(x) \zeta$. By (1) of Lemma 3.11, $\zeta = G(\theta)^{-1} \nabla_\theta H(\theta)$. Thus ODE (3.18) for our example is

$$\dot{\Gamma} = -\Sigma^{-1}\Gamma + D\Gamma^{-T}, \quad \Gamma_0 = \sqrt{\Sigma_0}, \tag{3.27}$$

$$\dot{b} = \Sigma^{-1}(\mu - b), \quad b_0 = \mu_0. \tag{3.28}$$

By (2) of Lemma 3.11, we know $\mathrm{grad}_{\mathrm{W}} \mathcal{H}(\rho_\theta)|_{\mathcal{P}_\Theta} = \mathrm{grad}_{\mathrm{W}} \mathcal{H}(\rho_\theta)$ for all $\theta \in \Theta$, which indicates that there is no error between our parametric Fokker–Planck and the original equations.

Following (3.27) and (3.28), we have the following corollary,

COROLLARY 3.12. *The solution of the Fokker–Planck equation* (2.2) *with condition* (3.26) *is a Gaussian distribution for all* $t > 0$.

Proof. If we denote $\{\Gamma_t, b_t\}$ as the solutions to (3.27), (3.28) and set $\theta_t = (\Gamma_t, b_t)$, then $\rho_t = T_{\theta_t\sharp} p$ solves the Fokker–Planck equation (2.2) with conditions (3.26). 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 $\rho_t = T_{\theta_t\sharp} p$ is always a Gaussian distribution. ∎

Remark 3.13. This is already a well-known property for the Ornstein–Uhlenbeck process [16]. We provide an alternative proof using our framework.

4. Numerical methods. In this section, we introduce our sampling efficient numerical method to compute the proposed parametric Fokker–Planck equations.

Before we start, we want to mention that, as stated in [35], when dimension $d = 1$, $G(\theta)$ has an explicit solution. Thus the pushforward approximation of the 1D Fokker–Planck equation can be directly computed by solving the ODE system

(3.18) with numerical methods, such as the forward Euler scheme. In this section, our focus is on numerical methods for (3.18) with dimension $d \geq 2$. It turns out to be very challenging to compute (3.18) by the forward Euler scheme directly. There are two reasons. One is that there is no known explicit formula for $G(\theta)$, and direct computation based on (3.5) can be expensive because it requires solving multiple differential equations. The other reason is incurred by the high dimensionality, which is the main goal of this paper. To overcome the challenge of dimensionality, we choose to use deep neural networks to construct our $T(\theta)$. However, directly evaluating $G(\theta)^{-1}\nabla_\theta H(\theta)$ is difficult, and alternative strategies must be sought.

There are a few papers investigating numerical methods for gradient flows on Riemannian manifolds, such as Fisher natural gradient [42] and Wasserstein gradient [12]. The well-known JKO scheme [23] calculates the time discrete approximation of the Wasserstein gradient flow using an optimization formulation,

$$\partial_t \rho_t = -\mathrm{grad}_W \mathcal{F}(\rho_t), \qquad \rho_{k+1} = \operatorname*{argmin}_{\rho \in \mathcal{P}} \left\{ \frac{W_2^2(\rho, \rho_k)}{2h} + \mathcal{F}(\rho) \right\}, \tag{4.1}$$

where h is the time step size, and $\mathcal{F}$ could be a suitable functional defined on $\mathcal{P}$. Along the line of the JKO scheme, there have been further developments in machine learning recently [34].

In our approach, we design schemes that compute the exact Wasserstein gradient flow directly with provable accuracy guarantee. Our algorithms are completely sample-based so that they can be run efficiently under a deep learning framework and can scale up to high-dimensional cases.

4.1. Normalizing flow as pushforward maps. We choose T_θ as the so-called normalizing flow [58]. Here is a brief sketch of its structure: T_θ is written as the composition of K invertible nonlinear transforms:

$$T_\theta = f_K \circ f_{K-1} \circ \cdots \circ f_2 \circ f_1,$$

where each f_k $(1 \leq k \leq K)$ takes the form

$$f_k(x) = x + \sigma(w_k^{\mathrm{T}} x + b_k) u_k.$$

Here $w_k, u_k \in \mathbb{R}^d$, $b_k \in \mathbb{R}$, and σ is a nonlinear function, which can be chosen as tanh, for example. In [58], it has been shown that f_k is invertible iff $w_k^{\mathrm{T}} u_k \geq -1$. Figure 2 shows several snapshots of how a normalizing flow T_θ with length equal to 10 pushes forward standard Gaussian distribution to a target distribution.

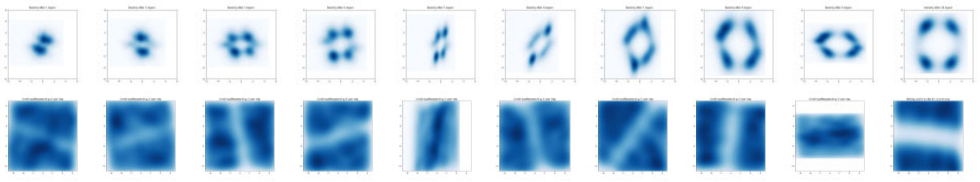


FIG. 2. *Top row from left to right are the probability densities of distributions* $f_{1\sharp}p, (f_2 \circ f_1)_\sharp p, \ldots, (f_{10} \circ f_9 \circ \ldots \circ f_1)_\sharp p$. *The last image displays our target distribution. Bottom row displays the pushforward effect of each single-layer transformation* f_k $(1 \leq k \leq 10)$.

In a normalizing flow, the parameters are $\theta = (w_1, u_1, b_1, \ldots, w_K, u_K, b_K)$. The determinant of the Jacobi matrix of T_θ, an important quantity for our schemes, can

be explicitly computed by

$$\mathbf{det}\left(\frac{\partial T_\theta(x)}{\partial x}\right)=\prod_{k=1}^{K}(1+\sigma'(w_k^{\mathrm{T}}x_k+b_k)w_k^{\mathrm{T}}u_k),$$

where $x_k=f_k\circ f_{k-1}\circ\cdots\circ f_1(x)$. Using the structure of normalizing flow, the logarithm of the density $\rho_\theta=T_{\theta\sharp}p$ can be written as

$$\begin{aligned}\log\rho_\theta(x)&=\log p\circ T_\theta^{-1}(x)-\sum_{k=1}^{K}\log(1+\sigma'(w_k^{\mathrm{T}}\tilde{x}_k)w_k^{\mathrm{T}}u_k),\\ \tilde{x}_k&=f_k\circ\cdots\circ f_1(T_\theta^{-1}(x))=f_{k+1}^{-1}\circ\cdots\circ f_K^{-1}(x).\end{aligned}\tag{4.2}$$

Then we can explicitly write the relative entropy functional $H(\theta)$ defined in (3.17) as

$$H(\theta)=\mathbb{E}_{\mathbf{X}\sim p}[V(T_\theta(\mathbf{X}))+\mathcal{L}_\theta(\mathbf{X})],\tag{4.3}$$

where $\mathcal{L}_\theta$ is defined by

$$\mathcal{L}_\theta(\cdot)=\log p(\cdot)-\sum_{k=1}^{K}\log(1+\sigma'(w_k^{\mathrm{T}}F_k(\cdot))w_k^{\mathrm{T}}u_k),\quad F_k(\cdot)=f_k\circ f_{k-1}\circ\cdots\circ f_1(\cdot).$$

Once $H(\theta)$ is computed explicitly, we can also compute the gradient $\nabla_\theta H(\theta)$ explicitly.

In summary, we choose the normalizing flow because it has sufficient expression power to approximate complicated distributions on $\mathbb{R}^d$ [58], and the relative entropy $H(\theta)$ has a very concise form (4.3), and its gradient can be conveniently computed.

Remark 4.1. We want to emphasize here that the normalizing flow is not the only choice for T_θ. One may choose other network structures as long as they have sufficient approximation power and can compute the gradient of relative entropy efficiently.

4.2. Numerical scheme. For the convenience of our presentation, we first introduce the following definition.

Definition 4.2 (orthogonal projection onto space of gradient fields). *Consider vector field $\vec{v}\in L^2(\mathbb{R}^d;\mathbb{R}^d,\rho)$. Define* $\mathrm{Proj}_\rho[\vec{v}]=\nabla\psi$ *as the $L^2(\rho)$-orthogonal projection of $\vec{v}$ onto the subspace of gradient fields, where ψ solves*

$$\min_\psi\left\{\int|\vec{v}(x)-\nabla\psi(x)|^2\rho(x)\,dx\right\},\tag{4.4}$$

or equivalently ψ solves $-\nabla\cdot(\rho(x)\nabla\psi(x))=-\nabla\cdot(\rho(x)\vec{v}(x))$.

4.2.1. Proposed double-minimization scheme. Our numerical scheme is inspired by the following semi-implicit scheme of (3.18):

$$\frac{\theta_{k+1}-\theta_k}{h}=-G^{-1}(\theta_k)\nabla_\theta H(\theta_{k+1}).$$

Equivalently, we can write it as a proximal algorithm:

$$\theta_{k+1}=\operatorname*{argmin}_\theta\left\{\frac{1}{2}\langle\theta-\theta_k,G(\theta_k)(\theta-\theta_k)\rangle+hH(\theta)\right\}.\tag{4.5}$$

Recall $\mathbf{\Psi}$ as defined in Theorem 3.4; if we denote $\psi = \mathbf{\Psi}^{\mathrm{T}}(\theta - \theta_k)$, we have $\langle(\theta - \theta_k), G(\theta)(\theta - \theta_k)\rangle = \int |\nabla\psi|^2 \rho_{\theta_k}\, dx$ with the constraint that ψ solves the equation

$$-\nabla\cdot(\rho_{\theta_k}\nabla\psi(x)) = -\nabla\cdot(\rho_{\theta_k}\partial_\theta T_{\theta_k}(T_{\theta_k}^{-1}(x))(\theta-\theta_k)). \tag{4.6}$$

By Definition 4.2, $\nabla\psi$ is the orthogonal projection of vector field $\partial_\theta T_{\theta_k}(T_{\theta_k}^{-1}(\cdot))(\theta-\theta_k)$. Equivalently, ψ can also be obtained by solving the least squares problem (4.4).

Based on the observation that $\nabla\psi$ is obtained via orthogonal projection after replacing $\partial_\theta T_{\theta_k}(\theta-\theta_k)$ by finite difference $T_\theta - T_{\theta_k}$, we end up with the following double-minimization scheme for solving (4.5):

$$\min_\theta \left\{ \left(\int \left(2\ \nabla\phi(x)\cdot((T_\theta - T_{\theta_k})\circ T_{\theta_k}^{-1}(x)) - |\nabla\phi(x)|^2\right)\rho_{\theta_k}(x)\ dx\right) + 2hH(\theta)\right\},$$
$$\text{with the constraint: } \phi \text{ solves } \min_\phi \left\{ \int |\nabla\phi(x) - ((T_\theta - T_{\theta_k})\circ T_{\theta_k}^{-1}(x))|^2 \rho_{\theta_k}(x)\ dx\right\}. \tag{4.7}$$

Scheme (4.7) has an equivalent saddle point optimization formulation,

$$\min_\theta \max_\phi \left\{ \left(\int (2\nabla\phi(x)\cdot((T_\theta - T_{\theta_k})\circ T_{\theta_k}^{-1}(x)) - |\nabla\phi(x)|^2)\rho_{\theta_k}(x)\ dx\right) + 2hH(\theta)\right\}, \tag{4.8}$$

which can be directly derived from (4.5) via the adjoint method. Their equivalence is explained in the next remark.

Remark 4.3. Here we briefly demonstrate the equivalence among the three schemes (4.5), (4.7), and (4.8). Our target function $\frac{1}{2}\langle\theta-\theta_k, G(\theta_k)(\theta-\theta_k)\rangle + hH(\theta)$ can be formulated as

$$\int \frac{1}{2}|\nabla\psi(x)|^2\rho_{\theta_k}(x)\ dx + hH(\theta) \quad \text{with the constraint: } \psi \text{ solves (4.6).}$$

By introducing the dual variable ϕ and applying the adjoint method, we obtain

$$\begin{aligned}
&\frac{1}{2}\langle\theta-\theta_k, G(\theta_k)(\theta-\theta_k)\rangle + hH(\theta)\\
&= \max_\phi \min_\psi \left\{ \int \frac{1}{2}|\nabla\psi(x)|^2\rho_{\theta_k}dx + hH(\theta) + \int \phi(x)(\nabla\cdot(\rho_{\theta_k}\nabla\psi(x))\right.\\
&\qquad\qquad \left. - \nabla\cdot(\rho_{\theta_k}\partial_\theta T_{\theta_k}(T_{\theta_k}^{-1}(x))(\theta-\theta_k)))\ dx\right\}\\
&= \max_\phi \min_\psi \left\{ \int \left(\frac{1}{2}|\nabla\psi(x)|^2 - \nabla\phi(x)\cdot\nabla\psi(x) + \nabla\phi(x)\cdot\partial_\theta T_{\theta_k}(T_{\theta_k}^{-1}(x))(\theta-\theta_k)\right)\right.\\
&\qquad\qquad \left. \times\, \rho_{\theta_k}(x)\ dx + hH(\theta)\right\}\\
&= \max_\phi \left\{ \int \left(-\frac{1}{2}|\nabla\phi(x)|^2 + \nabla\phi(x)\cdot\partial_\theta T_{\theta_k}(T_{\theta_k}^{-1}(x))(\theta-\theta_k)\right)\rho_{\theta_k}(x)\ dx + hH(\theta)\right\}.
\end{aligned} \tag{4.9}$$

In implementation, we substitute $\partial_\theta T_{\theta_k}(\theta-\theta_k)$ by $T_\theta - T_{\theta_k}$ since the latter is tractable in computation. As a consequence, by substituting (4.9) into (4.5) we obtain (by

multiplying the entire function by 2) the saddle scheme (4.8). To verify the equivalence between (4.8) and (4.7), we check the identity

$$\begin{aligned}&\int (2\nabla\phi(x)\cdot((T_\theta - T_{\theta_k})\circ T_{\theta_k}^{-1}(x)) - |\nabla\phi(x)|^2)\rho_{\theta_k}(x)\,dx\\ =&-\int |\nabla\phi(x) - (T_\theta - T_{\theta_k})\circ T_{\theta_k}^{-1}(x)|^2\rho_{\theta_k}(x)\,dx + \underbrace{\int |(T_\theta - T_{\theta_k})\circ T_{\theta_k}^{-1}(x)|^2\rho_{\theta_k}(x)\,dx}_{\text{Constant w.r.t. }\phi}.\end{aligned}$$

Thus the ϕ-minimization process of (4.7) is equivalent to the ϕ-maximization process of (4.8). This leads to the equivalence between (4.7) and (4.8).

Remark 4.4. Our proposed schemes (4.7), (4.8) can be viewed as an approximation to the JKO scheme (4.1), with $\mathcal{F}$ being the relative entropy $H(\theta)$. To see this, we denote

$$\mathcal{E}(\phi) = \int (2\nabla\phi(x)\cdot((T_\theta - T_{\theta_k})\circ T_{\theta_k}^{-1}(x)) - |\nabla\phi(x)|^2)\rho_{\theta_k}(x)\,dx$$

and set $\hat{\psi} = \operatorname{argmax}_\phi \mathcal{E}(\phi)$. We let $\vec{v}_h(x) = \frac{1}{h}(T_\theta \circ T_{\theta_k}^{-1}(x) - x)$. Under mild conditions, one can show

(4.10)
$$W_2^2(\rho_\theta, \rho_{\theta_k}) = W_2^2((\mathrm{Id} + h\vec{v}_h)_\sharp\rho_{\theta_k}, \rho_{\theta_k}) = \int |\nabla\hat{\psi}|^2\rho_{\theta_k}\,dx + o(h^2) = \max_\phi \mathcal{E}(\phi) + o(h^2).$$

By replacing $W_2^2(\rho_\theta, \rho_{\theta_k})$ in (4.1) by its approximation $\max_\phi \mathcal{E}(\phi)$, we obtain the scheme (4.7), (4.8).

Although (4.7) and (4.8) are mathematically equivalent, we use them for different purposes. The saddle scheme (4.8) is our main tool to investigate the theoretical properties of our proposed method in section 4.2.2, because it better reflects the nature of our approximation method. In our implementation, as discussed in section 4.2.3, we prefer the double minimization scheme (4.7). Our experience indicates that (4.7) makes our code run more efficiently and behave more stably than (4.8).

4.2.2. Local error of the proposed scheme. We now analyze the local error of scheme (4.8) as well as (4.7) compared with the semi-implicit scheme (4.5). Let us denote $\max_\phi \mathcal{E}(\phi)$ as $\widehat{W}_2^2(\theta, \theta_k)$ (here $\widehat{W}_2$ is treated as an approximation of L^2-Wasserstein distance (Remark 4.4)). It is straightforward to verify $\widehat{W}_2(\theta, \theta') \geq 0$ and $\widehat{W}_2(\theta, \theta) = 0$. Consider the following assumption:

$$\widehat{W}_2^2(\theta, \theta') \geq l(|\theta - \theta'|) \qquad \text{for any} \quad \theta, \theta' \in \Theta. \tag{4.11}$$

Here $l : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ satisfies $l(0) = 0$. $l(r)$ is continuous, strictly increasing when $r \leq r_0$ for a positive r_0 and is bounded below by $\lambda_0 > 0$ when $r > r_0$. Notice that this assumption generally guarantees positive definiteness of $\widehat{W}_2$. Clearly, (4.11) only depends on the structure of T_θ, and we expect that (4.11) holds for the neural networks used as pushforward maps, including those used in this paper.

THEOREM 4.5. *Suppose assumption* (4.11) *holds true for the class of pushforward maps* $\{T_\theta\}$. *Then the local error of scheme* (4.8) *is of order* h^2, *i.e., assuming that* θ_{k+1} *is the optimal solution to* (4.8), *then*

$$|\theta_{k+1} - \theta_k + hG(\theta_k)^{-1}\nabla_\theta H(\theta_{k+1})| \sim O(h^2), \tag{4.12}$$

or equivalently $\limsup_{h\to 0^+} \frac{|\theta_{k+1}-\theta_k+hG(\theta_k)^{-1}\nabla_\theta H(\theta_{k+1})|}{h^2} < +\infty$.

Before proving Theorem 4.5, we introduce some additional notation. We define ϵ to be a ball in parameter space as $B_\epsilon(\theta_k) = \{\theta \mid |\theta - \theta_k| \le \epsilon\}$, and we let $T_\theta^{(i)}$ be the ith component ($1 \le i \le d$) of map T_θ. For fixed θ_k and $\epsilon > 0$ small enough, we assume the following two quantities are finite:

$$
\begin{aligned}
L(\theta_k,\epsilon) &= \sum_{i=1}^{d} \mathbb{E}_{x\sim p} \sup_{\theta\in B_\epsilon(\theta_k)} \left\{ |\partial_\theta T_\theta^{(i)}(x)|^2 \right\}, \\
H(\theta_k,\epsilon) &= \sum_{i=1}^{d} \mathbb{E}_{x\sim p} \sup_{\theta\in B_\epsilon(\theta_k)} \left\{ \|\partial^2_{\theta\theta} T_\theta^{(i)}(x)\|_2^2 \right\}.
\end{aligned} \tag{4.13}
$$

To prove Theorem 4.5, we need the following three lemmas.

LEMMA 4.6. *Suppose we fix $\theta_0 \in \Theta$; for arbitrary $\theta \in \Theta$ and $\nabla\phi \in L^2(\mathbb{R}^d;\mathbb{R}^d,\rho_{\theta_0})$ we consider*

$$
F(\theta,\nabla\phi \mid \theta_0) = \left(\int (2\nabla\phi(x)\cdot(T_\theta - T_{\theta_0})\circ T_{\theta_0}^{-1}(x) - |\nabla\phi(x)|^2)\ \rho_{\theta_0}(x)\ dx \right) + 2hH(\theta). \tag{4.14}
$$

Then $F(\theta,\nabla\phi \mid \theta_0) < \infty$, and furthermore $F(\cdot,\nabla\phi \mid \theta_0) \in C^1(\Theta)$. We can compute

$$
\partial_\theta F(\theta,\nabla\phi \mid \theta_0) = 2\left(\int\ \partial_\theta T_\theta(T_{\theta_0}^{-1}(x))^T\ \nabla\phi(x)\ \rho_{\theta_0}(x)\ dx + h\ \nabla_\theta H(\theta) \right). \tag{4.15}
$$

LEMMA 4.7. *Suppose we fix $\theta_0 \in \Theta$ and define*

$$
J(\theta) = \sup_{\nabla\phi\in L^2(\mathbb{R}^d;\mathbb{R}^d,\rho_{\theta_0})} F(\theta,\nabla\phi \mid \theta_0).
$$

Then J is differentiable. If we denote $\hat{\psi}_\theta = \operatorname{argmax}_\phi \{F(\theta,\nabla\phi \mid \theta_0)\}$, then

$$
\nabla_\theta J(\theta) = \partial_\theta F(\theta,\nabla\hat{\psi}_\theta \mid \theta_0) = 2\left(\int\ \partial_\theta T_\theta(T_{\theta_0}^{-1}(x))^T\ \nabla\hat{\psi}_\theta(x)\ \rho_{\theta_0}(x)\ dx + h\ \nabla_\theta H(\theta) \right).
$$

This lemma is an analogy of the envelope theorem [1] under our problem setting.

LEMMA 4.8. *Under assumption* (4.11)*, the optimal solution of* (4.8) θ_{k+1} *satisfies*

$$
|\theta_{k+1} - \theta_k| \sim o(1), \quad \textit{i.e.,} \quad \lim_{h\to 0^+} |\theta_{k+1} - \theta_k| = 0.
$$

This lemma provides an a priori estimation of $|\theta_{k+1} - \theta_k|$.

We prove Lemmas 4.6, 4.7, and 4.8 in Appendix C.

Proof of Theorem 4.5. Let us consider $F(\theta,\nabla\phi \mid \theta_k)$. We denote

$$
\nabla\hat{\psi}_\theta = \operatorname*{argmax}_{\nabla\phi\in L^2(\mathbb{R}^d;\mathbb{R}^d,\rho_{\theta_k})} \{F(\theta,\nabla\phi \mid \theta_k)\}.
$$

Then we can set

$$
\nabla\hat{\psi}_\theta = \operatorname{Proj}_{\rho_{\theta_k}}[(T_\theta - T_{\theta_k})\circ T_{\theta_k}^{-1}] \quad \text{and} \quad J(\theta) = \sup_{\nabla\phi\in L^2(\mathbb{R}^d;\mathbb{R}^d,\rho_{\theta_k})} F(\theta,\nabla\phi \mid \theta_k).
$$

Applying Lemma 4.7, we obtain

$$\nabla_\theta J(\theta) = 2\left(\int \partial_\theta T_\theta(T_{\theta_k}^{-1}(x))^{\mathrm{T}}\ \nabla\hat{\psi}_\theta(x)\ \rho_{\theta_k}(x)\ dx + h\ \nabla_\theta H(\theta)\right).$$

Due to the differentiability of $J(\theta)$, at the optimizer θ_{k+1}, the gradient must vanish, i.e.,

$$\left(\int \partial_\theta T_{\theta_{k+1}}(T_{\theta_k}^{-1}(x))^{\mathrm{T}}\ \nabla\hat{\psi}_{\theta_{k+1}}(x)\ \rho_{\theta_k}(x)\ dx\right) + h\nabla_\theta H(\theta_{k+1}) = 0. \tag{4.16}$$

We use Taylor expansion at θ_{k+1} to get $T_{\theta_{k+1}} - T_{\theta_k} = \partial_\theta T_{\theta_k}(\theta_{k+1} - \theta_k) + R(\theta_{k+1}, \theta_k)$, in which $R(\theta, \theta')(\cdot) \in L^2(\mathbb{R}^d; \mathbb{R}^m, \rho_{\theta_k})$, and the ith entry of $R(\theta, \theta')$ is $R_i(\theta, \theta')(x) = \frac{1}{2}(\theta - \theta')^{\mathrm{T}}\partial^2_{\theta\theta}T^{(i)}_{\tilde{\theta}_i(x)}(x)(\theta - \theta')$, $1 \le i \le m$, where each $\tilde{\theta}_i(x) = \lambda_i(x)\theta + (1 - \lambda_i(x))\theta'$ for some $\lambda_i(x) \in [0, 1]$. Then we can write

$$\begin{aligned}\nabla\hat{\psi}_{\theta_{k+1}} &= \mathrm{Proj}_{\rho_{\theta_k}}[(T_{\theta_{k+1}} - T_{\theta_k}) \circ T_{\theta_k}^{-1}] \\ &= \mathrm{Proj}_{\rho_{\theta_k}}[\partial_\theta T_{\theta_k} \circ T_{\theta_k}^{-1}(\theta_{k+1} - \theta_k)] + \mathrm{Proj}_{\rho_{\theta_k}}[R(\theta_{k+1}, \theta_k) \circ T_{\theta_k}^{-1}].\end{aligned} \tag{4.17}$$

On the other hand,

$$\partial_\theta T_{\theta_{k+1}} = \partial_\theta T_{\theta_k} + r(\theta_{k+1}, \theta_k). \tag{4.18}$$

Here $r(\theta, \theta') \in L^2(\mathbb{R}^d; \mathbb{R}^{d\times m}, \rho_{\theta_k})$, and the (i, j)th entry of $r(\theta, \theta')(x)$ is $(\theta_{k+1} - \theta_k)^{\mathrm{T}}\partial_\theta(\partial_{\theta_j}T^{(i)}_{\tilde{\theta}_{ij}(x)}(x))$, $1 \le i \le d$, $1 \le j \le m$, where each $\tilde{\theta}_{ij}(x) = \mu_{ij}(x)\theta_{k+1} + (1 - \mu_{ij}(x))\theta_k$ for some $\mu_{ij}(x) \in (0, 1)$. Applying (4.18), (4.17) to (4.16), we obtain

$$\begin{aligned}&\int \partial_\theta T_{\theta_k}(T_{\theta_k}^{-1}(x))^{\mathrm{T}}\mathrm{Proj}_{\rho_{\theta_k}}[\partial_\theta T_{\theta_k} \circ T_{\theta_k}^{-1}(x)(\theta_{k+1} - \theta_k)]\ \rho_{\theta_k}(x)\ dx \\ &+ \int \partial_\theta T_{\theta_k}(T_{\theta_k}^{-1}(x))^{\mathrm{T}}\mathrm{Proj}_{\rho_{\theta_k}}[R(\theta_{k+1}, \theta_k) \circ T_{\theta_k}^{-1}](x)\ \rho_{\theta_k}(x)\ dx \\ &+ \int r(\theta_{k+1}, \theta_k)(T_{\theta_k}^{-1}(x))^{\mathrm{T}}\mathrm{Proj}_{\rho_{\theta_k}}[(T_{\theta_{k+1}} - T_{\theta_k}) \circ T_{\theta_k}^{-1}](x)\ \rho_{\theta_k}(x)\ dx = -h\nabla_\theta H(\theta_{k+1}).\end{aligned} \tag{4.19}$$

Recalling the definition of $\mathbf{\Psi}$ in Theorem 3.4, and using (3.3) in Lemma 3.3, we know that the first term on the left-hand side of (4.19) equals

$$\int \nabla\mathbf{\Psi}(x)\nabla\mathbf{\Psi}(x)^{\mathrm{T}}(\theta_{k+1} - \theta_k)\ \rho_{\theta_k}(x)\ dx = G(\theta_k)(\theta_{k+1} - \theta_k).$$

By applying the Cauchy–Schwarz inequality and (3.4) in Lemma 3.3, we bound the ith entry of the second term in (4.19) by

$$\begin{aligned}&\left(\int |\partial_\theta T^{(i)}_{\theta_k}(x)|^2\ dp(x) \cdot \int \sum_{i=1}^d |(\theta_{k+1} - \theta_k)\partial^2_{\theta\theta}T^{(i)}_{\tilde{\theta}_i(x)}(x)(\theta_{k+1} - \theta_k)|^2\ dp(x)\right)^{\frac{1}{2}} \\ &\le \left(\mathbb{E}_p|\partial_\theta T^{(i)}_{\theta_k}(x)|^2 \cdot \mathbb{E}_p\left[\sum_{i=1}^d \|\partial^2_{\theta\theta}T^{(i)}_{\tilde{\theta}_i(x)}(x)\|_2\right]\right)^{\frac{1}{2}} |\theta_{k+1} - \theta_k|^2 \overset{\text{denote as}}{=} A^{(i)}|\theta_{k+1} - \theta_k|^2.\end{aligned}$$

To bound the third term in (4.19), we consider the ith entry of $T_{\theta_{k+1}}(x) - T_{\theta_k}(x)$, which can be written as

$$T^{(i)}_{\theta_{k+1}}(x) - T^{(i)}_{\theta_k}(x) = \partial_\theta T_{\bar{\theta}_i(x)}(x)(\theta_{k+1} - \theta_k),$$

where $\bar{\theta}_i(x) = \zeta_i(x)\theta_{k+1} + (1 - \zeta_i(x))\theta_k$ for some $\zeta_i(x) \in (0, 1)$. The ith entry of the third term of (4.19) can be bounded by

$$\left(\int \sum_{i=1}^{d} |(\theta_{k+1} - \theta_k)^{\mathrm{T}} \partial_{\theta\theta} T^{(i)}_{\tilde{\theta}_{ij}(x)}(x)|^2 \, dp(x) \cdot \int |T^{(i)}_{\theta_{k+1}}(x) - T^{(i)}_{\theta_k}(x)|^2 \, dp(x) \right)^{\frac{1}{2}}$$

$$\leq \left(\mathbb{E}_p \left[\sum_{i=1}^{d} \|\partial^2_{\theta\theta} T_{\tilde{\theta}_{ij}(x)}(x)\|_2^2 \right] \cdot \mathbb{E}_p |\partial_\theta T^{(i)}_{\bar{\theta}_i(x)}(x)|^2 \right)^{\frac{1}{2}} |\theta_{k+1} - \theta_k|^2 \overset{\text{denote as}}{=} B^{(i)} |\theta_{k+1} - \theta_k|^2.$$

We denote $A \in \mathbb{R}^m$ with entries $A^{(i)}$, $1 \leq i \leq m$, and similarly $B \in \mathbb{R}^m$ with entries $B^{(i)}$, $1 \leq i \leq m$. Equation (4.19) leads to the following inequality:

$$|\theta_{k+1} - \theta_k + hG(\theta_k)^{-1} \nabla_\theta H(\theta_{k+1})| \leq \|G(\theta_k)^{-1}\|_2 (|A| + |B|) \, |\theta_{k+1} - \theta_k|^2.$$

As we have shown in Lemma 4.8 that $|\theta_{k+1} - \theta_k| \sim o(1)$ for any $\epsilon > 0$ when step size h is small enough, we always have $\theta_{k+1} \in B_\epsilon(\theta_k)$. Recalling the notation in (4.13), we have $|A|, |B| \leq \sqrt{L(\theta_k, \epsilon) H(\theta_k, \epsilon)}$. Thus we have

$$|\theta_{k+1} - \theta_k + hG(\theta_k)^{-1} \nabla_\theta H(\theta_{k+1})| \leq 2\sqrt{L(\theta_k, \epsilon) H(\theta_k, \epsilon)} \|G(\theta_k)^{-1}\|_2 |\theta_{k+1} - \theta_k|^2.$$

Denote $\theta_{k+1} - \theta_k = \eta$, $G(\theta_k)^{-1} \nabla_\theta H(\theta_{k+1}) = \xi$, and $C = 2\sqrt{L(\theta_k, \epsilon) H(\theta_k, \epsilon)} \|G(\theta_k)^{-1}\|_2$; the previous inequality is

$$|\eta \ - \ h \ \xi| \leq C |\eta|^2. \tag{4.20}$$

Since $|\eta - h\xi| \geq |\eta| - h|\xi|$, we have

$$C|\eta|^2 \geq |\eta| - h|\xi|. \tag{4.21}$$

Solving (4.21) gives

$$|\eta| \leq \frac{2|\xi|h}{1 + \sqrt{1 - 4C|\xi|h}} \quad \text{or} \quad |\eta| > \frac{1 + \sqrt{1 - 4Ch|\xi|}}{2C}.$$

The second inequality leads to $|\theta_{k+1} - \theta_k| > \frac{1}{2C}$ for any $h > 0$, which avoids $|\theta_{k+1} - \theta_k| \sim o(1)$. Thus, when h is sufficiently small, we have

$$|\eta| \leq \frac{2|\xi|h}{1 + \sqrt{1 - 4C|\xi|h}}. \tag{4.22}$$

Combining (4.22) and (4.20), we have

$$|\theta_{k+1} - \theta_k + hG(\theta_k)^{-1} \nabla_\theta H(\theta_{k+1})| \leq \frac{4 \ C \ |\xi|^2}{(1 + \sqrt{1 - 4C|\xi|h})^2} \ h^2 \leq 4C|\xi|^2 h^2. \tag{4.23}$$

This proves the result. ∎

Remark 4.9. One may be aware of the relation between the positive definite condition (4.11) and the positive definiteness of the metric tensor $G(\theta_k)$. A positive definite $G(\theta)$ guarantees the inequality $\widehat{W}_2^2(\theta,\theta') \geq C|\theta-\theta'|^2$ for $\theta' \in B_{r_0}(\theta)$ (r_0 depends on θ is small enough). However, we are not able to bound $\widehat{W}_2^2(\theta,\theta')$ from below when $|\theta-\theta'| > r_0$. On the other hand, (4.11) is a locally weaker condition than the positive definiteness of $G(\theta)$.

4.2.3. Implementation. As mentioned in section 4.2.1, we prefer double-minimization scheme (4.7) over saddle scheme (4.8). We will thus implement scheme (4.7). Let us denote

$$J(\theta) = \left(\int \left(2\, \nabla\hat{\psi}(T_{\theta_k}(x)) \cdot ((T_\theta(x) - T_{\theta_k}(x))) - |\nabla\hat{\psi}(T_{\theta_k}(x))|^2 \right) dp(x) \right) + 2hH(\theta), \tag{4.24}$$

$$\text{with} \quad \hat{\psi} = \operatorname*{argmin}_{\phi} \left\{ \int |\nabla\phi(T_{\theta_k}) - (T_\theta(x) - T_{\theta_k}(x))|^2 dp(x) \right\}. \tag{4.25}$$

We then solve ODE (3.18) at t_k by solving

$$\theta_{k+1} = \operatorname*{argmin}_{\theta} J(\theta). \tag{4.26}$$

Here we provide some detailed discussion on our implementation.

- In our numerical computation, we approximate ϕ by $\psi_\nu : M \to \mathbb{R}$, which is a ReLU neural network [17]. Here ν denotes the parameter vector of the network ψ_ν. We know that in this case ψ_ν is a piecewise affine function and its gradient $\nabla\psi_\nu(\cdot)$ forms a piecewise constant vector field.
- The entire procedure of solving (4.26) can be formulated as nested loops:
  - (inner loop) Every inner loop aims at solving (4.25) on ReLU functions ψ_ν, i.e., solving

    $$\min_{\nu} \left\{ \mathbb{E}_{\boldsymbol{X}\sim p} |\nabla\psi_\nu(T_{\theta_k}(\boldsymbol{X})) - (T_\theta(\boldsymbol{X}) - T_{\theta_k}(\boldsymbol{X}))|^2 \right\}. \tag{4.27}$$

    One can use stochastic gradient descent (SGD) methods like RMSProp [62] or Adam [26] with learning rate α_{in} to deal with this inner loop optimization. In our implementation, we will stop after M_{in} iterations. Let us denote the optimal ν in each inner loop as $\hat{\nu}$.
  - (outer loop) We apply a similar SGD method to $J(\theta)$: using Lemma 4.7, we are able to compute $\nabla_\theta J(\theta)$ as

    $$\nabla_\theta J(\theta) = \partial_\theta \left(\left(\int 2\nabla\hat{\psi}(x) \cdot (T_\theta \circ T_{\theta_k}^{-1}(x)) \rho_{\theta_k}(x)\, dx \right) + 2hH(\theta) \right).$$

    If we treat optimal $\hat{\psi}$ as $\psi_{\hat{\nu}}$, what we need to do in each outer loop is to consider

    $$\tilde{J}(\theta) = \mathbb{E}_{\boldsymbol{X}\sim p}\, 2[\nabla\psi_{\hat{\nu}}(T_{\theta_k}(\boldsymbol{X})) \cdot T_\theta(\boldsymbol{X})] + 2h[V(T_\theta(\boldsymbol{X})) + \mathcal{L}_\theta(\boldsymbol{X})] \tag{4.28}$$

    and update θ for one step by our chosen SGD method with learning rate α_{out} applied to optimize $\tilde{J}(\theta)$. In our actual computation, we will stop the outer loop after M_{out} iterations.

- We now present the entire algorithm for computing (3.18) based on the scheme (4.7) in Algorithm 4.1. This algorithm contains the following parameters: $T, N; M_{\text{out}}, K_{\text{out}}, \alpha_{\text{out}}; M_{\text{in}}, K_{\text{in}}, \alpha_{\text{in}}$. Recall that we set reference distribution p as standard Gaussian on $M = \mathbb{R}^d$.

Algorithm 4.1 Computing (3.18) by scheme (4.8) on the time interval $[0, T]$.

Initialize θ
for $i = 1, \dots, N$ **do**
  Save current parameter value to θ_0: $\theta_0 = \theta$
  for $j = 1, \dots, M_{\text{out}}$ **do**
    for $p = 1, \dots, M_{\text{in}}$ **do**
      Sample $\{\mathbf{X}_1, \dots, \mathbf{X}_{K_{\text{in}}}\}$ from p
      Apply one SGD (Adam) step with learning rate α_{in} to loss function of variable λ.

$$\frac{1}{K_{\text{in}}}\left(\sum_{k=1}^{K_{\text{in}}} |\nabla\psi_\nu(T_{\theta_0}(\mathbf{X}_k)) - (T_\theta(\mathbf{X}_k) - T_{\theta_0}(\mathbf{Y}_k))|^2\right)$$

    end for
    Sample $\{\mathbf{X}_1, \dots, \mathbf{X}_{K_{\text{out}}}\}$ from p
    Apply one SGD (Adam) step with learning rate α_{out} to loss function of variable θ.

$$\frac{1}{K_{\text{out}}}\left(\sum_{k=1}^{K_{\text{out}}} 2[\nabla\psi_\nu(T_{\theta_0}(\boldsymbol{X}_k)) \cdot T_\theta(\boldsymbol{X}_k)] + 2h[V(T_\theta(\boldsymbol{X}_k)) + \mathcal{L}_\theta(\boldsymbol{X}_k)]\right)$$

  end for
  Set $\theta_i = \theta$
end for
The sequence of probability densities $\{T_{\theta_0\sharp}p, T_{\theta_1\sharp}p, \dots, T_{\theta_N\sharp}p\}$ will be the numerical solution of $\{\rho_{t_0}, \rho_{t_1}, \dots, \rho_{t_N}\}$, where $t_i = i\frac{T}{N}$ $(i = 0, 1, \dots, N-1, N)$. Here ρ_t solves the original Fokker–Planck equation (2.2).

Remark 4.10 (rescaling). In our implementation, $T_\theta(\boldsymbol{X}) - T_{\theta_k}(\boldsymbol{X})$ is usually of order $O(\alpha_{\text{out}})$, which is a small quantity. We can rescale it so that each inner loop can be solved in a more stable way with larger step size (learning rate). That is to say, we choose some small $\epsilon \sim O(\alpha_{\text{out}})$ and consider

(4.29)
$$\min_\theta \max_\phi \left\{ \underbrace{\left(\int \left(2\nabla\phi(x)\cdot\left(\frac{1}{\epsilon}(T_\theta - T_{\theta_k})\circ T_{\theta_k}^{-1}(x)\right) - |\nabla\phi(x)|^2\right)\rho_{\theta_k}(x)\, dx\right)}_{\mathcal{E}_\epsilon(\phi)} + \frac{2h}{\epsilon^2}H(\theta)\right\}.$$

We can also check

$$\operatorname{argmax}\mathcal{E}_\epsilon(\phi) = \operatorname{Proj}_{\rho_{\theta_k}}\left[\frac{1}{\epsilon}(T_\theta - T_{\theta_k})\circ T_{\theta_k}^{-1}\right] = \frac{1}{\epsilon}\operatorname{Proj}_{\rho_{\theta_k}}[(T_\theta - T_{\theta_k})\circ T_{\theta_k}^{-1}] = \frac{1}{\epsilon}\operatorname{argmax}\mathcal{E}(\phi).$$

Using this, we are able to verify $\max_\phi \mathcal{E}_\epsilon(\phi) = \frac{1}{\epsilon^2}\max_\phi \mathcal{E}(\phi)$. Thus the optimal

solution of (4.29) is

$$\operatorname{argmin}_\theta \left\{ \frac{1}{\epsilon^2} \max_\phi \ \mathcal{E}(\phi) + \frac{2h}{\epsilon^2} H(\theta) \right\} = \operatorname{argmin}_\theta \left\{ \max_\phi \ \mathcal{E}(\phi) + 2hH(\theta) \right\}.$$

This shows the equivalence between the modified scheme (4.29) and the original scheme (4.8).

In our actual implementation, we still prefer the double-minimization scheme. We solve

$$\min_\nu \left\{ \mathbb{E}_{\boldsymbol{X}\sim p} \left| \nabla \psi_\nu(T_{\theta_k}(\boldsymbol{X})) - \left(\frac{T_\theta(\boldsymbol{X}) - T_{\theta_k}(\boldsymbol{X})}{\epsilon} \right) \right|^2 \right\} \tag{4.30}$$

instead of (4.27) in each inner loop and set

$$\tilde{J}(\theta) = \mathbb{E}_{\boldsymbol{X}\sim p} \ 2[\nabla \psi_{\hat{\nu}}(T_{\theta_k}(\boldsymbol{X})) \cdot T_\theta(\boldsymbol{X})] + \frac{2h}{\epsilon}[V(T_\theta(\boldsymbol{X})) + \mathcal{L}_\theta(\boldsymbol{X})] \tag{4.31}$$

in each outer loop. In actual experiments, we set $\epsilon = \alpha_{\text{out}}$.

Remark 4.11 (sufficiently large sample size). It is worth mentioning that the sample size $K_{\text{in}}, K_{\text{out}}$ in each SGD step (especially K_{in}) should be chosen reasonably large so that the inner optimization problem can be solved with enough accuracy. In practice, we usually choose $K_{\text{in}} = K_{\text{out}} = \max\{1000, 300d\}$. Here d is the dimension of sample space. This is very different from the small batch technique applied to training the neural network in deep learning [43].

Remark 4.12 (using fixed samples). Our numerical experiments indicate that the same samples can be used for both the inner and outer iterations, which may reduce the computational cost of our original algorithm.

5. Asymptotic properties and error estimations. In this section, we establish numerical analysis for the parametric Fokker–Planck equation (3.18).

5.1. An important quantity. Before our analysis, we introduce an important quantity that plays an essential role in our numerical analysis. Let us recall the optimal value of the least squares problem (3.19) in Theorem 3.8 of section 3.2, or equivalently (3.20) of section 3.2, and (3.25) of section 3.3. If we denote the upper bound of all possible values to be δ_0, i.e.,

$$\delta_0 = \sup_{\theta\in\Theta} \min_{\xi\in\mathbb{R}^m} \left\{ \int \left| \sum_{k=1}^M \xi_k \nabla \psi_k(x) - \nabla \left(V(x) + D \log \rho_\theta(x) \right) \right|^2 \rho_\theta(x) dx \right\}, \tag{5.1}$$

where ψ_k are solutions to (3.6) in Theorem 3.4, then this quantity provides a crucial error bound between our parametric equation and original equation in the forthcoming analysis. Ideally, we hope δ_0 to be sufficiently small. This can be guaranteed if the neural network we select has universal approximation power. δ_0 can be bounded by another constant with a more approachable form:

$$\hat{\delta}_0 = \sup_{\theta\in\Theta} \min_{\xi\in\mathbb{R}^m} \left\{ \int \left| \sum_{k=1}^M \xi_k \frac{\partial T_\theta(x)}{\partial \theta_k} - \nabla \left(V(x) + D \log \rho_\theta(x) \right) \right|^2 \rho_\theta(x) dx \right\}. \tag{5.2}$$

By (3.4) of Lemma 3.3, one can verify $\delta_0 \le \hat{\delta}_0$. From (5.2), we observe that $\hat{\delta}_0$ is determined by the optimal linear combination of $\{\frac{\partial T_\theta}{\partial \theta_k}\}_{k=1}^M$ to approximate the vector

field $\nabla(V + D\log\rho_\theta)$. One may understand this approximation from three different aspects:

- If T_θ is chosen as a linear combination of basis functions, i.e., $T_\theta(x) = \sum_{k=1}^{M}\theta_k\Phi_k(x)$, we can give an explicit estimate on $\hat{\delta}_0$. For example, if $\Phi_k(x)$ is picked as the Fourier basis and $\nabla(V + D\log\rho_\theta) \in H^s$ $(s > 1)$, the classical spectral method theory can be applied to obtain an estimate $\hat{\delta}_0 = O(M^{-s})$ [49, 66]. If the radial basis function is selected, a related approximation bounded can be obtained too [9].
- Having a small value for $\hat{\delta}_0$ as well as δ_0 is equivalent to finding a suitable T_θ such that a specific vector field $\nabla(V + D\log\rho_\theta)$ can be accurately approximated in our estimate. In other words, when neural networks are used for T_θ, one needs to pick a neural network structure such that it can approximate $\nabla(V + D\log\rho_\theta)$ well. This seems to be an easier question than the task for the so-called universal approximation theory for neural networks, which requires T_θ to approximate an arbitrary function in a space.
- In our implementation, we use normalizing flows, a special type of deep neural network. Our numerical examples seem to show promising performance. In the existing literature, although there are several references providing the universal approximation power of neural networks [72, 15], the results are mainly focused on general ReLU networks and on the approximation power of function value, which is different from our case. To the best of our knowledge, there is no existing study discussing explicit bounds for vector field approximation by deep neural networks. We believe that the question of how δ_0 or $\hat{\delta}_0$ explicitly depends on the structure of T_θ is a fundamental research problem that deserves careful investigations.

It is also worth mentioning that δ_0 is used for an a priori estimate in this section, because we don't know the exact trajectory of $\{\theta_t\}$ when solving ODE (3.18), and we take the supremum over Θ to obtain δ_0. Once solved for $\{\theta_t\}$, denote by $\mathcal{C}$ the set covering its trajectory, i.e.,

$$\mathcal{C} = \{\theta \mid \exists\ t \geq 0,\ \text{s.t. } \theta = \theta_t\}. \tag{5.3}$$

We define another quantity, δ_1:

$$\delta_1 = \sup_{\theta\in\mathcal{C}} \min_{\xi\in\mathcal{T}_\theta\Theta} \left\{ \int |\nabla\boldsymbol{\Psi}(T_\theta(x))^{\mathrm{T}}\xi - \nabla\left(V + D\log\rho_\theta\right)\circ T_\theta(x)|^2\ dp(x) \right\}. \tag{5.4}$$

Clearly, we have $\delta_1 \leq \delta_0$. We can obtain corresponding *posterior* estimates for the asymptotic convergence and error analysis by replacing δ_0 with δ_1.

5.2. Asymptotic convergence analysis. In this section, we consider the solution $\{\theta_t\}_{t\geq 0}$ of our parametric Fokker–Planck equation (3.18). We define

$$\mathcal{V} = \left\{ V \;\middle|\; \begin{array}{c} V \in \mathcal{C}^2(\mathbb{R}^d),\ V \text{ can be decomposed as } V = U + \phi, \text{ with } U, \phi \in \mathcal{C}^2(\mathbb{R}^d); \\ \nabla^2 U \succeq KI \text{ with } K > 0 \text{ and } \phi \in L^\infty(\mathbb{R}^d) \end{array} \right\}.$$

As we know, for the Fokker–Planck equation (2.2), when the potential $V \in \mathcal{V}$, $\{\rho_t\}$ will converge to the Gibbs distribution $\rho_* = \frac{1}{Z_D}e^{-V(x)/D}$ as $t \to \infty$ under the measure of KL divergence [20]. For (3.18), we wish to study its asymptotic convergence property. We come up with the following result.

THEOREM 5.1 (a priori estimation on asymptotic convergence). *Consider the Fokker–Planck equation* (2.2) *with the potential* $V \in \mathcal{V}$. *Suppose* $\{\theta_t\}$ *solves the parametric Fokker–Planck equation* (3.18), *and denote* δ_0 *as in* (5.1). *Let* $\rho_*(x) = \frac{1}{Z_D} e^{-V(x)/D}$ *be the Gibbs distribution of original equation* (2.2). *Then we have the inequality*

$$\mathcal{D}_{KL}(\rho_{\theta_t} \| \rho_*) \leq \frac{\delta_0}{\tilde{\lambda}_D D^2}(1 - e^{-D\tilde{\lambda}_D t}) + \mathcal{D}_{KL}(\rho_{\theta_0} \| \rho_*) e^{-D\tilde{\lambda}_D t}. \tag{5.5}$$

Here $\tilde{\lambda}_D > 0$ *is the constant associated to the logarithmic Sobolev inequality discussed in Lemma* 5.2 *with potential function* $\frac{1}{D}V$.

To prove Theorem 5.1, we need the following two lemmas.

LEMMA 5.2 (Holley–Stroock perturbation). *Suppose the potential* $V \in \mathcal{V}$ *is decomposed as* $V = U + \phi$, *where* $\nabla^2 U \succeq KI$ *and* $\phi \in L^\infty$. *Let* $\tilde{\lambda} = Ke^{-osc(\phi)}$, *where* $osc(\phi) = \sup \phi - \inf \phi$. *Then the following logarithmic Sobolev inequality holds for any probability density* ρ:

$$\mathcal{D}_{KL}(\rho \| \rho_*) \leq \frac{1}{\tilde{\lambda}} \mathcal{I}(\rho | \rho_*). \tag{5.6}$$

Here $\rho_* = \frac{1}{Z} e^{-V}$ and $\mathcal{I}(\rho|\rho_*)$ is the Fisher information functional defined as

$$\mathcal{I}(\rho|\rho_*) = \int \left| \nabla \log \left(\frac{\rho(x)}{\rho_*(x)} \right) \right|^2 \rho(x)\, dx.$$

Lemma 5.2 was first proved in [20].

LEMMA 5.3. *For any* $\theta \in \Theta$, *we have*

$$D^2\, \mathcal{I}(\rho_\theta | \rho_*) \leq \delta_0 + \nabla_\theta H(\theta) \cdot G(\theta)^{-1} \nabla_\theta H(\theta), \tag{5.7}$$

where δ_0 *is defined in* (5.1).

Proof of Lemma 5.3. Let us denote $\xi = G(\theta)^{-1}\nabla_\theta H(\theta)$ for convenience. Suppose $\{\theta_t\}$ solves (3.18) with $\theta_0 = \theta$. By Theorem 3.7, $\frac{d}{dt}\rho_{\theta_t}\big|_{t=0} = -(T_{\theta\sharp})_*\xi$ is an orthogonal projection of $-\mathrm{grad}_W \mathcal{H}(\rho_\theta)$ onto $\mathcal{T}_{\rho_\theta}\mathcal{P}$ with respect to metric g^W. Thus the orthogonal relation gives

$$\begin{aligned} g^W(-\mathrm{grad}_W \mathcal{H}(\rho_\theta), -\mathrm{grad}_W \mathcal{H}(\rho_\theta)) &= g^W(\mathrm{grad}_W \mathcal{H}(\rho_\theta) - (T_{\theta\sharp})_*\xi, \mathrm{grad}_W \mathcal{H}(\rho_\theta) - (T_{\theta\sharp})_*\xi) \\ &\quad + g^W((T_{\theta\sharp})_*\xi, (T_{\theta\sharp})_*\xi). \end{aligned} \tag{5.8}$$

One can verify that the left-hand side of (5.8) is

$$g^W(-\mathrm{grad}_W \mathcal{H}(\rho_\theta), -\mathrm{grad}_W \mathcal{H}(\rho_\theta)) = \int |\nabla(V(x) + D\log \rho_\theta(x))|^2 \rho(x)\, dx = D^2\, \mathcal{I}(\rho_\theta|\rho_*). \tag{5.9}$$

Recalling the equivalence between (3.19) and (3.20) and the definition of δ_0 in (5.1), we know that the first term on the right-hand side of (5.8) has an upper bound,

$$g^W(\mathrm{grad}_W \mathcal{H}(\rho_\theta) - (T_{\theta\sharp})_*\xi, \mathrm{grad}_W \mathcal{H}(\rho_\theta) - (T_{\theta\sharp})_*\xi) \leq \delta_0. \tag{5.10}$$

The second term on the right-hand side of (5.8) is

$$\begin{aligned} g^W((T_{\theta\sharp})_*\xi, (T_{\theta\sharp})_*\xi) = (T_{\theta\sharp})^* g^W(\xi, \xi) &= G(\theta)(G(\theta)^{-1}\nabla_\theta H(\theta),\ G(\theta)^{-1}\nabla_\theta H(\theta)) \\ &= \nabla_\theta H(\theta) \cdot G(\theta)^{-1} \nabla_\theta H(\theta). \end{aligned} \tag{5.11}$$

Combining (5.8), (5.9), (5.10), and (5.11) yields (5.7). ∎

Proof of Theorem 5.1. Let us recall the relationship between KL divergence and relative entropy,

$$\mathcal{D}_{\mathrm{KL}}(\rho\|\rho_*) = \frac{1}{D}\mathcal{H}(\rho) + \log(Z_D).$$

Actually, we can treat $\mathcal{D}_{\mathrm{KL}}(\rho_\theta\|\rho_*)$ as a Lyapunov function for our ODE (3.18), because by taking the time derivative of $\mathcal{D}_{\mathrm{KL}}(\rho_{\theta_t}\|\rho_*)$, we obtain

$$\frac{d}{dt}\mathcal{D}_{\mathrm{KL}}(\rho_{\theta_t}\|\rho_*) = \frac{1}{D}\frac{d}{dt}\mathcal{H}(\rho_{\theta_t}) = \frac{1}{D}\dot{\theta}_t \cdot \nabla H(\theta_t) = -\frac{1}{D}\nabla H(\theta_t)\cdot G^{-1}(\theta_t)\nabla H(\theta_t).$$

Using the inequality in Lemma 5.3, we are able to show that

$$\frac{d}{dt}\mathcal{D}_{\mathrm{KL}}(\rho_{\theta_t}\|\rho_*) \le \frac{\delta_0}{D} - D\ \mathcal{I}(\rho_{\theta_t}|\rho_*).$$

By Lemma 5.2, we have

$$\frac{d}{dt}\mathcal{D}_{\mathrm{KL}}(\rho_{\theta_t}\|\rho_*) \le \frac{\delta_0}{D} - D\ \tilde{\lambda}_D\ \mathcal{D}_{\mathrm{KL}}(\rho_{\theta_t}\|\rho_*).$$

Therefore we obtain, by Gronwall's inequality, the following estimate:

$$\mathcal{D}_{\mathrm{KL}}(\rho_{\theta_t}\|\rho_*) \le \frac{\delta_0}{\tilde{\lambda}_D D^2}(1 - e^{-D\tilde{\lambda}_D t}) + \mathcal{D}_{\mathrm{KL}}(\rho_{\theta_0}\|\rho_*)e^{-D\tilde{\lambda}_D t}. \qquad \square$$

Remark 5.4. Following the previous proof, we can show a similar convergence estimation for the solution $\{\rho_t\}_{t\ge 0}$ of (2.2). Such a result was first discovered in [7].

$$\mathcal{D}_{\mathrm{KL}}(\rho_t\|\rho_*) \le \mathcal{D}_{\mathrm{KL}}(\rho_0\|\rho_*)\ e^{-D\tilde{\lambda}_D t} \quad \forall\ t > 0. \tag{5.12}$$

A nominal modification of our proof for Theorem 5.1 leads to an *a posteriori* version of our asymptotic convergence analysis, which is stated in the following theorem.

THEOREM 5.5 (a posteriori estimation on asymptotic convergence).

$$\mathcal{D}_{KL}(\rho_{\theta_t}\|\rho_*) \le \frac{\delta_1}{\tilde{\lambda}_D D^2}(1 - e^{-D\tilde{\lambda}_D t}) + \mathcal{D}_{KL}(\rho_{\theta_0}\|\rho_*)e^{-D\tilde{\lambda}_D t},$$

where δ_1 *is defined in* (5.4).

5.3. Wasserstein error estimations. In this subsection, we establish our error bounds for both continuous and discrete versions of the parametric Fokker–Planck equation (3.18) as approximations to the original equation (2.2).

5.3.1. Wasserstein error for the parametric Fokker–Planck equation. The following theorem provides an upper bound between the solutions of (2.2) and (3.18).

THEOREM 5.6. *Assume that* $\{\theta_t\}_{t\ge 0}$ *solves* (3.18) *and* $\{\rho_t\}_{t\ge 0}$ *solves* (2.2). *If the Hessian of the potential function* V *in* (2.2) *is bounded below by a constant* λ, *i.e.,* $\nabla^2 V \succeq \lambda\ I$, *the* 2*-Wasserstein difference between* ρ_t *and* ρ_{θ_t} *can be bounded as*

$$W_2(\rho_{\theta_t}, \rho_t) \le \Omega_\lambda(t) = \begin{cases} \frac{\sqrt{\delta_0}}{\lambda}(1 - e^{-\lambda t}) + e^{-\lambda t}W_2(\rho_{\theta_0}, \rho_0) & \text{if } \lambda \ne 0, \\ \sqrt{\delta_0}t + W_2(\rho_{\theta_0}, \rho_0) & \text{if } \lambda = 0. \end{cases} \tag{5.13}$$

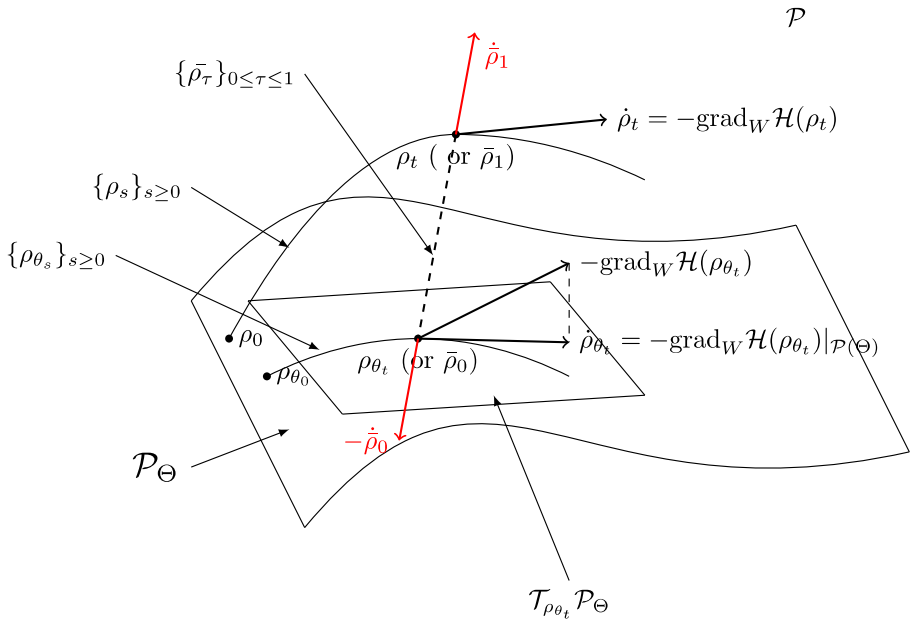


FIG. 3. *An illustrative diagram for the proof of Theorem* 5.6.

To prove this inequality, we need the following lemmas.

LEMMA 5.7 (constant speed of geodesic). *The geodesic connecting* $\rho_0, \rho_1 \in \mathcal{P}(M)$ *is described by*

$$\begin{cases} \frac{\partial \rho_t}{\partial t} + \nabla\cdot(\rho_t\nabla\psi_t) = 0, \\ \frac{\partial \psi_t}{\partial t} + \frac{1}{2}|\nabla\psi_t|^2 = 0, \end{cases} \qquad \rho_t|_{t=0} = \rho_0,\ \rho_t|_{t=1} = \rho_1. \tag{5.14}$$

Using the notation $\dot{\rho}_t = \partial_t\rho_t = -\nabla\cdot(\rho_t\nabla\psi_t) \in \mathcal{T}_{\rho_t}\mathcal{P}(M)$, $g^W(\dot{\rho}_t, \dot{\rho}_t)$ *is constant for* $0 \le t \le 1$ *and* $g^W(\dot{\rho}_t, \dot{\rho}_t) = W_2^2(\rho_0, \rho_1)$ *for* $0 \le t \le 1$.

LEMMA 5.8 (displacement convexity of relative entropy). *Suppose* $\{\rho_t\}$ *solves* (5.14)*, and the relative entropy* $\mathcal{H}$ *in* (2.8) *has potential* V *satisfying* $\nabla^2 V \succeq \lambda I$*; then we have* $\frac{d}{dt}g^W(\mathrm{grad}_W\mathcal{H}(\rho_t), \dot{\rho}_t) \ge \lambda W_2^2(\rho_0, \rho_1)$*, or equivalently* $\frac{d^2}{dt^2}\mathcal{H}(\rho_t) \ge \lambda W_2^2(\rho_0, \rho_1)$.

Lemma 5.7 originates from section 7.2 of [4]. A generalization of it has been proved in Lemma 5 of [40]. A more general version on the displacement convexity related to Lemma 5.8 was discussed in Chapters 16 and 17 of [68]. To be self-contained, we provide direct proofs of both Lemma 5.7 and Lemma 5.8 in Appendix D.

Proof of Theorem 5.6. Figure 3 provides a sketch of our proof: For a given time t, the geodesic $\{\bar{\rho}_\tau\}_{0\le\tau\le1}$ on Wasserstein manifold $\mathcal{P}(M)$ that connects ρ_{θ_t} and ρ_t satisfies the geodesic equations (5.14). If differentiating $W_2^2(\rho_{\theta_t}, \rho_t)$ with respect to time t according to Theorem 23.9 of [68], we are able to deduce that

$$\frac{d}{dt}W_2^2(\rho_{\theta_t}, \rho_t) = 2g^W(\dot{\rho}_{\theta_t}, -\dot{\bar{\rho}}_0) + 2g^W(\dot{\rho}_t, \dot{\bar{\rho}}_1), \tag{5.15}$$

in which $\dot{\bar{\rho}}_0 = \partial_\tau\bar{\rho}_\tau|_{\tau=0} = -\nabla\cdot(\bar{\rho}_0\nabla\psi_0)$, $\dot{\bar{\rho}}_1 = \partial_\tau\bar{\rho}_\tau|_{\tau=1} = -\nabla\cdot(\bar{\rho}_1\nabla\psi_1)$. Notice that

$$\dot{\rho}_{\theta_t} = (T_{\theta\sharp})_*\dot{\theta}_t, \quad \dot{\rho}_t = -\mathrm{grad}_W\mathcal{H}(\rho_t) = \nabla\cdot(\rho_t\nabla(V + D\log\rho_t)).$$

Using the definition (2.5) of Wasserstein metric, we can compute the following (recall that $\rho_{\theta_t} = \bar{\rho}_0$, $\rho_t = \bar{\rho}_1$):

$$g^W(\dot{\rho}_{\theta_t}, \dot{\bar{\rho}}_0) = \int \nabla(V + D\log\bar{\rho}_0)\cdot\psi_0\ \bar{\rho}_0\ dx \quad g^W(\dot{\rho}_t, \dot{\bar{\rho}}_1) = \int \nabla(V + D\log\bar{\rho}_1)\cdot\psi_1\ \bar{\rho}_1\ dx.$$

Now we can write (5.15) as

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}W_2^2(\rho_{\theta_t}, \rho_t) &= g^W((T_{\theta_t\sharp})_*\dot{\theta}_t + \mathrm{grad}_W\mathcal{H}(\rho_{\theta_t}), -\dot{\bar{\rho}}_0) + g^W(-\mathrm{grad}_W\mathcal{H}(\rho_{\theta_t}), -\dot{\bar{\rho}}_0)\\ &\quad + g^W(-\mathrm{grad}_W\mathcal{H}(\rho_t), \dot{\bar{\rho}}_1)\\ &\overset{\text{set: } \xi = -\dot{\theta}_t}{=} g^W(\mathrm{grad}_W\mathcal{H}(\rho_{\theta_t}) - (T_{\theta_t\sharp})_*\xi,\ \ -\dot{\bar{\rho}}_0)\\ &\quad - (g^W(\mathrm{grad}_W\mathcal{H}(\bar{\rho}_1),\ \dot{\bar{\rho}}_1) - g^W(\mathrm{grad}_W\mathcal{H}(\bar{\rho}_0),\ \dot{\bar{\rho}}_0)).\end{aligned} \tag{5.16}$$

For the first term in (5.16), we use the Cauchy–Schwarz inequality, (5.1), and Lemma 5.7, which implies $g(\dot{\bar{\rho}}_0, \dot{\bar{\rho}}_0) = W_2^2(\rho_{\theta_t}, \rho_t)$, to obtain

$$\begin{aligned}&g^W(\mathrm{grad}_W\mathcal{H}(\rho_{\theta_t}) - (T_{\theta_t\sharp})_*\xi, -\dot{\bar{\rho}}_0)\\ &\le \sqrt{g^W(\mathrm{grad}_W\mathcal{H}(\rho_{\theta_t}) - (T_{\theta_t\sharp})_*\xi,\ \mathrm{grad}_W\mathcal{H}(\rho_{\theta_t}) - (T_{\theta_t\sharp})_*\xi)}\sqrt{g^W(\dot{\bar{\rho}}_0, \dot{\bar{\rho}}_0)}\\ &\le \sqrt{\delta_0}W(\rho_{\theta_t}, \rho_t).\end{aligned} \tag{5.17}$$

For the second term in (5.16), we write it as

$$g^W(\mathrm{grad}_W\mathcal{H}(\bar{\rho}_1), \dot{\bar{\rho}}_1) - g^W(\mathrm{grad}_W\mathcal{H}(\bar{\rho}_0), \dot{\bar{\rho}}_0) = \int_0^1 \frac{d}{d\tau}g^W(\mathrm{grad}_W\mathcal{H}(\bar{\rho}_\tau), \dot{\bar{\rho}}_\tau)\ d\tau. \tag{5.18}$$

By Lemma 5.8, we have

$$g^W(\mathrm{grad}_W\mathcal{H}(\bar{\rho}_1), \dot{\bar{\rho}}_1) - g^W(\mathrm{grad}_W\mathcal{H}(\bar{\rho}_0), \dot{\bar{\rho}}_0) \ge \lambda\ W_2^2(\rho_{\theta_t}, \rho_t). \tag{5.19}$$

Combining inequalities (5.17), (5.19), and (5.16), we get

$$\frac{1}{2}\frac{d}{dt}W_2^2(\rho_{\theta_t}, \rho_t) \le -\lambda W_2^2(\rho_{\theta_t}, \rho_t) + \sqrt{\delta_0}\ W_2(\rho_{\theta_t}, \rho_t).$$

This is

$$\frac{d}{dt}W_2(\rho_{\theta_t}, \rho_t) \le -\lambda W_2(\rho_{\theta_t}, \rho_t) + \sqrt{\delta_0}.$$

When $\lambda \neq 0$, Gronwall's inequality gives

$$W_2(\rho_{\theta_t}, \rho_t) \le \frac{\sqrt{\delta_0}}{\lambda}(1 - e^{-\lambda t}) + e^{-\lambda t}W_2(\rho_{\theta_0}, \rho_0).$$

When $\lambda = 0$, the inequality is $\frac{d}{dt}W_2(\rho_{\theta_t}, \rho_t) \le \sqrt{\delta_0}$, and direct integration yields

$$W_2(\rho_{\theta_t}, \rho_t) \le \sqrt{\delta_0}t + W_2(\rho_{\theta_0}, \rho_0)\ . \qquad \square$$

When the potential V is strictly convex, i.e., $\lambda > 0$, (5.13) in Theorem 5.6 provides a nice estimation of the error term $W_2(\rho_{\theta_t}, \rho_t)$ at any time t that is always upper bounded by $\max\{\frac{\sqrt{\delta_0}}{\lambda}, W_2(\rho_{\theta_0}, \rho_0)\}$.

In the case that the potential V is not strictly convex, i.e., λ could be 0 or negative, the right-hand side in (5.13) may increase to infinity when time $t \to \infty$. However, (5.5) and (5.12) reveal that both ρ_{θ_t} and ρ_t stay in a small neighborhood of the Gibbs ρ_* when t is large. When taking this into account, we are able to show that the error term $W_2(\rho_{\theta_t}, \rho_t)$ doesn't get arbitrarily large. In the following theorem, we provide a uniform bound for the error depending on t.

THEOREM 5.9. *Suppose* $\{\rho_t\}_{t\geq 0}$ *solves* (2.2) *and* $\{\rho_{\theta_t}\}_{t\geq 0}$ *solves* (3.18)*, and the Hessian of the potential* $V \in \mathcal{V}$ *is bounded from below by* λ*, i.e.,* $\nabla^2 V \succeq \lambda I$*. Then* (5.20)

$$W_2(\rho_{\theta_t}, \rho_t) \leq \min\left\{\Omega_\lambda(t),\ \sqrt{\frac{2\delta_0}{\tilde{\lambda}_D^2 D^2}} + \left(\sqrt{\left|2K_1 - \frac{2\delta_0}{\tilde{\lambda}_D^2 D^2}\right|} + \sqrt{\frac{2K_2}{\tilde{\lambda}_D}}\right) e^{-\frac{\tilde{\lambda}_D}{2}Dt}\right\},$$

where the function $\Omega_\lambda(t)$ *is defined in* (5.13)*,* $E_0 = W_2(\rho_{\theta_0}, \rho_0)$*,* $K_1 = \mathcal{D}_{KL}(\rho_{\theta_0}\|\rho_*)$*, and* $K_2 = \mathcal{D}_{KL}(\rho_0\|\rho_*)$*.*

LEMMA 5.10 (Talagrand inequality [47, 68]). *If the Gibbs distribution* ρ_* *satisfies the logarithmic Sobolev inequality* (5.6) *with constant* $\tilde{\lambda} > 0$*,* ρ_* *also satisfies the Talagrand inequality:*

$$\sqrt{2\frac{\mathcal{D}_{KL}(\rho\|\rho_*)}{\tilde{\lambda}}} \geq W_2(\rho, \rho_*) \quad \text{for any } \rho \in \mathcal{P}. \tag{5.21}$$

Proof of Theorem 5.9. The first term is already provided in Theorem 5.6, and the second term is just a quick result of Theorem 5.1 and the Talagrand inequality: for t fixed, (5.5) together with the Talagrand inequality (5.21) gives

$$\begin{aligned} W_2(\rho_{\theta_t}, \rho_*) \leq \sqrt{2\frac{\mathcal{D}_{\mathrm{KL}}(\rho_{\theta_t}\|\rho_*)}{\tilde{\lambda}_D}} &\leq \sqrt{\frac{2\delta_0}{\tilde{\lambda}_D^2 D^2}(1 - e^{-\tilde{\lambda}_D Dt}) + 2K_1 e^{-\tilde{\lambda}_D Dt}} \\ &\leq \sqrt{\frac{2\delta_0}{\tilde{\lambda}_D^2 D^2}} + \sqrt{\left|2K_1 - \frac{2\delta_0}{\tilde{\lambda}_D^2 D^2}\right|} e^{-\frac{\tilde{\lambda}_D}{2}Dt}. \end{aligned}$$

Similarly, (5.12) and (5.21) give

$$W_2(\rho_t, \rho_*) \leq \sqrt{2\frac{\mathcal{D}_{\mathrm{KL}}(\rho_t\|\rho_*)}{\tilde{\lambda}_D}} \leq \sqrt{\frac{2K_2}{\tilde{\lambda}_D}} e^{-\frac{\tilde{\lambda}_D}{2}Dt}.$$

Applying the triangle inequality of Wasserstein distance $W_2(\rho_{\theta_t}, \rho_t) \leq W_2(\rho_{\theta_t}, \rho_*) + W_2(\rho_t, \rho_*)$, we get (5.20). ∎

Based on Theorem 5.9, we can obtain a uniform a priori error estimate.

THEOREM 5.11 (main theorem on a priori error analysis of the parametric Fokker–Planck equation). *Assume* $E_0 = W_2(\rho_{\theta_0}, \rho_0)$ *and* δ_0 *defined in* (5.1) *are sufficiently small in the sense that*

$$E_0 < A\sqrt{\delta_0} + B,\ \sqrt{\delta_0} + E_0 \leq Be^{-\mu_D(A+1)}. \tag{5.22}$$

Then the approximation error $W_2(\rho_{\theta_t}, \rho_t)$ *at any time* $t > 0$ *can be uniformly bounded by* E_0 *and* δ_0*:*

- *When* $\lambda > 0$*,* $W_2(\rho_{\theta_t}, \rho_t) \leq \max\{\sqrt{\delta_0}/\lambda, E_0\} \sim O(\sqrt{\delta_0} + E_0)$*.*

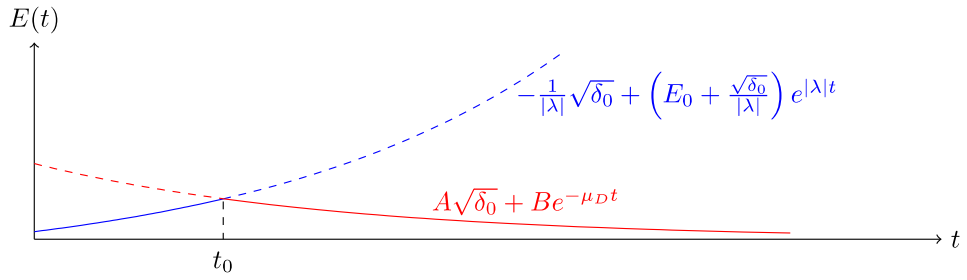


FIG. 4. *An illustrative diagram for the proof of Theorem* 5.11.

- *When* $\lambda=0$, $W_2(\rho_{\theta_t},\rho_t)\le\frac{\sqrt{\delta_0}}{\mu_D}\log\frac{B}{\sqrt{\delta_0}+E_0}+E_0\sim O(\sqrt{\delta_0}\log\frac{1}{\sqrt{\delta_0}+E_0}+E_0)$.
- *When* $\lambda<0$, $W_2(\rho_{\theta_t},\rho_t)\le A\sqrt{\delta_0}+B^{\frac{|\lambda|}{|\lambda|+\mu_D}}\left(E_0+\sqrt{\delta_0}/|\lambda|\right)^{\frac{\mu_D}{|\lambda|+\mu_D}}\sim O((E_0+\sqrt{\delta_0})^{\frac{\tilde\lambda_D D}{2|\lambda|+\tilde\lambda_D D}})$.

Here A,B,μ_D *are* $O(1)$ *constants depending on* V,D,ρ_0,θ_0. *Their values are given in* (5.24).

Proof of Theorem 5.11. When $\lambda>0$, by (5.20), we have $E(t)\le\frac{\sqrt{\delta_0}}{\lambda}+\left(E_0-\frac{\sqrt{\delta_0}}{\lambda}\right)e^{-\lambda t}$, and the right-hand side can be bounded by $\max\{E_0,\frac{\sqrt{\delta_0}}{\lambda}\}$.

When $\lambda<0$, we denote the right-hand side of (5.20) by

$$E(t)=\min\left\{-\frac{1}{|\lambda|}\sqrt{\delta_0}+\left(E_0+\frac{\sqrt{\delta_0}}{|\lambda|}\right)e^{|\lambda|t},A\sqrt{\delta_0}+Be^{-\mu_D t}\right\},\tag{5.23}$$

where

$$A=\frac{\sqrt{2}}{\tilde\lambda_D D},\quad B=\sqrt{\left|2K_1-\frac{2\delta_0}{\tilde\lambda_D^2 D^2}\right|}+\sqrt{\frac{2K_2}{\tilde\lambda_D}},\quad\text{and}\quad\mu_D=\frac{\tilde\lambda_D D}{2}\tag{5.24}$$

are all positive numbers. The first term in (5.23) is increasing as a function of time t, while the second term is decreasing; combining $E_0<A\sqrt{\delta_0}+B$, we know $t_0=\operatorname{argmax}_{t\ge0}E(t)$ is unique and satisfies

$$-\frac{1}{|\lambda|}\sqrt{\delta_0}+\left(E_0+\frac{\sqrt{\delta_0}}{|\lambda|}\right)e^{|\lambda|t_0}=A\sqrt{\delta_0}+Be^{-\mu_D t_0},\tag{5.25}$$

as indicated in Figure 4.

Since $A>0$, (5.25) leads to $\left(E_0+\frac{\sqrt{\delta_0}}{|\lambda|}\right)e^{|\lambda|t_0}>Be^{-\mu_D t_0}$, and thus

$$t_0>\frac{\log B-\log\left(E_0+\frac{\sqrt{\delta_0}}{|\lambda|}\right)}{|\lambda|+\mu_D}.\tag{5.26}$$

Using (5.26), we show that

$$\max_{t\ge0}E(t)=E(t_0)=A\sqrt{\delta_0}+B\,e^{-\mu_D t_0}<A\sqrt{\delta_0}+B^{\frac{|\lambda|}{|\lambda|+\mu_D}}\left(E_0+\frac{\sqrt{\delta_0}}{|\lambda|}\right)^{\frac{\mu_D}{|\lambda|+\mu_D}}.\tag{5.27}$$

As a result, $W_2(\rho_{\theta_t}, \rho_t)$ can be uniformly bounded by the right-hand side of (5.27). Since A, B are $O(1)$ coefficients, this uniform bound is dominated by the term $O\big(\big(E_0 + \frac{\sqrt{\delta_0}}{|\lambda|}\big)^{\frac{\mu_D}{|\lambda|+\mu_D}}\big) = O((E_0 + \sqrt{\delta_0})^{\frac{\tilde{\lambda}_D D}{2|\lambda|+\tilde{\lambda}_D D}})$.

At last, when $\lambda = 0$, by (5.20)

$$E(t) = \min\left\{\sqrt{\delta_0}t + E_0, A\sqrt{\delta_0} + Be^{-\mu_D t}\right\}.$$

Let us denote $f(t) = A\sqrt{\delta_0} + Be^{-\mu_D t} - \sqrt{\delta_0}t - E_0$. Similar to the analysis for the case $\lambda < 0$, we denote $t_0 = \operatorname{argmax}_{t\geq 0} E(t)$, and then t_0 is unique and solves $f(t_0) = 0$. Since $f(t)$ is decreasing with $f(A+1) > 0$, $t_0 > A + 1$. Then we have

$$\max_{t\geq 0} E(t) = E(t_0) = A\sqrt{\delta_0} + Be^{-\mu_D t_0} = \sqrt{\delta_0}t_0 + E_0 > \sqrt{\delta_0}(A+1) + E_0.$$

This leads to $Be^{-\mu_D t_0} > \sqrt{\delta_0} + E_0$, i.e., $t_0 < \frac{1}{\mu_D}\log\frac{B}{\sqrt{\delta_0}+E_0}$. Thus we have

$$\max_{t\geq 0} E(t) = E(t_0) = \sqrt{\delta_0}t_0 + E_0 < \frac{\sqrt{\delta_0}}{\mu_D}\log\frac{B}{\sqrt{\delta_0}+E_0} + E_0.$$

Therefore $W_2(\rho_{\theta_t}, \rho_t)$ can be uniformly bounded by the term $\frac{\sqrt{\delta_0}}{\mu_D}\log\frac{B}{\sqrt{\delta_0}+E_0} + E_0 \sim O(\sqrt{\delta_0}\log\frac{1}{\sqrt{\delta_0}+E_0} + E_0)$. ∎

Remark 5.12. In the case that $V \in \mathcal{V}$ is not convex, we can decompose V by $V = U + \phi$ with $\nabla^2 U \succeq KI$ $(K > 0)$ and $\nabla^2\phi \succeq K_\phi I$. We can still assume $\nabla^2 V \succeq \lambda I$, but λ may be negative. One can verify that $K_\phi < 0$ and $|K_\phi| - K \geq |\lambda|$. On the other hand, one can compute $\tilde{\lambda}_D = \frac{K}{D}e^{-\frac{\operatorname{osc}(\phi)}{D}}$. Combining them, we provide a lower bound for α:

$$\alpha \geq \gamma(D, U, \phi) = \frac{1}{1 + 2\left(\frac{|K_\phi|}{K} - 1\right)e^{\frac{\operatorname{osc}(\phi)}{D}}}.$$

One can verify that increasing the diffusion coefficient D or convexity K, or decreasing the oscillation $\operatorname{osc}(\phi)$ and convexity K_ϕ, can improve the lower bound $\gamma(D, U, \phi)$ for the order α.

In a similar way, we can establish the corresponding a posteriori error estimate for $W_2(\rho_{\theta_t}, \rho_t)$.

THEOREM 5.13 (a posteriori error analysis of the parametric Fokker–Planck equation). *Suppose* $E_0 = W_2(\rho_{\theta_0}, \rho_0)$ *and* δ_1 *defined in* (5.4) *satisfy the condition* (5.22) *with* δ_0 *replaced by* δ_1. *Then*

1. *when* $\lambda \geq 0$, $W_2(\rho_{\theta_t}, \rho_t)$ *can be uniformly bounded by* $O(E_0 + \sqrt{\delta_1})$;
2. *when* $\lambda = 0$, $W_2(\rho_{\theta_t}, \rho_t)$ *can be uniformly bounded by* $O(\sqrt{\delta_1}\log\frac{1}{\sqrt{\delta_1}+E_0} + E_0)$;
3. *when* $\lambda < 0$, $W_2(\rho_{\theta_t}, \rho_t)$ *can be uniformly bounded by* $O((E_0 + \sqrt{\delta_1})^{\frac{\tilde{\lambda}_D D}{2|\lambda|+\tilde{\lambda}_D D}})$.

5.3.2. Wasserstein error for the time discrete schemes. To solve (3.18) numerically, we need time discrete schemes, such as the one proposed in (4.8). In this subsection, we present the error estimate in Wasserstein distance for our scheme. We begin our analysis by focusing on the forward Euler scheme, meaning that we apply the forward Euler scheme to solve (3.18) and compute θ_k at each time step. We denote $\rho_{\theta_k} = T_{\theta_k}{}_\sharp p$. We estimate the W_2-error between ρ_{θ_k} and the real solution

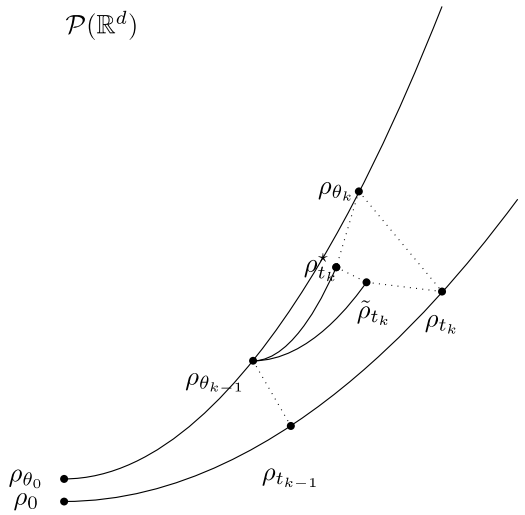


FIG. 5. *Trajectory of* $\{\rho_{\theta_k}\}_{k=0,\dots,N}$ *is our numerical solution; trajectory of* $\{\rho_t\}_{t\geq 0}$ *is the real solution of the Fokker–Planck equation;* $\{\tilde{\rho}_t\}_{t\geq t_{k-1}}$ *solves* (5.30)*;* $\{\rho^\star_t\}_{t\geq t_{k-1}}$ *solves* (5.31).

ρ_{t_k}. Then we analyze the W_2 distance between the solutions obtained by the forward Euler scheme and our scheme (4.8), respectively, which in turn gives us the W_2 error estimate for our scheme.

THEOREM 5.14 (a priori error analysis of forward Euler scheme). *Let* θ_k *(*$k = 0, 1, \dots, N$*) be the solution of forward Euler scheme applied to* (3.18) *at time* $t_k = kh$ *on* $[0, T]$ *with time step size* $h = \frac{T}{N}$, $\rho_{\theta_k} = T_{\theta_k}{}_\sharp p$, *and* $\{\rho_t\}_{t\geq 0}$ *solves the Fokker–Planck equation* (2.2) *exactly. Assume that the Hessian of the potential function* $V \in \mathcal{C}^2(\mathbb{R}^d)$ *can be bounded from above and below, i.e.,* $\lambda I \preceq \nabla^2 V \preceq \Lambda I$. *Then*

(5.28)
$$W_2(\rho_{\theta_k}, \rho_{t_k}) \leq (\sqrt{\delta_0}h + Ch^2)\frac{1-e^{-\lambda t_k}}{1-e^{-\lambda h}} + e^{-\lambda t_k} W_2(\rho_{\theta_0}, \rho_0) \quad \textit{for any } \ t_k = kh,\ 0 \leq k \leq N,$$

where C *is a constant whose direct formula is provided in* (5.45).

In order to estimate $W_2(\rho_{\theta_k}, \rho_{t_k})$, we use the triangle inequality of the W_2 distance [68] to separate it into three parts:

$$W_2(\rho_{\theta_k}, \rho_{t_k}) \leq W_2(\rho_{\theta_k}, \tilde{\rho}^\star_{t_k}) + W_2(\rho^\star_{t_k}, \tilde{\rho}_{t_k}) + W_2(\tilde{\rho}_{t_k}, \rho_{t_k}). \tag{5.29}$$

Here $\{\tilde{\rho}_t\}_{t_{k-1}\leq t\leq t_k}$ satisfies

$$\frac{\partial \tilde{\rho}_t}{\partial t} = \nabla \cdot (\tilde{\rho}_t \nabla V) + D\Delta \tilde{\rho}_t \ , \quad \tilde{\rho}_{t_{k-1}} = \rho_{\theta_{k-1}}, \tag{5.30}$$

and $\{\rho^\star_t\}_{t\geq t_{k-1}}$ satisfies

$$\frac{\partial \rho^\star_t}{\partial t} = \nabla \cdot (\rho^\star_t \nabla (V + D \log \rho_{\theta_{k-1}})) \ , \quad \rho^\star_{t_{k-1}} = \rho_{\theta_{k-1}}. \tag{5.31}$$

Figure 5 shows the relations of different items used in our proof. We present three lemmas that estimate three terms in (5.29), respectively.

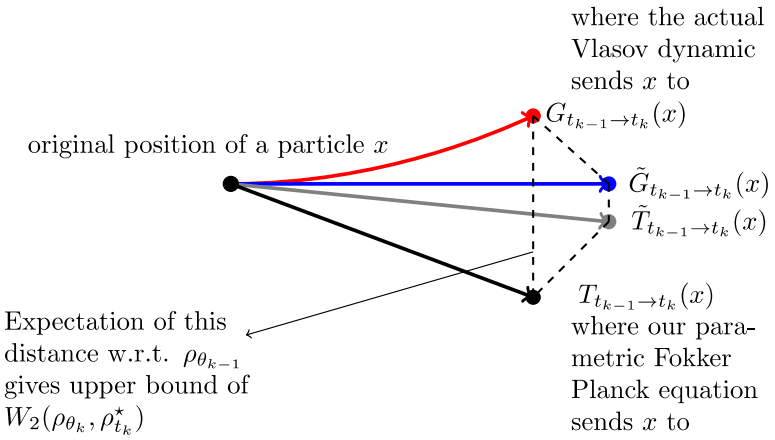


FIG. 6. *Illustration of proof strategy for Lemma* 5.15.

LEMMA 5.15. *$W_2(\rho_{\theta_k}, \rho^\star_{t_k})$ in* (5.29) *can be upper bounded by $\sqrt{\delta_0}h + O(h^2)$.*

An explicit formula for the coefficient of h^2 is included in the following proof.

Proof. We establish the desired estimation by introducing several different push-forward maps as shown in Figure 6 and then applying the triangle inequality.

(a) We know $\rho_{\theta_{k-1}} = T_{\theta_{k-1}\sharp}p$ and $\rho_{\theta_k} = T_{\theta_k\sharp}p$, and let us denote $T_{t_{k-1}\to t_k} = T_{\theta_k} \circ T^{-1}_{\theta_{k-1}}$. Then $\rho_{\theta_k} = T_{t_{k-1}\to t_k\sharp}\rho_{\theta_{k-1}}$.

(b) Let $\xi_{k-1} = \dot{\theta}_{k-1} = -G(\theta_{k-1})^{-1}\nabla_\theta H(\theta_{k-1})$, and by convention we denote $\boldsymbol{\Psi}$ as a solution of (3.6). We consider the map $\tilde{T}_{t_{k-1}\to t_k}(\cdot) = \mathrm{Id} + h\nabla\boldsymbol{\Psi}(\cdot)^{\mathrm{T}}\xi_{k-1}$.

(c) We denote $\zeta_\theta(\cdot) = V(\cdot) + D\log\rho_\theta(\cdot)$. The particle version (recall (2.3)) of (5.31) is

(5.32)
$$\dot{z}_t = -\nabla\zeta_{\theta_{k-1}}(z_t), \quad 0 \le t \le h, \quad \text{with initial condition } z_0 = x \sim \rho_{\theta_{k-1}}.$$

We denote the solution map of (5.32) by $G_{t_{k-1}\to t_k}(x) = z_{t_k}$. Then $\rho^\star_{t_k} = G_{t_{k-1}\to t_k\sharp}\rho_{\theta_{k-1}}$.

(d) The map $G_{t_{k-1}\to t_k}$ is obtained by solving an ODE in order to compare the difference with $T_{t_{k-1}\to t_k}$. We consider the ODE with fixed initial vector field:

$$\dot{\tilde{z}}_t = -\nabla\zeta_{\theta_{k-1}}(x) \quad 0 \le t \le h \quad \tilde{z}_0 = x \sim \rho_{\theta_{k-1}}. \tag{5.33}$$

This ODE will induce the solution map $\tilde{G}_{t_{k-1}\to t_k}(\cdot) = \mathrm{Id} - h\nabla\zeta_{\theta_{k-1}}(\cdot)$.

With the maps defined in (a), (b), (c), (d), and using the triangle inequality of W_2 distance, we have

$$\begin{aligned} W_2(\rho_{\theta_k}, \tilde{\rho}^\star_{t_k}) &= W_2(T_{t_{k-1}\to t_k\sharp}\rho_{\theta_{k-1}}, G_{t_{k-1}\to t_k\sharp}\rho_{\theta_{k-1}}) \\ &\le \underbrace{W_2(T_{t_{k-1}\to t_k\sharp}\rho_{\theta_{k-1}}, \tilde{T}_{t_{k-1}\to t_k\sharp}\rho_{\theta_{k-1}})}_{\text{(A)}} + \underbrace{W_2(\tilde{T}_{t_{k-1}\to t_k\sharp}\rho_{\theta_{k-1}}, \tilde{G}_{t_{k-1}\to t_k\sharp}\rho_{\theta_{k-1}})}_{\text{(B)}} \\ &\quad + \underbrace{W_2(\tilde{G}_{t_{k-1}\to t_k\sharp}\rho_{\theta_{k-1}}, G_{t_{k-1}\to t_k\sharp}\rho_{\theta_{k-1}})}_{\text{(C)}}. \end{aligned}$$

In the rest of the proof, we give upper bounds for distances (A), (B), and (C), respectively.

(A) Let us define $\xi(\theta) = -G(\theta)^{-1}\nabla H(\theta)$. Now we set $\theta(\tau) = \theta_{k-1} + \frac{\tau}{h}(\theta_k - \theta_{k-1}) = \theta_{k-1} + \tau\xi(\theta_{k-1})$. For any x, consider $x_\tau = T_{\theta(\tau)}(T^{-1}_{\theta_{k-1}}(x))$ with $0 \le \tau \le h$; then $\{x_\tau\}_{0\le\tau\le h}$ satisfies

$$\dot{x}_\tau = \partial_\theta T_{\theta(\tau)}(T^{-1}_{\theta(\tau)}(x_\tau))\xi(\theta_{k-1}), \quad 0 \le \tau \le h. \tag{5.34}$$

If $x_0 \sim \rho_{\theta_{k-1}}$ in (5.34), it is clear that $x_h \sim T_{t_{k-1}\to t_k\sharp}\rho_{\theta_{k-1}}$. Furthermore, we denote the distribution of x_τ as ρ_τ and $\{\psi_\tau\}$ satisfying

$$-\nabla\cdot(\rho_\tau(x)\partial_\theta T_{\theta(\tau)}(T^{-1}_{\theta(\tau)}(x))\xi_{k-1}) = -\nabla\cdot(\rho_\tau(x)\nabla\psi_\tau(x)), \quad 0 \le \tau \le h. \tag{5.35}$$

If we consider

$$\dot{y}_\tau = \nabla\psi_\tau(y_\tau), \quad 0 \le \tau \le h, \quad \text{with } y_0 \sim \rho_{\theta_{k-1}},$$

and denote ϱ_τ as the distribution of y_τ, by the continuity equation and (5.35), we know $\rho_\tau = \varrho_\tau$ for $0 \le \tau \le h$, and thus $y_h \sim T_{t_{k-1}\to t_k\sharp}\rho_{\theta_{k-1}}$. On the other hand, when $\tau = 0$, (5.35) shows $\nabla\psi_0(x) = \nabla\mathbf{\Psi}(x)^{\mathrm{T}}\xi_{k-1}$. Combining them, we bound term (A) as

$$\begin{aligned}
&W_2^2(T_{t_{k-1}\to t_k\sharp}\rho_{\theta_{k-1}}, \tilde{T}_{t_{k-1}\to t_k\sharp}\rho_{\theta_{k-1}})\\
&\le \mathbb{E}_{y_0\sim\rho_{\theta_{k-1}}}|y_h - (y_0 + h\nabla\psi_0(y_0))|^2 = \mathbb{E}_{y_0\sim\rho_{\theta_{k-1}}}\left|\int_0^h (\nabla\psi_\tau(y_\tau) - \nabla\psi_0(y_0))\, d\tau\right|^2\\
&= \mathbb{E}_{y_0}\left|\int_0^h\int_0^\tau \frac{d}{ds}(\nabla\psi_s(y_s))\, ds\, d\tau\right|^2 = \mathbb{E}_{y_0}\left|\int_0^h\int_s^h \frac{d}{ds}(\nabla\psi_s(y_s))\, d\tau\, ds\right|^2\\
&= \mathbb{E}_{y_0}\left|\int_0^h (h-s)\frac{d}{ds}(\nabla\psi_s(y_s))\, ds\right|^2 \le \mathbb{E}_{y_0}\int_0^h (h-s)^2\, ds\int_0^h\left|\frac{d}{ds}(\nabla\psi_s(y_s))\right|^2 ds\\
&= \frac{h^3}{3}\int_0^h \mathbb{E}_{y_0}\left|\frac{d}{ds}(\nabla\psi_s(y_s))\right|^2 ds\\
&= \frac{h^4}{3}\left(\frac{1}{h}\int_0^h \mathbb{E}_{y_s}\left|\frac{\partial\nabla\psi_s(y_s)}{\partial t} + \nabla^2\psi_s(y_s)\nabla\psi_s(y_s)\right|^2 ds\right).
\end{aligned}$$

Notice that y_s follows the distribution $\rho_s = (T_{\theta_{k-1}+s\xi(\theta_{k-1})}\circ T^{-1}_{\theta_{k-1}})_\sharp\rho_{\theta_{k-1}} = T_{\theta_{k-1}+s\xi(\theta_{k-1})\sharp}p$.

If we define

$$\begin{aligned}
\mathfrak{M}(\theta, s) = \int\left|\frac{\partial}{\partial t}\nabla\psi_s(T_{\theta(s)}(z)) + \nabla^2\psi_s(T_{\theta(s)}(z))\nabla\psi_s(T_{\theta(s)}(z))\right|^2 p(z)\, dz\\
\text{with } \theta(s) = \theta + s\xi(\theta),\\
\text{and } \psi_s \text{ solving } -\nabla\cdot(\rho_s\nabla\psi_s) = -\nabla\cdot(\rho_s\, \partial_\theta T_{\theta(s)}\circ T^{-1}_{\theta(s)}\, \xi(\theta)),\\
\text{where } \rho_s = T_{\theta+s\xi(\theta)\sharp}p,
\end{aligned} \tag{5.36}$$

we are able to derive

$$W_2^2(T_{t_{k-1}\to t_k\sharp}\rho_{\theta_{k-1}}, \tilde{T}_{t_{k-1}\to t_k\sharp}\rho_{\theta_{k-1}}) \le \frac{1}{3}\sup_{0\le s\le h}\mathfrak{M}(\theta_{k-1}, s)h^4. \tag{5.37}$$

(B) We have

$$
\begin{aligned}
&W_2^2(\tilde{T}_{t_{k-1}\to t_k\sharp}\rho_{\theta_{k-1}}, \tilde{G}_{t_{k-1}\to t_k\sharp}\rho_{\theta_{k-1}})\\
&\le \int |\tilde{T}_{t_{k-1}\to t_k}(x) - \tilde{G}_{t_{k-1}\to t_k}(x)|^2 \rho_{\theta_{k-1}}(x)\ dx\\
&= h^2 \left(\int |\nabla \mathbf{\Psi}(x)^{\mathrm{T}} \xi(\theta_{k-1}) - (-\nabla \zeta_{\theta_{k-1}}(x))|^2 \rho_{\theta_{k-1}}(x)\ dx \right)\\
&= h^2 \left(\int |\nabla \mathbf{\Psi}(T_{\theta_{k-1}}(x))^{\mathrm{T}} \xi(\theta_{k-1}) - (-\nabla (V + D\log \rho_{\theta_{k-1}}) \circ T_{\theta_{k-1}}(x))|^2\ dp(x) \right)\\
&\le \delta_0\ h^2.
\end{aligned}
$$

The last inequality is due to Theorem 3.8 and definition (5.1).

(C) Recalling that $\{z_t\}$ and $\{\tilde{z}_t\}$ solve (5.32) and (5.33) with initial condition $z_0 = \tilde{z}_0 = x$, respectively, similar to the analysis in (A), we can estimate term (C) as

$$
\begin{aligned}
&W_2^2(\tilde{G}_{t_{k-1}\to t_k\sharp}\rho_{\theta_{k-1}}, G_{t_{k-1}\to t_k\sharp}\rho_{\theta_{k-1}})\\
&\le \mathbb{E}_{x\sim\rho_{\theta_{k-1}}} |z_h - \tilde{z}_h|^2 = \mathbb{E}_{x\sim\rho_{\theta_{k-1}}} \left| \int_0^h \nabla\zeta_{k-1}(x) - \nabla\zeta_{k-1}(z_\tau)\ d\tau \right|^2\\
&= \mathbb{E}_x \left| \int_0^h \int_0^\tau \frac{d}{ds} \nabla\zeta_{\theta_{k-1}}(z_s)\ ds\ d\tau \right|^2 = \mathbb{E}_x \left| \int_0^h (h-s) \frac{d}{ds} \nabla\zeta_{\theta_{k-1}}(z_s)\ ds \right|^2\\
&\le \mathbb{E}_x \frac{h^3}{3} \int_0^h \left| \frac{d}{ds} \nabla\zeta_{\theta_{k-1}}(z_s) \right|^2 ds = \frac{h^4}{3} \left(\frac{1}{h} \int_0^h \mathbb{E}_{z_s} \left| \nabla^2 \zeta_{\theta_{k-1}}(z_s) \zeta_{\theta_{k-1}}(z_s) \right|^2\ ds \right).
\end{aligned}
$$

We define

$$
\begin{aligned}
&\mathfrak{N}(\theta, s) = \mathbb{E}_{z_s} \left| \nabla^2 \zeta_\theta(z_s) \zeta_\theta(z_s) \right|^2, \quad \text{with} \quad \zeta_\theta(\cdot) = V(\cdot) + D\log\rho_\theta(\cdot),\\
&\dot{z}_t = -\nabla\zeta_\theta(z_t), \quad z_0 \sim \rho_\theta.
\end{aligned}
$$

Similarly to (A), we have

$$
W_2^2(\tilde{G}_{t_{k-1}\to t_k\sharp}\rho_{\theta_{k-1}}, G_{t_{k-1}\to t_k\sharp}\rho_{\theta_{k-1}}) \le \frac{1}{3} \sup_{0\le s\le h} \mathfrak{N}(\theta_{k-1}, h) h^4.
$$

Combining the estimates for terms (A), (B), and (C) and defining

$$
M(\theta, h) = \sup_{0\le s\le h} \mathfrak{M}(\theta_{k-1}, s), \quad N(\theta, h) = \sup_{0\le s\le h} \mathfrak{N}(\theta_{k-1}, s), \tag{5.38}
$$

we obtain

$$
W_2(\rho_{\theta_k}, \tilde{\rho}^\star_{t_k}) \le \sqrt{\delta_0} h + \frac{M(\theta_{k-1}, h) + N(\theta_{k-1}, h)}{\sqrt{3}}\ h^2. \qquad \square
$$

LEMMA 5.16. *The second term in* (5.29) *can be upper bounded by* $O(h^2)$.

Proof. Recall that $\tilde{\rho}_t$ is defined by (5.30) and ρ_t^* is defined by (5.31). We can rewrite (5.31) as

$$
\frac{\partial \rho_t^\star}{\partial t} = \nabla \cdot (\rho_t^\star (\nabla V + D\nabla \log \rho_{\theta_{k-1}} - D\nabla \log \rho_t^\star)) + D\Delta \rho_t^\star, \quad t_{k-1} \le t \le t_k.
$$

We consider the following SDEs sharing the same trajectory of Brownian motion $\{\boldsymbol{B}_\tau\}_{0\le\tau\le h}$ and initial condition:

(5.39)
$$dx_\tau = -\nabla V(x_\tau)d\tau + \sqrt{2D}\; d\boldsymbol{B}_\tau,$$

(5.40)
$$dx^\star_\tau = -\nabla V(x^\star_\tau)d\tau + (D\nabla \log \rho^\star_{t_{k-1}+\tau}(x^\star_\tau) - D\nabla \log \rho_{\theta_{k-1}}(x^\star_\tau))d\tau + \sqrt{2D}\; d\boldsymbol{B}_\tau,$$
with initial condition: $x_0 = x^\star_0 \sim \rho_{\theta_{k-1}}$ and $0 \le \tau \le h$.

Subtracting (5.39) from (5.40), we get

$$x^\star_\tau - x_\tau = \int_0^\tau \nabla V(x_s) - \nabla V(x^\star_s) + \vec{r}(x^\star_s, s)\; ds,$$

in which we denote $\vec{r}(x,\tau) = D\nabla \log \rho^\star_{t_{k-1}+\tau}(x) - D\nabla \log \rho_{\theta_{k-1}}(x)$ for convenience. Hence,

$$\begin{aligned}
\mathbb{E}|x^\star_\tau - x_\tau|^2 &= \mathbb{E}\left|\int_0^\tau \nabla V(x_s) - \nabla V(x^\star_s) + \vec{r}(x^\star_s, s)\; ds\right|^2 \\
&\le 2\;\mathbb{E}\left|\int_0^\tau \nabla V(x_s) - \nabla V(x^\star_s)\; ds\right|^2 + 2\;\mathbb{E}\left|\int_0^\tau \vec{r}(x^\star_s, s)\; ds\right|^2 \\
&\le 2\;\mathbb{E}\left[\tau\;\int_0^\tau |\nabla V(x_s) - \nabla V(x^\star_s)|^2\; ds\right] + 2\;\mathbb{E}\left[\tau\;\int_0^\tau |\vec{r}(x^\star_s, s)|^2\; ds\right] \\
&= 2\tau\left(\int_0^\tau \mathbb{E}|\nabla V(x_s) - \nabla V(x^\star_s)|^2 + \mathbb{E}|\vec{r}(x^\star_s, s)|^2\; ds\right).
\end{aligned}$$

Since the Hessian of V is bounded from above by Λ, $|\nabla V(x) - \nabla V(y)| \le \Lambda|x-y|$ for any $x, y \in \mathbb{R}^d$, we have the inequality

$$\mathbb{E}|x^\star_\tau - x_\tau|^2 \le 2\tau\Lambda^2\;\int_0^\tau \mathbb{E}|x^\star_s - x_s|^2\; ds + 2\tau\;\int_0^\tau \mathbb{E}|\vec{r}(x^\star_s, s)|^2\; ds. \tag{5.41}$$

If we define $U_\tau = \int_0^\tau \mathbb{E}|x^\star_s - x_s|^2\; ds$ and $R_\tau = \int_0^\tau \mathbb{E}|\vec{r}(x^\star_s, s)|^2\; ds$, (5.41) becomes

$$U'_\tau \le 2\Lambda^2\tau U_\tau + 2\tau R_\tau.$$

By integrating this inequality, we have $U_\tau \le \int_0^\tau 2e^{\Lambda(\tau^2-s^2)}sR_s\; ds$ and $U'_\tau \le 4\Lambda^2\tau\int_0^\tau e^{\Lambda(\tau^2-s^2)}sR_s ds + 2\tau R_\tau$. Therefore

$$W_2(\rho^\star_{t_k}, \tilde{\rho}_{t_k}) \le \sqrt{\mathbb{E}|x^\star_h - x_h|^2} = \sqrt{U'_h} \le \sqrt{4\Lambda^2 h\int_0^h e^{\Lambda(h^2-s^2)}sR_s\; ds + 2hR_h}.$$

Since R_τ is increasing with respect to τ, we are able to estimate

$$W_2(\rho^\star_{t_k}, \tilde{\rho}_{t_k}) \le \sqrt{4\Lambda^2h^2\int_0^h e^{\Lambda(h^2-s^2)}sds + 2h}\;\sqrt{R_h} = \sqrt{2\Lambda(e^{\Lambda h^2}-1)h + 2h}\sqrt{R_h}. \tag{5.42}$$

Next we estimate R_h. Recalling $\rho^*_{t_{k-1}} = \rho_{\theta_{k-1}}$ as in (5.31), we have

$$\begin{aligned} R_h &= \int_0^h \mathbb{E}_{x^\star_s} |D \log \rho^\star_{t_{k-1}+s}(x^\star_s) - D \log \rho^\star_{t_{k-1}}(x^\star_s)|^2 \, ds \\ &= D^2 \int_0^h \mathbb{E}_{x^\star_s} \left| \int_0^s \frac{\partial}{\partial t} \nabla \log \rho^\star_{t_{k-1}+t}(x^\star_s) \, dt \right|^2 \, ds \\ &\le D^2 \int_0^h \mathbb{E}_{x^\star_s} \left[s \int_0^s \left| \frac{\partial}{\partial t} \nabla \log \rho^\star_{t_{k-1}+t}(x^\star_s) \right|^2 \, dt \right] \, ds \\ &= D^2 \int_0^h \int_0^s s \int \left| \frac{\partial}{\partial t} \nabla \log \rho^\star_{t_{k-1}+t} \right|^2 \rho^\star_{t_{k-1}+s} \, dx \, dt \, ds. \end{aligned}$$

By (5.31), one can further compute $\frac{\partial}{\partial t} \log \rho^\star_{t_{k-1}+t} = -\nabla \log \rho^\star_{t_{k-1}+t} \cdot \nabla \zeta_{\theta_{k-1}} - \Delta \zeta_{\theta_{k-1}}$. Let us define

$$\begin{aligned} \mathfrak{L}(\theta, t, s) = \int |\nabla(\nabla \log \rho_t \cdot \nabla \zeta_\theta + \Delta \zeta_\theta)|^2 \rho_s \, dx \quad \text{with} \quad \zeta_\theta = V + D \log \rho_\theta \\ \text{and} \quad \frac{\partial \rho_s}{\partial s} + \nabla \cdot (\rho_s \nabla \zeta_\theta) = 0, \quad \rho_0 = \rho_\theta. \end{aligned}$$

Then we have the estimation

$$R_h \le D^2 \int_0^h \int_0^s \, s \cdot \left(\sup_{0 \le t \le s \le h} \mathfrak{L}(\theta_{k-1}, t, s) \right) \, dt \, ds = \frac{D^2}{3} \sup_{0 \le t \le s \le h} \mathfrak{L}(\theta_{k-1}, t, s) \, h^3.$$

Let us also define

$$L(\theta, h) = \left(\sup_{0 \le t \le s \le h} \mathfrak{L}(\theta, t, s) \right)^{\frac{1}{2}}. \tag{5.43}$$

Thus (5.42) becomes $W_2(\rho^\star_{t_k}, \tilde{\rho}_{t_k}) \le \sqrt{\frac{2D^2}{3}(\Lambda(e^{\Lambda h^2} - 1) + 2)} L(\theta_{k-1}, h) \, h^2$. When the step size h is small enough, we have $e^{\Lambda h^2} < 2$. Let us denote $K(D, \Lambda) = \sqrt{\frac{2D^2}{3}(\Lambda + 2)}$. Thus we have $W_2(\rho^\star_{t_k}, \tilde{\rho}_{t_k}) \le K(D, \Lambda) L(\theta_{k-1}, h) \, h^2$. ∎

Remark 5.17. Analyzing the discrepancy of stochastic particles under different movements provides a natural upper bound for the W_2 distance. Both Lemma 5.15 and Lemma 5.16 are derived by making use of the particle version of their corresponding density evolution. Such proving strategy was motivated from section 3.3.

LEMMA 5.18. *The third term in* (5.29) *satisfies* $W_2(\rho_{t_k}, \tilde{\rho}_{t_k}) \le e^{-\lambda h} W_2(\rho_{t_{k-1}}, \rho_{\theta_{k-1}})$. *Here we recall that* λ *satisfies* $\nabla^2 V \succeq \lambda I$.

This lemma is a direct corollary of the following theorem.

THEOREM 5.19. *Suppose the potential* $V \in C^2(\mathbb{R}^d)$ *satisfying* $\nabla^2 V \succeq \lambda I$ *for a finite real number* λ, *i.e., the matrix* $\nabla^2 V(x) - \lambda I$, *is semi–positive definite for any* $x \in \mathbb{R}^d$. *Given* $\rho_1, \rho_2 \in \mathcal{P}$, *and denoting by* $\rho^{(1)}_t$ *and* $\rho^{(2)}_t$ *the solutions of the Fokker–Planck equation with different initial distributions* ρ_1 *and* ρ_2, *respectively, i.e.,*

$$\begin{aligned} \frac{\partial \rho^{(1)}_t}{\partial t} &= \nabla \cdot (\rho^{(1)}_t \nabla V) + D \Delta \rho^{(1)}_t, \quad \rho^{(1)}_0 = \rho_1, \\ \frac{\partial \rho^{(2)}_t}{\partial t} &= \nabla \cdot (\rho^{(2)}_t \nabla V) + D \Delta \rho^{(2)}_t, \quad \rho^{(2)}_0 = \rho_2, \end{aligned}$$

then

$$W_2(\rho_t^{(1)}, \rho_t^{(2)}) \le e^{-\lambda t} W_2(\rho_1, \rho_2). \tag{5.44}$$

This is a known stability result on Wasserstein gradient flows. One can find its proof in [4] or [68]. With the results in Lemmas 5.15, 5.16, and 5.18, we are ready to prove Theorem 5.14.

Proof of Theorem 5.14. For convenience, we write

$$\mathrm{Err}_k = W_2(\rho_{\theta_k}, \rho_{t_k}), \quad k = 0, 1, \dots, N.$$

Combining Lemmas 5.15, 5.16, and 5.18 and the triangle inequality (5.29), we obtain

$$\mathrm{Err}_k \le \sqrt{\delta_0}\, h + \left(\frac{1}{\sqrt{3}} M(\theta_{k-1}, h) + \frac{1}{\sqrt{3}} N(\theta_{k-1}, h) + K(D, \Lambda) L(\theta_{k-1}, h) \right) h^2 + e^{-\lambda h}\, \mathrm{Err}_{k-1}.$$

Let us denote the constant C depending on initial parameter θ_0, time step size h, and time steps N:

$$C(\theta_0, h, N) = \max_{0 \le k \le N-1} \left\{ \frac{1}{\sqrt{3}} M(\theta_{k-1}, h) + \frac{1}{\sqrt{3}} N(\theta_{k-1}, h) + K(D, \Lambda) L(\theta_{k-1}, h) \right\}. \tag{5.45}$$

In the following discussion, we will denote $C = C(\theta_0, h, N)$ for simplicity. By (5.45), we have

$$\mathrm{Err}_k \le \sqrt{\delta_0} h + Ch^2 + e^{-\lambda h} \mathrm{Err}_{k-1}. \tag{5.46}$$

Multiplying $e^{\lambda k h}$ on both sides of (5.46), we get

$$e^{\lambda k h} \mathrm{Err}_k \le (\sqrt{\delta_0}\, h + Ch^2) e^{\lambda k h} + e^{\lambda (k-1) h} \mathrm{Err}_{k-1}. \tag{5.47}$$

For any n, $1 \le n \le N$, summing (5.47) from 1 to n, we reach

$$e^{\lambda n h} \mathrm{Err}_n \le (\sqrt{\delta_0} h + Ch^2) \left(\sum_{k=1}^{n} e^{\lambda k h} \right) + \mathrm{Err}_0 = (\sqrt{\delta_0} h + Ch^2) \frac{e^{\lambda (n+1) h} - e^{\lambda h}}{e^{\lambda h} - 1} + \mathrm{Err}_0.$$

Recalling that $t_n = nh$ for $1 \le n \le N$, this leads to

$$\mathrm{Err}_n \le (\sqrt{\delta_0} h + Ch^2) \frac{1 - e^{-\lambda t_n}}{1 - e^{-\lambda h}} + e^{-\lambda t_n} \mathrm{Err}_0, \quad n = 1, \dots, N. \qquad \square$$

Theorem 5.14 indicates that the error $W_2(\rho_{\theta_k}, \rho_{t_k})$ is upper bounded by $O(\sqrt{\delta_0}) + O(Ch) + O(W_2(\rho_{\theta_0}, \rho_0))$. Here $O(\sqrt{\delta_0})$ is the essential error term that originates from the approximation mechanism of our parametric Fokker–Planck equation. The $O(Ch)$ error term is induced by the finite difference scheme, and the $O(W_2(\rho_{\theta_0}, \rho_0))$ term is the initial error.

It is worth mentioning that the error bound for the forward Euler scheme in (5.28) matches the error bound for the continuous scheme (5.13) as we reduce the effects introduced by finite difference. To be more precise, under the assumption

$\lim_{h\to 0} C(\theta_0, h, N)h = 0$, we have

$$\begin{aligned}
&\lim_{h\to 0}(\sqrt{\delta_0}h + Ch^2)\frac{1-e^{-\lambda t}}{1-e^{-\lambda h}} + e^{-\lambda t}W_2(\rho_{\theta_0}, \rho_0)\\
&= \lim_{h\to 0}(\sqrt{\delta_0} + Ch)(1-e^{-\lambda t})\frac{h}{1-e^{-\lambda h}} + e^{-\lambda t}W_2(\rho_{\theta_0}, \rho_0)\\
&= \frac{\sqrt{\delta_0}}{\lambda}(1-e^{-\lambda t}) + e^{-\lambda t}W_2(\rho_{\theta_0}, \rho_0).
\end{aligned}$$

This indicates that error bounds (5.28) and (5.13) are compatible as $h \to 0$.

Remark 5.20 ($O(h)$ error order). Under further assumptions that $\Theta = \mathbb{R}^m$, $T_\theta(x) \in C^3(\Theta \times \mathbb{R}^d)$, and

$$\lim_{\theta\to\infty} H(\theta) = +\infty, \tag{5.48}$$

we can show that the finite difference error term $O(Ch)$ is of order $O(h)$. In fact, the solution obtained from the forward Euler scheme is always restricted in a fixed bounded region of Θ. To be more precise, supposing the initial value is θ_0, we consider $\Theta_0 = \{\theta | H(\theta) \le H(\theta_0)\}$. By (5.48), one can verify Θ_0 is a bounded and closed set and thus compact. We set $l = \max_{\theta\in\Theta_0} |G(\theta)^{-1}\nabla_\theta H(\theta)|$. Then we consider a slightly larger set $\Theta_0^l = \{\theta \mid \text{there exists } \theta' \in \Theta_0, \text{ s.t. } |\theta - \theta'| \le l\}$. Notice that Θ_0^l is also bounded. We define

$$\sigma_{\min}^G = \min_{\theta\in\Theta_0^l} \sigma_{\min}(G(\theta)), \quad \sigma_{\max}^H = \max_{\theta\in\Theta_0^l} \sigma_{\max}(\nabla^2_{\theta\theta} H(\theta)).$$

Here $\sigma_{\max}(A), \sigma_{\min}(A)$ denotes the maximum and the minimum singular values of matrix A. We can show that for any time step size $h < \min\{\frac{2\sigma_{\min}^G}{\sigma_{\max}^H}, 1\}$, the numerical solution $\{\theta_k\}_{k=1}^N$ obtained by applying the forward Euler scheme to (3.18) is included in Θ_0. To prove this, we first show $\theta_1 \in \Theta_0$ and consider

$$\begin{aligned}
H(\theta_1) = H(\theta_0 - hG(\theta_0)^{-1}\nabla_\theta H(\theta_0)) &= H(\theta_0) - h\xi^{\mathrm{T}} G(\theta_0)\xi + \frac{h^2}{2}\xi^{\mathrm{T}}\nabla^2_{\theta\theta}H(\tilde{\theta})\xi\\
&\le H(\theta_0) - h\sigma_{\min}^G|\xi|^2 + \frac{h^2}{2}\sigma_{\max}^H|\xi|^2 \le H(\theta_0).
\end{aligned}$$

Here we denote $\xi = G(\theta_0)^{-1}\nabla_\theta H(\theta_0)$. The second equality is due to $T_\theta(x) \in C^3(\Theta \times \mathbb{R}^d)$ and thus $H(\cdot) \in C^2(\Theta)$. We notice that $\tilde{\theta} = \theta_0 + \tau(hG(\theta_0)^{-1}\nabla_\theta H(\theta_0))$ with $0 \le \tau \le 1$ and thus $\tilde{\theta} \in \Theta_0^l$. Since $H(\theta_1) \le H(\theta_0)$, we know $\theta_1 \in \Theta_0$. Applying a similar argument with θ_0 being replaced by θ_1, we can further prove $\theta_2 \in \Theta_0$. By induction, we can prove $\{\theta_k\}_{k=1}^N \subset \Theta_0$. Since $\mathfrak{M}(\theta, s), \mathfrak{N}(\theta, s), \mathfrak{L}(\theta, s)$ depend continuously on θ, s, their supreme values on compact set $\Theta_0 \times [0, 1]$ must be finite so we know $C(\theta_0, h, N)$ in (5.45) is upper bounded by a constant independent of h as well as N (recall $N = \frac{T}{h}$). Thus the error term $O(Ch)$ is of $O(h)$ order.

Similar to the discussion in previous sections, we can naturally extend Theorem 5.14 to an a posteriori estimate.

THEOREM 5.21 (a posteriori error analysis of forward Euler scheme).

$$\begin{aligned}
W_2(\rho_{\theta_k}, \rho_{t_k}) \le (\sqrt{\delta_1}h + Ch^2)\frac{1-e^{-\lambda t_k}}{1-e^{-\lambda h}} + e^{-\lambda t_k}W_2(\rho_{\theta_0}, \rho_0)\\
\textit{for any} \quad t_k = kh, \quad 0 \le k \le N.
\end{aligned}$$

The explicit definition of the constant C is in (5.45).

Up to this point, we have mainly analyzed the error term for the forward Euler scheme. In our numerical implementation, we adopt the scheme (4.8), which turns out to be a semi-implicit scheme with $O(h^2)$ local error. In the following discussion, we compare the difference between the numerical solutions of our semi-implicit scheme and the forward Euler scheme.

Recall that the parametric Fokker–Planck equation (3.18) is an ODE: $\dot{\theta} = -G(\theta)^{-1}\nabla_\theta H(\theta)$. We consider two numerical schemes:

(5.49)
$$\theta_{n+1} = \theta_n - hG(\theta_n)^{-1}\nabla_\theta H(\theta_n), \quad \theta_0 = \theta,\ n = 1, 2, \dots, N \quad \text{(forward Euler scheme)},$$

(5.50)
$$\hat{\theta}_{n+1} = \hat{\theta}_n - hG(\hat{\theta}_n)^{-1}\nabla_\theta H(\hat{\theta}_{n+1}), \quad \hat{\theta}_0 = \theta,\ n = 1, 2, \dots, N \text{ (semi-implicit Euler scheme)}.$$

We denote $F(\theta') = G(\theta')^{-1}\nabla_\theta F(\theta'')$ and set

$$L_1 = \max_{1\le n\le N}\left\{\|F(\theta_n) - F(\hat{\theta}_n)\|/\|\theta_n - \hat{\theta}_n\|\right\},$$
$$L_2 = \max_{1\le k\le N}\{\|\nabla_\theta H(\hat{\theta}_n) - \nabla_\theta H(\hat{\theta}_{n+1})\|/\|\hat{\theta}_n - \hat{\theta}_{n+1}\|\},$$
$$M_1 = \max_{1\le n\le N}\{\|G(\hat{\theta}_n)^{-1}\|\}, \quad M_2 = \max_{1\le n\le N}\{\|\nabla_\theta H(\hat{\theta}_n)\|\},$$

where $\|\cdot\|$ is a vector norm (or its corresponding matrix norm).

THEOREM 5.22 (relation between forward Euler and proposed semi-implicit schemes). *The numerical solutions θ_n and $\hat{\theta}_n$ of the forward Euler and semi-implicit schemes with time step size h and $Nh = T$ satisfy*

$$\|\theta_n - \hat{\theta}_n\| \le ((1 + L_1 h)^n - 1)\frac{M_1^2 M_2 L_2}{L_1}h, \quad n = 1, 2, \dots, N.$$

This result implies that $\|\theta_n - \hat{\theta}_n\|$ can be upper bounded by $(e^{L_1 T} - 1)\frac{M_1^2 M_2 L_2}{L_1}h$. When assuming the upper bounds $L_1, L_2, M_1, M_2 \sim O(1)$ as $h \to 0$ (or equivalently $N \to \infty$), the differences between our proposed semi-implicit scheme and the forward Euler scheme can be bounded by $O(h)$. As a consequence, we are able to establish an $O(h)$ error bound for our proposed scheme (4.8).

Proof of Theorem 5.22. If we subtract (5.50) from (5.49),

$$(\theta_{n+1} - \hat{\theta}_{n+1}) = (\theta_n - \hat{\theta}_n) - h(G(\theta_n)^{-1}\nabla_\theta H(\theta_n) - G(\hat{\theta}_n)^{-1}\nabla_\theta H(\hat{\theta}_{n+1})),$$

and denote $e_n = \theta_n - \hat{\theta}_n$ and $F(\theta) = G(\theta_n)^{-1}\nabla_\theta H(\theta)$, we may rewrite this equation as

$$e_{n+1} = e_n - h(F(\theta_n) - F(\hat{\theta}_n) + G(\hat{\theta}_n)^{-1}(\nabla_\theta H(\hat{\theta}_n) - \nabla_\theta H(\hat{\theta}_{n+1}))).$$

Recalling the definitions of L_1, L_2, M_1, we have

$$\|e_{n+1}\| \le \|e_n\| + hL_1\|e_n\| + hM_1L_2\|\hat{\theta}_{n+1} - \hat{\theta}_n\|.$$

By the semi-implicit scheme, we have

$$\hat{\theta}_{n+1} - \hat{\theta}_n = -hG(\hat{\theta}_n)^{-1}\nabla_\theta H(\hat{\theta}_{n+1}).$$

Thus $|\hat{\theta}_{n+1} - \hat{\theta}_n\| \leq hM_1M_2$. This gives us a recurrent inequality,

$$\|e_{n+1}\| \leq \|e_n\| + hL_1\|e_n\| + M_1^2M_2L_2h^2,$$

which implies

$$\left(\|e_{n+1}\| + \frac{M_1^2M_2L_2}{L_1}h\right) \leq (1+hL_1)\left(\|e_n\| + \frac{M_1^2M_2L_2}{L_1}h\right), \quad n = 0, 1, \ldots, N-1.$$

This leads to

$$\|e_n\| \leq ((1+hL_1)^n - 1)\frac{M_1^2M_2L_2}{L_1}h.$$

When we solve the ODE on $[0,T]$ with $h = T/N$, we have $(1+hL_1)^n \leq (1+hL_1)^N = \left(1 + \frac{L_1T}{N}\right)^N \leq e^{L_1T}$. This means all terms $\{\|e_n\|\}_{1\leq n\leq N}$ can be upper bounded by $(e^{L_1T}-1)\frac{M_1^2M_2L_2}{L_1}h$. ∎

Remark 5.23. In order to make our argument clear and concise, we omitted the errors introduced by the approximation of ReLU function ψ_ν. Careful analysis on how well $\nabla\psi_\nu$ can approximate a general gradient field is among our future research directions.

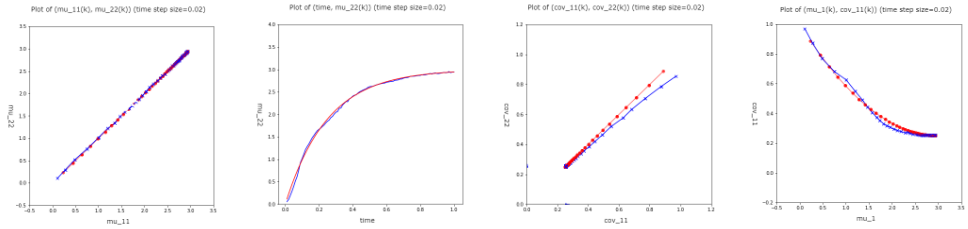
Remark 5.24. The convergence property of the SGD method (mainly the Adam method) used in our Algorithm 4.1 is not discussed in details. One can check its convergence analysis in the paper [26]. Based on our experiences, for most of the smooth potential functions $V \in \mathcal{V}$ with diffusion coefficient D not too small (i.e., $D > 0.1$), our algorithm shows convergent behavior and produces accurate results when checking against the true solution if it is possible.

6. Numerical examples. In this section, we consider solving the Fokker–Planck equation (2.2) on $\mathbb{R}^d$ with initial condition $\rho_0(x) = \mathcal{N}(0, I_d)$ by using Algorithm 4.1.[2] We demonstrate several numerical examples with different potential functions V. In the following experiments, unless specifically stated, we choose the length of normalizing flow T_θ as 60. We set $\psi_\nu : \mathbb{R}^d \to \mathbb{R}$ as the ReLU network with the number of layers equal to 6 and hidden dimension equal to 20. We use the Adam (adaptive moment estimation) SGD method [26] with default parameters $D_1 = 0.9, D_2 = 0.999; \epsilon = 10^{-8}$. For the parameters of Algorithm 4.1, we choose $\alpha_{\text{out}} = 0.005$, $\alpha_{\text{in}} = 0.0005$. We follow Remark 4.11 to choose $K_{\text{in}}, K_{\text{out}} = \max\{1000, 300d\}$. Based on our experience, we set $M_{\text{out}} = O(\frac{h}{\alpha_{\text{out}}})$. The suitable value of M_{in} can be chosen after several quick tests to make sure that every inner optimization problem (4.27) can be solved.

Our Python code can be downloaded from the Github website: https://github.com/LSLSliushu/Parametric-Fokker-Planck-Equation.

6.1. Quadratic potential. Our first set of examples uses quadratic potential V. In this case, we can compute the explicit solution of (2.2). These examples are used for verification purposes, because we can check the results with exact solutions.

[2]We can set initial value θ_0 so that $T_{\theta_0} = \text{Id}$, and thus $\rho_0 = T_{\theta_0\sharp}p$ is a standard Gaussian distribution.



(a) $\{\hat{\mu}^{(k)}\}$ (b) $\{\hat{\mu}_1^{(k)}\}$ (c) $\{(\hat{\Sigma}_{11}^{(k)}, \hat{\Sigma}_{22}^{(k)})\}$ (d) $\{(\hat{\mu}_{(k)}, \hat{\Sigma}_{11}^{(k)})\}$

FIG. 7. *Plot of empirical statistics (numerical solution: blue; real solution: red). (Color available online.)*

6.1.1. 2D cases. We take $d = 2$ and set $V(x) = \frac{1}{2}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)$, with $\mu = [3,3]^{\mathrm{T}}$ and $\Sigma = \mathrm{diag}([0.25, 0.25])$. The solution of (2.2) is

$$\rho_t = \mathcal{N}(\mu(t), \Sigma(t)) \quad \mu(t) = (1-e^{-4t})\mu, \ \Sigma(t) = \begin{bmatrix} \frac{1}{4} + \frac{3}{4}e^{-8t} & \\ & \frac{1}{4} + \frac{3}{4}e^{-8t} \end{bmatrix}, \quad t \geq 0.$$

We solve the equation in time interval $[0, 0.7]$ with time step size 0.01. We set $M_{\mathrm{out}} = 20$ and $M_{\mathrm{in}} = 100$.

To compare against the exact solution, we set $M = 6000$ and sample $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_M\} \sim T_{\theta_k \sharp} p$ at time t_k and use

$$\hat{\mu}^k = \frac{1}{M}\sum_{j=1}^{M} \boldsymbol{X}_j, \quad \hat{\Sigma}^k = \frac{1}{M-1}\sum_{j=1}^{M} (\boldsymbol{X}_j - \hat{\mu}_k)(\boldsymbol{X}_j - \hat{\mu}_k)^{\mathrm{T}}$$

to compute for its empirical mean and covariance of $\hat{\rho}_k$. We plot the blue curves $\{\hat{\mu}^{(k)}\}$, $\{\hat{\mu}_2^{(k)}\}$, $\{(\hat{\Sigma}_{11}^{(k)}, \hat{\Sigma}_{22}^{(k)})\}$, $\{(\hat{\mu}_1^{(k)}, \hat{\Sigma}_{11}^{(k)})\}$ in Figure 7; these plots properly capture the exponential convergence exhibited by the explicit solution (red curves) $\{\mu(t)\}$, $\{\mu_2(t)\}$, $\{(\Sigma_{11}(t), \Sigma_{22}(t))\}$, $\{(\mu_1(t), \Sigma_{11}(t))\}$.

We also examine the network $\psi_{\hat{\nu}}$ trained at the end of each outer iteration. Generally speaking, the gradient field $\nabla\psi_{\hat{\nu}}$ reflects the movements of the particles under the Vlasov-type dynamic (2.3) at every time step. Shown in Figures 8 and 9 are the graphs of $\psi_{\hat{\nu}}$ at $k = 10$ and $k = 140$, respectively. As we can see from these graphs, the gradient field is in the same direction, but judging from the variation of two $\psi_{\hat{\nu}}$'s, when $k = 10$, $|\nabla\psi_{\hat{\nu}}|$ is much greater than its value at $k = 140$. This is because when $t = 140$, the distribution is already close to the Gibbs distribution, and the particles no longer need to move for a long distance to reach their final destination.

In the next example, we apply our algorithm to the Fokker–Planck equation with nonisotropic potential

$$V(x) = \frac{1}{2}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu), \quad \mu = \begin{bmatrix} 3 \\ 3 \end{bmatrix}, \text{ and } \Sigma = \begin{bmatrix} 1 & \\ & \frac{1}{4} \end{bmatrix}.$$

One can verify that the solution to (2.2) is

$$\rho_t = \mathcal{N}(\mu_t, \Sigma_t), \quad \mu_t = \begin{bmatrix} 3(1-e^{-t}) \\ 3(1-e^{-4t}) \end{bmatrix}, \ \Sigma_t = \begin{bmatrix} 1 & \\ & \frac{1}{4}(1+3e^{-8t}) \end{bmatrix}.$$

We use the same parameters as before. We solve (2.2) at time interval $[0, 1.4]$ with time step size 0.005.

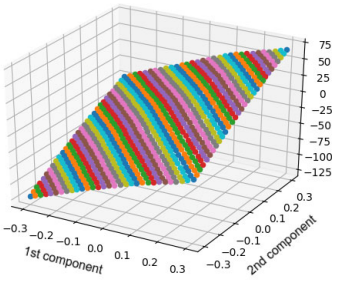


FIG. 8. *Graph of $\psi_{\hat{\nu}}$ after $M_{out} = 20$ outer iterations at $k = 10$th time step.*

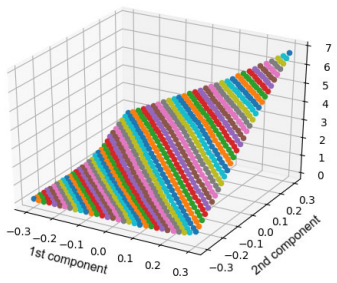


FIG. 9. *Graph of $\psi_{\hat{\nu}}$ after $M_{out} = 20$ outer iterations at $k = 140$th time step.*

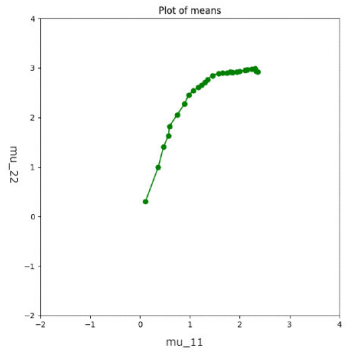


FIG. 10. *Mean trajectory of $\{\rho_{\theta_t}\}$ w.r.t. $\dot{\theta} = -G(\theta)^{-1}\nabla_\theta H(\theta)$.*

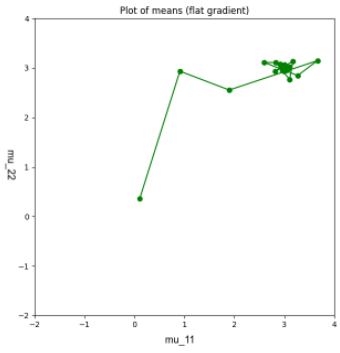


FIG. 11. *Mean trajectory of $\{\rho_{\theta_t}\}$ w.r.t. $\dot{\theta} = -\nabla_\theta H(\theta)$.*

Similarly, we also plot the empirical mean trajectory; one can compare it with the true solution $\mu(t) = (3(1-e^{-t}), 3(1-e^{-4t}))$. Both the curvature and the exponential convergence to μ are captured by our numerical result. To demonstrate the effectiveness of our formulation, we also compare the mean trajectory obtained by our result (Figure 10) with the mean trajectory obtained by computing the flat gradient flow $\dot{\theta} = -\nabla_\theta H(\theta)$ (Figure 11). It reveals very different behavior of the flat gradient (∇_θ) flow and Wasserstein gradient ($G(\theta)^{-1}\nabla_\theta$) flow. Clearly, our approximation based on Wasserstein gradient flow captures the exact mean function much more accurately. We compare the graph of trained $\psi_{\hat{\nu}}$ at different time steps $k = 10, 140$ (Figures 12 and 13). The directions of $\nabla\psi_{\hat{\nu}}$ at $k = 10$ and $k = 140$ are different from the previous example. This is caused by the nonisotropic quadratic (Gaussian) potential V used in this example.

6.1.2. Verification of the error estimate. We verify the $O(h)$ error estimation discussed in section 5.3.2 based on numerical experiments with quadratic potentials. We consider $V(x) = |x - \mu|^2$ defined on $\mathbb{R}^2$ with $\mu = (12.0, 12.0)$ and ρ_0 as standard Gaussian at time interval $[0, 1]$. We run our algorithm with several different time step sizes $h = 0.01, 0.05, 0.08, 0.1, 0.2, 0.3$ and record their corresponding mean trajectory $\{\hat{\mu}^{(k)}\}$ as defined in section 6.1.1. During this process, we need to adjust our hyperparameters $\alpha_{\text{in}}, \alpha_{\text{out}}, M_{\text{in}}, M_{\text{out}}$ correspondingly in order to guarantee the convergence of the Adam method. Denote $\{\mu(t_k)\}$ as the real solution. We

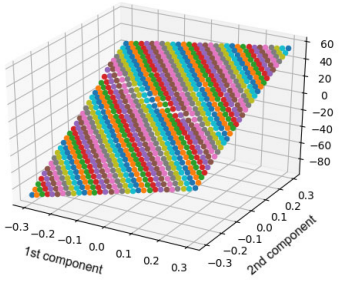


FIG. 12. *Graph of $\psi_{\hat{\rho}}$ after $M_{out} = 20$ outer iterations at $k = 10$th time step.*

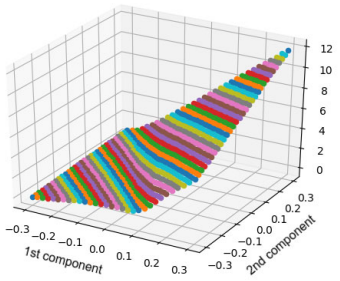


FIG. 13. *Graph of $\psi_{\hat{\rho}}$ after $M_{out} = 20$ outer iterations at $k = 140$th time step.*

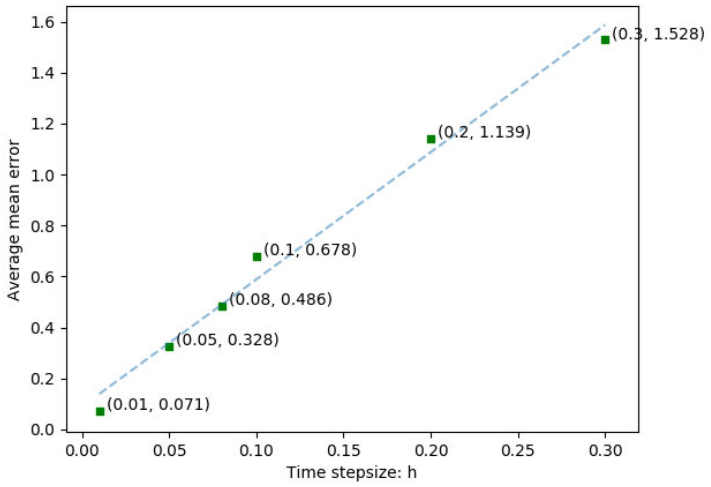
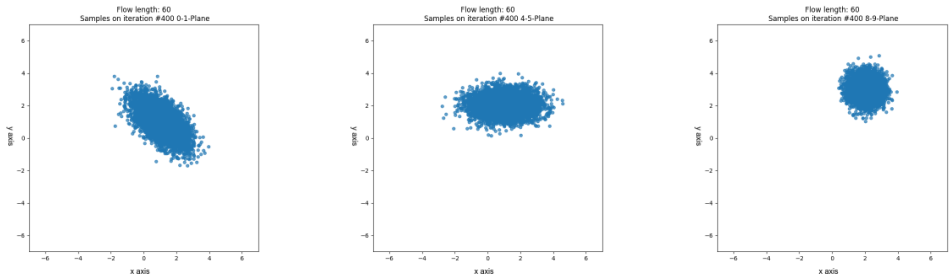


FIG. 14. *Numerical errors versus time step size h.*

compute the average l^2 error of mean values as $\text{AveErr}(h) = \frac{1}{N}\sum_k |\hat{\mu}^{(k)} - \mu(t_k)|$. We pick h in a range larger than 0.01 because when h is smaller, the influence from the approximation error δ_0 of normalizing flow T_θ as well as initial error $W_2(\rho_0, \rho_{\theta_0})$ start to dominate the overall error. Figure 14 exhibits the linear relationship between our numerical error $\text{AveErr}(h)$ and time step size h, which confirms our theoretical estimates.

Remark 6.1. The reason for choosing quadratic potential is because its corresponding Fokker–Planck equation has an explicit solution. The reason that we focus on the average error of mean vectors is mainly due to computational accuracy and convenience: one can approximate the error of the mean vector of a distribution by computing the arithmetic average of samples, which is faster and more accurate than computing for the L^2-Wasserstein error among two distributions.



(a) projection of samples on the 0-1 plane

(b) projection of samples on the 4-5 plane

(c) projection of samples on the 8-9 plane

FIG. 15. *Sample points of computed ρ_{θ_t} projected on different planes at $t = 2.0$.*

6.1.3. Higher dimension. We implement our algorithm in higher dimensional space. In the next example, we take $d = 10$ and consider the quadratic potential

$$V(x) = \frac{1}{2}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu),\ \Sigma = \mathrm{diag}(\Sigma_A, I_2, \Sigma_B, I_2, \Sigma_C),\ \mu = (1,1,0,0,1,2,0,0,2,3)^{\mathrm{T}}.$$

Here we set the diagonal blocks as

$$\Sigma_A = \begin{bmatrix} \frac{5}{8} & -\frac{3}{8} \\ -\frac{3}{8} & \frac{5}{8} \end{bmatrix},\quad \Sigma_B = \begin{bmatrix} 1 & \\ & \frac{1}{4} \end{bmatrix},\quad \Sigma_C = \begin{bmatrix} \frac{1}{4} & \\ & \frac{1}{4} \end{bmatrix}.$$

We solve the equation at time interval $[0, 0.7]$ with time step size 0.005. We set $M_{\text{out}} = 20$ and $M_{\text{in}} = 100$. To demonstrate the results, 6000 samples from the reference distribution p are drawn and pushed forward by using our computed map T_{θ_k}. We plot a few snapshots of the pushed forward points (from $t = 0.05$ to $t = 0.70$) in Figure 15. One can check that the distribution of our numerical computed samples gradually converges to the Gibbs distribution $\mathcal{N}(\mu, \Sigma)$.

We solve (2.2) at time interval $[0, 2]$ with time step size $h = 0.005$. We set $K_{\text{in}} = K_{\text{out}} = 3000$ and choose $M_{\text{out}} = 30$, $M_{\text{in}} = 100$. To demonstrate the results, 6000 samples from the reference distribution p are drawn and pushed forward by using our computed map T_{θ_k}. We exhibit the projection of the samples on the 0-1, 4-5, and 8-9 planes in Figure 15 at time $t = 2.0$. One can verify that the distribution of our numerical computed samples converges to the Gibbs distribution $\mathcal{N}(\mu, \Sigma)$. The explicit solution to the Fokker–Planck equation is always the Gaussian distribution $\mathcal{N}(\mu(t), \Sigma(t))$ with mean $\mu(t)$ and covariance matrix $\Sigma(t)$:

$$\begin{aligned}
&\mu(t) = (1-e^{-t}, 1-e^{-t}, 0, 0, 1-e^{-t}, 2(1-e^{-4t}), 0, 0, 2(1-e^{-4t}), 3(1-e^{-4t}))^{\mathrm{T}},\\
&\Sigma(t) = \mathrm{diag}(\Sigma_A(t), I, \Sigma_B(t), I, \Sigma_C(t)),\\
&\quad \text{with}\ \ \Sigma_A(t) = \begin{bmatrix} \frac{5}{8}+f(t) & -\frac{3}{8}+f(t) \\ -\frac{3}{8}+f(t) & \frac{5}{8}+f(t) \end{bmatrix},\ \Sigma_B(t) = \begin{bmatrix} 1 & \\ & \frac{1+3e^{-8t}}{4} \end{bmatrix},\\
&\qquad\quad \Sigma_C(t) = \begin{bmatrix} \frac{1+3e^{-8t}}{4} & \\ & \frac{1+3e^{-8t}}{4} \end{bmatrix},\\
&\quad \text{where}\ \ f(t) = -\frac{2}{7}e^{-t} + \frac{1}{3}e^{-2t} + \frac{55}{168}e^{-8t}.
\end{aligned}$$

To compare against the exact solution, we set sample size $M = 6000$ and compute the empirical mean $\hat{\mu}^k$ and covariance $\hat{\Sigma}^k$ of our numerical solution $\hat{\rho}_k$ at time t_k.

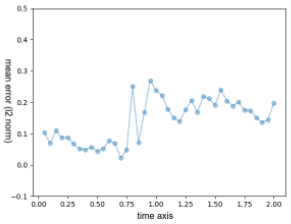


FIG. 16. *Mean error* (l_2).

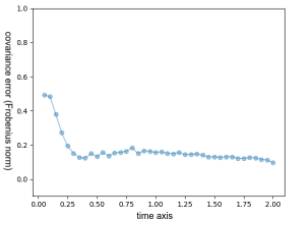


FIG. 17. *Covariance error* ($\|\cdot\|_F$).

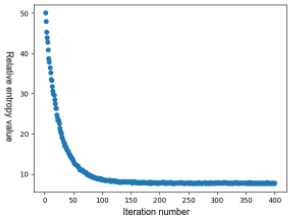
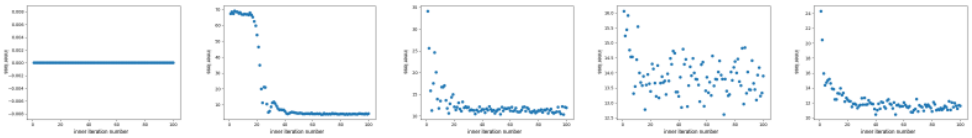


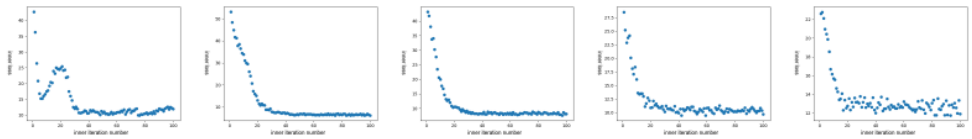
FIG. 18. *Plot of* $\{H(\theta)\}$.

We evaluate the error between $\hat{\mu}^{(k)}$ and $\mu(t_k)$, $\hat{\Sigma}^{(k)}$ and $\Sigma(t_k)$. We plot the error curves of $\|\hat{\mu}^{(k)} - \mu(t_k)\|_2$ (Figure 16) and $\|\hat{\Sigma}^{(k)} - \Sigma(t_k)\|_F$ (Figure 17). Here $\|\cdot\|_F$ is the matrix Frobenius norm. Figure 18 captures the exponential decay of H along its Wasserstein gradient flow; this verifies the entropy dissipation property of the Fokker–Planck equation with convex potential function V.

In this case, we take a closer look at the loss in the inner loops. Figure 19 shows the first 10 (out of 20) loss plots when applying the SGD method to solve (4.30) with $k = 200$ ($t = 200 \cdot h = 1.0$). The remaining loss plots from the 11th outer iteration to 20th iteration are similar to the plots in the second row. The situations are similar for other time steps k. We believe that $M_{\text{in}} = 100$ works well in this problem; the SGD method we used can thoroughly solve the variational problem (4.30) for each outer loop.



(a) 1st iteration (b) 2nd iteration (c) 3rd iteration (d) 4th iteration (e) 5th iteration



(f) 6th iteration (g) 7th iteration (h) 8th iteration (i) 9th iteration (j) 10th iteration

FIG. 19. *Plots of inner loop losses.*

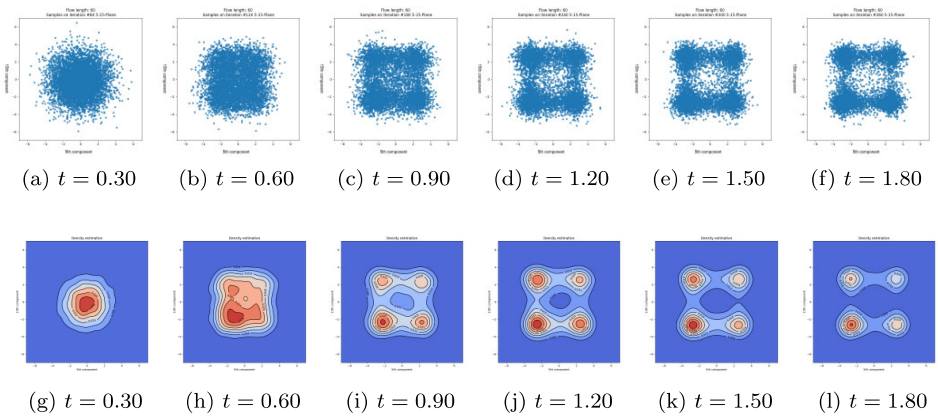
6.2. Experiments with more general potentials. In this section, we exhibit two examples with more general potentials in higher dimensional space.

6.2.1. Styblinski–Tang potential. In this example, we set dimension $d = 30$ and consider the Styblinski–Tang function [65]

$$V(x) = \frac{3}{50}\left(\sum_{i=1}^{d} x_i^4 - 16x_i^2 + 5x_i\right).$$

We solve (2.2) with potential V on time interval $[0, 3]$ with time step size $h = 0.005$. We set $K_{\text{in}} = K_{\text{out}} = 9000$ and $M_{\text{in}} = 100$, $M_{\text{out}} = 30$.

To exhibit sample results, due to the symmetric structure of the potential function, we project the sample points in $\mathbb{R}^{30}$ to some random plane, such as the 5-15 plane used in this paper. The sample plots and their estimated densities are presented in Figure 20.



(a) $t = 0.30$ (b) $t = 0.60$ (c) $t = 0.90$ (d) $t = 1.20$ (e) $t = 1.50$ (f) $t = 1.80$

(g) $t = 0.30$ (h) $t = 0.60$ (i) $t = 0.90$ (j) $t = 1.20$ (k) $t = 1.50$ (l) $t = 1.80$

FIG. 20. *Sample points and estimated densities of ρ_{θ_t} on the 5-15 plane at different time nodes.*

In this special example, the potential function is the direct addition of the same functions; we can exploit this property and show that any marginal distribution

$$\varrho_j(x_j, t) = \int \cdots \int \rho(x, t)\ dx_1 \dots dx_{j-1} dx_{j+1} \dots dx_d$$

of the solution ρ_t solves the following the 1D Fokker–Planck equation:

$$\frac{\partial \varrho(x,t)}{\partial t} = \frac{\partial}{\partial x}(\varrho(x,t)\ V'(x)) + D\Delta\varrho(x,t), \quad \varrho(\cdot, 0) = \mathcal{N}(0,1), \tag{6.1}$$
$$\text{with } V(x) = \frac{3}{50}(x^4 - 16x^2 + 5x).$$

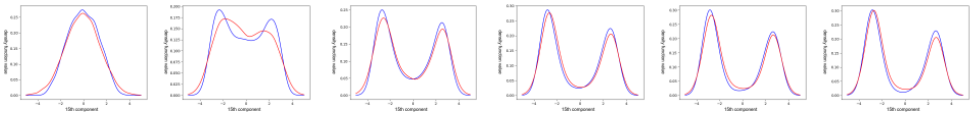
We then solve the SDE associated to (6.1):

$$dX_t = -V'(X_t)\ dt + \sqrt{2D} dB_t, \quad X_0 \sim \mathcal{N}(0,1). \tag{6.2}$$

Since (6.2) is an SDE in 1D space, we can solve it with high accuracy by the Euler–Maruyama scheme [28] and use it as a benchmark for our numerical solution. Figure 21 exhibits both the estimated densities for our numerical solutions (marginal distribution on the 15th component) and the solution of (6.2) given by the Euler–Maruyama scheme with step size 0.005. The sample sizes for both solutions equal 6000.

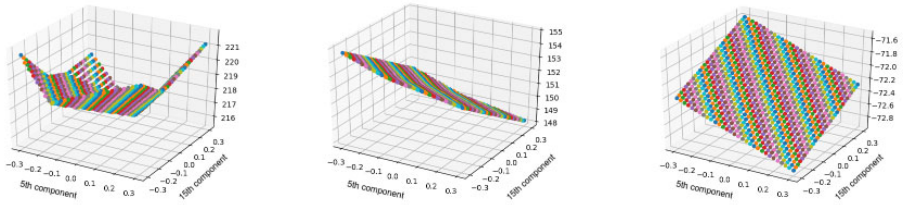
We also illustrate the graphs of $\psi_{\hat{\nu}}$ on the 5-15 plane trained at different time steps in Figure 22.

6.2.2. Affects of different initial distributions. Different initial conditions ρ_0 affect the behavior of solutions of the neural parametric Fokker–Planck equation differently, especially on the convergence speed to the Gibbs distribution. Here is an example. We consider V the Styblinski–Tang potential in $\mathbb{R}^2$. We compute the



(a) $t = 0.30$ (b) $t = 0.60$ (c) $t = 0.90$ (d) $t = 1.20$ (e) $t = 1.50$ (f) $t = 1.80$

FIG. 21. *Estimated densities of our numerical solution (red) (projected onto the* 15*th component) and the solution given by the Euler–Maruyama scheme (blue). (Color available online.)*



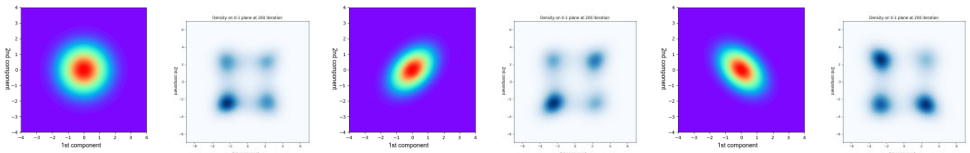
(a) $\psi_{\hat{\nu}}$ at $k = 30$ (b) $\psi_{\hat{\nu}}$ at $k = 150$ (c) $\psi_{\hat{\nu}}$ at $k = 360$

FIG. 22. *Graph of* $\psi_{\hat{\nu}}$ *on the* 5-15 *plane trained at different time steps.*

solutions with three different initial distributions given as Gaussian distributions with covariances

$$\Sigma_1 = \begin{bmatrix} 1 & \\ & 1 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} \frac{13}{8} & \frac{5}{8} \\ \frac{5}{8} & \frac{13}{8} \end{bmatrix}, \quad \Sigma_3 = \begin{bmatrix} \frac{13}{8} & -\frac{5}{8} \\ -\frac{5}{8} & \frac{13}{8} \end{bmatrix},$$

respectively. Although the solutions converge to the Gibbs distribution, as expected from the theory, regardless of the initial density, their convergence speed may be different. Figure 23 shows the initial distributions and the corresponding densities (which are the estimations of the samples obtained from our algorithm) at $t = 1.0$. As we can observe, the numerical result produced by $\rho_0 = \mathcal{N}(0, \Sigma_1)$ is already close to the Gibbs distribution at $t = 1.0$, while numerical results associated to Σ_2, Σ_3 still have noticeable differences from Gibbs. They seem to be trapped in intermediate metastable statuses that are clearly influenced by the orientations in initial distributions.



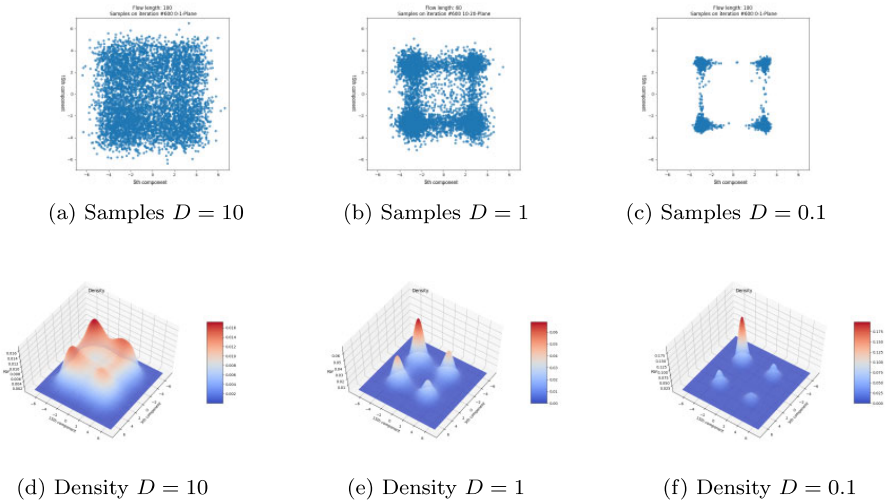
(a) $\rho_0 = \mathcal{N}(0, \Sigma_1)$ (b) $t = 1.0$ (c) $\rho_0 = \mathcal{N}(0, \Sigma_2)$ (d) $t = 1.0$ (e) $\rho_0 = \mathcal{N}(0, \Sigma_3)$ (f) $t = 1.0$

FIG. 23. *Different behaviors of numerical solution with different* ρ_0*'s.*

In general, we believe that the choice of ρ_0 affects the behavior of numerical solution. Choosing a suitable ρ_0 may shorten the computing time in the training process.

6.2.3. Solving the equation with different diffusion coefficients. The different behaviors of the Fokker–Planck equation caused by different diffusion coeffi-

cients D can be captured by our algorithm. As Figure 24 shows, we apply our method to solve the Fokker–Planck equation with the Styblinski–Tang potential function with $D = 0.1, 1.0, 10.0$ and exhibit samples points and estimated density surfaces at time $t = 3.0$.



(a) Samples $D = 10$ (b) Samples $D = 1$ (c) Samples $D = 0.1$

(d) Density $D = 10$ (e) Density $D = 1$ (f) Density $D = 0.1$

FIG. 24. *Samples and estimated densities at $t = 3.0$; from left to right: $D = 10$, $D = 1.0$, $D = 0.1$.*

6.2.4. Rosenbrock potential. In this example, we set dimension $d = 10$. We consider the Rosenbrock-type function [61]

$$V(x) = \frac{3}{50}\left(\sum_{i=1}^{d-1} 10(x_{k+1} - x_k^2)^2 + (x_k - 1)^2\right),$$

which involves interactions among its coordinates. We solve the corresponding (2.2) on time interval $[0, 1]$ with step size $h = 0.005$. We set the length of normalizing flow T_θ as 100. We set $K_{\text{in}} = K_{\text{out}} = 3000$ and $M_{\text{in}} = 100$, $M_{\text{out}} = 60$.

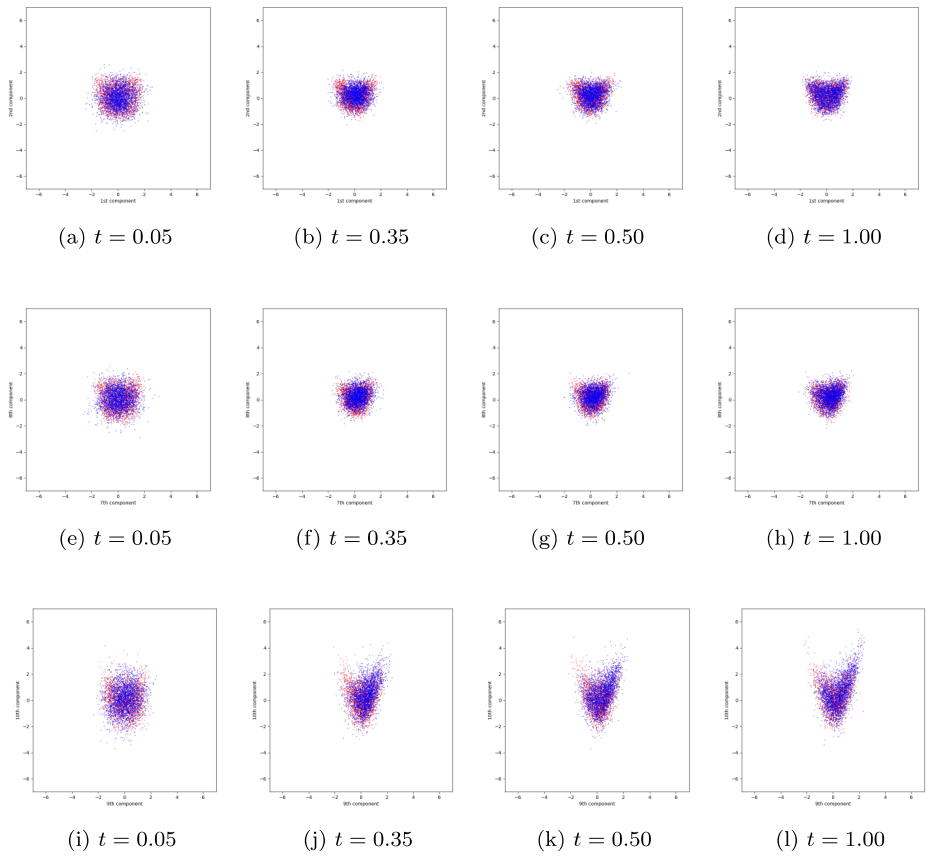
We exhibit the projection of sample points on the 1-2, 7-8, and 9-10 planes in Figure 25. Blue samples are obtained from our numerical solution, while red samples are obtained by applying the Euler–Maruyama scheme with the same step size.

6.3. Discussion on time consumption. We should point out that the running time of our algorithm depends on the following three aspects:

(i) Dimension d of the problem; potential function V.
(ii) The size of normalizing flow T_θ and fully connected neural network ψ_ν.
(iii) Number of time steps N; outer iterations M_{out}; inner iterations M_{in}; sample size K_{out} and K_{in}.

Among them, the networks in (ii) are selected according to (i). The hyperparameters $M_{\text{out}}, M_{\text{out}}, K_{\text{out}}, K_{\text{in}}$ in (iii) are chosen based on our trial and error as well as Remark 4.10 stated earlier in this paper.

All numerical examples reported in this paper are computed on a laptop with an Intel Core i5-8250U CPU @ 1.60GHz × 8 processor. For most of the high-dimensional examples ($d \geq 10$), we choose the length of T_θ between 60 and 100; for the ReLU network ψ_ν, we set its number of layers equal to 6 with hidden dimension 20. We set



(a) $t = 0.05$ (b) $t = 0.35$ (c) $t = 0.50$ (d) $t = 1.00$

(e) $t = 0.05$ (f) $t = 0.35$ (g) $t = 0.50$ (h) $t = 1.00$

(i) $t = 0.05$ (j) $t = 0.35$ (k) $t = 0.50$ (l) $t = 1.00$

FIG. 25. *Samples of our numerical solution (blue) and Euler–Maruyama (red) on different planes at different time nodes. (Color available online.)*

$M_{\text{out}} \sim 50, M_{\text{in}} \sim 100$ and choose sample sizes $K_{\text{out}}, K_{\text{in}}$ according to Remark 4.10. The total running time is ranged in 20–40 hours.

We observe that the running time of our algorithm is dominated by the inner loop of Algorithm 4.1, i.e., the part for optimizing over ψ_ν. The cost associated with this part can be estimated as $O(N \cdot M_{\text{out}} \cdot M_{\text{in}} \cdot (K_{\text{in}} t_a + t_b))$, where t_a denotes the time cost of using backpropagation to evaluate the gradient with respect to ν of each $|\nabla\psi_\nu(T_{\theta_0}(\mathbf{X}_k)) - (T_\theta(\mathbf{X}_k) - T_{\theta_0}(\mathbf{Y}_k))|^2$ in every inner loop of Algorithm 4.1, and t_b denotes the time for updating ν by the Adam method. Here t_a, t_b both depend on d, V and the sizes of networks T_θ, ψ_ν. According to our experiences, for most of the cases, t_a is of the order of magnitude around 10^{-5}s and t_b is around 10^{-2}s.

Although the cost for our current implementation of the train process is still high, we want to recall that there is a distinct advantage in the sampling application, namely, that the network training needs to be done only once. The trained network can be reused to generate samples, regardless of the sample size, from distribution ρ_t by pushing forward samples from the reference distribution p with negligible additional cost. This is in sharp contrast to the classical MCMC sampling techniques, which require one to solve the SDE associated with the Fokker–Planck equation by numerical methods, such as the Euler–Maruyama scheme, for every sample.

7. Discussion. In this paper, we design and analyze an algorithm for computing the high-dimensional Fokker–Planck equations. Our approach is based on transport information geometry with probability formulations stemming from deep learning generative models. We first introduce the parametric Fokker–Planck equations, a set of ODEs, to approximate the original Fokker–Planck equation. The ODE can be viewed as the "spatial" discretization of the PDE using neural networks. We propose a variational version of the semi-implicit Euler scheme and design a discrete time updating algorithm to compute the solution of the parametric Fokker–Planck equations. Our method is a sampling-based approach that is capable of handling high-dimensional cases. It can also be viewed as an alternative to the JKO scheme used in conjunction with neural networks. More importantly, we prove the asymptotic convergence and error estimates, both under the Wasserstein metric, for our proposed scheme.

We hope that our study may shed light on principally designing deep neural networks and other machine learning approaches to compute solutions of high-dimensional PDEs, and systematically analyzing their error bounds for understandable and trustworthy computations. Our parametric Fokker–Planck equations are derived by approximating the density function in free energy using neural networks, and then following the rules in calculus of variation to get its Euler–Lagrange equation. The energy law and principles in variational framework build a solid foundation for our "spatial" discretization that is able to inherit many desirable physical properties shared by the PDEs, such as relative entropy dissipation in a neural network setting. Our numerical scheme provides a systemic mechanism to design sampling efficient algorithms, which are critical for high-dimensional problems. One distinction of our method is that, contrary to the data-dependent machine learning studies in the literature, our approach does not require any knowledge of the "data" from the PDEs. In fact, we generate the "data" to compute the numerical solutions, just like the traditional numerical schemes do for PDEs. More importantly, we carried out the numerical analysis, using tools such as KL divergence and the Wasserstein metric from the transport information geometry, to study the asymptotic convergence and error estimates in probability space. We emphasize that the Wasserstein metric provides a suitable geometric structure to analyze the convergence behavior in generative models, which are widely used in machine learning. For this reason, we believe that our investigations can be adopted to understand many machine learning algorithms, and to design efficient sampling strategies based on pushforward maps that can generate flows of samples in generative models.

We also believe that the approaches in algorithm design and error analysis developed in this study can be extended to other equations, such as the porous media equation, the Schrödinger equation, the Schrödinger bridge system, and many more. These topics are worth further investigating in the future.

Appendix A. Proof of Lemma 3.3.

LEMMA 3.3. *Suppose $\vec{u}, \vec{v}$ are two vector fields defined on $\mathbb{R}^d$, and suppose φ, ψ solves $-\nabla\cdot(\rho\nabla\varphi) = -\nabla\cdot(\rho\vec{u})$ and $-\nabla\cdot(\rho\nabla\psi) = -\nabla\cdot(\rho\vec{v})$, or equivalently* $\mathrm{Proj}_\rho[\vec{u}] = \nabla\varphi$ *and* $\mathrm{Proj}_\rho[\vec{v}] = \nabla\psi$ *(cf. Definition* 4.2*). Then*

$$\int \vec{u}(x)\cdot\nabla\psi(x)\rho(x)\,dx = \int \nabla\varphi(x)\cdot\nabla\psi(x)\rho(x)\,dx, \tag{3.3}$$

$$\int |\nabla\psi(x)|^2\rho(x)\,dx \le \int |\vec{v}(x)|^2\rho(x)\,dx. \tag{3.4}$$

Proof of Lemma 3.3. For (3.3),

$$\int \vec{u}(x)\cdot\nabla\psi(x)\rho(x)\,dx = \int -\nabla\cdot(\rho(x)\vec{u}(x))\psi(x)\,dx = \int -\nabla\cdot(\rho(x)\nabla\varphi(x))\psi(x)\,dx$$
$$= \int \nabla\varphi(x)\cdot\nabla\psi(x)\rho(x)\,dx.$$

For (3.4),

$$\int |\vec{v}(x)|^2\rho(x)\,dx = \int (|\nabla\psi(x)|^2 + 2(\vec{v}(x)-\nabla\psi(x))\cdot\nabla\psi(x) + |\vec{v}(x)-\nabla\psi(x)|^2)\rho(x)\,dx$$
$$= \int (|\nabla\psi(x)|^2 + |\vec{v}(x)-\nabla\psi(x)|^2)\rho(x)\,dx \geq \int |\nabla\psi(x)|^2\rho(x)\,dx.$$

The second equality is due to (3.3). ∎

Appendix B. Proof of Theorem 3.7 .

THEOREM 3.7. *Suppose* $\{\theta_t\}_{t\geq 0}$ *solves* (3.18). *Then* $\{\rho_{\theta_t}\}$ *is the gradient flow of* $\mathcal{H}$ *on probability submanifold* $\mathcal{P}_\Theta$. *Furthermore, at any time* t, $\dot{\rho}_{\theta_t} = \frac{d}{dt}\rho_{\theta_t} \in \mathcal{T}_{\rho_{\theta_t}}\mathcal{P}_\Theta$ *is the orthogonal projection of* $-\mathrm{grad}_W\mathcal{H}(\rho_{\theta_t}) \in \mathcal{T}_{\rho_{\theta_t}}\mathcal{P}$ *onto the subspace* $\mathcal{T}_{\rho_{\theta_t}}\mathcal{P}_\Theta$ *with respect to the Wasserstein metric* g^W.

Theorem 3.7 easily follows from the following two general results about the manifold gradient.

THEOREM B.1. *Suppose* $(N, g^N), (M, g^M)$ *are Riemannian manifolds. Suppose* $\varphi : N \to M$ *is isometric. Consider* $\mathcal{F} \in \mathcal{C}^\infty(M)$, *and define* $F = \mathcal{F}\circ\varphi \in \mathcal{C}^\infty(N)$. *Suppose* $\{x_t\}_{t\geq 0}$ *is the gradient flow of* F *on* N*:*

$$\dot{x} = -\mathrm{grad}_N F(x).$$

Then $\{y_t = \varphi(x_t)\}_{t\geq 0}$ *is the gradient flow of* $\mathcal{F}$ *on* M. *That is,* $\{y_t\}$ *satisfies* $\dot{y} = -\mathrm{grad}_M\mathcal{F}(y)$.

Proof. Since we always have $\dot{y}_t = \varphi_*\dot{x}_t = -\varphi_*\mathrm{grad}_N F(x_t)$, we only need to show that $\varphi_*\mathrm{grad}_N F(x_t) = \mathrm{grad}_M\mathcal{F}(\varphi(x_t))$. Fix the time t, and consider any curve $\{\xi_\tau\}$ on N passing through x_t at $\tau = 0$; since φ is isometric, we have $g^N = \varphi^* g^M$, and thus

$$\frac{d}{d\tau}F(\xi_\tau)\Big|_{\tau=0} = g^N(\mathrm{grad}_N F(x_t), \dot{\xi}_0) = \varphi^* g^M(\mathrm{grad}_N F(x_t), \dot{\xi}_0)$$
$$= g^M(\varphi_*\mathrm{grad}_N F(x_t), \varphi_*\dot{\xi}_0).$$

On the other hand, denoting $\eta_\tau = \varphi(\xi_\tau)$, we have

$$\frac{d}{d\tau}F(\xi_\tau)\Big|_{\tau=0} = \frac{d}{d\tau}\mathcal{F}(\eta_\tau)\Big|_{\tau=0} = g^M(\mathrm{grad}_M\mathcal{F}(y_t), \dot{\eta}_0) = g^M(\mathrm{grad}_M\mathcal{F}(y_t), \varphi_*\dot{\xi}_0).$$

As a result, $g^M(\varphi_*\mathrm{grad}_N F(x_t) - \mathrm{grad}_M\mathcal{F}(y_t), \varphi_*\dot{\xi}_0) = 0$ for all $\dot{\xi}_0 \in T_{x_t}N$. Since φ_* is surjective, we have $\varphi_*\mathrm{grad}_N F(x_t) = \mathrm{grad}_M\mathcal{F}(\varphi(x_t))$. ∎

THEOREM B.2. *Suppose* (M, g^M) *is a Riemannian manifold, and* $M_{sub} \subset M$ *is the submanifold of* M. *Assume* M_{sub} *inherits metric* g^M, *i.e., define* $\iota : M_{sub} \to M$ *as the inclusion map, which induces a metric tensor on* M_{sub} *as* $g^{M_{sub}} = \iota^* g^M$. *For any* $\mathcal{F} \in \mathcal{C}^\infty(M)$, *denote the restriction of* $\mathcal{F}$ *on* M_{sub} *by* $\mathcal{F}^{sub}$. *Then the gradient* $\mathrm{grad}_{M_{sub}}\mathcal{F}^{sub}(x) \in T_x M_{sub}$ *is the orthogonal projection of* $\mathrm{grad}_M\mathcal{F}(x) \in T_x M$ *onto subspace* $T_x M_{sub}$ *with respect to the metric* g^M *for any* $x \in M_{sub}$.

Proof. For any $x \in M_{\text{sub}}$, consider any curve $\{\gamma_\tau\}$ on M_{sub} passing through x at $\tau = 0$. We have

$$\begin{aligned}\frac{d}{d\tau}\mathcal{F}^{\text{sub}}(\gamma_\tau)\Big|_{\tau=0} &= g^{M_{\text{sub}}}(\text{grad}_{M_{\text{sub}}}\mathcal{F}^{\text{sub}}(x), \dot{\gamma}_0) = g^M(\iota_* \text{grad}_{M_{\text{sub}}}\mathcal{F}^{\text{sub}}(x), \iota_*\dot{\gamma}_0)\\ &= g^M(\text{grad}_{M_{\text{sub}}}\mathcal{F}^{\text{sub}}(x), \dot{\gamma}_0).\end{aligned}$$

The last equality is because ι_* restricted on TM_{sub} is the identity. On the other hand, $\mathcal{F}^{\text{sub}}(\gamma_\tau) = \mathcal{F}(\gamma_\tau)$ for all τ. We also have

$$\frac{d}{d\tau}\mathcal{F}^{\text{sub}}(\gamma_\tau)\Big|_{\tau=0} = g^M(\text{grad}_M \mathcal{F}(x), \dot{\gamma}_0).$$

Combining them, we know

$$\begin{aligned} g^M(\text{grad}_{M_{\text{sub}}}\mathcal{F}^{\text{sub}}(x) - \text{grad}_M\mathcal{F}(x), v) = 0\\ \forall\, v \in T_x M_{\text{sub}} \Rightarrow \text{grad}_{M_{\text{sub}}}\mathcal{F}^{\text{sub}}(x) - \text{grad}_M\mathcal{F}(x) \perp_{g^M} T_x M_{\text{sub}},\end{aligned}$$

which proves this result. ∎

Proof of Theorem 3.7. To prove the first part of Theorem 3.7, we apply Theorem B.1 with $(N, g^N) = (\Theta, G)$, $M = \mathcal{P}_\Theta$ with its metric inherited from $(\mathcal{P}, g^W)$ and $\varphi = T_{(\cdot)\sharp}$. To prove the second part, we apply Theorem B.2 with $(M, g^M) = (\mathcal{P}, g^W)$, $M_{\text{sub}} = \mathcal{P}_\Theta$. ∎

Appendix C. Proof of Lemmas 4.6, 4.7, and 4.8.

LEMMA 4.6. *Suppose we fix $\theta_0 \in \Theta$; for arbitrary $\theta \in \Theta$ and $\nabla\phi \in L^2(\mathbb{R}^d; \mathbb{R}^d, \rho_{\theta_0})$ we consider*

(4.14)
$$F(\theta, \nabla\phi \mid \theta_0) = \left(\int (2\nabla\phi(x)\cdot(T_\theta - T_{\theta_0})\circ T_{\theta_0}^{-1}(x) - |\nabla\phi(x)|^2)\ \rho_{\theta_0}(x)\ dx\right) + 2hH(\theta).$$

Then $F(\theta, \nabla\phi \mid \theta_0) < \infty$, and furthermore $F(\cdot, \nabla\phi \mid \theta_0) \in C^1(\Theta)$. We can compute

$$\partial_\theta F(\theta, \nabla\phi \mid \theta_0) = 2\left(\int\ \partial_\theta T_\theta(T_{\theta_0}^{-1}(x))^T\ \nabla\phi(x)\ \rho_{\theta_0}(x)\ dx + h\ \nabla_\theta H(\theta)\right). \tag{4.15}$$

Proof. To show $F(\theta, \nabla\phi\ |\theta_0) < \infty$, we write

$$\begin{aligned}F(\theta, \nabla\ |\theta_0) = \underbrace{\int 2\nabla\phi\cdot T_\theta(T_{\theta_0}^{-1}(x))\rho_{\theta_0}\ dx}_{A} - \underbrace{\int 2\nabla\phi(T_{\theta_0}(x))\cdot x dp(x)}_{B}\\ - \underbrace{\int |\nabla\phi(x)|^2\rho_{\theta_0}(x)\ dx}_{C} + 2hH(\theta).\end{aligned}$$

By the Cauchy–Schwarz inequality, the first two terms can be estimated as

$$|A - B| \le 2\|\nabla\phi\|_{L^2(\rho_{\theta_0})}\left(\int |T_\theta(x)|^2 dp(x) + \int x^2 dp(x)\right).$$

Recalling (3.1) and p having finite second order moment, we know the first two terms are finite. In addition $C = \|\nabla\phi\|^2_{L^2(\rho_{\theta_0})} < \infty$. We thus have shown $F(\theta, \nabla\phi\ |\theta_0) < \infty$.

To show $F(\cdot, \nabla\phi \mid \theta_0) \in C^1(\Theta)$, recall $T_\theta(x) \in C^2(\Theta \times \mathbb{R}^d)$ as mentioned in section 3.1. We know the relative entropy $H(\cdot) \in C^1(\Theta)$, and thus we only need to prove for $\tilde{F}(\cdot, \nabla\phi \mid \theta_0) = F(\cdot, \nabla\phi \mid \theta_0) - 2hH(\theta)$. We consider $\xi \in \mathbb{R}^m$ with $|\xi|$ small enough and $\theta + \xi \in \Theta$. Then the difference is

$$\tilde{F}(\theta+\xi, \nabla\phi \mid \theta_0) - \tilde{F}(\theta, \nabla\phi \mid \theta_0) = \int 2\nabla\phi(x) \cdot (T_{\theta+\xi} - T_\theta) \circ T_{\theta_0}^{-1}(x)\ \rho_{\theta_0}(x)\ dx. \tag{C.1}$$

We denote the ith component of T_θ by $T_\theta^{(i)}$, $1 \le i \le d$. By Taylor expansion (with respect to θ), we have $T_{\theta+\xi}^{(i)}(x) - T_\theta^{(i)}(x) = \partial_\theta T_\theta^{(i)}(x)^{\mathrm{T}}\xi + \frac{1}{2}\xi^{\mathrm{T}}\partial_{\theta\theta}^2 T_{\theta+\lambda_i(x)\xi}(x)\xi$ with $\lambda_i(x) \in [0,1]$, and then the right-hand side of (C.1) is

$$\begin{aligned} &\underbrace{\left(\int 2\partial_\theta T_\theta(T_{\theta_0}^{-1}(x))^{\mathrm{T}}\nabla\phi(x)\rho_{\theta_0}\ dx\right)^{\mathrm{T}}\xi}_{\text{Denote by } \mathcal{J}(\theta)^{\mathrm{T}}\xi} \\ &+ \int \left(\sum_{i=1}^d \partial_{x_i}\phi \cdot (\xi^{\mathrm{T}}\partial_{\theta\theta}^2 T_{\theta+\lambda_i(x)\xi}^{(i)}(T_{\theta_0}^{-1}(x))\xi)\right)\rho_{\theta_0}\ dx. \end{aligned} \tag{C.2}$$

By the Cauchy–Schwarz inequality, the sum in the second term of (C.2) can be estimated as

$$\begin{aligned} &\left(\sum_{i=1}^d |\partial_{x_i}\phi|^2\right)^{\frac{1}{2}} \cdot \left(\sum_{i=1}^d |\xi^{\mathrm{T}}\partial_{\theta\theta}^2 T_{\theta+\lambda_i(x)\xi}^{(i)}(T_{\theta_0}^{-1}(x))\xi|^2\right)^{\frac{1}{2}} \\ &\le |\nabla\phi| \cdot \left(\sum_{i=1}^d \|\partial_{\theta\theta}^2 T_{\theta+\lambda_i(x)\xi}^{(i)}(T_{\theta_0}^{-1}(x))\|_2^2\right)^{\frac{1}{2}} |\xi|^2. \end{aligned}$$

Let us recall (4.13) and that the absolute value of the second term in (C.2) can be upper bounded by

$$\begin{aligned} &\left(\int |\nabla\phi|^2\rho_{\theta_0}\ dx\right)^{\frac{1}{2}} \cdot \left(\int \sum_{i=1}^d \|\partial_{\theta\theta}^2 T_{\theta+\lambda_i(x)\xi}^{(i)}(x)\|_2^2 dp(x)\right)^{\frac{1}{2}} |\xi|^2 \\ &\le \|\nabla\phi\|_{L^2(\rho_{\theta_0})}^2 \cdot \sqrt{H(\theta_0, |\xi|)}|\xi|^2. \end{aligned}$$

As a result, we have

$$\frac{|\tilde{F}(\theta+\xi, \nabla\phi \mid \theta_0) - \tilde{F}(\theta, \nabla \mid \theta_0) - \mathcal{J}(\theta)^{\mathrm{T}}\xi|}{|\xi|} \le \|\nabla\phi\|_{L^2(\rho_{\theta_0})}^2 \cdot \sqrt{H(\theta_0, |\xi|)}\ |\xi|. \tag{C.3}$$

Since $H(\theta_0, \epsilon)$ is increasing with respect to ϵ, when we send $|\xi| \to 0$, the upper bound in (C.3) approaches 0. This verifies the differentiability of $\tilde{F}(\cdot, \nabla\phi \mid \theta_0)$. Thus $F(\cdot, \nabla\phi \mid \theta_0)$ is also differentiable and $\partial_\theta F(\theta, \nabla\phi \mid \theta_0) = \mathcal{J}(\theta) + 2h\nabla_\theta H(\theta)$. At last, to show that $F(\cdot, \nabla\phi \mid \theta_0) \in C^1(\Theta)$, we only need to prove the continuity of $\mathcal{J}(\theta)$. One only need notice that

$$2\partial_\theta T_{\theta'}^{(i)}(T_{\theta_0}^{-1}(x))^{\mathrm{T}}\nabla\phi(x) \le |\partial_{\theta'} T_\theta^{(i)}(T_{\theta_0}^{-1}(x))|^2 + |\nabla\phi(x)|^2 \le L_2(T_{\theta_0}^{-1}(x)|\theta) + |\nabla\phi(x)|^2 \quad \forall\ \theta',\ |\theta' - \theta| < r(\theta).$$

The last inequality is due to condition (3.2). Since $L_2(T_{\theta_0}^{-1}(x)|\theta) + |\nabla\phi(x)|^2 \in L^1(\rho_{\theta_0})$, then by the dominated convergence theorem, we are able to prove the continuity of $\partial_\theta F(\theta, \nabla\phi \mid \theta_0)$. ∎

LEMMA 4.7. *Suppose we fix* $\theta_0 \in \Theta$ *and define*

$$J(\theta) = \sup_{\nabla\phi \in L^2(\mathbb{R}^d;\mathbb{R}^d,\rho_{\theta_0})} F(\theta, \nabla\phi \mid \theta_0).$$

Then J *is differentiable. If we denote* $\hat{\psi}_\theta = \operatorname{argmax}_\phi \{F(\theta, \nabla\phi \mid \theta_0)\}$, *then*

$$\nabla_\theta J(\theta) = \partial_\theta F(\theta, \nabla\hat{\psi}_\theta \mid \theta_0) = 2\left(\int \partial_\theta T_\theta(T_{\theta_0}^{-1}(x))^T \nabla\hat{\psi}_\theta(x)\ \rho_{\theta_0}(x)\ dx + h\ \nabla_\theta H(\theta)\right).$$

Proof. Let us denote $\Xi_\theta = (T_\theta - T_{\theta_0}) \circ T_{\theta_0}^{-1}$. Then for any $\xi \in \mathbb{R}^m$ such that $\theta + \xi \in \Theta$, we set $\hat{\psi}_{\theta+\xi} = \operatorname{argmax}_\phi\{F(\theta+\xi, \nabla\phi \mid \theta_0)\}$. Then according to Definition 4.2, $\hat{\psi}_\theta, \hat{\psi}_{\theta+\xi}$ solves

$$-\nabla\cdot(\rho_{\theta_0}\nabla\hat{\psi}_\theta) = -\nabla\cdot(\rho_{\theta_0}\Xi_\theta), \quad -\nabla\cdot(\rho_{\theta_0}\nabla\hat{\psi}_{\theta+\xi}) = -\nabla\cdot(\rho_{\theta_0}\Xi_{\theta+\xi}). \tag{C.4}$$

Subtracting the two equations, then multiplying $\hat{\psi}_{\theta+\xi} - \hat{\psi}_\theta$ on both sides and integrating yields

$$\int |\nabla\hat{\psi}_{\theta+\xi} - \nabla\hat{\psi}_\theta|^2 \rho_{\theta_0}\ dx = \int (\nabla\hat{\psi}_{\theta+\xi} - \nabla\hat{\psi}_\theta)\cdot(\Xi_{\theta+\xi} - \Xi_\theta)\rho_{\theta_0}\ dx.$$

Then by the Cauchy–Schwarz inequality, we derive

$$\int |\nabla\hat{\psi}_{\theta+\xi} - \nabla\hat{\psi}_\theta|^2 \rho_{\theta_0}\ dx \le \int |\Xi_{\theta+\xi} - \Xi_\theta|^2 \rho_{\theta_0}\ dx.$$

Now since $\Xi_{\theta_\xi}(x) - \Xi_\theta(x) = (T_{\theta+\xi} - T_\theta)\circ T_{\theta_0}^{-1}(x)$, by the mean value theorem, the ith component of $\Xi_{\theta+\xi}(x) - \Xi_\theta(x)$ can be written as $\partial_\theta T^{(i)}_{\theta+\lambda_i(x)\xi}(T_{\theta_0}^{-1}(x))^{\mathrm{T}}\xi$ with $\lambda_i(x) \in [0,1]$. Then, recalling the definition of $L(\theta,\epsilon)$ in (4.13), we can verify

$$\int |\Xi_{\theta+\xi} - \Xi_\theta|^2 \rho_{\theta_0}\ dx = \int |T_{\theta+\xi}(x) - T_\theta(x)| dp(x) \le L(\theta, |\xi|)|\xi|^2.$$

Thus we have the following estimation:

$$\int |\nabla\hat{\psi}_{\theta+\xi} - \nabla\hat{\psi}_\theta|^2 \rho_{\theta_0}\ dx \le L(\theta, |\xi|)|\xi|^2. \tag{C.5}$$

Now let us consider $J(\theta+\xi) - J(\theta)$:

$$\begin{aligned} J(\theta+\xi) - J(\theta) &= F(\theta+\xi, \nabla\hat{\psi}_{\theta+\xi} \mid \theta_0) - F(\theta, \nabla\hat{\psi}_\theta \mid \theta_0) \\ &= \underbrace{F(\theta+\xi, \nabla\hat{\psi}_{\theta+\xi} \mid \theta_0) - F(\theta, \nabla\hat{\psi}_{\theta+\xi} \mid \theta_0)}_{A} \\ &\quad + \underbrace{F(\theta, \nabla\hat{\psi}_{\theta+\xi} \mid \theta_0) - F(\theta, \nabla\hat{\psi}_\theta \mid \theta_0)}_{B}. \end{aligned} \tag{C.6}$$

Now according to Lemma 4.6, $F(\cdot, \nabla\phi \mid \theta_k) \in C^1(\Theta)$. By the mean value theorem,

term A can be written as

$$\begin{aligned}A &= F(\theta+\xi, \nabla\hat{\psi}_{\theta+\xi} \mid \theta_0) - F(\theta, \nabla\hat{\psi}_{\theta+\xi} \mid \theta_0) = \partial_\theta F(\theta+\tau\xi, \nabla\hat{\psi}_{\theta+\xi} \mid \theta_0)\xi \quad \text{with } \tau\in[0,1]\\ &= \partial_\theta F(\theta, \nabla\hat{\psi}_\theta \mid \theta_0)^{\mathrm{T}}\xi + \underbrace{(\partial_\theta F(\theta+\tau\xi, \nabla\hat{\psi}_\theta \mid \theta_0) - \partial_\theta F(\theta, \nabla\hat{\psi}_\theta \mid \theta_0))^{\mathrm{T}}}_{r_1(\theta,\xi)}\xi\\ &\quad + \underbrace{(\partial_\theta F(\theta+\tau\xi, \nabla\hat{\psi}_{\theta+\xi} \mid \theta_0) - \partial_\theta F(\theta+\tau\xi, \nabla\hat{\psi}_\theta \mid \theta_0))^{\mathrm{T}}}_{r_2(\theta,\xi)}\xi.\end{aligned}$$

Term B can be computed as

$$\begin{aligned}B &= F(\theta, \nabla\hat{\psi}_{\theta+\xi} \mid \theta_0) - F(\theta, \nabla\hat{\psi}_\theta \mid \theta_0)\\ &= \int (2(\nabla\hat{\psi}_{\theta+\xi} - \nabla\hat{\psi}_\theta)\cdot\Xi_\theta - (|\nabla\hat{\psi}_{\theta+\xi}|^2 - |\nabla\hat{\psi}_\theta|^2))\rho_{\theta_0}\, dx\\ &= 2\int (\nabla\hat{\psi}_{\theta+\xi} - \nabla\hat{\psi}_\theta)\cdot(\Xi_\theta - \nabla\hat{\psi}_\theta)\rho_{\theta_0}\, dx - \int |\nabla\hat{\psi}_{\theta+\xi} - \nabla\hat{\psi}_\theta|^2\rho_{\theta_0}\, dx\\ &= -\int |\nabla\hat{\psi}_{\theta+\xi} - \nabla\hat{\psi}_\theta|^2\rho_{\theta_0}\, dx.\end{aligned}$$

The last equality is due to integration by parts and (C.4).

Now substituting A and B in (C.6) yields

$$J(\theta+\xi) - J(\theta) = \partial_\theta F(\theta, \nabla\hat{\psi}_\theta \mid \theta_0) + r_1(\theta,\xi)^{\mathrm{T}}\xi + r_2(\theta,\xi)^{\mathrm{T}}\xi - \|\nabla\hat{\psi}_{\theta+\xi} - \nabla\hat{\psi}_\theta\|^2_{L^2(\rho_{\theta_0})}.$$

We can estimate

(C.7)
$$\frac{\left|J(\theta+\xi) - J(\theta) - \partial_\theta F(\theta, \nabla\hat{\psi}_\theta \mid \theta_0)^{\mathrm{T}}\xi\right|}{|\xi|} \le |r_1(\theta,\xi)| + |r_2(\theta,\xi)| + \frac{1}{|\xi|}\|\nabla\hat{\psi}_{\theta+\xi} - \nabla\hat{\psi}_\theta\|^2_{L^2(\rho_{\theta_0})}.$$

Now we prove the right-hand side of (C.7) approaches 0 as $\xi\to 0$. Since $\partial_\theta F(\cdot, \nabla\hat{\psi}_\theta \mid \theta_0) \in C^1(\Theta)$, using continuity, we know $\lim_{\xi\to 0} r_1(\theta,\xi) = 0$. For $r_2(\theta,\xi)$, when $|\xi|$ is sufficiently small, we have

$$\begin{aligned}|r_2(\theta,\xi)| &= \left|\int \partial_\theta T_{\theta+\tau\xi}(T_{\theta_0}^{-1}(x))^{\mathrm{T}}(\nabla\hat{\psi}_{\theta+\xi}(x) - \nabla\hat{\psi}_\theta(x))\rho_{\theta_0}(x)\, dx\right|\\ &\le \left(\int \|\partial_\theta T_{\theta+\tau\xi}(x)\|_F^2 dp(x)\right)^{\frac{1}{2}}\left(\int |\nabla\hat{\psi}_{\theta+\xi} - \nabla\hat{\psi}_\theta|^2\rho_{\theta_0}\, dx\right)^{\frac{1}{2}}\\ &\le \sqrt{\|L_2(\cdot|\theta)\|_{L^1(p)}}\sqrt{L(\theta,|\xi|)}|\xi|.\end{aligned}$$

The last inequality is due to (3.2) (when $|\xi|$ is small enough so that $|\xi| < r(\theta)$) and (C.5). Using this we are able to show $\lim_{\xi\to 0} r_2(\theta,\xi) = 0$. Using (C.5) again, we can verify $\frac{1}{|\xi|}\|\nabla\hat{\psi}_{\theta+\xi} - \nabla\hat{\psi}_\theta\|^2_{L^2(\rho_{\theta_0})} \le L(\theta,|\xi|)|\xi| \to 0$ as $\xi\to 0$. Thus J is differentiable at θ and we know $\nabla_\theta J(\theta) = \partial_\theta F(\theta, \nabla\hat{\psi}_\theta \mid \theta_0)$. We complete the proof by applying (4.15) of Lemma 4.6. ∎

LEMMA 4.8. *Under assumption* (4.11)*, the optimal solution of* (4.8) θ_{k+1} *satisfies*

$$|\theta_{k+1} - \theta_k| \sim o(1), \quad \textit{i.e.,} \quad \lim_{h\to 0^+} |\theta_{k+1} - \theta_k| = 0.$$

Proof of Lemma 4.8. Recall that the function to be minimized in (4.8) is $J(\theta) = \widehat{W}_2^2(\theta, \theta_k) + 2hH(\theta)$. If choosing $\theta = \theta_k$ in (4.8), we have $J(\theta_k) = 2hH(\theta_k)$. Thus $J(\theta_{k+1}) \le J(\theta_k) = 2hH(\theta_k)$. Since $H(\theta_k) \ge 0$, this leads to $\widehat{W}_2^2(\theta_{k+1}, \theta_k) \le 2hH(\theta_k)$. When h is small enough, $|\theta_{k+1} - \theta_k| \le l^{-1}(2hH(\theta_k))$, where l^{-1} is the inverse function of l defined on $[0, l(r_0)]$. We know $l^{-1}(0) = 0$ and l^{-1} is also a continuous and increasing function. This leads to $\lim_{h\to 0^+} |\theta_{k+1} - \theta_k| \le \lim_{h\to 0^+} l^{-1}(2hH(\theta_k)) = 0$. ∎

Appendix D. Proofs for Lemmas 5.7 and 5.8.

LEMMA 5.7. *The geodesic connecting $\rho_0, \rho_1 \in \mathcal{P}(M)$ is described by*

$$\begin{cases} \frac{\partial \rho_t}{\partial t} + \nabla \cdot (\rho_t \nabla \psi_t) = 0, \\ \frac{\partial \psi_t}{\partial t} + \frac{1}{2}|\nabla \psi_t|^2 = 0, \end{cases} \qquad \rho_t|_{t=0} = \rho_0, \ \rho_t|_{t=1} = \rho_1. \tag{5.14}$$

Using the notation $\dot{\rho}_t = \partial_t \rho_t = -\nabla \cdot (\rho_t \nabla \psi_t) \in \mathcal{T}_{\rho_t}\mathcal{P}(M)$, $g^W(\dot{\rho}_t, \dot{\rho}_t)$ is constant for $0 \le t \le 1$ and $g^W(\dot{\rho}_t, \dot{\rho}_t) = W_2^2(\rho_0, \rho_1)$ for $0 \le t \le 1$.

Proof. Recall definition (2.5) of Wasserstein metric g^W, $g^W(\dot{\rho}_t, \dot{\rho}_t) = \int |\nabla \psi_t|^2 \rho_t \, dx$. Since $\{\rho_t\}$ is the geodesic on $(\mathcal{P}(M), g^W)$, the speed $g^W(\sigma_t, \sigma_t)$ remains constant. To directly verify this, we compute the time derivative:

$$\frac{d}{dt} g^W(\dot{\rho}_t, \dot{\rho}_t) = \frac{d}{dt}\left(\int |\nabla \psi_t|^2 \rho_t \, dx \right) = \int \frac{\partial}{\partial t} |\nabla \psi_t|^2 \rho_t \, dx + \int |\nabla \psi_t|^2 \partial_t \rho_t \, dx.$$

Using the first equation in (5.14), we obtain

$$\int |\nabla \psi_t|^2 \partial_t \rho_t \, dx = \int |\nabla \psi_t|^2 \cdot (-\nabla \cdot (\rho_t \nabla \psi_t)) \, dx = \int \nabla(|\nabla \psi_t|^2) \cdot \nabla \psi_t \rho_t \, dx.$$

Taking the spatial gradient of the second equation in (5.14), we have

$$\partial_t (\nabla \psi_t) = -\nabla\left(\frac{1}{2} |\nabla \psi_t|^2 \right).$$

Then

$$\int \frac{\partial}{\partial t} |\nabla \psi_t|^2 \rho_t \, dx = \int 2 \partial_t (\nabla \psi_t) \cdot \nabla \psi_t \rho_t \, dx = \int -\nabla(|\nabla \psi_t|^2) \cdot \nabla \psi_t \rho_t \, dx.$$

Adding them together, we verify $\frac{d}{dt} g^W(\dot{\rho}_t, \dot{\rho}_t) = 0$; hence $\int_0^1 g^W(\dot{\rho}_t, \dot{\rho}_t) \, dt = W_2^2(\rho_0, \rho_1)$. Thus we know $g^W(\dot{\rho}_t, \dot{\rho}_t) = W_2^2(\rho_0, \rho_1)$ for any $0 \le t \le 1$. ∎

LEMMA 5.8. *Suppose $\{\rho_t\}$ solves (5.14), and the relative entropy $\mathcal{H}$ in (2.8) has potential V satisfying $\nabla^2 V \succeq \lambda I$. Then we have $\frac{d}{dt} g^W(\mathrm{grad}_W \mathcal{H}(\rho_t), \dot{\rho}_t) \ge \lambda W_2^2(\rho_0, \rho_1)$, or equivalently $\frac{d^2}{dt^2} \mathcal{H}(\rho_t) \ge \lambda W_2^2(\rho_0, \rho_1)$.*

Proof. Let us write

$$g^W(\mathrm{grad}_W \mathcal{H}(\rho_t), \dot{\rho}_t) = \int \nabla(V + D \log \rho_t) \cdot \nabla \psi_t \ \rho_t \, dx.$$

Then

$$\begin{aligned} \frac{d}{dt} g^W(\mathrm{grad}_W \mathcal{H}(\rho_t), \dot{\rho}_t) &= \frac{d}{dt}\left(\int \nabla(V + D \log \rho_t) \cdot \nabla \psi_t \ \rho_t \, dx \right) \\ &= \int (\nabla \psi_t^{\mathrm{T}} \nabla^2 V \nabla \psi_t + \mathrm{Tr}(\nabla^2 \psi_t \nabla^2 \psi_t)) \ \rho_t \, dx. \end{aligned}$$

The second equality can be carried out by direct calculations. One can check [67] or [68] for its complete derivation. Using $\nabla^2 V \succeq \lambda I$, we get

$$\frac{d}{dt} g^W(\mathrm{grad}_W \mathcal{H}(\rho_t), \dot{\rho}_t) \geq \int \lambda |\nabla \psi_t|^2 \rho_t \, dx = \lambda \, g^W(\dot{\rho}_t, \dot{\rho}_t) = \lambda W_2^2(\rho_0, \rho_1).$$

The last equality is due to Lemma 5.7. By the definition of Wasserstein gradient (2.7), we have $\frac{d}{dt}\mathcal{H}(\rho_t) = g^W(\mathrm{grad}_W \mathcal{H}(\rho_t), \dot{\rho}_t)$, and we also proved $\frac{d^2}{dt^2}\mathcal{H}(\rho_t) \geq \lambda W_2^2(\rho_0, \rho_1)$. ∎

REFERENCES

[1] S. N. Afriat, *Theory of maxima and the method of Lagrange*, SIAM J. Appl. Math., 20 (1971), pp. 343–357, https://doi.org/10.1137/0120037.

[2] S. Amari, *Natural gradient works efficiently in learning*, Neural Comput., 10 (1998), pp. 251–276.

[3] S. Amari, *Information Geometry and Its Applications*, Appl. Math. Sci. 194, Springer, Tokyo, 2016.

[4] L. Ambrosio, N. Gigli, and G. Savaré, *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*, Springer Science & Business Media, 2008.

[5] M. Arjovsky, S. Chintala, and L. Bottou, *Wasserstein generative adversarial networks*, in International Conference on Machine Learning, PMLR, 2017, pp. 214–223.

[6] N. Ay, J. Jost, H. V. Lê, and L. J. Schwachhöfer, *Information Geometry*, Ergeb. Math. Grenzgeb. (3) 64, Springer, Cham, 2017.

[7] D. Bakry and M. Émery, *Diffusions hypercontractives*, in Séminaire de Probabilités XIX 1983/84, Springer, 1985, pp. 177–206.

[8] A. Brock, J. Donahue, and K. Simonyan, *Large Scale GAN Training for High Fidelity Natural Image Synthesis*, preprint, https://arxiv.org/abs/1809.11096, 2018.

[9] D. A. C. Cabrera, P. Gonzalez-Casanova, C. Gout, L. H. Juárez, and L. R. Reséndiz, *Vector field approximation using radial basis functions*, J. Comput. Appl. Math., 240 (2013), pp. 163–173.

[10] J. A. Carrillo, Y.-P. Choi, and M. Hauray, *The derivation of swarming models: Mean-field limit and Wasserstein distances*, in Collective Dynamics from Bacteria to Crowds, Springer, 2014, pp. 1–46.

[11] J. A. Carrillo, K. Craig, and F. S. Patacchini, *A blob method for diffusion*, Calc. Var. Partial Differential Equations, 58 (2019), 53.

[12] J. A. Carrillo, K. Craig, L. Wang, and C. Wei, *Primal Dual Methods for Wasserstein Gradient Flows*, preprint, https://arxiv.org/abs/1901.08081, 2019.

[13] J. A. Carrillo, M. DiFrancesco, A. Figalli, T. Laurent, and D. Slepčev, *Global-in-time weak measure solutions and finite-time aggregation for nonlocal interaction equations*, Duke Math. J., 156 (2011), pp. 229–271.

[14] J. Chang and G. Cooper, *A practical difference scheme for Fokker-Planck equations*, J. Comput. Phys., 6 (1970), pp. 1–16.

[15] I. Daubechies, R. DeVore, S. Foucart, B. Hanin, and G. Petrova, *Nonlinear Approximation and (Deep) ReLU Networks*, preprint, https://arxiv.org/abs/1905.02199, 2019.

[16] J. L. Doob, *The Brownian movement and stochastic equations*, Ann. of Math. (2), 43 (1942), pp. 351–369.

[17] X. Glorot, A. Bordes, and Y. Bengio, *Deep sparse rectifier neural networks*, in Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, PMLR, 2011, pp. 315–323.

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, *Generative adversarial nets*, in Advances in Neural Information Processing Systems, Curran Associates, 2014, pp. 2672–2680.

[19] U. Grenander and M. I. Miller, *Representations of knowledge in complex systems*, J. Roy. Statist. Soc. Ser. B, 56 (1994), pp. 549–581.

[20] R. Holley and D. Stroock, *Logarithmic Sobolev inequalities and stochastic Ising models*, J. Statist. Phys., 46 (1987), pp. 1159–1194.

[21] P.-E. Jabin, *A review of the mean field limits for Vlasov equations*, Kinet. Relat. Models, 7 (2014), pp. 661–711.

[22] A. Jacot, F. Gabriel, and C. Hongler, *Neural tangent kernel: Convergence and generaliza-*

tion in neural networks, in Advances in Neural Information Processing Systems, Curran Associates, 2018, pp. 8571–8580.

[23] R. Jordan, D. Kinderlehrer, and F. Otto, *The variational formulation of the Fokker–Planck equation*, SIAM J. Math. Anal., 29 (1998), pp. 1–17, https://doi.org/10.1137/S0036141096303359.

[24] Y. Khoo, J. Lu, and L. Ying, *Solving Parametric PDE Problems with Artificial Neural Networks*, preprint, https://arxiv.org/abs/1707.03351, 2017.

[25] Y. Khoo, J. Lu, and L. Ying, *Solving for high-dimensional committor functions using artificial neural networks*, Res. Math. Sci., 6 (2019), 1.

[26] D. P. Kingma and J. Ba, *Adam: A Method for Stochastic Optimization*, preprint, https://arxiv.org/abs/1412.6980, 2014.

[27] A. Klar and S. Tiwari, *A multiscale meshfree method for macroscopic approximations of interacting particle systems*, Multiscale Model. Simul., 12 (2014), pp. 1167–1192, https://doi.org/10.1137/130945788.

[28] P. E. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*, Appl. Math. 23, Springer Science & Business Media, 2013.

[29] P. Kumar and S. Narayanan, *Solution of Fokker-Planck equation by finite element and finite difference methods for nonlinear systems*, Sadhana, 31 (2006), pp. 445–461.

[30] J. D. Lafferty, *The density manifold and configuration space quantization*, Trans. Amer. Math. Soc., 305 (1988), pp. 699–741.

[31] T. Lelièvre and G. Stoltz, *Partial differential equations and stochastic methods in molecular dynamics*, Acta Numer., 25 (2016), 681–880, https://doi.org/10.1017/S0962492916000039.

[32] A. J. Leverentz, C. M. Topaz, and A. J. Bernoff, *Asymptotic dynamics of attractive-repulsive swarms*, SIAM J. Appl. Dyn. Syst., 8 (2009), pp. 880–908, https://doi.org/10.1137/090749037.

[33] W. Li, *Geometry of Probability Simplex via Optimal Transport*, preprint, https://arxiv.org/abs/1803.06360, 2018.

[34] W. Li, A. T. Lin, and G. Montúfar, *Affine natural proximal learning*, in Geometric Science of Information, Lecture Notes in Comput. Sci. 11712, F. Nielsen and F. Barbaresco, eds., Springer, Cham, 2019, pp. 705–714.

[35] W. Li, S. Liu, H. Zha, and H. Zhou, *Parametric Fokker-Planck equation*, in Geometric Science of Information, Lecture Notes in Comput. Sci. 11712, F. Nielsen and F. Barbaresco, eds., Springer, Cham, 2019, pp. 715–724.

[36] W. Li and G. Montufar, *Natural Gradient via Optimal Transport*, preprint, https://arxiv.org/abs/1803.07033, 2018.

[37] W. Li and G. Montufar, *Ricci Curvature for Parametric Statistics via Optimal Transport*, preprint, https://arxiv.org/abs/1807.07095, 2018.

[38] A. T. Lin, W. Li, S. Osher, and G. Montufar, *Wasserstein proximal of GANs*, in Geometric Science of Information, Lecture Notes in Comput. Sci. 12829, F. Nielsen and F. Barbaresco, eds., Springer, Cham, 2021, pp. 524–533.

[39] Q. Liu and D. Wang, *Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm*, preprint, https://arxiv.org/abs/1608.04471, 2016.

[40] J. Lott, *Some geometric calculations on Wasserstein space*, Comm. Math. Phys., 277 (2008), pp. 423–437.

[41] D. Maoutsa, S. Reich, and M. Opper, *Interacting Particle Solutions of Fokker-Planck Equations through Gradient-Log-Density Estimation*, preprint, https://arxiv.org/abs/2006.00702, 2020.

[42] J. Martens and R. Grosse, *Optimizing neural networks with Kronecker-factored approximate curvature*, in International Conference on Machine Learning, JMLR, 2015, pp. 2408–2417.

[43] D. Masters and C. Luschi, *Revisiting Small Batch Training for Deep Neural Networks*, preprint, https://arxiv.org/abs/1804.07612, 2018.

[44] E. Nelson, *Quantum Fluctuations*, Princeton Series in Physics, Princeton University Press, Princeton, NJ, 1985.

[45] N. Nüsken and L. Richter, *Solving High-Dimensional Hamilton-Jacobi-Bellman PDEs Using Neural Networks: Perspectives from the Theory of Controlled Diffusions and Measures on Path Space*, preprint, https://arxiv.org/abs/2005.05409, 2020.

[46] F. Otto, *The geometry of dissipative evolution equations: The porous medium equation*, Commun. Partial Differential Equations, 26 (2001), pp. 101–174.

[47] F. Otto and C. Villani, *Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality*, J. Funct. Anal., 173 (2000), pp. 361–400.

[48] E. Pardoux and A. Y. Veretennikov, *On the Poisson equation and diffusion approximation.* I, Ann. Probab., 29 (2001), pp. 1061–1085.

[49] A. T. Patera, *A spectral element method for fluid dynamics: Laminar flow in a channel expansion*, J. Comput. Phys., 54 (1984), pp. 468–488.
[50] S. Pathiraja and S. Reich, *Discrete Gradients for Computational Bayesian Inference*, preprint, https://arxiv.org/abs/1903.00186, 2019.
[51] G. A. Pavliotis, *Stochastic Processes and Applications: Diffusion Processes, the Fokker-Planck and Langevin Equations*, Texts Appl. Math. 60, Springer, 2014.
[52] M. Pavon, E. G. Tabak, and G. Trigila, *The Data-Driven Schroedinger Bridge*, preprint, https://arxiv.org/abs/1806.01364, 2018.
[53] P. Petersen and F. Voigtlaender, *Optimal approximation of piecewise smooth functions using deep ReLU neural networks*, Neural Networks, 108 (2018), pp. 296–330.
[54] L. Pichler, A. Masud, and L. A. Bergman, *Numerical solution of the Fokker–Planck equation by finite difference and finite element methods—a comparative study*, in Computational Methods in Stochastic Dynamics, Springer, 2013, pp. 69–85.
[55] D. Qi and A. J. Majda, *Low-dimensional reduced-order models for statistical response and uncertainty quantification: Barotropic turbulence with topography*, Phys. D, 343 (2017), pp. 7–27.
[56] M. Raissi, P. Perdikaris, and G. E. Karniadakis, *Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations*, J. Comput. Phys., 378 (2019), pp. 686–707.
[57] S. Reich and S. Weissmann, *Fokker-Planck Particle Systems for Bayesian Inference: Computational Approaches*, preprint, https://arxiv.org/abs/1911.10832, 2020.
[58] D. J. Rezende and S. Mohamed, *Variational Inference with Normalizing Flows*, preprint, https://arxiv.org/abs/1505.05770, 2015.
[59] H. Risken, *The Fokker-Planck Equation*, Springer Ser. Synergetics 18, Springer, Berlin, Heidelberg, 1989.
[60] G. O. Roberts and R. L. Tweedie, *Exponential convergence of Langevin distributions and their discrete approximations*, Bernoulli, 2 (1996), pp. 341–363.
[61] H. Rosenbrock, *An automatic method for finding the greatest or least value of a function*, Computer J., 3 (1960), pp. 175–184.
[62] S. Ruder, *An Overview of Gradient Descent Optimization Algorithms*, preprint, https://arxiv.org/abs/1609.04747, 2016.
[63] T. Schlick, *Molecular Modeling and Simulation: An Interdisciplinary Guide*, Interdiscip. Appl. Math. 21, Springer Science & Business Media, 2010.
[64] J. Sirignano and K. Spiliopoulos, *Mean Field Analysis of Neural Networks*, preprint, https://arxiv.org/abs/1805.01053, 2018.
[65] S. Surjanovic and D. Bingham, *Virtual Library of Simulation Experiments: Test Functions and Datasets*, http://www.sfu.ca/~ssurjano (accessed 8 February 2020).
[66] B. Szabó and I. Babuška, *Finite Element Analysis*, John Wiley & Sons, 1991.
[67] C. Villani, *Topics in Optimal Transportation*, American Mathematical Society, 2003.
[68] C. Villani, *Optimal Transport: Old and New*, Grundlehren Math. Wiss. 338, Springer Science & Business Media, 2008.
[69] V. A. Volpert, *Elliptic Partial Differential Equations*, Vol. 1, Springer, 2011.
[70] E. Weinan, J. Han, and A. Jentzen, *Deep learning-based numerical methods for high-dimensional parabolic partial differential equations and backward stochastic differential equations*, Commun. Math. Statist., 5 (2017), pp. 349–380.
[71] M. Welling and Y. W. Teh, *Bayesian learning via stochastic gradient Langevin dynamics*, in Proceedings of the 28th International Conference on Machine Learning (ICML-11), Omnipress, 2011, pp. 681–688.
[72] D. Yarotsky, *Error bounds for approximations with deep ReLU networks*, Neural Networks, 94 (2017), pp. 103–114.
[73] Y. Zang, G. Bao, X. Ye, and H. Zhou, *Weak adversarial networks for high-dimensional partial differential equations*, J. Comput. Phys., 411 (2020), 109409.