Notes for Unbalanced Optimal Transport Flow

Jiaheng Chen

February 15, 2022

1 Problems and some ideas

1.1 About KSD

Proposition 3.3 in [9] says that, if we assume $k(x, y)$ is integrally strictly positive definite, and p, q are **continuous densities** with $\|p(x)(s_q(x) - s_p(x))\|_2^2 < \infty$, we have $S(p, q) \geq 0$ and $S(p, q) = 0$ if and only if $p = q$.

For 2d toy models in [11], the density function are not continuous, which leads to that $S(p, q)$ cannot detect the non-convergence for these examples. In particular, if we take some fixed part (not all the support) of standard normal distribution and rescale it as $p(x)$, standard normal distribution as $q(x)$, then $S(p, q) = 0$ in (29), but obviously $p \neq q$. Therefore, when discontinuous $\rho_0(x)$ is transported by flow, since our network is Lipschitz in some sense, it tends to be transported to some part of standard normal distribution but not all and get trapped there. Although KSD value is small, the mapping we obtain in this way is unsatisfactory. Following examples in 2d illustrate this problem clearly.

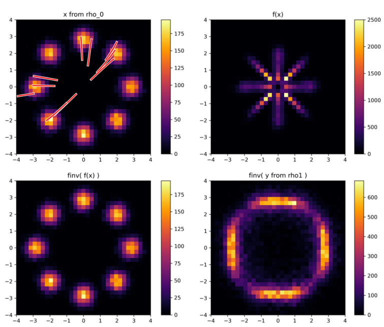

Figure 1: 8gaussians

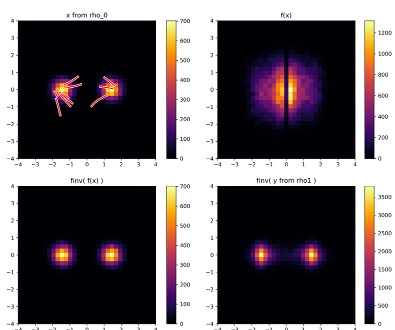

Figure 2: 2gaussians

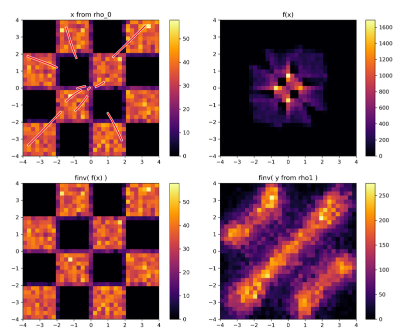

Figure 3: checkerboard

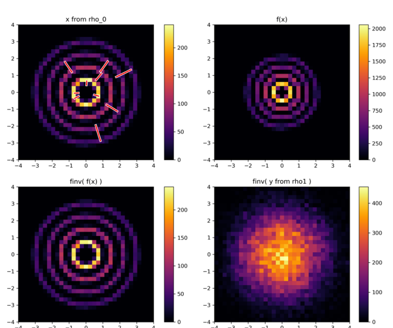

Figure 4: rings

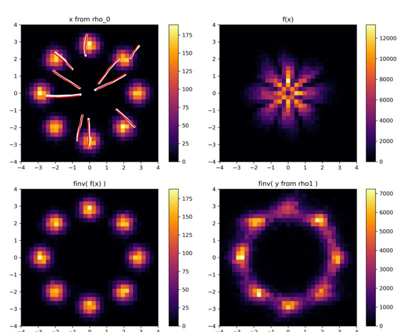

Figure 5: 8gaussians noise

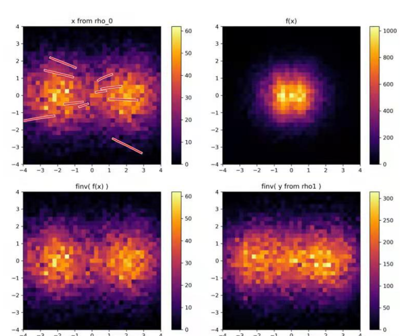

Figure 6: 2gaussians near

From the Figure 4, we should enforce the module of final paticles to obey $\chi^2(d)$ distribution

($d \geq 2$). Applying $q = \chi^2(d)$ to KSD:

$$s_q(x) = \frac{d-2}{2x} - \frac{1}{2} \tag{1}$$

From the "8gaussians" example, we know that we should enforce the final distribution to be isotropic. To overcome the obstacle of discontinuity, another possible method is to add noise to the flow trajectory as "diffusion" at the cost of some invertibility.

Another possible angle is that, enforcing the mutual information among different dimensions to be very small to achieve standard normal distribution when we have a discontinuous density distribution.

1.2 Some ideas about term C in (43)

Let $\rho_T(x) = \rho_0(x)e^{-\frac{1}{\alpha}\int_0^T(\Phi(z(x,s))-\bar{\Phi}(s))ds}$, it is to see ρ_T is probability density dunction, and then $\int_{\mathbb{R}^d}\log(\rho_0(x))\rho_0(x)e^{-\frac{1}{\alpha}\int_0^T(\Phi(z(x,s))-\bar{\Phi}(s))ds}dx$ is actually the cross entropy between ρ_0 and ρ_T. However, we want to raise the following point. In fact, the support of ρ_T is just the same as ρ_0, which is especially important when ρ_0 is discontinuous. The ρ_T can be viewed as a scale version on the support of the initial distribution ρ_0. Furthermore, $\int_{\mathbb{R}^d}\log(\rho_T(x))\rho_T(x)dx$ is the negative entropy of ρ_T, which appears in the KL divergence in (14). If Φ is zero, this term equals to $\int_{\mathbb{R}^d}\log(\rho_0(x))\rho_0(x)dx$, a unknown constant needs no consideration during our training. On the effect of minimize $\int_{\mathbb{R}^d}\log(\rho_T(x))\rho_T(x)dx$, which equals to maximize $-\int_{\mathbb{R}^d}\log(\rho_T(x))\rho_T(x)dx$. It's clear to see that, when $e^{-\frac{1}{\alpha}\int_0^T(\Phi(z(x,s))-\bar{\Phi}(s))ds} = \frac{1}{\rho_0(x)\cdot\text{measure}(\text{supp of}\rho_0)}$, the entropy of ρ_T takes its maximum.

1.3 Mixed initial distribution

Another point of view, if we consider adding some noise to initial distribution ρ_0, which makes it become $(1-\varepsilon)\rho_0(x) + \varepsilon\rho_1(x), 0 \leq \varepsilon < 1$, the negative entropy of it:

$$S_\varepsilon = \int_{\mathbb{R}^d}\log((1-\varepsilon)\rho_0(x)+\varepsilon\rho_1(x))((1-\varepsilon)\rho_0(x)+\varepsilon\rho_1(x))dx \tag{2}$$

By the Taylor's expansion, one can get this formula

$$S_\varepsilon = (1-\epsilon)S_0 + \varepsilon\int_{\mathbb{R}^d}\log(\rho_0(x))\rho_1(x)dx + O(\varepsilon^2) \tag{3}$$

After adding some noise to ρ_0, the support of $(1-\varepsilon)\rho_0(x)+\varepsilon\rho_1(x)$ becomes the whole space, which may help our training process.

1.4 Smoothing initial distribution

If we can take a transformation which is invertible, then we can first transform ρ_0 to a (continuous) distribution, then transport this smooth one to the standard normal distribution (in this way the KSD measure will be more effective). For the inverse generation process, we first sample from standard gaussian and then transport back through the flow, and finally transform it (inversely) to the original ρ_0. The most important thing here is that the transformation is invertible first, and it can play a role as smoothing the discontinuous distribution over the whole space.

Goldfeld, Z. and Greenewald, K. studies Smooth Wasserstein Distance[1] [2][3]. [10] investigates the structural and statistical behavior of the Gaussian-smoothed p-Wasserstein distance.

1.5 FSSD

[7] proposed a linear-time kernel goodness-of-fit test called Finite Set Stein Discrepancy (FSSD), which uses a set of random vectors $\{v_i\}_{i=1}^n$ in the domain χ to evaluate the stein witness function. On the one hand, the computational cost will be reduced to $O(n)$, and more importantly, the key idea of FSSD may help us detect the discontinuous density function with the help of $\{v_i\}_{i=1}^n$.

1.6 Some other ideas

One problem is that approximate $\nabla \log(\rho)$, maybe we can bring some methods from Pseudo differential operator. On the other hand, to achieve the target standard normal distribution, maybe we can control the module distribution $\chi^2(d)$ and a spherical uniform distribution. The latter one is studied also over a various fields, include using a complete basis of symmetric polynomials, which is also used in interatomic potentials simulations...

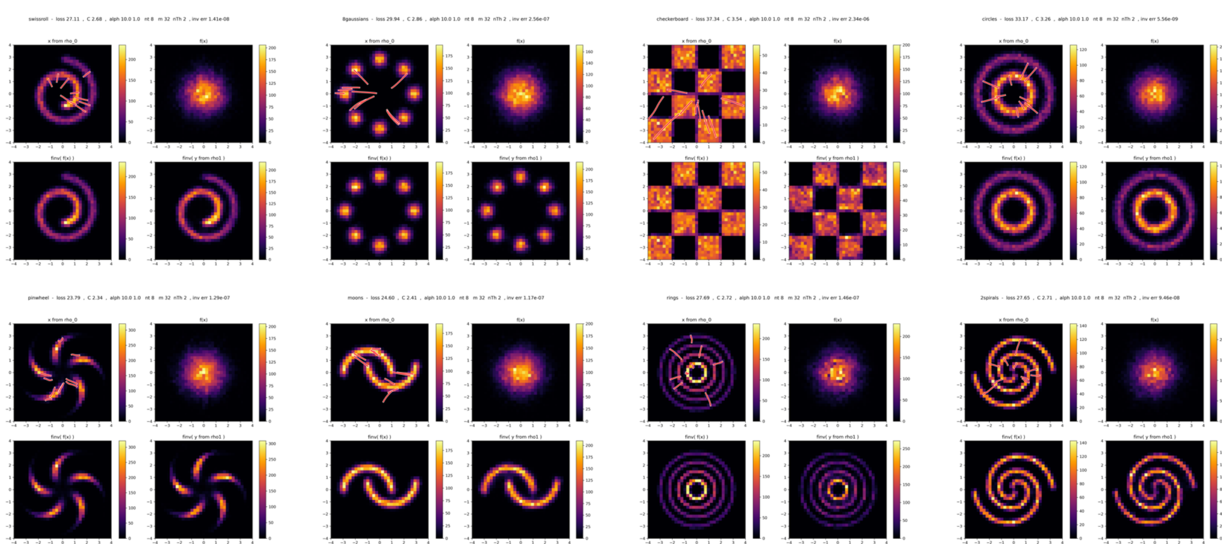

Figure 7: OT flow 2d toys models

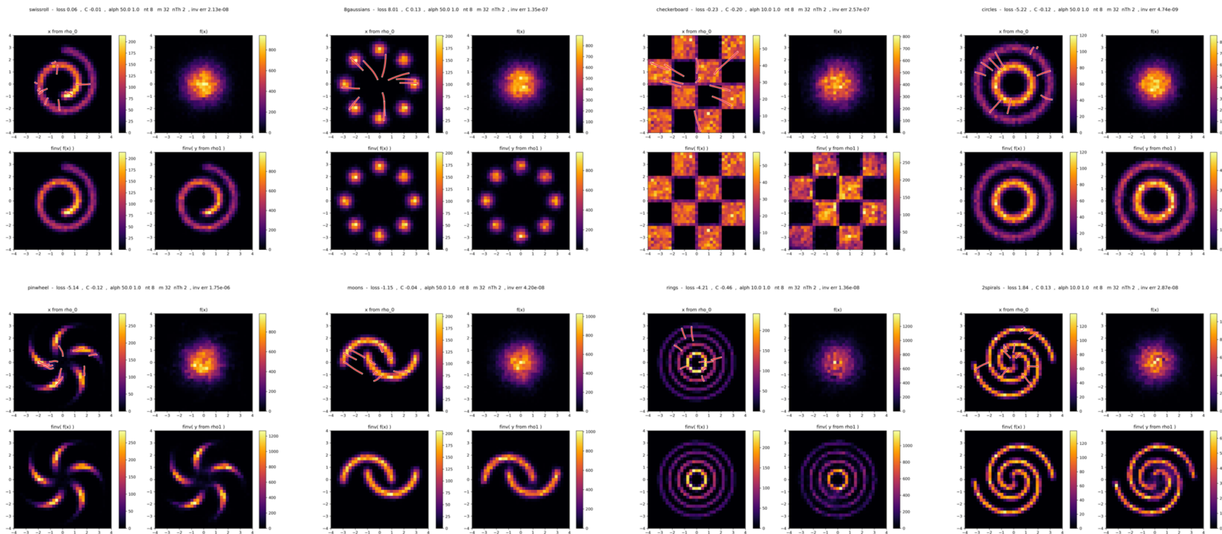

Figure 8: UOT 2d toys models

We compare the known samples to the generated samples via maximum mean discrepancy (MMD)

$$MMD(X,Q) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} k(x_i,x_j) + \frac{1}{M^2}\sum_{i=1}^{M}\sum_{j=1}^{M} k(q_i,q_j) - \frac{2}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} k(x_i,q_j) \tag{4}$$

for Gaussian kernel $k(x_i, q_j) = \exp(-\frac{1}{2}\|x_i - q_j\|^2)$. A low MMD value means that the two sets of samples are likely to have been drawn from the same distribution.

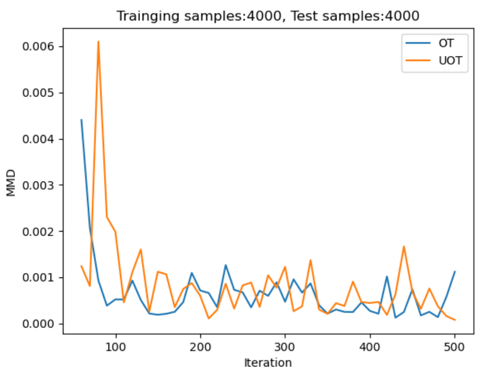

Figure 9: Train/Test:4000/4000

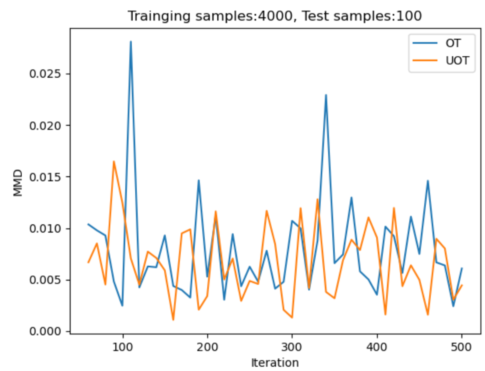

Figure 10: Train/Test:4000/100

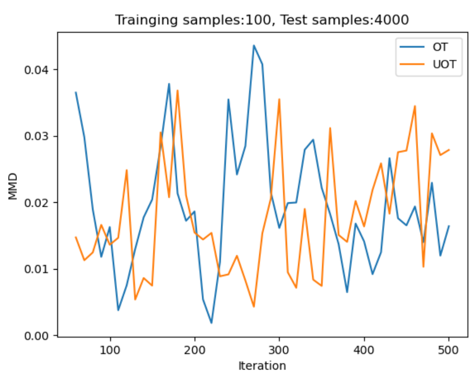

Figure 11: Train/Test:100/4000

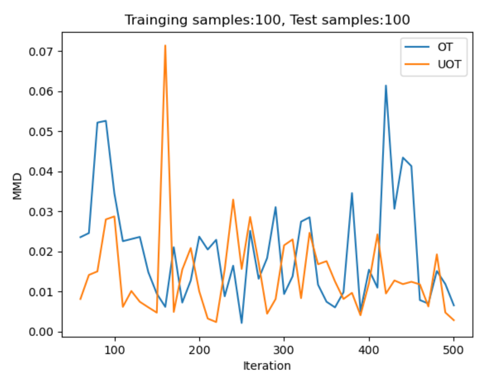

Figure 12: Train/Test:100/100

$T : \rho \to \rho_1$

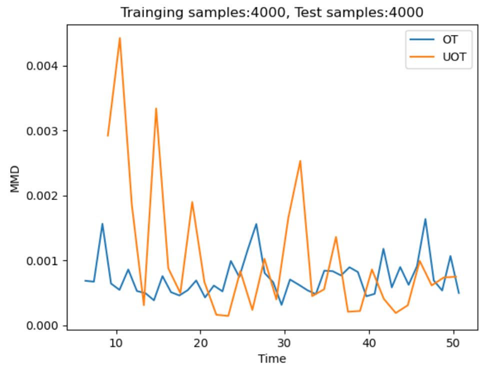

Figure 13: Train/Test:4000/4000

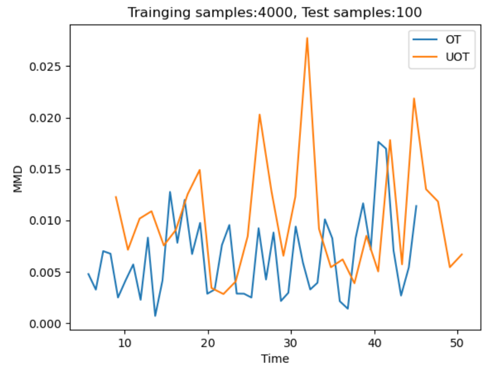

Figure 14: Train/Test:4000/100

1.7 Traditional Method

We also try to reduplicate the traditional method to solve normalizing flow problem. Considering the kinetic energy minimization problem:

$$\min \left\{ \int_0^1 \int_\Omega |\mathbf{v}_t|^p \; \mathrm{d}\varrho_t \; \mathrm{d}t : \partial_t \varrho_t + \nabla \cdot (\mathbf{v}_t \rho_t) = 0, \varrho_0 = \mu, \varrho_1 = v \right\} \tag{5}$$

We need to solve the variable pair $(\varrho_t, \mathbf{v}_t)$. However the constraint is nonlinear and the function is non-convex. It is sufficient to switch $(\varrho_t, \mathbf{v}_t)$ into $(\varrho_t, \mathbf{E}_t)$ where $\mathbf{E}_t = \varrho_t \mathbf{v}_t$. Then the problem can be converted to another optimize problem:

$$\min \left\{ \mathscr{B}_p(\varrho, E) : \partial_t \varrho_t + \nabla \cdot E_t = 0, \varrho_0 = \mu, \varrho_1 = v \right\} \tag{6}$$

where $\mathscr{B}_p(\varrho, E) = \int_0^1 \mathscr{B}_p(\varrho_t, E_t) \, \mathrm{d}t = \int_0^1 \int_\Omega f_p(\varrho_t(x), E_t(x)) \, \mathrm{d}x \; \mathrm{d}t$. f_p is defined as following:

$$f_p(t, x) := \sup_{(a,b) \in K_q} (at + b \cdot x) = \begin{cases} \frac{1}{p} \frac{|x| p}{p^{-1}} & \text{if } t > 0 \\ 0 & \text{if } t = 0, x = 0 \\ +\infty & \text{if } t = 0, x \neq 0, \text{ or } t < 0 \end{cases} \tag{7}$$

where $K_q := \left\{ (a, b) \in \mathbb{R} \times \mathbb{R}^d : a + \frac{1}{q} |b|^q \leq 0 \right\}$.

We will use tradition optimize method to solve the minimize problem instead of network.First of all, we will write the constraint in a weak form:

$$\min_{e,E} \mathscr{B}_p(\varrho, E) + \sup_\phi \left(- \int_0^1 \int_\Omega \left((\partial_t \phi) \, \varrho_t + \nabla \phi \cdot E_t \right) + G(\phi) \right) \tag{8}$$

where

$$G(\phi) := \int_\Omega \phi(1, x) \mathrm{d}v(x) - \int_\Omega \phi(0, x) \mathrm{d}\mu(x) \tag{9}$$

In particular we will focus on $p = 2$:

$$\min_{(E,\varrho):e\geq 0} \int_0^1 \int_\Omega \frac{|E|^2}{2\varrho} + \sup_\phi - \int_0^1 \int_\Omega ((\partial_t \phi)\, \varrho + \nabla\phi \cdot E) + G(\phi), \tag{10}$$

By using $f_p(t, x)$, we can rewrite the problem:

$$\min_{\varrho,E} \sup_{(a,b)\in K_{q,\phi}} \int_0^1 \int_\Omega (a(t,x)\mathrm{d}\varrho + b(t,x)\cdot \mathrm{d}E - \partial_t\phi \mathrm{d}\varrho - \nabla\phi \cdot \mathrm{d}E) + G(\phi) \tag{11}$$

Denote $m = (\varrho, E)$ and $\xi = (a, b)$, then problem becomes:

$$\min_m \sup_{\xi,\phi:\xi\in K_q} \langle \xi - \nabla_{t,x}\phi, m\rangle + G(\phi) \tag{12}$$

We use augmented Lagrangian method to solve above problem. Considering the following form :

$$\min_m \sup_{\xi,\phi:\xi\in K_q} \langle \xi - \nabla_{t,x}\phi, m\rangle + G(\phi) - \frac{\tau}{2}\left|\xi - \nabla_{t,x}\phi\right|^2 \tag{13}$$

The algorithm as following, suppose we have a triplet (m_k, ξ_k, ϕ_k):

- Given m_k and ξ_k, find the optimal ϕ_{k+1}, by solving :

  $$\max_\phi \quad -\langle \nabla_{t,x}\phi, m_k\rangle + G(\phi) - \frac{\tau}{2}\left\|\xi_k - \nabla_{t,x}\phi\right\|^2$$

  The solution can be found as the solution of a Laplace equation

  $$\tau\Delta_{t,x}\phi = \nabla \cdot (\tau\xi_k - m_k)$$

  Boundary condition can derived from variation w.r.t ϕ

- Given m_k and ϕ_{k+1}, find the optimal ξ_{k+1}, by solving :

  $$\max_{\xi\in K_q} \langle \xi, m_k\rangle - \frac{\tau}{2}\left|\xi - \nabla_{t,x}\phi_{k+1}\right|^2$$

  this problem is equivalent to the projection of $\nabla_{t,x}\phi_{k+1} + \frac{1}{\tau}m_k$ in the convex set K_q mentioned above.

- Finally we update m by

  $$m_{k+1} = m_k - \tau\left(\xi_{k+1} - \nabla_{t,x}\phi_{k+1}\right)$$

2 Paper Draft

2.1 Abstract

Flow model is a family of models that build an invertible mapping between two distributions. Normally one distribution is standard normal distribution and the other is arbitrary. Such model is named normalizing flow. Flow model can be used for generating samples, density estimation and Bayesian inference. Continuous normalizing flows (CNFs) solves an neural ordinary differential equation (ODE) to obtain the mapping. The density and velocity field satisfy transport equations. By adding a source term in the transport equation, we can obtain a weighted flow model. We design a new loss function to train the network, avoiding to estimate the original density in the formulation. We also introduce new regularization to restrict velocity field base on the weight change of particles.

2.2 Introduction

We will introduce traditional continuous normalizing flows (CNFs) first. CNFs aim to build a continuous and invertible mapping between an arbitrary distribution ρ_0 and a standard normal distribution ρ_1. Alternatively, for a given time T, we are trying to obtain a mapping $z : \mathbf{R}^d \times [0,T] \to \mathbf{R}^d$. The mapping z defines a continuous change process of every $x \in \mathbf{R}^d$, which is known as flow or trajectory of particles. Then the density $\rho(z(x,t),t)$ satisfies:

$$\log \rho_0(x) = \log \rho(z(x,t),t) + \log|\det \nabla z(x,t)| \quad \text{for all} \quad x \in \mathbf{R}^d \tag{14}$$

Especially at time T we have $\log \rho_0(x) = \log \rho_1(z(x,T),T) + \log|\det \nabla z(x,T)|$. $z(x,T)$ is also known as normalizing flow. $z(x,t)$ satisfies following ODE:

$$\partial_t \begin{bmatrix} z(x,t) \\ \ell(x,t) \end{bmatrix} = \begin{bmatrix} v(z(x,t),t;\boldsymbol{\theta}) \\ \mathrm{tr}(\nabla v(z(x,t),t;\boldsymbol{\theta})) \end{bmatrix}, \quad \begin{bmatrix} z(x,0) \\ \ell(x,0) \end{bmatrix} = \begin{bmatrix} x \\ 0 \end{bmatrix} \tag{15}$$

In the the second ODE, $\ell(x,t) = \log\det z(x,t)$. The second ODE in 15 can be formulated from first ODE. We solve it together to gain the change of ρ for convenience. It will lead to a quick estimation of density with fewer computational cost. Following is the formulation of second ODE:

$$\begin{aligned} \frac{\partial \ell(x,t)}{\partial t} &= \frac{1}{\det(\nabla z(x,t))} \frac{\partial \det(\nabla z(x,t))}{\partial t} \\ &= \frac{1}{\det(\nabla z(x,t))} \cdot \det(\nabla z(x,t)) \cdot \mathrm{tr}\left[(\nabla z(x,t))^{-1} \frac{\partial \nabla z(x,t)}{\partial t}\right] \\ &= \frac{1}{\det(\nabla z(x,t))} \cdot \det(\nabla z(x,t)) \cdot \mathrm{tr}\left[(\nabla z(x,t))^{-1} \nabla z(x,t) \nabla v(z(x,t),t)\right] \\ &= \mathrm{tr}\left[(\nabla v(z(x,t),t)\right] \end{aligned} \tag{16}$$

Above we use following properties:

$$\frac{\partial \det(A)}{\partial t} = \det A \cdot \operatorname{tr}\left[A^{-1}\frac{\partial A}{\partial t}\right]$$
$$\operatorname{tr}(AB) = \operatorname{tr}(BA) \tag{17}$$

From such ODE system we can see that if we have the velocity field, then we can push forward the ODE system and obtain the final distribution at time T. On the other hand, different velocity field can lead to same final distribution. We hope to find an invertible and smooth solution. In OT-flow, they design the following cost function to train the velocity field:

$$J = \mathrm{D}_{\mathrm{KL}}\left[\rho(x,T)\|\rho_1(x)\right] + \mathbb{E}_{\rho_o(x)}\left[\int_0^T \frac{1}{2}\|v(z(x,t),t)\|^2 \mathrm{d}t\right] \tag{18}$$

The first part in 18 is the KL divergence between $\rho(x,T)$ and normal distribution ρ_1. This term will lead to final distribution solved by ODE getting closed to normal distribution. The second term is based on optimal transport theorem, which can be regarded as a penalty of the squared arc-length of the trajectories v, in order to encourage straight trajectory. We will not explain it in details but look at the calculation of KL divergence. We will use similar calculation in our weight model.

$$\rho_0(x) = \rho(z(x,t)) \cdot \det(\nabla z(x,t)) \tag{19}$$

$$\begin{aligned}
\mathrm{D}_{\mathrm{KL}}\left[\rho(z(x,T))\|\rho_1(x)\right] &= \int_{\mathbb{R}^d} \log\left(\frac{\rho(z(x,T))}{\rho_1(z(x,T)}\right)\rho(z(x,T))\mathrm{d}z \\
&= \int_{\mathbb{R}^d} \log\left(\frac{\rho(z(x,T))}{\rho_1(z(x,T))}\right)\rho(z(x,T))\det(\nabla z(x,T))\mathrm{d}x \\
&= \int_{\mathbb{R}^d} \log\left(\frac{\rho_0(x)}{\rho_1(z(x,T))\det(\nabla z(x,T))}\right)\rho_0(x)\mathrm{d}x \\
&= \int_{\mathbb{R}^d} \left[\log\left(\rho_0(x)\right) - \log\left(\rho_1(z(x,T))\right) - \log\det(\nabla z(x,T))\right]\rho_0(x)\mathrm{d}x
\end{aligned} \tag{20}$$

Since ρ_1 is normal distribution, thus

$$\log\left(\rho_1(\boldsymbol{z}(\boldsymbol{x},T))\right) = -\frac{1}{2}\|\boldsymbol{z}(\boldsymbol{x},T)\|^2 - \frac{d}{2}\log(2\pi) \tag{21}$$

Then KL divergence can be written as

$$\begin{aligned}
\mathrm{D}_{\mathrm{KL}}\left[\rho(z(x,T))\|\rho_1(x)\right] &= \int_{\mathbb{R}^d}\left[\log\left(\rho_0(\boldsymbol{x})\right) - \log\det(\nabla\boldsymbol{z}(\boldsymbol{x},T)) + \frac{1}{2}\|\boldsymbol{z}(\boldsymbol{x},T)\|^2 + \frac{d}{2}\log(2\pi)\right]\rho_0(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&= \mathbb{E}_{\rho_o(x)}\left[\log(\rho_0(x) - \ell(x,T) + \frac{1}{2}\|\boldsymbol{z}(\boldsymbol{x},T)\|^2 + \frac{d}{2}\log(2\pi)\right]
\end{aligned} \tag{22}$$

Since ρ_0 is already known, we can just drop it when training. The whole cost can be written as following form:

$$
\begin{aligned}
J &= \mathbb{E}_{\rho_o(x)}\left[C(x,t) + L(x,t)\right] \\
C(x,T) &= -\ell(x,T) + \frac{1}{2}\|\boldsymbol{z}(\boldsymbol{x},T)\|^2 + \frac{d}{2}\log(2\pi) \\
L(x,T) &= \int_0^T \frac{1}{2}\|v(z(x,t),t)\|^2 \mathrm{d}t
\end{aligned}
\tag{23}
$$

2.3 Weight Model

Now we consider a more general form of transport flow:

$$
\partial_t \rho + \nabla \cdot (\rho v) = \rho g \tag{24}
$$

where $g : \mathbb{R}^d \times [0,T] \to \mathbb{R}$ is a source.

And the cost will be:

$$
J = \mathrm{D}_{\mathrm{KL}}\left[\rho(x,T)\|\rho_1(x)\right] + \mathbb{E}_{\rho_o(x)}\left[\int_0^T \frac{1}{2}\|v(z(x,t),t)\|^2 \mathrm{d}t + \int_0^T \frac{1}{2}\alpha g^2(z(x,t),t)\mathrm{d}t\right] \tag{25}
$$

Where α is a hyper-parameter to control the influence of the source.For fixed $\rho(t)$,there are still infinite pairs (v,g) that can achieve it. We need to minimize the cost w.r.t v and g .

Consider the following optimization problem and its Lagrangian function (1/2 is added so that the objective function connects to the kinetic energy)

$$
\min\{\int_0^1 \int_{\mathbb{R}^d} \rho|v|^2 + \alpha\rho g^2 dzdt, \partial_t\rho + \nabla\cdot(\rho v) = \rho g\}
$$

$$
\mathcal{L} = \frac{1}{2}\int_0^1 \int_{\mathbb{R}^d} \rho|v|^2 + \alpha\rho g^2 dzdt - \int_0^1 \int_{\mathbb{R}^d} \Phi(z,t)(\partial_t\rho + \nabla\cdot(\rho v) - \rho g)dzdt
$$

Taking the variation $\frac{\delta\mathcal{L}}{\delta\rho} = 0$, $\frac{\delta\mathcal{L}}{\delta g} = 0$, $\frac{\delta\mathcal{L}}{\delta v} = 0$, one obtains

$$
\begin{cases}
\frac{1}{2}|v|^2 + \frac{1}{2}\alpha g^2 + \partial_t\Phi + \Phi g + \nabla\Phi\cdot v = 0 \\
\qquad v = -\nabla\Phi \\
\qquad \alpha g = -\Phi
\end{cases}
$$

it follows that $\begin{cases} v = -\nabla\Phi = \alpha\nabla g \\ \partial_t\Phi = \frac{1}{2}|\nabla\Phi|^2 + \frac{1}{2\alpha}\Phi^2 \end{cases}$

The original PDE can be written as:

$$\partial_t \rho - \nabla \cdot (\rho \nabla \Phi) = -\frac{1}{\alpha}\rho\Phi \tag{26}$$

To keep ρ as a measure, which is equivalent $\int \Phi d\rho = 0$, we add a term $\bar{\Phi} = \int \Phi d\rho$:

$$\partial_t \rho - \nabla \cdot (\rho \nabla \Phi) = -\frac{1}{\alpha}\rho(\Phi - \bar{\Phi}) \tag{27}$$

(Note: $\bar{\Phi} = \bar{\Phi}(t) = \int \Phi d\rho$ is just a function of t.)
For the velocity term, one has

$$\begin{aligned} -\partial_t v &= \partial_t(\nabla \Phi) \\ &= \nabla(\frac{1}{2}|\nabla\Phi|^2 + \frac{1}{2\alpha}\Phi^2) \\ &= |\nabla\Phi|\nabla(|\nabla\Phi|) + \frac{1}{\alpha}\Phi\nabla\Phi \\ &= |\nabla\Phi| \cdot \frac{\nabla\Phi \cdot \nabla^2\Phi}{|\nabla\Phi|} + \frac{1}{\alpha}\Phi\nabla\Phi \\ &= \nabla\Phi \cdot \nabla^2\Phi + \frac{1}{\alpha}\Phi\nabla\Phi \\ &= v \cdot \nabla v - \frac{1}{\alpha}\Phi v \end{aligned} \tag{28}$$

Replace Φ with $\Phi - \bar{\Phi}$, one has

$$\partial_t v + v \cdot \nabla v - \frac{1}{\alpha}(\Phi - \bar{\Phi})v = 0 \tag{29}$$

We will use the method of characteristic lines. $v(t,x)$ denotes the velocity at position and time (x,t). Let $\gamma(s;t,x)$ be the characteristic line which satisfies

$$\begin{cases} \dfrac{d\gamma(s;t,x)}{ds} = v(s;\gamma(s;t,x)) \\ \gamma(t;t,x) = x \end{cases}$$

then $U(s) := v(s,\gamma(s;t,x))$ satisfies

$$U'(s) = \partial_t v + v \cdot \nabla v = \frac{1}{\alpha}(\Phi - \bar{\Phi})v(s;\gamma(s;t,x)) = \frac{1}{\alpha}(\Phi - \bar{\Phi})U(s) \tag{30}$$

It follows that:

$$v(t,x) = U(t) = U(0)e^{\frac{1}{\alpha}\int_0^t (\Phi(s;\gamma(s;t,x)) - \bar{\Phi}(s))ds} \tag{31}$$

Note that $\gamma(s; t, z(x,t))) = z(x,s)$ and $\gamma(0; t, z(x,t)) = x$, hence

$$v(t, z(x,t)) = v(0,x)e^{\frac{1}{\alpha}\int_0^t(\Phi(s;z(x,s))-\bar{\Phi}(s))ds} \tag{32}$$

Thus we can impose

$$\text{Cost\_v} := \int_0^T \int_{\mathbb{R}^d} |v(t, z(x,t)) - v(0,x)e^{\frac{1}{\alpha}\int_0^t(\Phi(s;z(x,s))-\bar{\Phi}(s))ds}|^2 \rho_0(x)e^{-\frac{1}{\alpha}\int_0^t(\Phi(s;z(x,s))-\bar{\Phi}(s))ds}dxdt \tag{33}$$

as one of the loss terms. Such term penalizes the velocity field along the trajectory, which can lead to a better velocity field suited for our weight model in trainging.

On the other hand, consider empirical distribution for particle system:

$$\rho(x,t) = \sum_{i=1}^n w_i(t)\delta(x - x_i(t)) \tag{34}$$

where $w_i(t)$ denotes the weight of particle x_i at time t. The weights satisfy $w_i(t) \geq 0$ and $\sum_{i=1}^n w_i(t) = 1$. Then $\bar{\Phi}(t) = \sum_{i=1}^n w_i(t)\Phi(x_i(t))$.

Take (3) into (2), one has:

$$w_i'(t) = -\frac{1}{\alpha}(\Phi(x_i(t)) - \bar{\Phi}(t))w_i(t) \tag{35}$$

$$x_i'(t) = -\nabla\Phi(x_i(t)) \tag{36}$$

Similar to

$$\rho_0(x) = \rho(z(x,t)) \cdot \det(\nabla z(x,t)) \tag{37}$$

In this formulation we have

$$\rho_0(x)e^{-\frac{1}{\alpha}\int_0^t(\Phi(z(x,s))-\bar{\Phi}(s))ds} = \rho(z(x,t)) \cdot \det(\nabla z(x,t)) \tag{38}$$

Then the KL divergence term can be computed as

$$
\begin{aligned}
D_{KL}[\rho(z(x,T))\|\rho_1(z)] &= \int_{\mathbb{R}^d} \log\Big(\frac{\rho(z(x,T))}{\rho_1(z(x,T))}\Big)\rho(z(x,T))dz \\
&= \int_{\mathbb{R}^d} \log\Big(\frac{\rho(z(x,T))}{\rho_1(z(x,T))}\Big)\rho(z(x,T))\text{det}\nabla z(x,T)dx \\
&= \int_{\mathbb{R}^d} \log\Big(\frac{\rho_0(x)e^{-\frac{1}{\alpha}\int_0^T(\Phi(z(x,s))-\bar{\Phi}(s))ds}}{\rho_1(z(x,T))\text{det}(\nabla z(x,T))}\Big)\rho_0(x)e^{-\frac{1}{\alpha}\int_0^T(\Phi(z(x,s))-\bar{\Phi}(s))ds}dx \\
&= \underbrace{\int_{\mathbb{R}^d}[-\text{log}(\rho_1(z(x,T)) - \text{log}(\text{det}(\nabla z(x,T))]\rho_0(x)e^{-\frac{1}{\alpha}\int_0^T(\Phi(z(x,s))-\bar{\Phi}(s))ds}dx}_{A} \\
&+ \underbrace{\int_{\mathbb{R}^d}\log\Big(\rho_0(x)e^{-\frac{1}{\alpha}\int_0^T(\Phi(z(x,s))-\bar{\Phi}(s))ds}\Big)\rho_0(x)e^{-\frac{1}{\alpha}\int_0^T(\Phi(z(x,s))-\bar{\Phi}(s))ds}dx}_{B}
\end{aligned}
\tag{39}
$$

Since ρ_1 is normal distribution, thus

$$
\text{log}(\rho_1(z(x,T))) = -\frac{1}{2}|z(x,T)|^2 - \frac{d}{2}\text{log}(2\pi) \tag{40}
$$

Denote $l(x,t) = \text{log}(\text{det}(\nabla z(x,t)))$, one can find that

$$
\partial_t l(x,t) = \text{tr}(\nabla v(z(x,t),t)) = -\text{tr}(\nabla^2\Phi(z(x,t),t)) \tag{41}
$$

In discrete sense the term A can be written as:

$$
\frac{d}{2}\text{log}(2\pi) + \sum_{i=1}^{n}\Big(\frac{1}{2}|z(x_i,T)|^2 - l(x_i,T)\Big)w_i(T) \tag{42}
$$

We can verify $\rho_t(x) = \rho_0(x)e^{-\frac{1}{\alpha}\int_0^t(\Phi(z(x,s))-\bar{\Phi}(s))ds}$ is a probability density function, but this is not $\rho(z(x,t))$. And in fact, the term B in KL divergence is the entropy of $\rho_t(x)$.

To compute B, since

$$
B = \underbrace{\int_{\mathbb{R}^d}\text{log}(\rho_0(x))\rho_0(x)e^{-\frac{1}{\alpha}\int_0^T(\Phi(z(x,s))-\bar{\Phi}(s))ds}dx}_{C} \tag{43}
$$

$$
\underbrace{-\frac{1}{\alpha}\int_{\mathbb{R}^d}\Big(\int_0^T(\Phi(z(x,s))-\bar{\Phi}(s))ds\Big)\rho_0(x)e^{-\frac{1}{\alpha}\int_0^T(\Phi(z(x,s))-\bar{\Phi}(s))ds}dx}_{D} \tag{44}
$$

D can be written as $-\frac{1}{\alpha}\sum_{i=1}^n \varphi_i(T)w_i(T)$, where $\varphi_i(T) = \int_0^T(\Phi(x_i(s)) - \bar{\Phi}(s))ds$ can be computed by $\partial_t\varphi_i(t) = \Phi(x_i(t)) - \bar{\Phi}(t)$.

However, the term C is difficult to compute, since ρ_0 is unknown and we only know some samples. Note that in OT-flow we have a similar term $\rho_0 \log \rho_0$. We can drop it in training since it is a constant. We cannot do that in weight model since weight is related to the network. We develop several ways to deal with this problem. The fisrt is using some tricks to estimate the initial density. The second is replacing KL divergence with another weak metric, kernelized discrepancy distance (KSD). KSD avoids estimating initial density and use discrete samples to evaluate the distance between two distribution. The third is more dedicated. Instead of using KL divergence between final distribution and normal distribution, we design a different KL divergence that avoid estimating ρ_0 but in some sense reflects how good is velocity field is trained. We will introduce them in detail in the following sections.

3 Estimating density

we adopt the clever method in [8] to approximate $\log(\rho_0(x))$, which is based on the following observation:

$$\mathcal{D}(x) := \log[\rho_0/\rho] = \text{argmin}_D[\mathbb{E}_{x\sim\rho_0}\log(1+e^{-D(x)}) + \mathbb{E}_{x\sim\rho}\log(1+e^{D(x)})] \tag{45}$$

Hence, if we take $\rho(x)$ as standard normal distribution, in practice $\log(\rho_0(x))$ can be computed as:

$$\log(\rho_0(x)) = -\frac{1}{2}|x|^2 - \frac{d}{2}\log(2\pi) + D(x) \tag{46}$$

where $D(x)$ is obtained by minimizing

$$\text{argmin}_{D\in\mathcal{C}}[\frac{1}{|S_*|}\sum_{x\in S_*}\log(1+e^{-D(x)}) + \frac{1}{|S|}\sum_{x\in S}\log(1+e^{D(x)})] \tag{47}$$

here $\mathcal{C}$ is some function class, S_* is sampled from ρ_0 and S is sampled from standard normal distribution. Intuitively, approximation improves with larger sample and more universal $\mathcal{C}$. We exploit a fully-connected neural network as the function class $\mathcal{C}$.

4 Kernelized Stein Discrepancy (KSD)

The stein's method is a general theoretical tool for obtaining bounds on distances between distributions. Roughly speaking, it relies on the basic fact that two smooth densities $p(x)$ and $q(x)$ supported on $\mathbb{R}$ are identical if and only if

$$\mathbb{E}_p[s_q(x)f(x) + \nabla_x f(x)] = 0 \tag{48}$$

for smooth functions $f(x)$ with proper zero-boundary conditions, where $s_q(x) = \nabla_x \log q(x)$ is the (Stein) score function of $q(x)$. When $p = q$, (29) is known as stein's identity.
As a result, one can define a Stein discrepancy measure between p and q via

$$\mathbb{S}(p,q) = \max_{f\in F}\Big(\mathbb{E}_p[s_q(x)f(x) + \nabla_x f(x)]\Big)^2 \tag{49}$$

[5] propose LSD to use neural network to maximize stein discrepancy[4] and train unnormalized models through a min-max process. [9] introduces **kernelized Stein discrepancy** (KSD) with an elementary definition and establish its connection with Stein's method and RKHS.

The kernelized stein discrepancy $\mathbb{S}(p, q)$ between distribution p and q is defined as

$$\mathbb{S}(p, q) = \mathbb{E}_{x,y\sim p}\Big[(s_q(x) - s_p(x))^T k(x, y)(s_q(y) - s_p(y))\Big] \tag{50}$$

where $s_p(x) = \nabla_x \text{log} p(x)$ is the (Stein) score function of p and $k(x, y)$ is integrally strictly positive positive definition. Theorem 3.6 in [9] defines

$$u_q(x, y) = s_q(x)^T k(x, y)s_q(y) + s_q(x)^T \nabla_y k(x, y) + \nabla_x k(x, y)^T s_q(y) + \text{trace}(\nabla_{x,y} k(x, y)) \tag{51}$$

then,

$$\mathbb{S}(p, q) = \mathbb{E}_{x,y\sim p}[u_q(x, y)] \tag{52}$$

Take RBF kernel $k(x, y) = e^{-\frac{1}{2h^2}\|x-y\|_2^2}$, and set the distribution $q(x)$ as standard normal distribution, we can claculate $u_q(x, y)$:

$$u_q(x, y) = e^{-\frac{1}{2h^2}\|x-y\|_2^2}\left(x^T y + \frac{d}{h^2} - \Big(\frac{1}{h^2} + \frac{1}{h^4}\Big)\|x - y\|_2^2\right) \tag{53}$$

Here d is the dimension and $x, y \in \mathbb{R}^d$.

If we take inverse multiquadric (IMQ) kernel suggested by [4] :

$$k(x, y) = (c^2 + \frac{\|x - y\|_2^2}{l^2})^\beta \tag{54}$$

for some $\beta \in (-1, 0), c > 0, l > 0$. Set the distribution $q(x)$ as standard normal distribution, then we can calculate $u_q(x, y)$:

$$u_q(x, y) = kx^T y + \frac{2\beta}{l^2} k^{\frac{\beta-1}{\beta}} \|x - y\|_2^2 - \frac{2\beta d}{l^2} k^{\frac{\beta-1}{\beta}} - \frac{4\beta(\beta - 1)}{l^4} k^{\frac{\beta-2}{\beta}} \|x - y\|_2^2 \tag{55}$$

Here $x, y \in \mathbb{R}^d, k = k(x, y)$.

Then we can use (30) to measure the discrepancy between learned distribution and target distribution. In discrete sense, the KSD is

$$\hat{\mathbb{S}}(p, q) = \frac{1}{n(n-1)} \sum_{1\le i\ne j\le n} u_q(x_i, x_j), \quad \{x_i\}_{i=1}^n \sim p \tag{56}$$

In our settings, we rewrite it as

$$\hat{\mathbb{S}}(\rho_T, \rho_1) = \sum_{1\le i\ne j\le n} w_i(T)w_j(T)u_q(x_i(T), x_j(T)) \tag{57}$$

We replace KL divergence with KSD and do experiments but the results are not satisfactory for discontinuous initial density. [6] argues that when KSD is small, it means that within the area of generated samples, the score function of s_p matches the target score function s_q well. An almost-zero empirical KSD does not necessarily imply capturing all the modes or recovering all the support of the true density. We clarify it in the next section with some toy examples and propose some possible plans.

4.1 Inverse KL

From another point of view, if we consider the inverse flow from Gaussian distribution to the target distribution, it seems to avoid the computation $log(\rho_0)$.

Note that in our previous formulation,

$$\rho_0(x)e^{-\frac{1}{\alpha}\int_0^t(\Phi(z(x,s),s)-\bar{\Phi}(s))ds} = \rho(z(x,t)) \cdot \det(\nabla z(x,t)) \tag{58}$$

We replace $\rho(z(x,t))$ with $\rho_1(z(x,t))$, $\rho_0(x)$ with $\widetilde{\rho_0}(x)$ in the above formula.

$$\widetilde{\rho_0}(x)e^{-\frac{1}{\alpha}\int_0^t(\Phi(z(x,s),s)-\bar{\Phi}(s))ds} = \rho_1(z(x,t)) \cdot \det(\nabla z(x,t)) \tag{59}$$

We turn to minimize the inverse version of KL divergence

$$D_{KL}[\rho_0(x)\|\widetilde{\rho_0}(x)] = \int_{\mathbb{R}^d} \log(\frac{\rho_0(x)}{\widetilde{\rho_0}(x)})\rho_0(x)dx = \text{const.} - \int_{\mathbb{R}^d} \log(\widetilde{\rho_0}(x))\rho_0(x)dx \tag{60}$$

In discrete sense, we first sample $\{x_i\}_{i=1}^n$ from $\rho_0(x)$, the KL loss term is

$$D_{KL}[\rho_0(x)\|\widetilde{\rho_0}(x)] \approx -\frac{1}{n}\log(\widetilde{\rho_0}(x_i)) \tag{61}$$

Using (69), this term can be rewritten as

$$-\frac{1}{n}\log(\widetilde{\rho_0}(x_i)) = \frac{d}{2}\log(2\pi)+\frac{1}{2n}\sum_{i=1}^n |z(x_i,T)|^2-\frac{1}{n}\sum_{i=1}^n \log(\det\nabla z(x_i,T))-\frac{1}{n}\Big(\frac{1}{\alpha}\int_0^T \Phi(z(x_i,t),t)-\bar{\Phi}(t)dt\Big) \tag{62}$$

Note that the term $\frac{1}{n\alpha}\int_0^T \bar{\Phi}(t)dt$ is computed from the inverse flow (from ρ_1 to $\widetilde{\rho_0}$)

More detail: In the origin formulation, we have the PDE

$$\begin{cases} \partial_t\rho(x,t) - \nabla\cdot(\rho(x,t)\nabla\Phi(x,t)) = -\frac{1}{\alpha}\rho(x,t)(\Phi(x,t)-\bar{\Phi}(t)) \\ \rho(x,0) = \rho_0(x) \geq 0, \int \rho_0(x)dx = 1 \\ \bar{\Phi}(t) = \int \Phi(x,t)\rho(x,t)dx \end{cases}$$

We invert the time and consider $\widetilde{\rho}(x,T-t) := \rho(x,t)$, then $\tilde{\rho}(x,T) := \rho(x,0) = \rho_0(x)$. $\widetilde{\rho}(x,t)$ satisfies:

$$\begin{cases} \partial_t \widetilde{\rho}(x,t) + \nabla \cdot (\widetilde{\rho}(x,t)\nabla \Phi(x,T-t)) = \dfrac{1}{\alpha}\widetilde{\rho}(x,t)(\Phi(x,T-t) - \bar{\Phi}(T-t)) \\ \widetilde{\rho}(x,0) = \rho(x,T) \geq 0, \displaystyle\int \rho(x,T)dx = 1 \\ \bar{\Phi}(T-t) = \displaystyle\int \Phi(x,T-t)\widetilde{\rho}(x,t)dx \end{cases}$$

For example, if we consider transport $\rho_1(x)$ back, then $\widetilde{\rho}(x,0) = \rho_1(x)$. Consider empirical distribution for particle system:

$$\widetilde{\rho}(x,t) = \sum_{i=1}^{n} w_i(t)\delta(x - x_i(t)) \tag{63}$$

where $w_i(t)$ denotes the weight of particle x_i at time t. The weights satisfy $w_i(t) \geq 0$ and $\sum_{i=1}^{n} w_i(t) = 1$. Then $\bar{\Phi}(T-t) = \sum_{i=1}^{n} w_i(t)\Phi(x_i(t), T-t)$.

Substitute it into the PDE:

$$w_i^{'}(t) = \frac{1}{\alpha}(\Phi(x_i(t), T-t) - \bar{\Phi}(T-t))w_i(t) \tag{64}$$

$$x_i^{'}(t) = \nabla \Phi(x_i(t), T-t) \tag{65}$$

Moreover, we have

$$\rho_1(x)e^{\frac{1}{\alpha}\int_0^T(\Phi(z(x,t),T-t)-\bar{\Phi}(T-t))dt} = \widetilde{\rho}(z(x,T),T) \cdot \text{det}(\nabla z(x,T)) \tag{66}$$

Here note that $z(x_i(0),t) = x_i(t)$. It's easy to see $\widetilde{\rho}(x,t)$ is a probability measure for any $0 \leq t \leq T$. We want to minimize the KL divergence between $\widetilde{\rho}(x,T)$ and $\rho_0(x)$. To see it clearly, we firstly change the variable along the trajectory. Denote $x(z(x_0,T), T-t) = z(x_0,t)$, then $x(z(x_0,T),0) = z(x_0,T)$ and $x(z(x_0,T),T) = z(x_0,0) = x_0$. (76) can be rewritten as

$$\rho_1(x(z(x,T),T))e^{\frac{1}{\alpha}\int_0^T(\Phi(x(z(x,T),T-t))-\bar{\Phi}(T-t))dt} = \widetilde{\rho}(z(x,T),T)\frac{1}{\text{det}\nabla_{z(x,T)}x(z(x,T),T)} \tag{67}$$

Let $z(x,T) = x_0$ and change the name of trajectory ($x \to z$):

$$\rho_1(z(x_0,T))e^{\frac{1}{\alpha}\int_0^T(\Phi(z(x_0,T-t),T-t)-\bar{\Phi}(T-t))dt} = \widetilde{\rho}(x_0,T)\frac{1}{\text{det}\nabla_{x_0}z(x_0,T)} \tag{68}$$

It follows that

$$\widetilde{\rho}(x,T)e^{-\frac{1}{\alpha}\int_0^T(\Phi(z(x,t),t)-\bar{\Phi}(t))dt} = \rho_1(z(x,T))\text{det}\nabla z(x,T) \tag{69}$$

Note that this derivation guarantees that $\widetilde{\rho}(x,T)$ is a probability measure. The KL divergence between $\widetilde{\rho}(x,T)$ and $\rho_0(x)$ is

$$D_{KL}[\rho_0(x)\|\widetilde{\rho}(x,T)] = \int_{\mathbb{R}^d} \log(\frac{\rho_0(x)}{\widetilde{\rho}(x,T)})\rho_0(x)dx = \text{const.} - \int_{\mathbb{R}^d} \log(\widetilde{\rho}(x,T))\rho_0(x)dx \tag{70}$$

which could be computed by using (79). Here $\bar{\Phi}(T-t) = \int_{\mathbb{R}^d}(\Phi(x,T-t)\widetilde{\rho}(x,t))dx$.

4.2 Calculation

$$\begin{cases} \partial_t\rho(x,t) - \nabla\cdot(\rho(x,t)\nabla\Phi(x,t)) = -\dfrac{1}{\alpha}\rho(x,t)(\Phi(x,t)-\bar{\Phi}(t)) \\ \rho(x,0) = \rho_0(x) \geq 0, \displaystyle\int \rho_0(x)dx = 1 \\ \bar{\Phi}(t) = \displaystyle\int \Phi(x,t)\rho(x,t)dx \end{cases}$$

We invert the time and consider $\widetilde{\rho}(z,T-t) := \rho(z,t)$, then $\tilde{\rho}(z,T) := \rho(z,0) = \rho_0(z)$. $\widetilde{\rho}(z,t)$ satisfies:

$$\begin{cases} \partial_t\widetilde{\rho}(z,t) + \nabla\cdot(\widetilde{\rho}(z,t)\nabla\Phi(z,T-t)) = \dfrac{1}{\alpha}\widetilde{\rho}(z,t)(\Phi(z,T-t)-\hat{\Phi}(T-t)) \\ \widetilde{\rho}(z,0) = \rho_1(z) \geq 0, \displaystyle\int \rho_1(z)dz = 1 \\ \hat{\Phi}(T-t) = \displaystyle\int \Phi(z,T-t)\widetilde{\rho}(z,t)dz \end{cases}$$

Equivalently,

$$\rho_0(x)e^{-\frac{1}{\alpha}\int_0^T(\Phi(z(x,t),t)-\bar{\Phi}(t))dt} = \rho(z(x,T),T)\cdot\det(\nabla z(x,T)) \tag{71}$$

$$\rho_1(z)e^{\frac{1}{\alpha}\int_0^T(\Phi(x(z,t),T-t)-\hat{\Phi}(T-t))dt} = \widetilde{\rho}(x(z,T),T)\cdot\det(\nabla x(z,T)) \tag{72}$$

Since

$$\int \rho_0(x)e^{-\frac{1}{\alpha}\int_0^T(\Phi(z(x,t),t)-\bar{\Phi}(t))dt}dx = 1 \tag{73}$$

$$\int \rho_1(z)e^{\frac{1}{\alpha}\int_0^T(\Phi(x(z,t),T-t)-\hat{\Phi}(T-t))dt}dz = 1 \tag{74}$$

one has

$$-\frac{1}{\alpha}\int_0^T \bar{\Phi}(t)dt = \log\Big(\int_{\mathbb{R}^d}\rho_0(x)e^{-\frac{1}{\alpha}\int_0^T\Phi(z(x,t),t)dt}dx\Big) \tag{75}$$

$$-\frac{1}{\alpha}\int_0^T \hat{\Phi}(T-t)dt = -\log\Big(\int_{\mathbb{R}^d}\rho_1(z)e^{\frac{1}{\alpha}\int_0^T\Phi(x(z,t),T-t)dt}dz\Big) \tag{76}$$

Let $E(\Phi,\rho_0,\rho_1) = (-\frac{1}{\alpha}\int_0^T\bar{\Phi}(t)dt) - (-\frac{1}{\alpha}\int_0^T\hat{\Phi}(T-t)dt) = \log(A(\Phi,\rho_0)) + \log(B(\Phi,\rho_1))$, where

$$A(\Phi,\rho_0) = \int_{\mathbb{R}^d}\rho_0(x)e^{-\frac{1}{\alpha}\int_0^T\Phi(z(x,t),t)dt}dx \tag{77}$$

$$B(\Phi,\rho_1) = \int_{\mathbb{R}^d}\rho_1(z)e^{\frac{1}{\alpha}\int_0^T\Phi(x(z,t),T-t)dt}dz \tag{78}$$

Let Φ_∞ satisfies

$$\rho_0(x)e^{-\frac{1}{\alpha}\int_0^T(\Phi_\infty(z(x,t),t)-\bar{\Phi}_\infty(t))dt} = \rho_1(z(x,T))\cdot\det(\nabla z(x,T)) \tag{79}$$

It's easy to see that $E(\Phi_\infty,\rho_0,\rho_1)=0$. Our goal is to show that first order variation of E at Φ_∞ is equal to zero. Let

$$Q(\delta\Phi)=\frac{\partial E(\Phi_\infty+\epsilon\delta\Phi,\rho_0,\rho_1)}{\partial\epsilon}\Big|_{\epsilon=0}=\lim_{\epsilon\to 0}\frac{1}{\epsilon}\Big(E(\Phi_\infty+\epsilon\delta\Phi,\rho_0,\rho_1)-E(\Phi_\infty,\rho_0,\rho_1)\Big) \tag{80}$$

We want to show that $Q(\delta\Phi)\equiv 0$. For simplicity we replace Φ_∞ with Φ.

$Q(\delta\Phi)=\frac{A^{'}(\delta\Phi)}{A(\Phi_\infty,\rho_0)}+\frac{B^{'}(\delta\Phi)}{B(\Phi_\infty,\rho_1)}$, where

$$\begin{aligned}A^{'}(\delta\Phi)&=\lim_{\epsilon\to 0}\int_{\mathbb{R}^d}\rho_0(x)e^{-\frac{1}{\alpha}\int_0^T\Phi(z(x,t),t)dt}\Big[-\frac{1}{\alpha}\frac{(\Phi+\epsilon\delta\Phi)(z_1(x,t),t)-\Phi(z(x,t),t)}{\epsilon}dt\Big]dx\\&=-\frac{1}{\alpha}\int_{\mathbb{R}^d}\rho_0(x)e^{-\frac{1}{\alpha}\int_0^T\Phi(z(x,t),t)dt}\Big[\int_0^T\delta\Phi(z(x,t),t)dt+\int_0^T\nabla\Phi(z(x,t),t)\cdot l(x,t)dt\Big]dx\end{aligned} \tag{81}$$

$z_1(x,t)$ is defined by $\begin{cases}\dfrac{dz_1(x,t)}{dt}=-\nabla(\Phi+\epsilon\delta\Phi)(z_1(x,t),t)\\ z_1(x,0)=x\end{cases}$

Recall that $z(x,t)$ satisfies $\begin{cases}\dfrac{dz(x,t)}{dt}=-\nabla\Phi(z(x,t),t)\\ z(x,0)=x\end{cases}$, then let $l(x,t):=\lim_{\epsilon\to 0}\frac{1}{\epsilon}\Big(z_1(x,t)-z(x,t)\Big)$, one has

$$\begin{cases}\dfrac{dl(x,t)}{dt}=-\nabla^2\Phi(z(x,t),t)l(x,t)-\nabla(\delta\Phi)(z(x,t),t)\\ l(x,0)=0\end{cases}$$

Similarly,

$$B^{'}(\delta\Phi)=\frac{1}{\alpha}\int_{\mathbb{R}^d}\rho_1(z)e^{\frac{1}{\alpha}\int_0^T\Phi(x(z,t),T-t)dt}\Big[\int_0^T\delta\Phi(x(z,t),T-t)dt+\int_0^T\nabla\Phi(x(z,t),T-t)\cdot\widetilde{l}(z,t)dt\Big]dz \tag{82}$$

where $\widetilde{l}$ satisfies $\begin{cases}\dfrac{d\widetilde{l}(z,t)}{dt}=\nabla^2\Phi(x(z,t),T-t)\widetilde{l}(z,t)+\nabla(\delta\Phi)(x(z,t),T-t)\\ \widetilde{l}(z,0)=0\end{cases}$

Let $\widetilde{\rho}_0(x) = \rho_0(x)e^{-\frac{1}{\alpha}\int_0^T (\Phi(z(x,t),t)-\bar{\Phi}(t))dt}$. Then

$$\begin{aligned}
Q(\delta\Phi) &= \frac{A'(\delta\Phi)}{A(\Phi_\infty,\rho_0)} + \frac{B'(\delta\Phi)}{B(\Phi_\infty,\rho_1)} \\
&= -\frac{1}{\alpha}\int_{\mathbb{R}^d} \rho_0(x)e^{-\frac{1}{\alpha}\int_0^T \Phi(z(x,t),t)-\bar{\Phi}(t)dt}\Big[\int_0^T \delta\Phi(z(x,t),t)dt + \int_0^T \nabla\Phi(z(x,t),t)\cdot l(x,t)dt\Big]dx \\
&+ \frac{1}{\alpha}\int_{\mathbb{R}^d} \rho_1(z)e^{\frac{1}{\alpha}\int_0^T \Phi(x(z,t),T-t)-\hat{\Phi}(T-t)dt}\Big[\int_0^T \delta\Phi(x(z,t),T-t)dt + \int_0^T \nabla\Phi(x(z,t),T-t)\cdot \widetilde{l}(z,t)dt\Big]dz \\
&= -\frac{1}{\alpha}\int_{\mathbb{R}^d} \widetilde{\rho}_0(x)\Big[\int_0^T \delta\Phi(z(x,t),t)dt + \int_0^T \nabla\Phi(z(x,t),t)\cdot l(x,t)dt\Big]dx \\
&+ \frac{1}{\alpha}\int_{\mathbb{R}^d} \rho_0(x(z,T))\cdot \det(\nabla x(z,T))\Big[\int_0^T \delta\Phi(x(z,t),T-t)dt + \int_0^T \nabla\Phi(x(z,t),T-t)\cdot \widetilde{l}(z,t)dt\Big]dz \\
&= -\frac{1}{\alpha}\int_{\mathbb{R}^d} \widetilde{\rho}_0(x)\Big[\int_0^T \delta\Phi(z(x,t),t)dt + \int_0^T \nabla\Phi(z(x,t),t)\cdot l(x,t)dt\Big]dx \\
&+ \frac{1}{\alpha}\int_{\mathbb{R}^d} \rho_0(x)\Big[\int_0^T \delta\Phi(z(x,T-t),T-t)dt + \int_0^T \nabla\Phi(z(x,T-t),T-t)\cdot \widetilde{l}(z(x,T),t)dt\Big]dx \\
&= -\frac{1}{\alpha}\int_{\mathbb{R}^d} \widetilde{\rho}_0(x)\Big[\int_0^T \delta\Phi(z(x,t),t)dt + \int_0^T \nabla\Phi(z(x,t),t)\cdot l(x,t)dt\Big]dx \\
&+ \frac{1}{\alpha}\int_{\mathbb{R}^d} \rho_0(x)\Big[\int_0^T \delta\Phi(z(x,t),t)dt + \int_0^T \nabla\Phi(z(x,t),t)\cdot \widetilde{l}(z(x,T),T-t)dt\Big]dx
\end{aligned} \tag{83}$$

Hence our goal is that

$$\int_{\mathbb{R}^d}(\widetilde{\rho}_0(x)-\rho_0(x))\Big(\int_0^T \delta\Phi(z(x,t),t)dt\Big)dx = \int_{\mathbb{R}^d}\rho_0(x)\Big[\int_0^T \nabla\Phi(z(x,t),t)\cdot\widetilde{l}(z(x,T),T-t)dt\Big]dx - \int_{\mathbb{R}^d}\widetilde{\rho}_0(x)\Big[\int_0^T \nabla\Phi(z(x,t),t)\cdot l(x,t)dt\Big]dx \tag{*}$$

where $\begin{cases} \dfrac{dl(x,t)}{dt} = -\nabla^2\Phi(z(x,t),t)l(x,t) - \nabla(\delta\Phi)(z(x,t),t) \\ l(x,0) = 0 \end{cases}$ and

$$\begin{cases} \dfrac{d\widetilde{l}(z,t)}{dt} = \nabla^2\Phi(x(z,t),T-t)\widetilde{l}(z,t) + \nabla(\delta\Phi)(x(z,t),T-t) \\ \widetilde{l}(z,0) = 0 \end{cases}$$

We can use Duhamel's principle to duduce that

$$l(x,t) = -\int_0^t J_{t,s}(\Phi)\nabla(\delta\Phi)(z(x,s),s)ds \tag{84}$$

where $J_{t,s}(\Phi)$ is the solution to $\begin{cases} \dfrac{d}{dt}J_{t,s}(\Phi) = -\nabla^2\Phi(z(x,t),t)J_{t,s}(\Phi) \\ J_{s,s} = Id \end{cases}$

Simliarly,

$$\widetilde{l}(z,t) = \int_0^t \hat{J}_{t,s}(\Phi)\nabla(\delta\Phi)(x(z,s),T-s)ds \tag{85}$$

where $\hat{J}_{t,s}(\Phi)$ is the solution to $\begin{cases} \dfrac{d}{dt}\hat{J}_{t,s}(\Phi) = \nabla^2\Phi(x(z,t),T-t)\hat{J}_{t,s}(\Phi) \\ J_{s,s} = Id \end{cases}$

Then

$$\begin{aligned}
\text{RHS of}(*) &= \int_{\mathbb{R}^d}\rho_0(x)\Big[\int_0^T \nabla\Phi(z(x,t),t)\cdot\widetilde{l}(z(x,T),T-t)dt\Big]dx - \int_{\mathbb{R}^d}\widetilde{\rho}_0(x)\Big[\int_0^T \nabla\Phi(z(x,t),t)\cdot l(x,t)dt\Big]d \\
&= \int_{\mathbb{R}^d}\rho_0(x)\Big[\int_0^T \nabla\Phi(z(x,t),t)\cdot\Big(\int_0^{T-t}\hat{J}_{T-t,s}(\Phi)\nabla(\delta\Phi)(x(z,s),T-s)ds\Big)dt\Big]dx \\
&- \int_{\mathbb{R}^d}\widetilde{\rho}_0(x)\Big[\int_0^T \nabla\Phi(z(x,t),t)\cdot\Big(-\int_0^t J_{t,s}(\Phi)\nabla(\delta\Phi)(z(x,s),s)ds\Big)dt\Big]dx \\
&= \int_{\mathbb{R}^d}\rho_0(x)\Big[\int_0^T\Big(\int_0^{T-s}\nabla\Phi(z(x,t),t)\hat{J}_{T-t,s}(\Phi)dt\Big)\cdot\nabla\delta\Phi(x(z,s),T-s)ds\Big]dx \\
&+ \int_{\mathbb{R}^d}\widetilde{\rho}_0(x)\Big[\int_0^T\Big(\int_s^T\nabla\Phi(z(x,t),t)J_{t,s}(\Phi)dt\Big)\cdot\nabla\delta\Phi(z(x,s),s)ds\Big]dx \\
&= \int_{\mathbb{R}^d}\rho_0(x)\Big[\int_0^T\Big(\int_0^s\nabla\Phi(z(x,t),t)\hat{J}_{T-t,T-s}(\Phi)dt\Big)\cdot\nabla\delta\Phi(z(x,s),s)ds\Big]dx \\
&+ \int_{\mathbb{R}^d}\widetilde{\rho}_0(x)\Big[\int_0^T\Big(\int_s^T\nabla\Phi(z(x,t),t)J_{t,s}(\Phi)dt\Big)\cdot\nabla\delta\Phi(z(x,s),s)ds\Big]dx
\end{aligned} \tag{86}$$

We claim that $\hat{J}_{T-t,T-s} = J_{t,s}$. Then

$$\begin{aligned}
\text{RHS of}(*) &= \int_{\mathbb{R}^d}\rho_0(x)\Big[\int_0^T\Big(\int_0^s\nabla\Phi(z(x,t),t)J_{t,s}(\Phi)dt\Big)\cdot\nabla\delta\Phi(z(x,s),s)ds\Big]dx \\
&+ \int_{\mathbb{R}^d}\widetilde{\rho}_0(x)\Big[\int_0^T\Big(\int_s^T\nabla\Phi(z(x,t),t)J_{t,s}(\Phi)dt\Big)\cdot\nabla\delta\Phi(z(x,s),s)ds\Big]dx
\end{aligned} \tag{87}$$

where $J_{t,s}(\Phi)$ is the solution to $\begin{cases} \dfrac{d}{dt}J_{t,s}(\Phi) = -\nabla^2\Phi(z(x,t),t)J_{t,s}(\Phi) \\ J_{s,s} = Id \end{cases}$

Note that $\nabla_x z(x,t)$ satisfies $\begin{cases} \dfrac{d}{dt}\nabla_x z(x,t) = -\nabla^2\Phi(z(x,t),t)\nabla_x z(x,t) \\ \nabla_x z(x,0) = Id \end{cases}$

$\nabla\delta\Phi(z(x,s),s) = ((\nabla_x z(x,s))^T)^{-1}\nabla_x\delta\Phi(z(x,s),s), (\nabla\Phi(z(x,t),t))^T = \nabla_x\Phi(z(x,t),t)(\nabla_x z(x,t))^{-1},$

then

$$\text{RHS of}(*) = \int_{\mathbb{R}^d} \rho_0(x)\Big[\int_0^T \Big(\int_0^s \nabla_x\Phi(z(x,t),t)(\nabla_x z(x,t))^{-1}J_{t,s}(\Phi)dt\Big)\cdot((\nabla_x z(x,s))^T)^{-1}\nabla_x\delta\Phi(z(x,s),s)d$$
$$+\int_{\mathbb{R}^d} \widetilde{\rho}_0(x)\Big[\int_0^T \Big(\int_s^T \nabla_x\Phi(z(x,t),t)(\nabla_x z(x,t))^{-1}J_{t,s}(\Phi)dt\Big)\cdot((\nabla_x z(x,s))^T)^{-1}\nabla_x\delta\Phi(z(x,s),s)d \tag{88}$$

Using integration by parts and comparing (*), our goal is

$$\widetilde{\rho}_0(x)-\rho_0(x) = -\nabla_x\cdot\Big(\rho_0(x)\int_0^t e^{2\int_0^t \nabla^2\Phi(z(x,s),s)}\nabla_x\Phi(z(x,s),s)ds\Big)-\nabla_x\cdot\Big(\widetilde{\rho}_0(x)\int_t^T e^{2\int_0^t \nabla^2\Phi(z(x,s),s)}\nabla_x\Phi(z(x,s \tag{89}$$

In particular if we take $t=0$

$$\text{RHS of}(89) = -\nabla_x\cdot\Big(\widetilde{\rho}_0(x)\int_0^T \nabla_x\Phi(z(x,s),s)ds\Big) \tag{90}$$

If our final conclusion (*) is right, we must have

$$\widetilde{\rho}_0(x)-\rho_0(x) = -\nabla_x\cdot\Big(\widetilde{\rho}_0(x)\int_0^T \nabla_x\Phi(z(x,s),s)ds\Big) \tag{91}$$

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[2] Ziv Goldfeld, Kristjan Greenewald, and Kengo Kato. Asymptotic guarantees for generative modeling based on the smooth wasserstein distance. *arXiv preprint arXiv:2002.01012*, 2020.

[3] Ziv Goldfeld, Kristjan Greenewald, Jonathan Niles-Weed, and Yury Polyanskiy. Convergence of smoothed empirical measures with applications to entropy estimation. *IEEE Transactions on Information Theory*, 66(7):4368–4391, 2020.

[4] Jackson Gorham and Lester Mackey. Measuring sample quality with kernels, 2020.

[5] Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, and Richard Zemel. Learning the stein discrepancy for training and evaluating energy-based models without sampling. In *International Conference on Machine Learning*, pages 3732–3747. PMLR, 2020.

[6] Tianyang Hu, Zixiang Chen, Hanxi Sun, Jincheng Bai, Mao Ye, and Guang Cheng. Stein neural sampler. *arXiv preprint arXiv:1810.03545*, 2018.

[7] Wittawat Jitkrittum, Wenkai Xu, Zoltán Szabó, Kenji Fukumizu, and Arthur Gretton. A linear-time kernel goodness-of-fit test. *arXiv preprint arXiv:1705.07673*, 2017.

[8] Rie Johnson and Tong Zhang. A framework of composite functional gradient methods for generative adversarial models. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):17–32, 2019.

[9] Qiang Liu, Jason Lee, and Michael Jordan. A kernelized stein discrepancy for goodness-of-fit tests. In *International conference on machine learning*, pages 276–284. PMLR, 2016.

[10] Sloan Nietert, Ziv Goldfeld, and Kengo Kato. Smooth p-wasserstein distance: Structure, empirical approximation, and statistical applications. In *International Conference on Machine Learning*, pages 8172–8183. PMLR, 2021.

[11] Derek Onken, Samy Wu Fung, Xingjian Li, and Lars Ruthotto. Ot-flow: Fast and accurate continuous normalizing flows via optimal transport. *arXiv preprint arXiv:2006.00104*, 2020.