Numer. Math. Theor. Meth. Appl.
doi: 10.4208/nmtma.OA-2024-0114

Vol. **18**, No. 2, pp. 325-352
May 2025

Convergence Analysis of OT-Flow for Sample Generation

Yang Jing[1,*] and Lei Li[1,2,3]

[1] *School of Mathematical Sciences, Shanghai Jiao Tong University, Shanghai 200240, P.R.China*
[2] *Institute of Natural Sciences, MOE-LSC, Shanghai Jiao Tong University, Shanghai 200240, P.R.China*
[3] *Shanghai Artificial Intelligence Laboratory, P.R.China*

Received 27 September 2024; Accepted (in revised version) 5 January 2025

Abstract. Deep generative models aim to learn the underlying distribution of data and generate new ones. Despite the diversity of generative models and their high-quality generation performance in practice, most of them lack rigorous theoretical convergence proofs. In this work, we aim to establish some convergence results for OT-Flow, one of the deep generative models. First, by reformulating the framework of OT-Flow model, we establish the Γ-convergence of the formulation of OT-Flow to the corresponding optimal transport (OT) problem as the regularization term parameter α goes to infinity. Second, since the loss function will be approximated by Monte Carlo method in training, we established the convergence between the discrete loss function and the continuous one when the sample number N goes to infinity as well. Meanwhile, the approximation capability of the neural network provides an upper bound for the discrete loss function of the minimizers. The proofs in both aspects provide convincing assurances for the stability of OT-Flow.

AMS subject classifications: 49Q22, 68T07

Key words: Generative models, continuous normalizing flows, OT-Flow, Benamou-Brenier functional, Γ-convergence.

1. Introduction

Deep generative models [15, 21, 23, 33] are increasingly being adopted as the preferred methodology across various tasks due to their impressive performance, including solving inverse problems [7], image generation [10], text-to-image [32] and video generation [25]. The widely-used frameworks include diffusion probabilistic models

*Corresponding author. *Email addresses:* sharkjingyang@sjtu.edu.cn (Y. Jing), leili2010@sjtu.edu.cn (L. Li)

(DPMs) [3, 17, 36], continuous normalizing flows (CNFs) [8, 16], variational auto-encoders (VAEs) [21, 23] and generative adversarial networks (GANs) [2, 15]. Among above four popular frameworks, CNFs are characterized by continuous-time ordinary differential equations (ODEs), and DPMs utilize stochastic differential equations (SDEs) as their backbone. Through DPMs and CNFs, samples evolve from data points to Gaussian distribution in the forward process and gradually remove noise to generate samples in the backward process. In comparison with GANs and VAEs, samples of DPMs and CNFs are generated in smoother ways, not only achieving superior sample quality but also enabling exact likelihood computation. Despite the diversity of generative models and their outstanding performance in downstream tasks, the mathematical principles behind the models and rigorous convergence proofs are developed far behind the rapid iteration of the models. In this paper, our focus lies in establishing convergence results for OT-Flow, which stands as one of the practical CNFs. Such convergence analysis ensures stability during the training and aids in comprehending the underlying mechanisms of the model.

The continuous normalizing flows (CNFs) are a class of sample generative models based on particle transportation purely. The CNFs aim to build continuous and invertible mappings between an arbitrary distribution ρ_0 and standard normal distribution ρ_1 by setting the velocity field as an output of neural network. In particular, for a given time T, one is trying to obtain a mapping $z : \mathbb{R}^d \times [0, T] \to \mathbb{R}^d$, which defines a continuous evolution $x \mapsto z(x, t)$ of every $x \in \mathbb{R}^d$. Then the density $\rho(z(x, t), t)$ satisfies

$$\log \rho_0(x) = \log \rho\big(z(x,t),t\big) + \log|\det \nabla z(x,t)| \quad \text{for all} \quad x \in \mathbb{R}^d. \tag{1.1}$$

Especially at time T we have

$$\log \rho_0(x) = \log \rho_1\big(z(x,T),T\big) + \log|\det \nabla z(x,T)|.$$

Define

$$\ell(x,t) := \log|\det \nabla z(x,t)|,$$

then $z(x, t)$ and $\ell(x, t)$ satisfy the following ODE system:

$$\partial_t \begin{bmatrix} z(x,t) \\ \ell(x,t) \end{bmatrix} = \begin{bmatrix} v(z(x,t),t;\boldsymbol{\theta}) \\ \operatorname{tr}(\nabla v(z(x,t),t;\boldsymbol{\theta})) \end{bmatrix}, \quad \begin{bmatrix} z(x,0) \\ \ell(x,0) \end{bmatrix} = \begin{bmatrix} x \\ 0 \end{bmatrix}. \tag{1.2}$$

To train the dynamics, CNFs minimize the expected negative log-likelihood given by the right-hand-side in (1.1), or equivalently the KL divergence between target distribution and final distribution under the constraint (1.2) [16, 31, 33]

$$J = \mathbb{KL}\big[\rho(\boldsymbol{z}(\boldsymbol{x},T))\|\rho_1(\boldsymbol{z}(\boldsymbol{x},T))\big]. \tag{1.3}$$

For convenience we solve (1.2) together to obtain the change of ρ, which will lead to a more efficient estimation of density.

From the ODE system (1.2), we can see that once the velocity field is learned, one can track the evolution of density and invert the transport map by running the ODE

backward. The invertibility of CNFs grants us access to estimate the density of the sample space, which can be employed for Bayesian inference tasks [11, 22, 30].

In general, the velocity field that transforms a given probability measure to a target one is not unique in the formulation of CNFs. One should observe the parallels between CNFs and the dynamic formulation of Wasserstein distance, where both involve the evolution of probability distributions through mass transportation under a velocity field. It is natural to incorporate optimal transport concepts into CNFs to enhance algorithm performance. Finlay *et al.* [13] pioneered the introduction of optimal transportation regularization in normalizing flows, while Onken *et al.* [29] subsequently proposed OT-Flow as an enhanced version of CNFs, which leverages optimal transport theory to regularize the CNFs and enforce straight trajectories that are easier for numerical integration. OT-Flow may be preferred in applications due to the particles' straight-line paths and trajectories avoiding intersections. Consequently, such a model is expected to improve its invertibility and generation efficiency.

The optimal transport was first introduced by Monge [28] in 1781 and was relaxed by Kantorovich [19]. The optimal transport theory actually provides a specific way to transform the measure μ to ν with minimum transportation cost. In particular, let $\Omega \subset \mathbb{R}^d$. Given two distributions $\mu, \nu \in \mathcal{P}(\Omega)$, where $\mathcal{P}(\Omega)$ is the set of all probability measures on Ω, one can define the Wasserstein p-distance ($p \geq 1$) between μ and ν

$$W_p(\mu, \nu) = \left(\inf_{\gamma \in \Pi(\mu,\nu)} \int |x - y|^p d\gamma \right)^{1/p}, \tag{1.4}$$

where $\Pi(\mu, \nu)$ is the set of transport plans, i.e. a joint measure on $X \times Y$, with marginal distribution μ and ν. Define the space

$$\mathcal{W}_p := \left\{ \mu \in \mathcal{P}(\Omega) : \int_\Omega |x|^p \mu(dx) < \infty \right\}. \tag{1.5}$$

Then $(\mathcal{W}_p, W_p)$ is a complete metric space. Wasserstein distance stands out as a prominent choice among metrics due to its ability to quantify dissimilarity between two distributions, even in cases where one or both distributions consist of discrete data samples with disjoint supports, which can be applied to traditional models for improvements [29, 34]. On a convex and compact domain Ω, the Wasserstein distance W_p admits the following dynamic Benamou-Brenier formulation [35, Chapter 5, Theorem 5.28]. Let μ and ν are two probability distribution on Ω and are absolutely continuous with respect to the Lebesgue measure, and v_t is a vector field on Ω,

$$\begin{aligned} &W_p^p(\mu, \nu) \\ &= \min_{\rho, v} \left\{ \int_0^1 \|v_t\|^p_{L^p(\rho)} \, dt : \partial_t \rho_t + \nabla \cdot (\rho_t v_t) = 0, v \cdot n|_{\partial\Omega} = 0, \rho_0 = \mu, \rho_1 = \nu \right\}, \end{aligned} \tag{1.6}$$

where

$$\|v_t\|^p_{L^p(\rho_t)} = \int_\Omega |v_t(x)|^p \rho_t(dx).$$

Clearly, the optimal velocity field v_t in this problem is one of the candidates for CNFs and could be optimal in certain sense. The cost in the OT-Flow to train the velocity field is then given as follows:

$$J = \mathbb{KL}\big[\rho(x,T)\|\rho_1(x)\big] + \frac{2}{\alpha}\mathbb{E}_{\rho_0}\left[\int_0^T \frac{1}{2}|v(z(x,t),t)|^2 dt\right]. \tag{1.7}$$

Here α is a hyperparameter to balance KL divergence and trajectory penalty. Note that even when Ω is not compact or not convex when there is some possibility that dynamic formulation is not strictly equal to the Wasserstein distance, such a formulation could still be beneficial since the dynamical formulation itself corresponds to some metric.

The KL divergence term in (1.7) serves as a soft terminal constraint, which enforces the terminal distribution $\rho(x,T)$ transported by velocity field to get close to ρ_1. The second term is related to the dynamic Benamou-Brenier formulation of W_2 distance in optimal transport theory, which can also be regarded as a penalty of the squared arc-length of the trajectories. Ideally if the KL divergence term is zero, minimizing the cost function is equivalent to minimizing W_2 distance and solving the optimal velocity field, which will encourage straight trajectory.

OT-Flow proves practical in sample generation, gaining an advantage over previous CNFs by integrating optimal transport concepts. One may find it interesting to explore the relationship between the velocity field solved by neural networks in OT-Flow and the classical solutions of OT problems. In our work, we conduct a convergence analysis for OT-Flow. More specifically, our convergence analysis mainly contain two parts. In the first part, we reformulate OT-Flow and classical OT problems into continuous optimization problems with similar form. Subsequently, we demonstrate that OT-Flow Γ-converges to OT as the regularization coefficient $\alpha \to \infty$, indicating the minimizers of OT-Flow will converges to ones of classical OT problems. One should notice that OT-Flow uses data samples to approximate the equivalent loss functional. In the second part, we illustrate that as with an increasing dataset size, the minimizers of whose loss functional is approximated through Monte Carlo method, will eventually converge to theoretical solutions with a sufficient neural network approximation capability.

The rest of the paper is organized as follows. Section 2 provides some setup and notations to formulate the problem. We provide an overview of Γ-convergence as well, which serves as a fundamental tool in our analysis. Then we establish a convergence analysis between OT-Flow and classical OT problems in Section 3. In Section 4, we consider the convergence of the minimizers with respect to sample number $N \to \infty$. We conclude the work and make a discussion in Section 5.

2. Setup and notations

Recall the optimization problem of OT-Flow in (1.7). The KL divergence (relative entropy) between two probability measure μ and ν on $\mathbb{R}^d$ is defined by

$$\mathbb{KL}\left[\mu||\nu\right] = \begin{cases} \displaystyle\int_D \log\left(\frac{d\mu}{d\nu}\right) d\mu, & \text{if} \quad \mu \ll \nu, \\ \infty, & \text{otherwise}, \end{cases} \tag{2.1}$$

where $d\mu/d\nu$ denotes the Radon-Nikodym derivative of μ with respect to ν. Note that the KL divergence is non-negative by Jensen's inequality and achieves zero only if $\mu = \nu$. Moreover, it is a convex functional with respect to either argument.

Assume $D \subset \mathbb{R}^d$ is a bounded domain with smooth boundary. To obtain the full description of the mathematical problems, one also needs to specify the no-flux boundary condition $v \cdot n = 0$ on ∂D by the physical significance. Hence, the optimization problem of OT-Flow becomes

$$\begin{aligned} \min_{\rho,v} \quad & \mathbb{KL}\big[\rho(x,T)||\rho_1(x)\big] + \frac{1}{\alpha}\int_0^T \int_D \rho|v|^2 dxdt, \\ \text{s.t.} \quad & \partial_t \rho + \nabla \cdot (\rho v) = 0 \quad \text{in } D \times [0,T], \\ \text{with} \quad & \rho(x,0) = \rho_0(x), \\ & v(x,t) \cdot n = 0 \qquad \text{on } \partial D \times [0,T]. \end{aligned} \tag{2.2}$$

However, the minimization problem above in the variables (ρ, v) has nonlinear constraints. It has been deemed advantageous to transition the variables from (ρ, v) into (ρ, m), where $m = \rho v$. Thus, the full description of the optimization problem associated with OT-Flow is presented as follows:

$$\begin{aligned} \min_{\rho,m} \quad & \mathbb{KL}\big[\rho(x,T)||\rho_1(x)\big] + \frac{1}{\alpha}\int_0^T \int_D \frac{|m|^2}{\rho} dxdt, \\ \text{s.t.} \quad & \partial_t \rho + \nabla \cdot m = 0 \quad \text{in } D \times [0,T], \\ \text{with} \quad & \rho(x,0) = \rho_0(x), \\ & m(x,t) \cdot n = 0 \qquad \text{on } \partial D \times [0,T]. \end{aligned} \tag{2.3}$$

Correspondingly, the optimal transport problem then becomes

$$\begin{aligned} \min_{\rho,m} \quad & \int_0^T \int_D \frac{|m|^2}{\rho} dxdt, \\ \text{s.t.} \quad & \partial_t \rho + \nabla \cdot m = 0 \quad \text{in } D \times [0,T], \\ \text{with} \quad & \rho(x,0) = \rho_0(x), \quad \rho(x,T) = \rho_1(x), \\ & m(x,t) \cdot n = 0 \qquad \text{on } \partial D \times [0,T]. \end{aligned} \tag{2.4}$$

In Section 3, we will first study the convergence of (2.3) to (2.4). We will allow m to be measures and the exact meaning of $|m|^2/\rho$ will be explained in Section 3.2. Moreover, the working space we choose in this work is

$$\rho \in L^1\big([0,T]; \mathcal{W}_2(D)\big), \quad m \in L^1\big([0,T]; \mathscr{M}(D)\big). \tag{2.5}$$

Then, the problems (2.3)-(2.4) become convex optimization problems, and the corresponding functionals have the lower semi-continuity with respect to the weak convergence (see Section 3.2 for details).

The main analytical tool employed in our study is the Γ-convergence [5]. Here, let us briefly introduce the relevant concepts. We will collect some basics of Γ-convergence, which is commonly used to investigate the convergence of the optimization problems and their optimizers. We first recall the definitions of the usual Γ-convergence.

Definition 2.1. *Let X be a topological space. Let (f_n) be a sequence of functionals on X. Define*

$$\begin{aligned} \Gamma\text{-}\limsup_{n\to\infty} f_n(x) &= \sup_{N_x} \limsup_{n\to\infty} \inf_{y\in N_x} f_n(y), \\ \Gamma\text{-}\liminf_{n\to\infty} f_n(x) &= \sup_{N_x} \liminf_{n\to\infty} \inf_{y\in N_x} f_n(y), \end{aligned} \tag{2.6}$$

where N_x ranges over all the neighborhoods of x. If there exists a functional f defined on X such that

$$\Gamma\text{-}\limsup_{n\to\infty} f_n = \Gamma\text{-}\liminf_{n\to\infty} f_n = f, \tag{2.7}$$

then we say the sequence (f_n) Γ-converges to f.

Proposition 2.1. *Any cluster point of the minimizers of a Γ-convergent sequence (f_n) is a minimizer of the corresponding Γ-limit functional f.*

Thus, Γ-convergence serves as an ideal tool to characterize the convergence of the minimizers of a sequence of optimization problems. Particularly, one can employ Γ-convergence to investigate the asymptotic behavior of neural network solutions as the regularization coefficient approaches infinity or approaches zero.

Direct verification of Γ-convergence using Definition 2.1 often proves to be challenging in many instances. Some refined versions of the Γ-convergence were proposed in [6], which are known as Γ_{seq}-convergence. There exist many notions of Γ_{seq}-convergence. In this discussion, we introduce two such formulations, which will be instrumental in establishing a Γ-convergence analysis from OT-Flow to OT, as detailed in Section 3. In particular, we will consider

$$\Gamma_{\text{seq}}(\mathbb{N}^+, X^-)\lim_n f_n := \inf_{x^n\to x} \limsup_{n\to\infty} f_n(x_n), \tag{2.8}$$

$$\Gamma_{\text{seq}}(\mathbb{N}^-, X^-)\lim_n f_n := \inf_{x^n\to x} \liminf_{n\to\infty} f_n(x_n). \tag{2.9}$$

Here, the $\inf_{x^n\to x}$ means the infimum is taken with respect to all sequences $\{x_n\}$ that converge to x. The following proposition gives the closed relationship between Γ_{seq}-convergence and the usual Γ-convergence. We will omit the proof and a more rigorous proof can be found in [40].

Proposition 2.2. *Suppose that X is a first-countable topological space. It holds that*

$$\begin{aligned} \inf_{x^n\to x} \limsup_{n\to\infty} f_n(x_n) &= \Gamma\text{-}\limsup_{n\to\infty} f_n, \\ \inf_{x^n\to x} \liminf_{n\to\infty} f_n(x_n) &= \Gamma\text{-}\liminf_{n\to\infty} f_n. \end{aligned} \tag{2.10}$$

Consequently, if

$$\inf_{x^n\to x}\limsup_{n\to\infty} f_n(x_n) = \inf_{x^n\to x}\liminf_{n\to\infty} f_n(x_n) =: f \tag{2.11}$$

exists, then (f_n) Γ*-converges to* f.

As mentioned in the introduction, OT-Flow utilizes a neural network to parameterize the velocity field and optimize the cost functional in (1.7). However, training the model directly with the cost functional is intractable as we only have data samples from ρ_0. To address this, one is supposed to rewrite cost functional in the form of expectation over ρ_0 and approximate the cost functional using Monte Carlo method. According to [29], one may simplify the KL divergence term with density relationship (1.1) and drop constant in the formulation. The final cost function J of OT-Flow gives as follows:

$$\begin{aligned} J &= \mathbb{E}_{\rho_0(x)}\left[C(x,T) + \frac{2}{\alpha}L(x,T)\right], \\ C(x,T) &= -\ell(x,T) + \frac{1}{2}|\boldsymbol{z}(\boldsymbol{x},T)|^2 + \frac{d}{2}\log(2\pi), \\ L(x,T) &= \int_0^T \frac{1}{2}|v(z(x,t),t)|^2 dt. \end{aligned} \tag{2.12}$$

In particular, from the Pontryagin maximum principle [12], there exists a potential function $\Phi : \mathbb{R}^d \times [0,T] \to \mathbb{R}$ such that

$$v(x,t) = -\nabla\Phi(x,t) \tag{2.13}$$

OT-Flow parameterize Φ with a neural network instead of v in practice. We will concentrate on using neural networks to directly approximate the velocity field for the sake of notation convenience, which will have no impact on our analysis. One should remember that the discretion of cost function will introduce variations to the minimizers as well. Consequently, we will demonstrate that as the sample number $N \to \infty$, the optimal solutions of neural network will converges to the theoretical minimizers of the cost functional in Section 4.

3. Convergence from OT-Flow problem to optimal transport problem

In this section, we will consider the OT-Flow and the optimal transport problems for given $\rho_0 \in \mathcal{P}(D)$ and $\rho_1 \in \mathcal{P}(D)$. For the purpose of facilitating the proof, we reformulate the optimal transport problem (2.4) into the following form to ensure consistency of the constraints:

$$\begin{aligned} \min_{\rho,m} \quad & \int_0^T \int_D \frac{|m|^2}{\rho} dx dt + \mathbf{1}_E, \\ \text{s.t.} \quad & \partial_t \rho + \nabla \cdot m = 0 \quad \text{in } D \times [0,T], \\ \text{with} \quad & \rho(x,0) = \rho_0(x), \\ & m(x,t)\cdot n = 0 \quad \text{on } \partial D \times [0,T], \end{aligned} \tag{3.1}$$

DOI https://doi.org/10.4208/nmtma.OA-2024-0114

where E is the set of the terminal constraints $\rho(x,T)=\rho_1(x)$. Since ρ is required to be integrable so far, rigorous definition of E will be given later. Moreover, $\mathbf{1}_E(x)$ is the indicator function for a set E in X

$$\mathbf{1}_E(x)=\begin{cases}0, & \text{if} \quad x\in E,\\ +\infty, & \text{otherwise.}\end{cases} \tag{3.2}$$

The OT-Flow problem (2.3) can be equivalently written as

$$\begin{aligned}\min_{\rho,m} \quad & \int_0^T\int_D \frac{|m|^2}{\rho}dxdt+\alpha\mathbb{KL}\big[\rho(x,T)||\rho_1(x)\big],\\ \text{s.t.} \quad & \partial_t\rho+\nabla\cdot m=0 \quad \text{in } D\times[0,T],\\ \text{with} \quad & \rho(x,0)=\rho_0(x),\\ & m(x,t)\cdot n=0 \quad \text{on } \partial D\times[0,T].\end{aligned} \tag{3.3}$$

We adopt this formulation for its closer alignment with the optimal transport problem (3.1).

Our goal is to prove that the minimization problem (3.3) is $\Gamma-$convergent to the minimization problem (3.1) when $\alpha\to\infty$. Subsequently, it is necessary to rigorously define the framework of the problems, including the topology of the space and the meaning of the constraints etc.

3.1. The topological properties of the working spaces

We will concentrate on the properties of the spaces we will work on

$$\begin{aligned}X&=L^1\big([0,T];\mathcal{W}_2(D)\big):=\left\{\rho:[0,T]\to\mathcal{W}_2(D)\middle|\int_0^T\int_D|x|^2\rho_t(dx)dt<\infty\right\},\\ Y&=L^1\big([0,T];\mathscr{M}^d(D)\big):=\left\{m:[0,T]\to\mathscr{M}^d(D)\middle|\int_0^T\|m\|_D dt<\infty\right\}.\end{aligned} \tag{3.4}$$

Here, $\mathscr{M}(D)$ is the set of signed measures on D and $\mathscr{M}^d(D)$ is the set of vector-valued signed measures on D with d components. The notation $\|m\|_D$ indicates the total variation norm which we explain below.

Let C_b be the set of all continuous functions on D (the subindex "b" indicates that it is bounded, which is natural since continuous functions on bounded set are bounded). Let C_b^1 be the set of all continuously differentiable functions on D and the first order derivatives are bounded on D. Without explicitly stated, C_b is equipped with the uniform convergence norm. Under the uniform convergence norm

$$\|\varphi\|:=\|\varphi\|_\infty=\sup_{x\in D}|\varphi(x)|, \tag{3.5}$$

C_b^1 is a dense subset of C_b (of course, if one considers the norm $\|\varphi\|_\infty+\|D\varphi\|_\infty$, C_b^1 itself is complete).

For a measure $\mu \in \mathscr{M}(D)$, the total variation norm is the dual operator norm over C_b

$$\|\mu\|_D := \sup_{\varphi \in C_b, \|\varphi\| \le 1} \int_D \varphi \mu(dx). \tag{3.6}$$

Similarly, if f is a vector-valued measure, the total variation norm is defined by

$$\|f\|_D := \sup \left\{ \int_D g \cdot f(dx) : g : D \to \mathbb{R}^d \text{ continuous}, \ \sup_x |g(x)|_2 \le 1 \right\}. \tag{3.7}$$

Here, $|g(x)|_2$ denotes the Euclidean length of g. If μ and f are absolutely continuous to some non-negative measure λ (for example, the Lebesgue measure dx), then the total variation norms for the scalar measure μ and vector valued measure f on D are as follows:

$$\|\mu\|_D = \int_D |\mu(x)| \, d\lambda, \quad \|f\|_D = \int_D |f(x)|_2 \, d\lambda, \tag{3.8}$$

where we abuse the notations for convenience and understand $\mu(x) = d\mu/d\lambda$ and $f(x) = df/d\lambda$ as the densities with respect to λ. The notation $|f(x)|_2$ is the Euclidean length of the vector $f(x)$.

In our case, when ρ and m are curves in $\mathcal{W}_2(D)$ and $\mathscr{M}^d(D)$, the total variation norms could be taken for each $t \in [0,T]$. In our case ρ is a probability measure, so $\|\rho_t\|_D = 1$ for each t, and thus $\int |x|^2 \rho(dx) < \infty$ since D is bounded. Then, for every measurable curve taking values in $\mathcal{P}(D)$, one then has

$$\|\rho\| = \int_0^T \|\rho(t)\|_D dt = T, \quad \int_0^T \int_D |x|^2 \rho(dx) \, dt < \infty. \tag{3.9}$$

The norm for $m \in Y$ is then

$$\|m\| = \int_0^T \|m(t)\|_D dt. \tag{3.10}$$

Similar as in [40], we equip X and Y with the weak topology respectively: we say $\rho_n \Rightarrow \rho$ in X if

$$\int_0^T \int_D \varphi d\rho_n \ \to \ \int_0^T \int_D \varphi d\rho, \quad \forall \varphi \in C_b(\bar{D} \times [0,T]; \mathbb{R}), \tag{3.11}$$

and $m_n \Rightarrow m$ in Y if

$$\int_0^T \int_D g \cdot dm_n \ \to \ \int_0^T \int_D g \cdot dm, \quad \forall g \in C_b(\bar{D} \times [0,T]; \mathbb{R}^d). \tag{3.12}$$

(Note that there is a typo in [40, p. 758, Eq. (3.18)] where C_b^1 should be C_b.) The weak topology used here as the dual of the class of C_b functions is standard in literature. If both m_n and ρ_n uniformly bounded in total variation, then one can replace the test functions from C_b to C_b^1 in (3.13) since the set C_b^1 is dense in C_b under the topology of uniform convergence.

With the weak topology introduced, we observe the following properties of the spaces.

Proposition 3.1. *Both X and Y are closed subspaces of the signed measures and vector-valued signed measures with respect to the weak topology in the sense that*

(i) If $\rho^{(n)} \in X$ and there is some $\rho : [0,T] \to \mathscr{M}(D)$ such that $\rho^{(n)} \Rightarrow \rho$, then $\rho \in X$.

(ii) If $m^{(n)} \in Y$ and there is some $m : [0,T] \to \mathscr{M}^d(D)$ such that $m^{(n)} \Rightarrow m$, then $m \in Y$.

Proof. Consider a sequence $\{\rho^{(n)}\} \in X$ such that $\rho_n \Rightarrow \rho$. First, since $C_b(\bar{D} \times [0,T])$ is a complete metric space, one can obtain by the Banach-Steinhaus theorem that

$$\sup_n \|\rho^{(n)}\| < \infty.$$

Then, one finds that

$$\left| \int_0^T \int_D \varphi \rho(dx)dt \right| = \lim_{n\to\infty} \left| \int_0^T \int_D \varphi \rho^{(n)}(dx)dt \right| \leq \liminf_{n\to\infty} \|\rho^{(n)}\| \|\varphi\|.$$

This indicates that $\|\rho\|_D < +\infty$. Since $C_b(\bar{D} \times [0,T])$ is separable, we can then find a countable dense set $\{\varphi_m\} \subset C_b$ such that for all φ_m and almost every $t \in [0,T]$ such that

$$\int_D \varphi_m(t)\rho_t^{(n)}(dx) \;\; \to \;\; \int_D \varphi_m(t)\rho_t(dx).$$

This indicates that in fact for all $C_b(D)$ functions, this weak convergence holds for these t. Consequently, $\rho_t \in \mathcal{P}(D)$ for a.e. $t \in [0,T]$. Hence, we find a version of ρ such that $\rho_t \in \mathcal{P}(D)$ for all t. Since D is bounded, one then finds that

$$\int_0^T \int_D |x|^2 \rho(dx)dt < +\infty.$$

Thus, $\rho \in L^1([0,T]; \mathcal{W}_2(D))$. For the space Y, it is straightforward by the Banach-Steinhaus theorem and the argument similarly above. Hence, both X and Y are closed under weak topology.

3.2. Lower semi-continuity of the functionals

First of all, we equip the product space $X \times Y$ for (ρ, m) with the product topology and then the weak topology $(\rho_n, m_n) \Rightarrow (\rho, m)$ is understood as: $\forall f \in C_b(\bar{D} \times [0,T]; \mathbb{R})$, $g \in C_b(\bar{D} \times [0,T]; \mathbb{R}^d)$, one has

$$\int_0^T \int_D f d\rho_n + \int_0^T \int_D g \cdot dm_n \;\; \to \;\; \int_0^T \int_D f d\rho + \int_0^T \int_D g \cdot dm. \tag{3.13}$$

Next, we also rewrite the continuity equation in the weak sense as dual of C_b^1, thereby incorporating the constraints into the working subspace. It should be noted

that, as of now, ρ is only guaranteed to be integrable and $\rho(x,T)$ is not well-defined. In particular, we can define the subspace $\mathcal{H}$ of $X \times Y$ as following:

$$\mathcal{H} := \Bigg\{(\rho,m) \in X \times Y : -\int_0^T \left(\int_D (\partial_t \varphi)\rho(dx) + \nabla\varphi \cdot m(dx)\right) dt \\ - \int_D \varphi(x,0)\rho_0(dx) = 0, \forall \varphi \in C_b^1(\bar{D} \times [0,T]), \varphi(\cdot,T) = 0\Bigg\}. \tag{3.14}$$

Clearly $\mathcal{H}$ is closed due to the fact that the constraints are linear. Similarly, E can now be rigorously defined as follows:

$$E := \Bigg\{(\rho,m) \in X \times Y : -\int_0^T \left(\int_D (\partial_t \varphi)\rho(dx) + \nabla\varphi \cdot m(dx)\right) dt \\ + \int_D \varphi(x,T)\rho_1(dx) - \int_D \varphi(x,0)\rho_0(dx) = 0, \forall \varphi \in C_b^1(\bar{D} \times [0,T])\Bigg\}. \tag{3.15}$$

One should note that E is a closed subset of $\mathcal{H}$.

If there is a version of $t \mapsto \rho_t$ that is continuous at T, one can define the limit

$$\bar{\rho}(x,T) := \lim_{\delta \to 0} \frac{1}{\delta} \int_{T-\delta}^T \rho_s \, ds, \tag{3.16}$$

and the functional

$$G(\rho,\rho_1) := \begin{cases} \mathbb{KL}\big[\bar{\rho}(x,T)||\rho_1(x)\big], & \text{if there is a version of } t \mapsto \rho_t \\ & \text{that is continuous at } T, \\ \infty, & \text{otherwise.} \end{cases} \tag{3.17}$$

$G(\rho,\rho_1)$ actually corresponds to the KL divergence term if we later focus on the feasible points, which ensures the continuity of ρ. Corresponding to the optimization problems (3.1) and (3.3), one naturally aims to introduce the functionals F_α and F_∞ on $\mathcal{H}$ by

$$\begin{aligned} F_\alpha(\rho,m) &= \int_0^T \int_D \frac{|m|^2}{\rho} dxdt + \alpha G(\rho,\rho_1), \\ F_\infty(\rho,m) &= \int_0^T \int_D \frac{|m|^2}{\rho} dxdt + \mathbf{1}_E. \end{aligned} \tag{3.18}$$

Here we have the expressions like $|m|^2/\rho$ for measures ρ and m, which must be defined. For this purpose, we recall the Benamou-Brenier functionals for $\rho \in \mathscr{M}(\Omega)$ and $m \in \mathscr{M}^d(\Omega)$

$$\mathscr{B}_p(\rho,m) := \sup\left\{\int_X a d\rho + \int_X b \cdot dm : (a,b) \in C_b(\Omega; K_q)\right\}, \tag{3.19}$$

where

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

It has the following characterizations.

Lemma 3.1. *The functional $\mathscr{B}_p$ is convex and lower semi-continuous on the space $\mathscr{M}(\Omega)\times \mathscr{M}^d(\Omega)$ for the weak convergence. Moreover, the following properties hold:*

- $\mathscr{B}_p(\rho, m) \geq 0$,
- *if both ρ and m are absolutely continuous with respect to a same positive measure λ on Ω, we can write*

$$\mathscr{B}_p(\rho, m) = \int_X f_p\big(\rho(x), m(x)\big) d\lambda(x),$$

where we identify $\rho(x)$ and $m(x)$ are the densities with respect to λ, and $f_p : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is defined as

$$f_p(t,x) := \sup_{(a,b)\in K_q} (at + b\cdot x) = \begin{cases} \dfrac{1}{p}\dfrac{|x|^p}{t^{p-1}}, & \text{if} \quad t > 0, \\ 0, & \text{if} \quad t = 0, \quad x = 0, \\ +\infty, & \text{if} \quad t = 0, \quad x \neq 0 \quad \text{or} \quad t < 0, \end{cases} \tag{3.20}$$

- $\mathscr{B}_p(\rho, m) < +\infty$ *only if* $\rho \geq 0$ *and* $m \ll \rho$,
- *for* $\rho \geq 0$ *and* $m \ll \rho$, *we have* $m = v \cdot \rho$ *and*

$$\mathscr{B}_p(\rho, m) = \int \frac{1}{p}|v|^p d\rho.$$

The detailed proof can be found in [35, Chapter 5, Proposition 5.18].

With the Benamou-Brenier functionals, we find that the rigorous definitions of the goal functionals should be

$$\begin{aligned} F_\alpha(\rho, m) &= \int_0^T \mathscr{B}_2(\rho, m)dt + \alpha G(\rho, \rho_1), \quad \forall (\rho, m) \in \mathcal{H}, \\ F_\infty(\rho, m) &= \int_0^T \mathscr{B}_2(\rho, m)dt + \mathbf{1}_E, \qquad\quad \forall (\rho, m) \in \mathcal{H}. \end{aligned} \tag{3.21}$$

Then, the OT-Flow problem (2.3) is formulated by

$$\min_{(\rho,m)\in\mathcal{H}} F_\alpha(\rho, m), \tag{3.22}$$

and the OT problem (2.4) is given by

$$\min_{(\rho,m)\in\mathcal{H}} F_\infty(\rho, m). \tag{3.23}$$

We note that the functionals just introduced enjoy good properties.

Proposition 3.2. *Both F_α and F_∞ are convex and lower semi-continuous.*

Proof. These two properties follow directly from those for the Benamou-Brenier functional and the KL divergence.

The convexity of these functionals are well-known. Here, we sketch the verification of the lower semi-continuity for the convenience of the readers (as the lower semi-continuity is essential in this work). We take F_α as the example.

If $\int_0^T \mathscr{B}_2(\rho, m)dt = +\infty$, then for each $M > 0$, there is some $(\bar{a}, \bar{b}) \in C_b(\bar{D} \times [0, T])$ such that

$$\int_0^T \int_D (\bar{a}\rho + \bar{b} \cdot m)\, dxdt > M.$$

Consider any sequence $(\rho^n, m^n) \Rightarrow (\rho, m)$, by the weak convergence, one then has

$$\lim_{n\to\infty} \int_0^T \int_D \left(\bar{a}\rho^n(dx) + \bar{b} \cdot m^n(dx)\right) dt > M.$$

This indicates that

$$\lim_{n\to\infty} \int_0^T \mathscr{B}_2(\rho^n, m^n)dt \geq \lim_{n\to\infty} \int_0^T \int_D \left(\bar{a}\rho^n(dx) + \bar{b} \cdot m^n(dx)\right) dt > M.$$

Since M is arbitrary, one then has

$$\lim_{n\to\infty} \int_0^T \mathscr{B}_2(\rho^n, m^n)dt = +\infty.$$

If

$$M := \int_0^T \mathscr{B}_2(\rho, m)dt < +\infty,$$

then for any $\epsilon > 0$, one may take $(\bar{a}, \bar{b}) \in C_b(\bar{D} \times [0, T])$ such that

$$\int_0^T \int_D (\bar{a}\rho + \bar{b} \cdot m)\, dxdt < M + \epsilon.$$

Consider any sequence $(\rho^n, m^n) \Rightarrow (\rho, m)$, by the weak convergence, one then has

$$\begin{aligned}
\int_0^T \mathscr{B}_2(\rho, m)dt &< \int_0^T \int_D \left(\bar{a}\rho(dx) + \bar{b} \cdot m(dx)\right)\, dt + \epsilon \\
&\leq \liminf_{n\to\infty} \int_0^T \int_D \left(\bar{a}\rho^n(dx) + \bar{b} \cdot m^n(dx)\right)\, dt + \epsilon \\
&\leq \liminf_{n\to\infty} \sup_{(a,b)\in C_b((D\times[0,T];K_2)} \int_0^T \int_D \left(a\rho^n(dx) + b \cdot m^n(dx)\right)\, dt + \epsilon \\
&= \liminf_{n\to\infty} \int_0^T \mathscr{B}_2(\rho^n, m^n)dt + \epsilon.
\end{aligned}$$

Since ϵ is arbitrary, the lower semi-continuity of the first part in F_α is verified.

The lower semi-continuity of KL divergence follows from a similar argument. The central equation to derive the lower semi-continuity of KL divergence is the (Fenchel) dual formulation of the KL [1, Lemma 9.4.4]

$$\mathbb{KL}[\mathbb{P}||\mathbb{Q}] = \sup_{h\in C_b(\mathbb{R}^d)} \left\{1 + \int h d\mathbb{P} - \int e^h d\mathbb{Q}\right\}. \tag{3.24}$$

Thus, KL divergence is lower semi-continuous with respect to $\mathbb{P}$ is a direct consequence of the fact that it is expressed as a supremum of linear functional. Suppose $\mathbb{P}^n \Rightarrow \mathbb{P}$, then for all ϵ, there exits some $\bar{h} \in C_b(\mathbb{R}^d)$ satisfies

$$\begin{aligned}
\mathbb{KL}[\mathbb{P}||\mathbb{Q}] &= \sup_{h\in C_b(\mathbb{R}^d)} \left\{1 + \int h d\mathbb{P} - \int e^h d\mathbb{Q}\right\} \\
&< 1 + \int \bar{h} d\mathbb{P} - \int e^{\bar{h}} d\mathbb{Q} + \epsilon \\
&\leq \liminf_{n\to\infty} \left\{1 + \int \bar{h} d\mathbb{P}^n - \int e^{\bar{h}} d\mathbb{Q}\right\} + \epsilon \\
&\leq \liminf_{n\to\infty} \sup_{h\in C_b(\mathbb{R}^d)} \left\{1 + \int h d\mathbb{P}^n - \int e^h d\mathbb{Q}\right\} + \epsilon.
\end{aligned}$$

Thus, we have $\mathbb{KL}[\mathbb{P}||\mathbb{Q}] \leq \liminf_{n\to\infty} \mathbb{KL}[\mathbb{P}^n||\mathbb{Q}]$, i.e. lower semi-continuity of KL divergence with respect to the first argument. Consequently, the convexity and the lower semi-continuity of G with respect to the first argument follow easily. □

Though the space we consider is simply L^1 in time. We find that the time regularity of ρ is actually good for feasible points.

Proposition 3.3. *Fix $\alpha > 0$ or $\alpha = \infty$. If (ρ, m) is a feasible solution of the optimization problem* (3.22) *or* (3.23)*, then there is a version of ρ such that $t \mapsto \rho_t$ is absolutely continuous in $\mathcal{W}_2(D)$. Moreover, $m_t \ll \rho_t$ and the Radon-Nikodym derivative*

$$v_t = \frac{dm_t}{d\rho_t}$$

satisfies that $v_t \in L^1([0,T]; L^2(\rho_t))$.

Proof. If (ρ, m) is a feasible solution for either F_α or F_∞, then

$$\int_0^T \mathscr{B}_2(\rho, m) dt < +\infty.$$

By Lemma 3.1, for there is a set E with Lebesgue measure zero such that $t \in [0,T] \setminus E$, it holds that

$$m_t \ll \rho_t,$$

and we can construct the corresponding velocity field by

$$v_t = \frac{dm_t}{d\rho_t}.$$

Then, by Lemma 3.1, if we take $\lambda = \rho_t$, one then has for these t that

$$\mathscr{B}_2(\rho, m) = \int_D |v_t|^2 \rho_t(dx).$$

Recall that $m \in L^1([0,1]; \mathscr{M}^d(D))$, then we can safely set $v = 0$ for $t \in E$ and modify the value of m on E freely. Then, $m = \rho v$ holds for all $t \in [0,T]$. This means that $v_t \in L^1([0,T]; L^2(\rho_t))$ and the constraint $\partial_t \rho + \nabla \cdot m = 0$ still holds for the modified version of m.

For the absolutely continuity of the feasible solution ρ then follows from the constraint $\partial_t \rho + \nabla \cdot m = 0$ in the weak sense (dual of C_b^1). One can find the rigorous proof in [35, Chapter 5, Theorem 5.14]. Here we provide a sketch of the proof. One can consider a segment of the curve, i.e. from ρ_t to ρ_{t+h}. A natural transport plan from ρ_t to ρ_{t+h} can be given by the curve driven by v_t

$$\gamma = (T_t, T_{t+h})_{\#}\rho_0,$$

where T_t is defined by

$$\frac{d}{dt}T_t(x) = v_t\big(T_t(x)\big).$$

This plan provides an upper bound for $W_2(\rho_t, \rho_{t+h})$

$$W_2\left(\rho_t, \rho_{t+h}\right) \leq \left(\int_{D\times D} |x-y|^2 d\gamma\right)^{1/2} = \left(\int_D \left|T_t(x) - T_{t+h}(x)\right|^2 d\rho_0\right)^{1/2}.$$

Using the fact that

$$\left|T_t(x) - T_{t+h}(x)\right|^2 \leq h \int_t^{t+h} \left|v_s\left(T_s(x)\right)\right|^2 ds,$$

one has

$$\frac{W_2\left(\rho_t, \rho_{t+h}\right)}{h} \leq \left(\frac{1}{h}\int_t^{t+h} \|v_s\|^2_{L^2(\rho_s)}\, ds\right)^{1/2}. \tag{3.25}$$

Since

$$\int_0^T \mathscr{B}_2(\rho, m) dt < +\infty$$

automatically indicates

$$\int_0^T \|v_t\|^2_{L^2(\rho_t)} dt < +\infty,$$

then (3.25) provides Lipschitz behavior for ρ_t, which directly leads to the fact ρ_t is absolutely continuous in $\mathcal{W}_2(D)$. □

3.3. Γ-convergence to optimal transport problems

In this section, we aim to prove the Γ-convergence of OT-Flow to the optimal transport problem as $\alpha \to \infty$ so that the solution of OT-Flow could converge to the solutions of the optimal transport problem. We first show that the solutions (minimizers) indeed exist.

Proposition 3.4. *Assume $D \subset \mathbb{R}^d$ is a bounded domain with smooth boundary, and F_α (resp. F_∞) has at least one feasible point over $\mathcal{H}$. Then,* (3.22) *(resp.* (3.23)*) has global minimizers over $\mathcal{H}$. Moreover, the minimizers, as feasible points, satisfy that ρ is absolutely continuous in $\mathcal{W}_2(D)$, and $m \ll \rho$.*

Proof. First of all, one should note that F_α is non-negative. Hence, it is clear that

$$F_\alpha^* = \inf_{(\rho,m)\in\mathcal{H}} F_\alpha(\rho, m) \geq 0.$$

With the assumption that one feasible point exists, $F_\alpha^* \in [0, \infty)$.

Consider a feasible minimizing sequence (ρ_n, m_n) such that $F_\alpha(\rho_n, m_n) \to F_\alpha^*$. Let $\rho_n(x)$ and $m_n(x)$ be the densities with respect to some common non-negative measure λ_t^n (for example, one can take $\lambda_t^n = \rho_n$ and then $\rho_n(x) = 1$). Then, $\sup_n F_\alpha(\rho_n, m_n) < \infty$ implies that

$$\sup_n \int_0^T \int_D \frac{|m_n(x)|^2}{\rho_n(x)} \lambda_t^n(dx)dt < +\infty.$$

By the Hölder inequality one then has that

$$\begin{aligned}\sup_n \|m_n\| &= \sup_n \int_0^T \int_D |m_n(x)| \lambda_t^n(dx)dt \\ &\leq \sup_n \left(\int_0^T \int_D \rho_n(x)\lambda_t^n(dx)dt \int_0^T \int_D \frac{|m_n(x)|^2}{\rho_n(x)} \lambda_t^n(dx)dt \right)^{1/2} < \infty.\end{aligned}$$

Hence, one has

$$\sup_n \|m_n\| + \|\rho_n\| < +\infty.$$

The Banach-Alaoglu thoeorem (the closed unit ball of the dual space of a normed vector space is compact in the weak star topology) tells us that there must be a weakly convergent subsequence. Together with the lower semi-continuity of F_α, the minimizer of F_α exists. Moreover, by Proposition 3.3, we conclude that the minimizer (ρ_α, m_α) of the functional F_α satisfies: $(\rho_\alpha)_t$ is absolutely continuous in $\mathcal{W}_2(D)$ and $m_\alpha \ll \rho_\alpha$. The argument for F_∞ is similar. □

Theorem 3.1. *Assume $D \subset \mathbb{R}^d$ is a bounded domain with smooth boundary. The following results hold about the OT-Flow and optimal transport problems:*

(i) For any $\rho_0, \rho_1 \in \mathcal{P}(D)$, F_α is Γ-convergent to F_∞ as $\alpha \to \infty$.

(ii) Suppose that both F_α and F_∞ have feasible points. Let (ρ^α, m^α) be an optimal solution to the corresponding minimization problem (3.22)*. Then, for any increasing sequence $\{\alpha_I\}$ going to infinity, where I is an index set, there exists a convergent subsequence $(\rho^{\alpha_k}, m^{\alpha_k}) \in \mathcal{H}$ with $\alpha_k \to +\infty$ such that the limit (ρ^∞, m^∞) is a solution of the minimization problem* (3.23)*.*

Proof. Since we consider feasible points, $t \mapsto \rho_t$ is continuous, one can replace G in F_α with KL divergence explicitly.

(i) We use Γ_{seq}-convergence mentioned in Proposition 2.2 to prove the Γ-convergence from F_α to F_∞. By definition, we need to verify: for any weakly convergent sequence $(\rho^\alpha, m^\alpha) \Rightarrow (\rho, m)$, one has

$$\begin{aligned} &\inf_{(\rho^\alpha, m^\alpha)\to(\rho,m)} \limsup_{\alpha\to\infty} F_\alpha(\rho^\alpha, m^\alpha) \\ &\le F_\infty(\rho, m) \le \inf_{(\rho^\alpha, m^\alpha)\to(\rho,m)} \liminf_{\alpha\to\infty} F_\alpha(\rho^\alpha, m^\alpha). \end{aligned} \tag{3.26}$$

Above statement is equivalent to prove

- $\exists(\rho^\alpha, m^\alpha) \Rightarrow (\rho, m)$, $F_\infty(\rho, m) \ge \limsup_{\alpha\to\infty} F_\alpha(\rho^\alpha, m^\alpha)$.
- $\forall(\rho^\alpha, m^\alpha) \Rightarrow (\rho, m)$, $F_\infty(\rho, m) \le \liminf_{\alpha\to\infty} F_\alpha(\rho^\alpha, m^\alpha)$.

For the first, one could take the constant sequence

$$(\rho^\alpha, m^\alpha) = (\rho, m).$$

Then one has

$$\limsup_{\alpha\to\infty} F_\alpha(\rho^\alpha, m^\alpha) = \limsup_{\alpha\to\infty} F_\alpha(\rho, m) \le F_\infty(\rho, m).$$

The above follows from the fact

$$\limsup_{\alpha\to\infty} \alpha \mathbb{KL}\big[\rho(x,T)||\rho_1(x)\big] \le \mathbf{1}_E,$$

which holds since

$$\mathbb{KL}\big[\rho(x,T)||\rho_1(x)\big] \ge 0$$

and the equality holds only when $\rho(x,T) = \rho_1(x)$.

For the second, for any sequence $(\rho^\alpha, m^\alpha) \Rightarrow (\rho, m)$, one needs to show that

$$F_\infty(\rho, m) \le \lim_{\alpha\to\infty} F_\alpha(\rho^\alpha, m^\alpha).$$

We consider the two parts in F_α and F_∞ separately. The required relation for the first part $\int_0^T \mathscr{B}_2(\rho, m)dt$ follows directly from the lower semi-continuity.

Considering the second part for the KL divergence, we can fix n. The lower semi-continuity of KL divergence gives

$$n\mathbb{KL}\big[\rho(x,T)||\rho_1(x)\big] \leq n\liminf_{\alpha\to\infty}\mathbb{KL}\big[\rho^\alpha(x,T)||\rho_1(x)\big].$$

Let $n\to+\infty$, the left side becomes the indicator function we want

$$\mathbf{1}_E \leq \lim_{n\to\infty} n\liminf_{\alpha\to\infty}\mathbb{KL}\big[\rho^\alpha(x,T)||\rho_1(x)\big] \leq \liminf_{\alpha\to\infty}\alpha\mathbb{KL}\big[\rho^\alpha(x,T)||\rho_1(x)\big].$$

Combining these two components consequently yields that

$$F_\infty(\rho,m) \leq \liminf_{\alpha\to\infty} F_\alpha(\rho^\alpha,m^\alpha).$$

Thus, the $\Gamma-$convergence from F_α to F_∞ has then been established.

(ii) Suppose (ρ^*,m^*) is a feasible solution of problem (3.23) and thus also a feasible point of (3.22). With the existence of feasible points, both F_α and F_∞ have global minimizers. It is then clear that

$$F_\alpha(\rho^\alpha,m^\alpha) \leq F_\alpha(\rho^*,m^*) = F_\infty(\rho^*,m^*).$$

This means that $F_\alpha(\rho^\alpha,m^\alpha)$ is uniformly bounded in α. Again, by Hölder inequality as demonstrated in Proposition 3.4, one has

$$\sup_\alpha \|m^\alpha\| + \|\rho^\alpha\| < +\infty.$$

By the Banach-Alaoglu theorem, we conclude that any bounded set in $X\times Y$ is pre-compact. Consequently, there is a subsequence $(\rho^{\alpha_k},m^{\alpha_k})\Rightarrow(\rho,m)$. Combining with $\Gamma-$convergence of the functionals, it follows that (ρ,m) is a minimizer of F_∞. □

For the existence of feasible points, one assumption is that both ρ_0 and ρ_1 are absolutely continuous with respect to the Lebesgue measure. Then according to (1.6), there is feasible point (ρ,m) with $\rho_T=\rho_1$ such that

$$W_2(\rho_0,\rho_1)^2 = \int_0^T \mathscr{B}_2(\rho,m)dt < \infty. \tag{3.27}$$

Hence, the feasible point exists. In the application, ρ_1 is a normal distribution, which is clearly absolutely continuous with respect to the Lebesgue measure. However, the data distribution ρ_0 is often singular (which may concentrate on some low-dimensional manifolds). The point is that one can always find a distribution $\tilde{\rho_0}$, which is absolutely continuous with respect to Lebesgue measure to approximate ρ_0. In specific tasks, we only have samples from ρ_0 for optimization. The training process later can learn a velocity field that automatically generates the approximation $\tilde{\rho_0}$ that is absolutely continuous with respect to the Lebesgue measure [18].

4. Convergence of Monte Carlo approximation in the large data limit

In practical applications, one only has access to data samples from ρ_0 to train the optimal velocity field. Hence, the loss functional (2.12) in the OT-Flow is not known exactly, where the expectation is replaced by the Monte Carlo method (empirical mean over the samples). Our goal in this part is to investigate the convergence of minimizers as the data size N increases to infinity, provided sufficient approximation capability of neural networks. We will neglect the error brought by training process and assume that the minimization problem can be solved exactly for convenience.

The research on the asymptotic behavior of minimizers in the large data limit is crucial in machine learning. It helps us gain a better understanding of the influence on the minimizers resulted from approximating loss functional by using data samples. As long as we are capable of providing an accurate estimation of the bound to control the error beyond the training sets, then we can guarantee some reliability on the generalisation of the neural network minimizers.

4.1. Setup and the Monte Carlo approximation

We first introduce some notations to define the finite dimensional spaces for the deep neural networks (see [26, 27] for similar discussions). A deep neural network $u : \mathbb{R}^d \to \mathbb{R}^{d'}$ with L layers is of the following form:

$$u(x;\theta) := F_L \circ \sigma \circ F_{L-1} \circ \cdots \circ \sigma \circ F_1(x). \tag{4.1}$$

The map $u(\cdot;\theta)$ is characterized by the hidden layers F_k, which are affine maps of the form

$$F_k x = W_k x + b_k, \quad \text{where} \quad W_k \in \mathbb{R}^{d_{k+1}\times d_k}, \quad b_k \in \mathbb{R}^{d_{k+1}}, \tag{4.2}$$

and σ is the activation function which we assume to be Lipschitz continuous. We use θ to represent collectively all the parameters of the network, namely $\{W_k, b_k\}$, $k = 1, \ldots, L$. The network is trained using an optimizer such as stochastic gradient descent (SGD) [24] or ADAM [20] commonly. In this paper, we will discuss the case of a multi-layer perceptron, but the results of convergence only rely on the approximation properties of the network, regardless of its precise architecture.

We denote the set of all functions u which can be given by networks with a given structure (fixed depth L and width in each layer) of the form (4.1) as

$$V_{\mathcal{N}} = \left\{ u : D \to \mathbb{R}^{d'} \mid \text{There exists a network (4.1) such that } u(x) = u(x;\theta) \right\}. \tag{4.3}$$

Let d_1 be the dimension of θ. Then, for each $\theta \in \mathbb{R}^{d_1}$, there is a corresponding $u \in V_{\mathcal{N}}$ such that $u(x) = u(x;\theta)$.

Lemma 4.1. *Assume D is a bounded domain and the neural network space $V_{\mathcal{N}}$ is fixed. If the activation function σ is assumed to be continuous, then given a sequence of network parameters $\{\theta_n\}_{n=1}^{\infty}$ that converges to θ in $\mathbb{R}^{d_1}$, $u(x;\theta_n)$ converges to $u(x;\theta)$ uniformly, or*

$$\lim_{n\to\infty} \sup_{x\in D} |u(x;\theta_n) - u(x;\theta)| = 0. \tag{4.4}$$

DOI https://doi.org/10.4208/nmtma.OA-2024-0114

Proof. By the specific form of the neural networks, it is clear that u is continuous with respect to (x,θ). If $\theta_n \to \theta$, then there exists $R>0$ such that

$$\sup_n |\theta_n| \le R, \quad |\theta| \le R.$$

Since the set $D\times B(0,R)$ is compact (bounded closed set in finitely dimensional space), then u is uniformly continuous on $D\times B(0,R)$. By the uniform continuity, it is easy to see that $u(\cdot;\theta_n)$ converges uniformly to $u(\cdot;\theta)$ for D bounded. □

Let us come back to the optimization problem for the OT-Flow problem (2.2). To solve this, one seeks the optimal field v using the neural networks (as mentioned, one may parametrize a scalar function Φ using the networks and set $v=-\nabla\Phi$). We will enforce the neural network for v to satisfy

$$v(x,t)\cdot n = 0, \quad x\in\partial D. \tag{4.5}$$

This condition may be obtained by post-processing of the output of the neural network. Using the network for v, one may find the trajectory of a particle $t\mapsto z(x,t)$ by solving (1.2). Using the flow map (trajectory), our goal is then to solve the following problem:

$$\min_{v\in V_{\mathcal{N}}} J, \tag{4.6}$$

where J is given as in (2.12). The training of the algorithm assumes knowledge of data only at a finite set of points $\{x_i\}_{i=1}^N$. Then, the expectation is replaced by the Monte Carlo approximation

$$J_N := \frac{1}{N}\sum_{i=1}^N \left[C(x_i,T)+\frac{2}{\alpha}L(x_i,T)\right]. \tag{4.7}$$

This formulation is then used to train v (find the optimal parameter $\theta\in\mathbb{R}^{d_1}$).

We are then concerned with the convergence as the number of data goes to infinity. However, studying the convergence of v is not convenient. Instead, the loss in (2.12) naturally defines a function

$$F_v(x) = \frac{1}{2}|\boldsymbol{z}(\boldsymbol{x},T)|^2 - \ell(x,T) + \frac{1}{\alpha}\int_0^T |v(z(x,t),t)|^2 dt + \frac{d}{2}\log(2\pi), \tag{4.8}$$

for a given velocity field v, where ℓ and z can be solved from ODE system (1.2). We will then investigate the convergence of F_v for the trained v. For this purpose, consider the set of all functions F_v with all possible v

$$F_{\mathcal{N}} := \{F_v : v\in V_{\mathcal{N}}\}. \tag{4.9}$$

The OT-Flow problem then corresponds to

$$\min_{F_v\in F_{\mathcal{N}}} \mathcal{E}(F_v) := \int_D F_v(x)\rho_0(dx), \tag{4.10}$$

where F_v denotes the loss functional for velocity v, ρ_0 is the data distribution.

The discrete loss functional can be defined as follows:

$$\min_{F_v \in F_{\mathcal{N}}} \mathcal{E}_N(F_v) = \frac{1}{N} \sum_{i=1}^{N} F_v(x_i). \tag{4.11}$$

Our goal is to investigate the convergence of the optimizers for (4.11) to that for (4.10) as $N \to \infty$.

The unboundedness of the parameters space still brings trouble. In practice, parameter clipping is a commonly employed technique in neural networks aimed at preventing the function $u(x;\theta)$ from exploding. Define the clipping parameter space

$$\Theta_R := \big\{\theta \in \mathbb{R}^{d_1} : |\theta| \le R\big\}. \tag{4.12}$$

With the clipped parameters, the function F_v defined in (4.8) has good features as stated below.

Lemma 4.2. *Assume D is a bounded domain. Fix the structure of the neural network space $V_{\mathcal{N}}$ with the activation function σ to be twice continuously differentiable such that the condition* (4.5) *is satisfied. Suppose the parameters $\theta \in \Theta_R$ for some fixed large $R > 0$. Then $v(x,t;\theta)$ is Lipschitz continuous and the Hessian $\nabla^2 v$ is bounded in the sense that*

$$\sup_{x\in D, \theta\in\Theta_R} \left|\frac{\partial^2 v}{\partial x_i \partial x_j}\right| < \infty, \quad \forall i,j. \tag{4.13}$$

Consequently, F_v is uniformly bounded and uniformly Lipschitz continuous with respect to x (uniform for $\theta \in \Theta_R$).

Proof. Since the architecture of the neural networks is fixed as in (4.2), both D and Θ_R are bounded, the Lipschitz continuity of $v(x,t;\theta)$ and the boundedness of the Hessian $\nabla^2 v$ are straightforward since σ is twice continuously differentiable, due to the fact that continuous functions on compact sets are bounded.

For the velocity field v, since $v(x,t)$ is Lipschitz continuous and the boundary condition (4.5) holds, then $z(x,T)$ is Lipschitz continuous with respect to x. Moreover, $z(x,T)$ is bounded. In fact, by the classical ODE theory, there is one and only one solution curve passing through each point since v is Lipschitz. We find that the solution curve passing through a point $x_0 \in \partial D$ stays on ∂D due to the boundary condition. More specifically, parameterize the boundary ∂D by $\gamma(s)$ using the arc length parameter $s : [-\delta, \delta] \to \mathbb{R}^d$ for some $\delta > 0$. Consider a curve of the force $x(t) = \gamma(s(t))$, then

$$x'(t) = \tau(s)s'(t),$$

where τ is a unit tangent vector. Since $v(x,t) \cdot n = 0$, one finds that the equation

$$x'(t) = v\big(x(t),t\big) \quad \Leftrightarrow \quad s'(t) = v\big(\gamma(s(t)),t\big) \cdot \tau\big(s(t)\big).$$

This is a well-defined ODE for s. Solving this ODE gives a solution $t \mapsto s(t)$ and thus giving a curve $x(t) = \gamma(s(t))$ solving the ODE $\dot{x} = v(x(t),t)$. By the uniqueness of the

solutions, there is no solution curve passing through ∂D to the exterior of the domain. Hence, $z(x,t) \in D$, $\forall t \in [0,T]$, which implies $z(x,T)$ is bounded. Hence, we conclude that $|\boldsymbol{z}(\boldsymbol{x},T)|^2/2$ is Lipschitz continuous with respect to x, and the Lipschitz constant is uniform for $\theta \in \Theta_R$.

We recall that ℓ satisfies the equation

$$\partial_t \ell(x,t) = \nabla \cdot v(x,t;\theta), \quad \ell(x,0) = 0.$$

Since $\nabla^2 v$ is bounded, $\mathrm{tr}(\nabla v)$ is Lipschitz continuous, which leads to the uniform Lipschitz continuity of $\ell(x,T)$.

The Lipschitz continuity of $\int_0^T |v(z(x,t),t)|^2/2\,dt$ is a direct result of the Lipschitz continuity of v and Lipscthitz continuity of $z(\cdot;t)$. The constant is uniform in θ. Hence, $F_v(\cdot;\theta)$ is uniformly Lipschitz continuous with respect to x.

Lastly, the uniform boundedness of F_v is straightforward as we have shown that $z(x,t)$ is uniformly bounded and v is continuous. □

With the clipped parameters and the help of Lemma 4.2, we will the investigate the convergence of the optimizers for (4.11) to that for (4.10) as $N \to \infty$ in the next section.

4.2. Large data limit

In this subsection, we explore the limit behavior of the minimizers as the sample size N tends to ∞, i.e., provided sufficient data samples, for fixed structure of the neural networks and clipped parameters with $\theta \in \Theta_R$. Although we neglect the effect of error brought by the approximation of the minimization problem (through, for example, stochastic gradient methods), the research on the behaviour of the minimizers as $N \to \infty$ is beneficial to understanding how machine learning algorithms work. If one can establish a bound to control the generalization error beyond the training sets, it offers theoretical stability and reliability for the generalization of the neural networks.

Suppose that we have a collection of data samples $\{X_i\}_{i=1}^N$ taking values in D, which are i.i.d. sampled from the data distribution ρ_0. Here, we would like to investigate the limit as $N \to \infty$. Since the samples are random samples, we put this into a probabilistic framework. In particular, by the Kolmogorov extension theorem [37], there is a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ for the i.i.d. sampled D-valued random variables $\{X_i\}_{i=1}^\infty$, with a common law $\rho_0 \in \mathcal{P}(D)$. The corresponding optimization problem (4.11) becomes a probabilistic problem

$$\min_{F_v \in F_{\mathcal{N}}} \mathcal{E}_{N,\omega}(F_v) := \frac{1}{N}\sum_{i=1}^N F_v\big(X_i(\omega)\big) = \int_D F_v(x)\rho_{N,X(\omega)}(dx). \tag{4.14}$$

Here,

$$\rho_N = \frac{1}{N}\sum_{i=1}^N \delta_{X_i(\omega)} \tag{4.15}$$

is the empirical measure and $\omega \in \Omega$ means one elementary event (a concrete realization of the N samples). The following is obvious.

Lemma 4.3. *The problem* (4.14) *always has a solution* $F_{N,\omega} \in V_{\mathcal{N}}$.

This holds because the parameter space Θ_R is a compact set and the functional is continuous in θ when we fix $X_1, \cdots, X_N$.

Since the samples are i.i.d. sampled from ρ_0, one has

$$\mathbb{E}[\mathcal{E}_{N,\omega}(F_v)] = \int_D F_v(x)\rho_0(dx). \tag{4.16}$$

Moreover, by the Law of Large numbers [38, 39], the empirical measure (4.15) converges weakly to ρ_0. Here, we cite a stronger result for the convergence of empirical measures under W_1 distance.

Lemma 4.4. *Assume* $D \subset \mathbb{R}^d$ *is a bounded domain. Let* $\mu \in \mathcal{P}(D)$. *Consider an i.i.d. sequence* $(X_k)_{k\geq 1}$ *of* μ*-distributed random variables and the empirical measure* $\mu_N := (\sum_{k=1}^N \delta_{X_k})/N$. *Then there exist a constant* C *such that for all* $N \geq 1$,

$$\mathbb{E}\left[W_1\left(\mu_N, \mu\right)\right] \leq CN^{-1/d}. \tag{4.17}$$

Note that D is a bounded domain, thus μ has q-th order moments for all q. Then this is a direct result of [14, Theorem 1] by setting $p = 1$ and $q \gg 1$.

Applying Lemma 4.4 to our problem, we conclude that.

Corollary 4.1. *Let* $X_1, X_2, \cdots$ *be a sequence of i.i.d. random variables with common law* ρ_0. *Consider the empirical measure in* (4.15). *Then*

$$\lim_{N\to\infty} W_1(\rho_N, \rho_0) \ \to \ 0 \tag{4.18}$$

as $N \to +\infty$ *almost surely.*

With the preparation above, we now show for a.s. $\omega \in \Omega$, the sequence of the minimizers $F_{N,\omega}$ has convergent subsequences and the limits would be a global minimizer of the continuous OT-Flow problem (4.10). The tool we use here is again the Γ-convergence. Moreover, due to the simple structure of the fixed neural network, we can obtain much stronger results compared to previous section.

Theorem 4.1. *Assume* D *is a bounded domain with smooth boundary and fix the neural network space* $V_{\mathcal{N}}$. *Consider* $F_v \in F_{\mathcal{N}}$ *and* $\theta \in \Theta_R$. *Then the following holds:*

(i) Almost surely, the functional $\mathcal{E}_{N,\omega}(F_v)$ *in* (4.14) *is* Γ*-convergent to the functional* $\mathcal{E}(F_v)$ *in* (4.10), *where the convergence of* F_v *is the uniform convergence in* $C_b(D)$.

(ii) Consider the minimizer $F_{N,\omega} \in V_{\mathcal{N}}$ *to* (4.14). *Then, for almost surely* $\omega \in \Omega$, *any subsequence of the sequence* $F_{N,\omega}$ *has a convergent subsequence, with the limit* $F_\omega^{\mathcal{N}}$ *being a global minimizer to* (4.10), *and it holds that*

$$\lim_{N\to\infty} \mathcal{E}_{N,\omega}\left(F_{N,\omega}\right) = \mathcal{E}\left(F_\omega^{\mathcal{N}}\right) = \inf_{\varphi\in F_{\mathcal{N}}} \mathcal{E}(\varphi). \tag{4.19}$$

We remark that the topology for $F_v \in F_{\mathcal{N}}$ is not important due to the correspondence to $\theta \in \Theta_R$, which is a subset of a finite dimensional space.

Proof of Theorem 4.1. (i) This is in fact straightforward. For any F_N that converges uniformly to F, and since ρ_N (defined in (4.15)) converges weakly converges to ρ_0, thus

$$\begin{aligned}
&\lim_{N\to\infty} \int_D F_N \rho_N(dx) \\
&= \lim_{N\to\infty} \int_D (F_N - F)\rho_N(dx) + \lim_{N\to\infty} \int_D F\rho_N(dx) \\
&= \int_D F\rho_0(dx). \qquad (4.20)
\end{aligned}$$

The first term here goes to zero because

$$\left| \int_D (F_N - F)\rho_N(dx) \right| \leq \|F_N - F\| \quad \to \quad 0.$$

The second term converges due to the weak convergence of ρ_N almost surely by Corollary 4.1.

(ii) First, for each $F_{N,\omega} \in V_{\mathcal{N}}$, we denote its corresponding neural network parameters in Θ_R as $\theta_{N,\omega}$. Since $|\theta_{N,\omega}| \leq R$, the sequence $\{\theta_{N,\omega}\}$ has a subsequence that converges to some $\theta_\omega^{\mathcal{N}}$. We denote the corresponding neural network function of $\theta_\omega^{\mathcal{N}}$ as $F_\omega^{\mathcal{N}}$. By Lemma 4.1, $F_{N,\omega}$ has a subsequence $\{F_{N_k,\omega}\}$ that uniformly converges to $F_\omega^{\mathcal{N}}$. By the Γ-convergence in part (i), $F_\omega^{\mathcal{N}}$ is in fact a global minimizer of (4.10)

$$\mathcal{E}\left(F_\omega^{\mathcal{N}}\right) = \inf_{\varphi \in F_{\mathcal{N}}} \mathcal{E}(\varphi). \qquad (4.21)$$

To show the claim that the optimal value of the loss function also converges, our core idea is to prove the following inequalities for any convergent subsequence with limit $F_\omega^{\mathcal{N}} \in V_{\mathcal{N}}$:

$$\mathcal{E}\left(F_\omega^{\mathcal{N}}\right) \leq \liminf_{N\to\infty} \mathcal{E}_{N,\omega}\left(F_{N,\omega}\right) \leq \limsup_{N\to\infty} \mathcal{E}_{N,\omega}\left(F_{N,\omega}\right) \leq \inf_{\varphi \in F_{\mathcal{N}}} \mathcal{E}(\varphi). \qquad (4.22)$$

Due to (4.21), the limit in fact exists along this subsequence. Then, for any subsequence, there is a further subsequence such that the loss function value converges to $\inf_{\varphi \in F_{\mathcal{N}}} \mathcal{E}(\varphi)$, which is the same limit. Hence, the whole sequence converges.

Next, we show (4.22). By the uniform convergence of the functions, it is clear (similar to (4.20)) that

$$\liminf_{N\to\infty} \mathcal{E}_{N,\omega}\left(F_{N,\omega}\right) = \mathcal{E}\left(F_\omega^{\mathcal{N}}\right).$$

For the second inequality, the minimising property of $F_{N,\omega}$ implies that

$$\mathcal{E}_{N,\omega}\left(F_{N,\omega}\right) \leq \mathcal{E}_{N,\omega}(F) \quad \text{for all} \quad F \in V_{\mathcal{N}}.$$

Next, take a sequence $(F_n)_n$ in $F_{\mathcal{N}}$ that realises the inf in (4.22), i.e.,

$$\lim_{n\to\infty} \mathcal{E}(F_n) = \inf_{\varphi\in F_{\mathcal{N}}} \mathcal{E}(\varphi).$$

We recall the dual description of the W_1 distance [35]

$$W_1(\mu,\nu) = \sup_{\varphi:\|\varphi\|_{\mathrm{lip}}\le 1} \int \varphi\, d(\mu-\nu),$$

where $\|\cdot\|_{\mathrm{lip}}$ means the largest Lipschitz constant for the function. It then follows that

$$\int_D F_n d(\rho_N - \rho_0) \le C_L \cdot W_1(\rho_N, \rho_0),$$

where C_L is the upper bound of Lipschitz constant for all functionals in $F_{\mathcal{N}}$ indicated in Lemma 4.2. By Lemma 4.1, $W_1(\rho_N,\rho_0)\to 0$ when $N\to\infty$ almost surely. Thus,

$$\limsup_{N\to\infty} \mathcal{E}_{N,\omega}(F_{N,\omega}) \le \limsup_{N\to\infty} \mathcal{E}_{N,\omega}(F_n) = \mathcal{E}(F_n)$$

holds for all n. Then we have

$$\limsup_{N\to\infty} \mathcal{E}_{N,\omega}(F_{N,\omega}) \le \inf_n \mathcal{E}(F_n) = \inf_{\varphi\in F_{\mathcal{N}}} \mathcal{E}(\varphi).$$

The proof is thus complete. □

4.3. Training the optimal velocity field

To enhance the approximation ability of neural networks, we can increase the complexity of neural networks and relax the clipping constant R for parameters. Then the classical theory of universal approximation [4,9] ensures that we can select a sequence of neural network space $\{V_\ell\}$ and clipping parameter space $\{\Theta_{R_\ell}\}$ with increasing clipping constant $R_\ell\to\infty$, such that for each $v\in L^1([0,T];L^2(\rho_t))$, there exists a $v_\ell\in V_\ell$ such that

$$\int_0^T \int_D |v-v_\ell|^2 \rho_t dx dt \;\to\; 0 \quad \text{as} \quad \ell \;\to\; \infty, \tag{4.23}$$

which indicates that with increasing complexity of architecture, neural network will gradually have enough express capability to approximate the theoretical solutions of the OT problem.

In real training, one can let $\ell\to\infty$ and $N\to\infty$ simultaneously. With delicately selected optimizers and hyperparameters, OT-Flow can find a minimizer close to the theoretical solution of problem (2.2). Moreover, as $\alpha\to\infty$, the neural network minimizers of OT-Flow with different α will converges to classical OT problem solutions as we mentioned in Section 3.

5. Conclusion and discussion

In summary, we have conducted two primary convergence analyses for OT-Flow, one of the deep generative models. The first part employs Γ-convergence to establish the convergence of minimizers from OT-Flow to classical OT as the regularization coefficient $\alpha \to \infty$. The second part leverages the liminf-limsup framework to demonstrate the convergence of minimizers as the number of training samples $N \to \infty$. Furthermore, if we provide the neural network with sufficient approximation capability, the minimizers of OT-Flow will theoretically converges to classical OT ones. Our work enhances the understanding of the convergence properties of CNFs models with regularization, providing theoretical assurances for stability during training. Future research directions may involve applying similar methodologies to develop convergence analyses for other deep generative models, such as DPMs, and generative models associated with optimal control.

Acknowledgements

The authors thank Ling Guo for helpful discussions and comments.

This work is partially supported by the National Key R&D Program of China (Nos. 2020YFA0712000, 2021YFA1002800). The work of L. Li was partially supported by the NSFC (Nos. 12371400, 12031013), by the Shanghai Municipal Science and Technology Major Project (No. 2021SHZDZX0102), and by the Shanghai Science and Technology Commission (Nos. 21JC1403700, 21JC1402900).

References

[1] L. AMBROSIO, N. GIGLI, AND G. SAVARÉ, *Gradient flows: In metric spaces and in the space of probability measures*, Springer Science & Business Media, 2005.

[2] M. ARJOVSKY, S. CHINTALA, AND L. BOTTOU, *Wasserstein generative adversarial networks*, in: International Conference on Machine Learning, PMLR, (2017), 214–223.

[3] F. BAO, C. LI, J. ZHU, AND B. ZHANG, *Analytic-DPM: An analytic estimate of the optimal reverse variance in diffusion probabilistic models*, in: The Tenth International Conference on Learning Representations, ICLR, (2022), `https://openreview.net/forum?id=0xiJLKH-ufZ`

[4] A. R. BARRON, *Universal approximation bounds for superpositions of a sigmoidal function*, IEEE Trans. Inf. Theory 39 (1993), 930–945.

[5] A. BRAIDES, *Gamma-Convergence for Beginners*, Oxford Lecture Series in Mathematics and Its Applications, Vol. 22, Oxford University Press, 2002.

[6] G. BUTTAZZO AND G. DAL MASO, *Γ-convergence and optimal control problems*, J. Optim. Theory Appl. 38 (1982), 385–407.

[7] C. CHEN, *Spatiotemporal imaging with diffeomorphic optimal transportation*, Inverse Probl. 37 (2021), 115004.

[8] R. T. CHEN, Y. RUBANOVA, J. BETTENCOURT, AND D. K. DUVENAUD, *Neural ordinary differential equations*, Adv. Neural Inf. Process. Syst. 31 (2018).

[9] G. CYBENKO, *Approximation by superpositions of a sigmoidal function*, Math. Control Signals Syst. 2 (1989), 303–314.
[10] P. DHARIWAL AND A. NICHOL, *Diffusion models beat gans on image synthesis*, Adv. Neural Inf. Process. Syst. 34 (2021), 8780–8794.
[11] L. DINH, J. SOHL-DICKSTEIN, AND S. BENGIO, *Density estimation using real NVP*, in: 5th International Conference on Learning Representations, ICLR, (2017), https://openreview.net/forum?id=HkpbnH9lx.
[12] L. C. EVANS, *An Introduction to Mathematical Optimal Control Theory*, https://math.berkeley.edu/~evans/control.course.pdf.
[13] C. FINLAY, J.-H. JACOBSEN, L. NURBEKYAN, AND A. OBERMAN, *How to train your neural ODE: The world of Jacobian and kinetic regularization*, in: International Conference on Machine Learning, PMLR, (2020), 3154–3164.
[14] N. FOURNIER AND A. GUILLIN, *On the rate of convergence in Wasserstein distance of the empirical measure*, Probab. Theory Relat. Fields 162 (2015), 707–738.
[15] I. GOODFELLOW, J. POUGET-ABADIE, M. MIRZA, B. XU, D. WARDE-FARLEY, S. OZAIR, A. COURVILLE, AND Y. BENGIO, *Generative adversarial nets*, Adv. Neural Inf. Process. Syst. 27 (2014).
[16] W. GRATHWOHL, R. T. Q. CHEN, J. BETTENCOURT, I. SUTSKEVER, AND D. DUVENAUD, *FFJORD: Free-form continuous dynamics for scalable reversible generative models*, in: 7th International Conference on Learning Representations, ICLR, (2019), https://openreview.net/forum?id=rJxgknCcK7.
[17] J. HO, A. JAIN, AND P. ABBEEL, *Denoising diffusion probabilistic models*, Adv. Neural Inf. Process. Syst. 33 (2020), 6840–6851.
[18] Y. JING, J. CHEN, L. LI, AND J. LU, *A machine learning framework for geodesics under spherical Wasserstein-Fisher-Rao Metric and its application for weighted sample generation*, J. Sci. Comput. 98 (2024), 5.
[19] L. V. KANTOROVICH, *On a problem of Monge*, J. Math. Sci. 133 (2006), 1383–1383.
[20] D. P. KINGMA AND J. BA, *Adam: A method for stochastic optimization*, in: International Conference on Learning Representations, 2015.
[21] D. P. KINGMA ET AL., *An introduction to variational autoencoders*, Found. Trends Mach. Learn. 12 (2019), 307–392.
[22] D. P. KINGMA, T. SALIMANS, R. JOZEFOWICZ, X. CHEN, I. SUTSKEVER, AND M. WELLING, *Improved variational inference with inverse autoregressive flow*, Adv. Neural Inf. Process. Syst. 29 (2016).
[23] D. P. KINGMA AND M. WELLING, *Auto-encoding variational bayes*, in: International Conference on Learning Representations, 2014.
[24] Y. LECUN, L. BOTTOU, Y. BENGIO, AND P. HAFFNER, *Gradient-based learning applied to document recognition*, Proceedings of the IEEE 86(11) (1998), 2278–2324.
[25] Y. LIU ET AL., *Sora: A review on background, technology, limitations, and opportunities of large vision models*, arXiv:2402.17177, 2024.
[26] M. LONGO, S. MISHRA, T. K. RUSCH, AND C. SCHWAB, *Higher-order quasi-Monte Carlo training of deep neural networks*, SIAM J. Sci. Comput. 43 (2021), A3938–A3966.
[27] M. LOULAKIS AND C. G. MAKRIDAKIS, *A new approach to generalisation error of machine learning algorithms: Estimates and convergence*, arXiv:2306.13784, 2023.
[28] G. MONGE, *Mémoire sur la théorie des déblais et des remblais*, Mem. Math. Phys. Acad. Royale Sci. (1781), 666–704.
[29] D. ONKEN, S. W. FUNG, X. LI, AND L. RUTHOTTO, *OT-Flow: Fast and accurate continuous normalizing flows via optimal transport*, in: Proceedings of the AAAI Conference on

Artificial Intelligence, 35 (2021), 9223–9232.
[30] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan, *Normalizing flows for probabilistic modeling and inference*, J. Mach. Learn. Res. 22 (2021), 1–64.
[31] G. Papamakarios, T. Pavlakou, and I. Murray, *Masked autoregressive flow for density estimation*, Adv. Neural Inf. Process. Syst. 30 (2017).
[32] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, *Hierarchical text-conditional image generation with clip latents*, arXiv:2204.06125, 2022.
[33] D. Rezende and S. Mohamed, *Variational inference with normalizing flows*, in: International Conference on Machine Learning, PMLR, (2015), 1530–1538.
[34] T. Salimans, H. Zhang, A. Radford, and D. N. Metaxas, *Improving GANs using optimal transport*, in: International Conference on Learning Representations, 2018.
[35] F. Santambrogio, *Optimal Transport for Applied Mathematicians*, Birkäuser, 2015.
[36] J. Song, C. Meng, and S. Ermon, *Denoising diffusion implicit models*, in: 9th International Conference on Learning Representations, ICLR, (2021), `https://openreview.net/forum?id=St1giarCHLP`.
[37] T. Tao, *An Introduction to Measure Theory*, AMS, 2011.
[38] H. G. Tucker, *A generalization of the Glivenko-Cantelli theorem*, Ann. Math. Stat. 30 (1959), 828–830.
[39] V. S. Varadarajan, *On the convergence of sample probability distributions*, Sankhyā: The Indian Journal of Statistics (1933-1960), 19, (1958), 23–26.
[40] Z. Xiong, L. Li, Y.-N. Zhu, and X. Zhang, *On the convergence of continuous and discrete unbalanced optimal transport models for 1-Wasserstein distance*, SIAM J. Numer. Anal. 62 (2024), 749–774.