1. Models and Path formulations.

1.1. Background. Let Ω be a smooth manifold without boundary. For simplicity, we assume $\Omega=\mathbb{R}^d$ throughout the present work, while generalization to general manifolds is straightforward. Denote the density function space defined on Ω as

$$\mathscr{P}(\Omega)=\left\{\rho:\Omega\to\mathbb{R}\colon \int\rho(x)dx=1,\ \rho(x)\geq 0,\ \int|x|^2\rho(x)\,dx<\infty\right\}. \tag{1.1}$$

Suppose $\mathscr{P}(\Omega)$ is equipped with the Wasserstein-2 distance:

$$\mathrm{d}_{\mathscr{W}}(\rho_1,\rho_2)=\left(\inf_{\pi\in\Pi(\rho_1,\rho_2)}\iint|x-y|^2d\pi(x,y)\right)^{1/2} \tag{1.2}$$

for any $\rho_1,\rho_2\in\mathscr{P}(\Omega)$, where $\Pi(\rho_1,\rho_2)$ is the joint density space with ρ_1 and ρ_2 as marginals. Then $\mathscr{P}(\Omega)$ becomes an infinite-dimensional Riemannian manifold $\mathscr{M}$, with $\mathrm{d}_{\mathscr{W}}$ inducing its Riemannian metric. The tangent space $\mathscr{T}_\rho\mathscr{M}$ is equipped with the negative Sobolev H_ρ^{-1} norm, defined as $\langle f_1,f_2\rangle_{-1,\rho}:=\int\nabla\psi_1(x)\cdot\nabla\psi_2(x)\,d\rho(x)$ for $f_1,f_2\in\mathscr{T}_\rho\mathscr{M}$ and ψ_1,ψ_2 satisfying $f_i=-\nabla\cdot(\rho\nabla\psi_i)$ with non-flux boundary condition $\rho(x)\nabla\psi_i(x)\to 0$ as $x\to\infty$.

Consider the free energy functional $\mathscr{F}$ defined on $\mathscr{P}(\Omega)$, As discovered by Otto, given an energy $\mathscr{F}:\mathscr{P}(\Omega)\to\mathbb{R}$, we may formally define its gradient with respect to the Wasserstein metric $\mathrm{d}_{\mathscr{W}}$ using the formula

$$\nabla_{\mathrm{d}_{\mathscr{W}}}\mathscr{F}(\rho)=-\nabla\cdot\left(\rho\nabla\frac{\delta\mathscr{F}}{\delta\rho}\right).$$

In this way, gradient flows of $\mathscr{F}$, $\partial_t\rho=-\nabla_{\mathrm{d}_{\mathscr{W}}}\mathscr{F}(\rho)$, correspond to solutions of the continuity equation with velocity $v=-\nabla\frac{\delta\mathscr{F}}{\delta\rho}$. In particular, the gradient flow of the energy

$$\mathscr{F}(\rho)=\int_\Omega\left[\beta^{-1}U(\rho(x))+V(x)\rho(x)\right]\mathrm{d}x+\frac{1}{2}\int_{\Omega\times\Omega}W(x-y)\rho(x)\rho(y)\,\mathrm{d}x\,\mathrm{d}y. \tag{1.3}$$

is

$$\begin{cases}\partial_t\rho=-\nabla\cdot(\rho v):=\nabla\cdot\left[\rho\nabla\left(\beta^{-1}U_m'(\rho)+V+W*\rho\right)\right],\\ \rho(x,0)=\rho_0(x),\end{cases}\qquad U_m(s)=\begin{cases}s\ln(s) & \text{for } m=1,\\ \frac{s^m}{m-1} & \text{for } m>1.\end{cases} \tag{1.4}$$

When $m=1$, the probabilistic interpretation of (1.4) is that the solution $\rho(t)$ describes the probability evolution of

$$dX_t=-\nabla V(X_t)\,dt-\mathbb{E}_{Y_t\sim X_t}\left[\nabla W(X_t,Y_t)\right]dt+\sqrt{2\beta^{-1}}\,dB_t, \tag{1.5}$$

where $Y_t\sim X_t$ means the independent and identical distribution as X_t. (1.5) can be viewed as the large N limit for N particles with mean field interaction:

$$dX_i=-\nabla V(X_i)\,dt-\frac{1}{N}\sum_{j=1}^N\nabla W(X_i-X_j)\,dt+\sqrt{2\beta^{-1}}\,dB_t^i. \tag{1.6}$$

As $N\to\infty$, the empirical distribution $\rho_t^N:=\frac{1}{N}\sum_{i=1}^N\delta_{X_i}$ converges to the distribution of X_t in (1.5) and the independence comes from the propagation of chaos.

A stationary measure ρ^* of the SDE (1.5) is a minimum state of the free energy $\mathscr{F}$, which satisfies

$$-\log\rho^*=\mathrm{const}+\beta\left(V(x)+\int W(x,y)\rho^*(dy)\right). \tag{1.7}$$

1.2. Paths and Variational Forms. For many W, at certain parameter regime of β, the stationary measure is not unique and the free energy $\mathscr{F}$ has multiple local minimizers in the probability space. This research proposal studies numerical algorithms related to rare events and metastability when the SDE exhibits multiple stationary measures. From the perspective of energy landscape, we are interested in the situation that $\mathscr{F}$ has multiple local minimum states in $\mathscr{P}$ and we aim to provide numerical tools to describe how the finite particle system make transitions between these equilibrium distributions.

Minimum Energy Path. Consider two local minimum ρ_a and ρ_b. We first regard the Minimum Energy Path (MEP) with the minimax principle in the Mountain Pass Theorem , which provides a mathematical justification for the existence of saddle point under certain conditions. Consider the following variational problem arising from the mountain pass theorem:

$$\min_{p\in AC_{a,b}} \|\mathscr{F}\circ p\|_\infty = \min_{p\in AC_{a,b}} \max_{0\le s\le 1} \mathscr{F}(p(s)). \tag{1.8}$$

where $AC_{\rho_a,\rho_b}([0,1];\mathscr{P}(\Omega))$ be the space of absolute continuous function $p(s)$ satisfying $p(0)=\rho_a$ and $p(1)=\rho_b$, and $AC_{a,b}$ is a short notation. This is a minimax problem and is very challenging to solve directly due to the maximum norm. In addition, the minimizing path in Equation (1.8) does not exhibit uniqueness. Any path that traverses the lowest saddle point between ρ_a and ρ_b, and subsequently descends in potential value when away from the saddle point, can be considered a minimizing path. This descent path does not need to strictly adhere to the steepest descent path. There is clearly an infinite number of such minimizing paths.

The minimum energy path in computational chemistry represents one particular type of minimizing path for the minimax problem (1.8), which, in addition, adheres to the gradient descent direction when moving away from the saddle point. The string method computes the MEP by evolving the gradient descent dynamics for every points on the path until the following condition is achieved:

$$\nabla_{\mathrm{d}_\mathscr{W}}\mathscr{F}(p^*)^\perp \equiv 0. \tag{1.9}$$

The notation "$\perp$" means the projection of the vector onto the normal hyperplane at the MEP p^*. In other words, the parallel condition for the tangent, $\nabla_{\mathrm{d}_\mathscr{W}}\mathscr{F}(p^*(s)) \parallel \partial_s p^*(s)$, holds.

Minimum Action Path. Consider two local minimum ρ_a and ρ_b. The random measure P_t^N for finite N in the infinite dimensional space $\mathscr{P}$ can transit from being near ρ_a to ρ_b in finite time. This type of collective transition refers to the distribution of all particles, instead of the trajectory of individual particles. The frequency of such transitions is exponentially small w.r.t. the number of particles and this probability can be well described by large deviation theory:

$$-N\log\mathbb{P}(\rho^N \approx p) \approx S_T[p] \quad \text{as } N\to\infty.$$

The Dawson-Gartner large deviation theory states that the rate function (aka action functional) for within a given time interval $[0,T]$ is given by:

$$S_{[0,T]}(p) = \int_0^T \|\dot p(t) + \nabla_{\mathrm{d}_\mathscr{W}}\mathscr{F}(p(t))\|^2_{-1,p(t)} dt.$$

for all absolute continuous paths $p\in AC_{\rho_a,\rho_b}([0,T];\mathscr{P}(\Omega))$. The geometric minimum action method further focuses on minimizing the following geometric action

$$\widehat{S}[p] := \int_0^1 \|\nabla_{\mathrm{d}_\mathscr{W}}\mathscr{F}(p(s))\|_{-1,p(s)}\|p'(s)\|_{-1,p(s)} + \left\langle \nabla_{\mathrm{d}_\mathscr{W}}\mathscr{F}(p(s)), p'(s)\right\rangle_{-1,p(s)} \mathrm{d}s \tag{1.10}$$

This manuscript is for review purposes only.

over all possible $p \in AC_{\rho_a,\rho_b}([0,1];\mathscr{P}(\Omega))$.

Notice that the second term

$$\int \left\langle \nabla_{\mathrm{d}_{\mathscr{W}}} \mathscr{F}(p(s)), p'(s) \right\rangle_{-1,p(s)} \mathrm{d}s = \int \frac{\mathrm{d}}{\mathrm{d}t} \mathscr{F}(p(s))\,\mathrm{d}s = \mathscr{F}(\rho_a) - \mathscr{F}(\rho_b)$$

is independent of the path. Then we only minimize the first term

$$I_g[p] := \int_0^1 \|\nabla_{\mathrm{d}_{\mathscr{W}}} \mathscr{F}(p(s))\|_{-1,p(s)} \|p'(s)\|_{-1,p(s)}\,\mathrm{d}s \tag{1.11}$$

where

$$\|\nabla_{\mathrm{d}_{\mathscr{W}}} \mathscr{F}(p(s))\|^2_{-1,p(s)} = \int \|\nabla \frac{\delta \mathscr{F}}{\delta \rho}\|^2\, p(s)(\mathrm{d}x)$$

Maximum Flux Path. Berkowitz and co-workers in 1983 showed the diffusive “flux” along a given dominant path is proportional to the line integral $\frac{1}{\int_L \exp(\beta \mathscr{F}(p))\,\mathrm{d}p}$, and thus designated the optimum reaction path as the minimum of the integral

$$\int_L \exp(\beta \mathscr{F}(p))\,\mathrm{d}p = \int_0^1 \exp(\beta \mathscr{F}(p(s))) \|p'(s)\|_{-1,p(s)}\,\mathrm{d}s. \tag{1.12}$$

The minimizer of (1.12) is referred to as “minimum resistance path” or “maximum flux path”.

We use the latter name here and call this minimizing path as “max-flux path” or “MFP” in short.

1.3. MFP approximates the MEP at low temperature. Based on the functional in (1.12) for the temperature-dependent max-flux path, we introduce

$$I_\beta[p] := \frac{1}{\beta} \log \int_0^1 \exp\left(\beta \mathscr{F}(p(s))\right) \|p'(s)\|_{-1,p(s)}\,\mathrm{d}s \tag{1.13}$$

Minimizing (1.12) is equivalent to minimizing the functional I_β. We note that (1.13) is the continuum version of the so-called “LogSumExp” function, a smooth approximation to the maximum function. This “log-sum-exp trick” is widely used in many machine learning algorithms, including the softmax function for classification. A similar form of (1.13) is also common in the literature of large deviation theories. And we have that at large β, $I_\beta[p] \approx \max\limits_{0\le s\le 1} \mathscr{F}(p(s)) + O(\beta^{-1})$. This is the point-wise convergence of $I_\beta \to I_\infty$ where

$$I_\infty[p] := \max_{0\le s\le 1} \mathscr{F}(p(s)) = \|\mathscr{F} \circ p\|_\infty.$$

It is obvious that $I_\beta[p] \le I_\infty[p]$. It holds that the original minimax problem (1.8) for the MEP is approximated by $\min_p \max_s \mathscr{F}(p) \approx \min_p I_\beta[p]$.

2. Numerical methods.

2.1. Neural network architecture. For any absolute continuous curve $p(s,\cdot)_{0\le s\le 1}$ in the Wasserstein space $\mathscr{M}$, there exists a vector field $v_s(x)_{0\le s\le 1}$ such that (p,v) jointly satisfies the continuity equation $\partial_s p + \nabla \cdot (pv) = 0$ and v has the minimal value of the norm (kinetic energy) $\|v_s\|_{L^2(\rho_s)} = \|\partial_s p\|_{-1,p(s)}$. The flow map defined by v: $\Phi(s) : X_0 \mapsto X_s$ where $dX_s/ds = v_s(X_s)$, then satisfies the push-forward condition of the curve p, i.e., $p(s) = \Phi(s)_{\#}p(0)$. Numerically, the flow map $\Phi(s)_{0\le s\le 1}$ can be chosen as a neural network $\Phi_{NN}(s)$.

We consider first parameterizing the flow map using a series of normalizing flows Ψ^i, which induces a sequence of density functions corresponding to $p(s_i)$. Here, s_i represents the grid points for numerical integration on the interval $[0,1]$. More precisely, we assume that

$$\Phi^n = \Psi^n \circ \Psi^{n-1} \circ \cdots \circ \Psi^1$$

then $p(s_n) = \Phi^n_{\#}p(0) = \Psi^n_{\#}p(s_{n-1})$

This manuscript is for review purposes only.

2.2. Loss function based on least squared tangent condition for minimum energy path. The minimax formulation for the MEP in (1.8) cannot be directly applied in computations due to the maximum norm. Alternatively, we consider to solve the first order condition (1.9) for the MEP. Let $\eta(s)$ denote the angle between the gradient $\nabla_{\mathrm{d}_{\mathscr{W}}}\mathscr{F}(p(s))$ and the tangent $\partial_s p(s)$. Equation (1.9) implies that $\eta(s) = 0$ or π, i.e., $\sin\eta(s) \equiv 0$. This condition can be addressed by the following minimization problem for the mean squared sine value, with an arbitrary weigh function $w(s) > 0$,

$$
\begin{aligned}
\ell_{\|}[\theta] &:= \int_0^1 (\sin\eta(s))^2 \, w(s)\mathrm{d}s \\
&= \int_0^1 \Big(1 - \frac{\big\langle \nabla_{\mathrm{d}_{\mathscr{W}}}\mathscr{F}(p_\theta(s)), \partial_s p_\theta(s)\big\rangle^2_{-1.p_\theta(s)}}{\|\nabla_{\mathrm{d}_{\mathscr{W}}}\mathscr{F}(p_\theta(s)\|^2_{-1.p_\theta(s)} \|\partial_s p_\theta(s))\|^2_{-1.p_\theta(s)}}\Big) w(s)\mathrm{d}s, \\
&\approx \frac{1}{M}\sum_{i=1}^{M} \Big(1 - \frac{\big\langle \nabla_{\mathrm{d}_{\mathscr{W}}}\mathscr{F}(p_\theta(s_i)), \partial_s p_\theta(s_i)\big\rangle^2_{-1.p_\theta(s_i)}}{\|\nabla_{\mathrm{d}_{\mathscr{W}}}\mathscr{F}(p_\theta(s_i))\|^2_{-1.p_\theta(s_i)} \|\partial_s p_\theta(s_i)\|^2_{-1.p_\theta(s_i)}}\Big) w(s_i), \qquad (2.1)
\end{aligned}
$$

For simplicity, we simply let w be constant. It is noted that this method only addresses the necessary condition (1.9). In addition, there is a numerical challenge near the critical value s where $\nabla_{\mathrm{d}_{\mathscr{W}}}\mathscr{F}(p(s)) = 0$.

2.3. Loss function for minimum action path in gradient systems. By the geometrical functional (1.11). Therefore, we consider the following loss function to find the mininum action path for the gradient system:

$$
\begin{aligned}
\ell_g(\theta) &:= \int_0^1 \|\nabla_{\mathrm{d}_{\mathscr{W}}}\mathscr{F}(p_\theta(s))\|_{-1,p_\theta(s)} \|\partial_s p_\theta(s)\|_{-1,p_\theta(s)} \,\mathrm{d}s \\
&\approx \frac{1}{M}\sum_{i=1}^{M} \|\nabla_{\mathrm{d}_{\mathscr{W}}}\mathscr{F}(p_\theta(s_i))\|_{-1,p_\theta(s_i)} \|\partial_s p_\theta(s_i)\|_{-1,p_\theta(s_i)}.
\end{aligned} \tag{2.2}
$$

2.4. Loss functions for max-flux path. The loss function for the max-flux path is from the variational problem (1.13):

$$
\begin{aligned}
\ell_\beta(\theta) &:= I_\beta[p_\theta] = \frac{1}{\beta}\log\int_0^1 \exp\left(\beta\mathscr{F}(p_\theta(s)\right) \|\partial_s p_\theta(s_i)\|_{-1,p_\theta(s_i)} \mathrm{d}s \\
&\approx \frac{1}{\beta}\log\left[\frac{1}{M}\sum_{i=1}^{M} \exp(\beta\mathscr{F}(p_\theta(s_i)))\|\partial_s p_\theta(s_i)\|_{-1,p_\theta(s_i)}\right].
\end{aligned} \tag{2.3}
$$

This optimization problem is usually quite efficient because, unlike other loss functions discussed later, only the potential function appears in this loss function – there is no ∇U term in ℓ_β. The parameter β here is the inverse temperature, characterizing the impact of the finite size of the noise. If one wants to approximate the MEP for the zero temperature limit, a large β is preferred. In practice, we find that the numerical MFP at a moderate β is a very good initial guess for the algorithms for the MEP.

2.5. Loss function for arc length parametrization. In any numerical computation of the path, it is very important to enforce the correct parametrization of the path geometrically. In the chain-of-state methods such as the string method and NEB, the neighboring configurations on the path should be equidistant so that the parameter s is the arc-length parameter. In our setting here, the arc-length parametrization is implemented in a natural way by the following penalty function

$$\ell_{arc}(\theta) \approx \frac{1}{M-1}\sum_{i=2}^{M}\left[\mathrm{d}_{\mathscr{W}}(p_\theta(s_i), p_\theta(s_{i-1})) - \mathrm{d}_{\mathscr{W}}(p_\theta(s_{i-1}), p_\theta(s_{i-2}))\right]^2. \tag{2.4}$$

where s_i are the grid points for numerical integration on $[0,1]$.

2.6. The sum of loss functions with weights tuning, and pre-training for MEP. To summarize, for the convenience of practical computation, we present a general form by summing each loss function together:

$$\min_\theta J(p_\theta) \tag{2.5}$$

where

$$J(p_\theta) := \alpha_1\ell_\beta + \alpha_2\ell_{arc} + \alpha_3\ell_{\|} + \alpha_4\ell_g. \tag{2.6}$$

By setting various values of weights α_i, we can achieve different goals.

In order to calculate the above four losses, we further examine their approximation. These four losses involve the following key components at i-th Image:

Free energy term

$$\mathscr{F}(p_i) = \mathscr{F}(p_\theta(s_i)) = \mathbb{E}_{x\sim p_i}\left[\beta^{-1}\frac{U(p_i(x))}{p_i(x)} + V(x) + \frac{1}{2}\mathbb{E}_{y\sim p_i}[W(x-y)]\right] \tag{2.7}$$

Inner Product term:

$$\left\langle \nabla_{\mathrm{d}_{\mathscr{W}}}\mathscr{F}(p_\theta(s_i)), \partial_s p_\theta(s_i)\right\rangle_{-1.p_\theta(s_i)} = \frac{\mathrm{d}}{\mathrm{d}t}\mathscr{F}(p_\theta(s_i)) \approx \frac{\mathscr{F}(p_i) - \mathscr{F}(p_{i-1})}{\Delta s} \tag{2.8}$$

Wasserstein Gradient Norm:

$$\|\nabla_{\mathrm{d}_{\mathscr{W}}}\mathscr{F}(p_\theta(s_i))\|^2_{-1,p_\theta(s_i)} = \mathbb{E}_{x\sim p_i}\left[\|\nabla(\frac{\delta}{\delta\rho}\mathscr{F}(p_i))\|^2\right] \tag{2.9}$$

Time derivative Norm:

$$\|\partial_s p_\theta(s_i)\|^2_{-1,p_\theta(s_i)} \approx \frac{\mathrm{d}^2_{\mathscr{W}}(p_i, p_{i-1})}{(\Delta s)^2}$$

where p_i is a short notation for $p_\theta(s_i)$, and $x = \Phi^i(z)$ follows the distribution p_i by definition. Moreover, by the Monge formulation of the Wasserstein-2 distance between p_i and p_{i-1} as $\mathrm{d}^2_{\mathscr{W}}(p_i, p_{i-1}) = \min_{T:T_\# p_{i-1}=p_i} \mathbb{E}_{x\sim p_{i-1}}\|x - T(x)\|^2$, solving for the discrete density path $p = \{p_i\}_{i=1}^M$ is equivalent to solving for the transport $\Phi = \{\Phi^i\}_{i-1}^M$ by making the following substitutions during the optimization process:

$$\|\partial_s p_\theta(s_i)\|^2_{-1,p_\theta(s_i)} \approx \frac{\mathrm{d}^2_{\mathscr{W}}(p_i, p_{i-1})}{(\Delta s)^2} = \frac{\mathbb{E}_{x\sim p_{i-1}}\|x - \Psi^i(x)\|^2}{(\Delta s)^2} \tag{2.10}$$

The equivalence is proved in following theorem:

THEOREM 2.1. *If $L \geq 0$, then the following two optimization problems*

$$\min_p L_p[p] = \sum_{i=2}^{M} L(p_i) * \mathrm{d}_{\mathscr{W}}(p_i, p_{i-1}) \tag{2.11}$$

$$\tag{2.12}\begin{aligned}\min_{\Phi} L_T[\phi] &= \sum_{i=2}^{M} L(\Phi^i_{\#} p_0) * (\mathbb{E}_{z\sim p_0}\|\Phi^i(z) - \Phi^{i-1}(z)\|^2)^{1/2}\\ &= \sum_{i=2}^{M} L(\Phi^i_{\#} p_0) * (\mathbb{E}_{z\sim \Phi^{i-1}_{\#} p_0}\|\Psi^i(z) - z\|^2)^{1/2}\end{aligned}$$

have the same minimum and

(a) If Φ^* is a minimizer of (2.12), then $p^* = (\Phi^*)_{\#} p_0$ is a minimizer of (2.11).

(b) If p^* is a minimizer of (2.11), then the optimal transport Ψ^i from p^*_{i-1} to p^*_i minimizes (2.12).

Proof. Let the minimum of (2.12) be L^*_T, and that of (2.11) be L^*_p.

Proof of (a): Suppose L_T achieves minimum at Φ^*, and it induced density path $p^* = (\Phi^*)_{\#} p_0$. Then Ψ^{i^*} is the optimal transport from p^*_{i-1} to p^*_i because otherwise L_T can be further improved. By definition of L_p, we have

$$L^*_T = L_T[T^*] = L_p[p^*] \geq L^*_p.$$

We claim that $L^*_T = L^*_p$, otherwise, there is another p' such that

$$L^*_p = L_p[p'] < L^*_T.$$

Let $\Psi^{i'}$ be the optimal transport from p'_{i-1} to p'_i, and then $L_T[\Phi'] = L_p[p'] < L^*_T$, contradicting with that L^*_T is the minimum of L_T. This also shows that $L_p[p^*] = L^*_T = L^*_p$, that is, p^* is a minimizer of L_p.

Proof of (b): Suppose L_p achieves minimum at p^*. Let $(\Psi^i)^*$ be the OT from p^*_{i-1} to p^*_i, then

$$\mathbb{E}_{z\sim p_0}\|\Phi^i(z) - \Phi^{i-1}(z)\|^2 = \mathbb{E}_{z\sim \Phi^{i-1}_{\#} p_0}\|\Psi^i(z) - z\|^2 = \mathrm{d}^2_{\mathscr{W}}(p^*_{i-1}, p^*_i),$$

and then $L_T[T^*] = L_p[\rho^*] = L^*_p$ which equals L^*_T since $L^*_T \geq L^*_p$ as proved in (a). This shows that Ψ^* is a minimizer of L_T. □

Algorithm 2.1 Generative String Method

```
Inputs: Initial flow map $\Phi^M(\theta)$, the base distribution $\rho_0$; the hyper-parameters
for $n = 1 : N_I$ do
    Sample $\{z_i\}_{i=1}^N$from $\rho_0$;
    for $m = 1 : M$ do
        Get samples in $m$-th image $\{x_i^m\}_{i=1}^N = \{\Phi^m(z_i)\}_{i=1}^N$ and calculate the corresponding
        density $p_\theta(s_m)$;
        Compute the loss function by (2.7), (2.8), (2.9) and (2.10)
    end for
    Update the parameters $\theta$ using the Adam optimizer;
end for
return The optimal transport map $\Phi^M(\theta)$
```

3. Numerical Tests.

In this section, we explore the numerical performance by testing the Generative String NET over several numerical examples. We use RealNVP to parameterize the generator Φ in all the examples. For each affine coupling layer in Real NVP, we use the fully connected neural networks with five hidden layers and the LeakyReLU as the activation function to parameterize the translation and scaling functions.

This manuscript is for review purposes only.

3.1. Wasserstein gradient flow. Based on the definition of the minimum energy path, it is understood that the trajectory of any point on the path to its corresponding local minimum must be a Wasserstein gradient flow. Let us first consider such a problem to verify the efficacy of our model.

In addition to using the cosine values corresponding to each image to assess the quality of this path, we can also compare the action loss and the free energy difference, because in the optimal case:

$$
\begin{aligned}
\ell_g(\theta) &:= \int_0^1 \|\nabla_{\mathrm{d}_{\mathscr{W}}} \mathscr{F}(p_\theta(s))\|_{-1,p_\theta(s)} \|\partial_s p_\theta(s)\|_{-1,p_\theta(s)} \,\mathrm{d}s \\
&= \int \left\langle \nabla_{\mathrm{d}_{\mathscr{W}}} \mathscr{F}(p(s)), p'(s) \right\rangle_{-1,p(s)} \mathrm{d}s \\
&= \int \frac{\mathrm{d}}{\mathrm{d}t} \mathscr{F}(p(s))\,\mathrm{d}s = \mathscr{F}(\rho_a) - \mathscr{F}(\rho_b)
\end{aligned}
$$

3.1.1. Fokker-planck equation.

- Initial distribution: standard normal distribution ρ_0
- Free energy:

$$
\mathscr{F}(\rho) = \int_\Omega \left[\beta^{-1} \log \rho(x) + V(x)\right] \rho(x)\,\mathrm{d}x \tag{3.1}
$$

  with $\beta = 0.25$ and $V(x) = (x_1 - 1)^2 + (x_2 - 1)^2$
- At first we trained with only terminal loss $\mathscr{F}(\Phi_\#^M \rho_0)$: it can learn the invariant distribution well, but in the absence of the transport cost, one cannot expect existing NF models to approximate the OT trajectory see Fig.1. Besides, the action loss is 11.789223 while the free energy difference is 2.7291.
- Then we added the ℓ_g, the results are shown in Fig.2. Besides, the action loss is 2.755086 while the free energy difference is 2.7260. It can be seen that the cosine values of the latter images are poor. We can examine the velocity fields corresponding to the latter layers, as shown in Fig.3. The mappings of these layers resemble rotations, which are not the maps corresponding to optimal transport.
- Afterwards, we added a penalty term for the total length: $\log\left[\frac{1}{M}\sum_{i=1}^M \|\partial_s p_\theta(s_i)\|_{-1,p_\theta(s_i)}\right]$, and the results are shown in Fig.4. The velocity field is shown in Fig.5. Besides, the action loss is 2.494124 while the free energy difference is 2.4892.

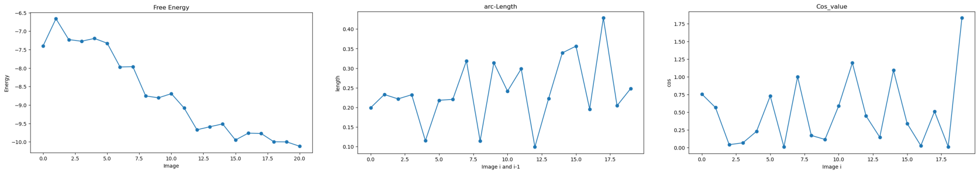

FIG. 1.

3.1.2. Mixture Gaussian. From the experiment just now, we found that the 'total length' also constrains the entire trajectory, which is intuitive. Next, we will continue to validate this point with a data-driven example.

- We set the initial density to be $\rho_0(x) = N(x; 0, 0.3I)$ and the terminal density to be $\rho_1(x) = \frac{1}{8}\sum_{i=1}^8 N(x; \mu_i, 0.3I)$, where $\mu_i = 4\cos\left(\frac{\pi i}{4}\right) e_1 + 4\sin\left(\frac{\pi i}{4}\right) e_2$, $i = 1, \ldots, 8$, and e_1, e_2 are the first two standard basis vectors, see Fig.6.

This manuscript is for review purposes only.

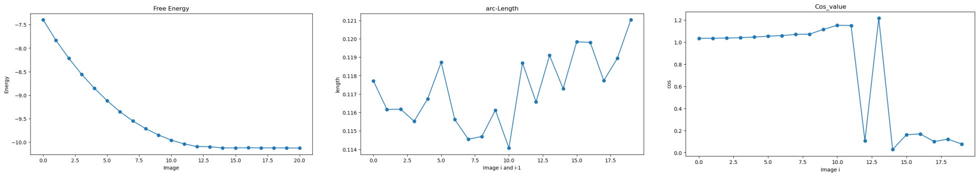

FIG. 2.

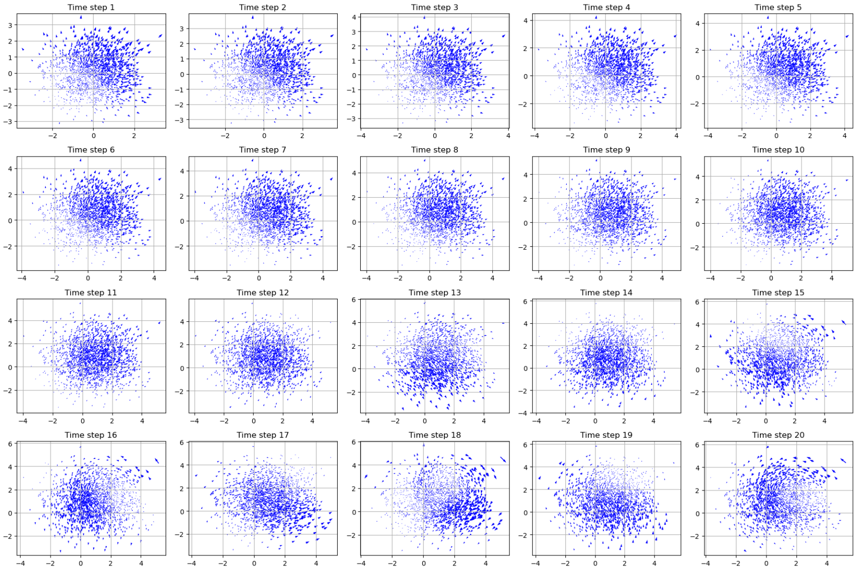

FIG. 3.

- In this example, there is no concept of energy, and the loss function is the Kullback-Leibler divergence between the distribution of the last layer $\Phi_{\#}^{M}\rho_0$ and the data distribution. The results without and with the addition of the total length penalty are shown in Fig.7 and Fig.8, respectively. It can be observed that this term indeed helps the trajectory to converge towards an optimum.

3.1.3. Aggregation-drift equation. Next, we compute solutions of aggregation-drift equations

$$\partial_t \rho = \nabla \cdot (\rho \nabla W * \rho) + \nabla \cdot (\rho \nabla V),$$

where $W(x) = |x|^2/2 - \ln(|x|)$ and $V(x) = -\frac{\alpha}{\beta}\ln(|x|)$. As shown in several analytical and numerical results, the steady state is a characteristic function on a torus or "milling profile", with inner and outer radius given by

$$R_i = \sqrt{\frac{\alpha}{\beta}}, \quad R_o = \sqrt{\frac{\alpha}{\beta}+1}\,.$$

we simulate the long time behavior of a solution of the aggregation-drift equation with $\alpha = 1$ and $\beta = 4$ and Gaussian initial data $\rho_0(x) = N(x; 0, 0.25)$.

- Initially, we only included the terminal cost. For this problem, we use a loss that combines physical information and data-driven elements. This includes both the

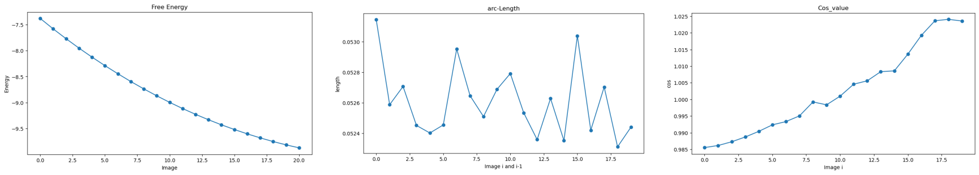

FIG. 4.

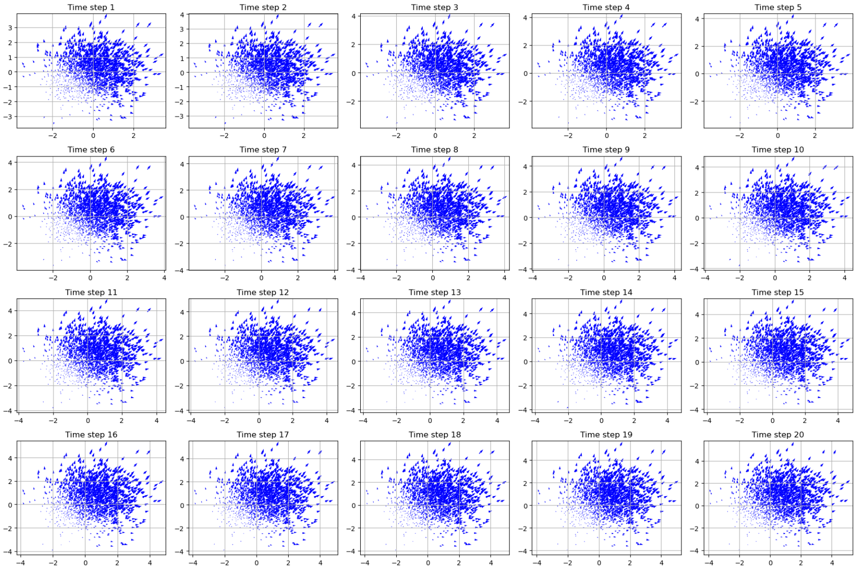

FIG. 5.

free energy and the Kullback-Leibler divergence. The results are shown in Fig.9 and Fig.10. Besides, the action loss is 331.29456 while the free energy difference is 0.1987.

- The experimental results of adding ℓ_g and $\ell_g + \ell_{len}$ are very similar, as shown in Figures 11, 12, 13, and 14. However, the action 1.226353 is still much larger than the free energy difference 0.1956 at this point.

REFERENCES

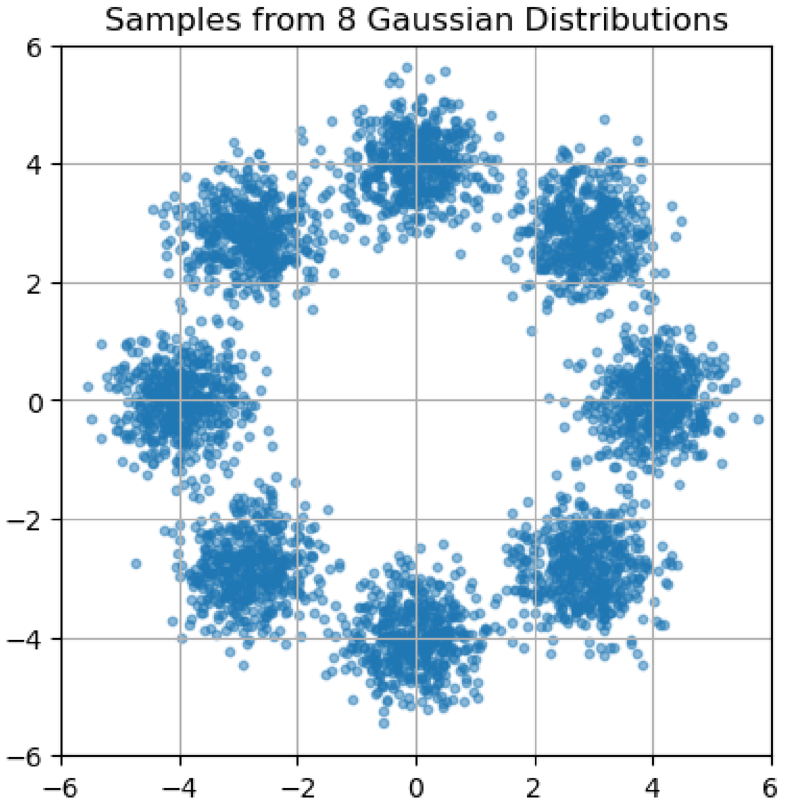
Samples from 8 Gaussian Distributions
6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6

FIG. 6.

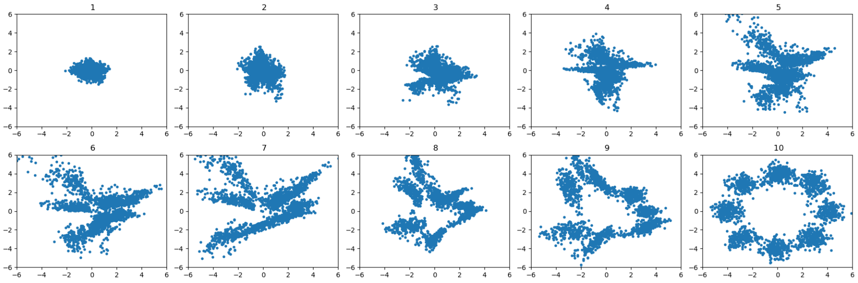
1
2
3
4
5
6
7
8
9
10
6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6

FIG. 7.

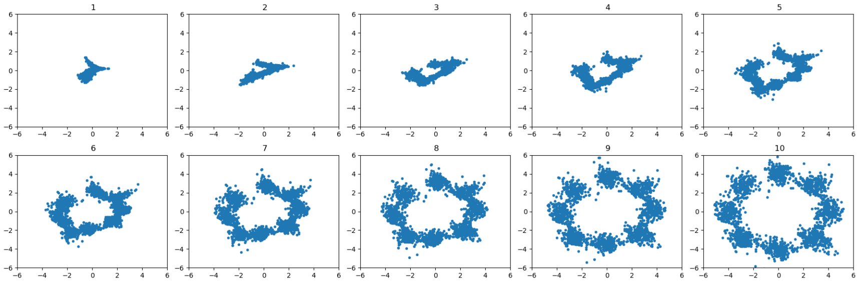

FIG. 8.

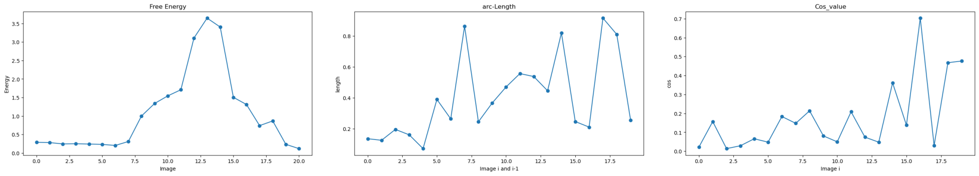

FIG. 9.

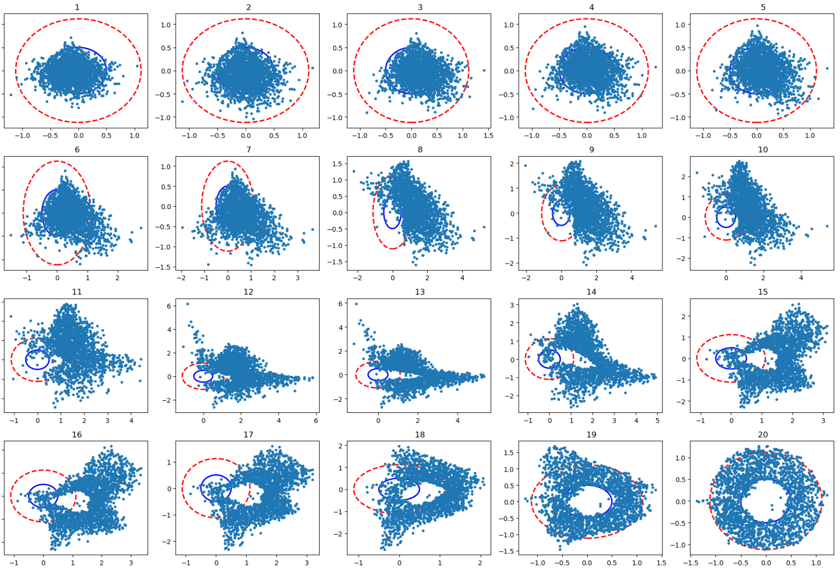

FIG. 10.

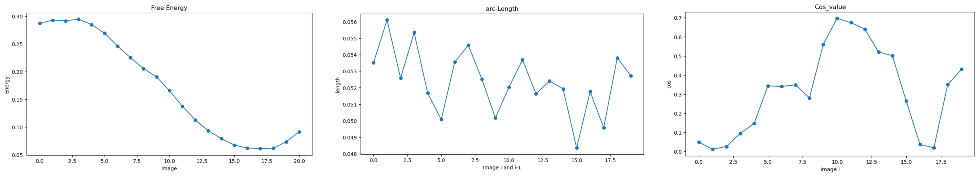

FIG. 11.

This manuscript is for review purposes only.

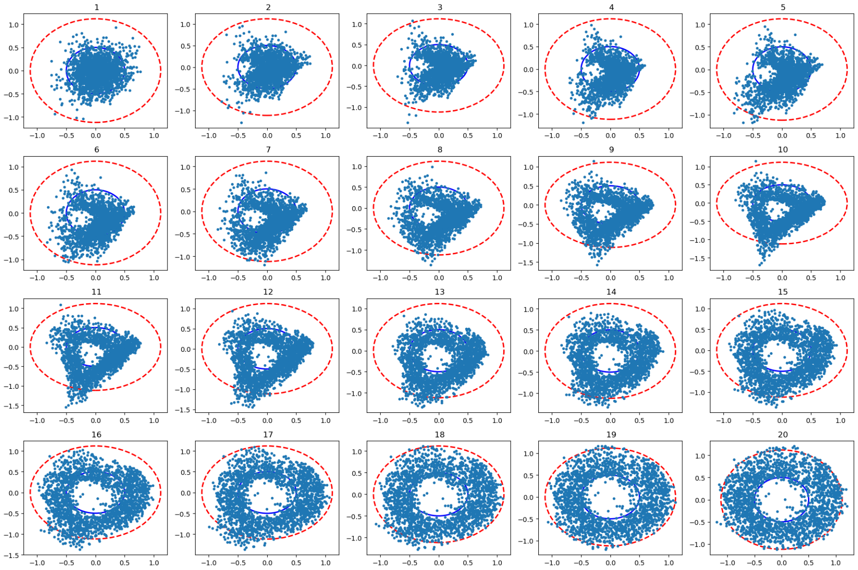

FIG. 12.

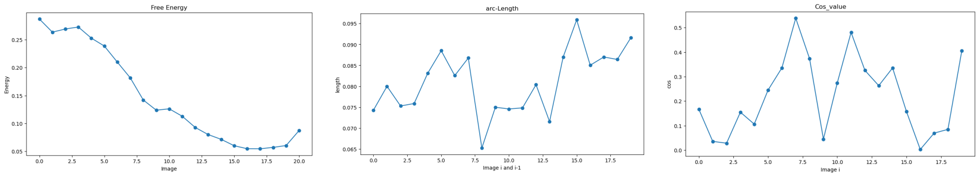

FIG. 13.

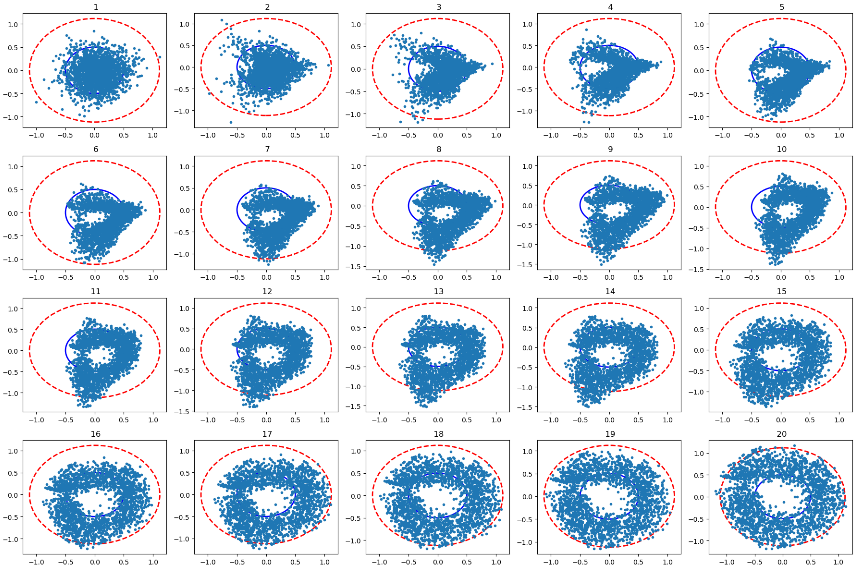

FIG. 14.

This manuscript is for review purposes only.