Learning normalizing flows from Entropy-Kantorovich potentials

Chris Finlay[*1] **Augusto Gerolin**[*2] **Adam M Oberman**[1] **Aram-Alexandre Pooladian**[*1]

[1]Department of Mathematics and Statistics, McGill University, Montréal, Canada
[2]Department of Theoretical Chemistry, Vrije Universiteit Amsterdam, Amsterdam, Netherlands

Abstract

We approach the problem of learning continuous normalizing flows from a dual perspective motivated by entropy-regularized optimal transport, in which continuous normalizing flows are cast as gradients of scalar potential functions. This formulation allows us to train a dual objective comprised only of the scalar potential functions, and removes the burden of explicitly computing normalizing flows during training. After training, the normalizing flow is easily recovered from the potential functions.

1 Introduction

Normalizing flows [31, 36, 37] are a popular mechanism for probabilistic modeling and inference, whereby an unknown distribution is parameterized by a transformation of the standard Normal distribution. Normalizing flows provide a general framework for defining flexible probability distributions over continuous random variables, and have been applied throughout a wide variety of fields, including density estimation (e.g. [18, 36]), generative modelling (e.g. [8, 21, 40]), and variational inference (e.g. [4, 22, 31, 38]).

Continuous normalizing flows (CNFs) [18] construct normalizing flows through a continuous time-dependent transformation of the data, in which the transformation is given as the solution operator of a neural ordinary differential equation (ODE) [8, 9, 20, 33]. In this framework, normalizing flows are parameterized as a flow generated by a (learned) vector field. Training this vector field can be difficult, and significant regularization may be necessary to learn well-behaved flows [13, 27, 29, 42].

Here we take a step back and frame CNFs within the lens of Optimal Transport (OT) theory [11, 41]. This is a natural connection, due to a correspondence between vector fields and the dynamical formulation of OT [3], which we explore below. Indeed, this direct connection was exploited in [13] and [29] to speed the training of vector fields for CNFs. While this direct approach to linking CNFs with OT theory has yielded promising improvements to CNF training, the problems of discretizing an ODE and training the CNF's vector field remain.

In this work, we will take an indirect route to constructing CNFs, and completely sidestep the need to solve an ODE generated by a vector field during CNF training. Instead we will use the property that the flows encountered in OT are *gradients of scalar potential functions*. Optimization will be done only in terms of these potential functions, without discretizing an ODE during training. Afterwards, a CNF will be recovered from the learned potential functions. In a sense, this is an energy-based modeling perspective [24] on CNFs.

To be more explicit, the link between CNFs and OT studied herein relies on the *entropy-regularized dynamical formulation* of OT, which will allow us to construct a time dependent curve ρ_t of densities

[*] Equal contribution. Correspondance to `christopher.finlay@mcgill.ca`

Preprint. Under review.

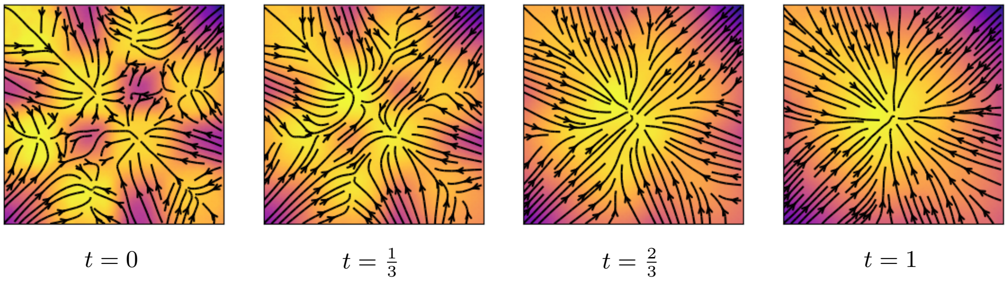

Figure 1: Using the learned Entropy-Kantorovich potentials and (4), the vector field v_t (black arrows) recovered from the potentials creates a CNF between the checkerboard distribution (at $t = 0$) and the standard Normal distribution (at $t = 1$). Log-densities of the distributions along the flow are shown with the heat map.

acting as the displacement of μ, the data measure, to ν, the Gaussian measure. More precisely, the entropy-regularized dynamical OT problem seeks the pair (ρ_t, v_t) of density-flow ρ_t and vector field v_t minimizing the variational problem

$$\inf_{(\rho_t, v_t)} \int_0^1 \int_{\mathbb{R}^d} \left(\frac{\|v_t\|^2}{2} + \frac{\varepsilon^2}{8} \|\nabla \log \rho_t\|^2 \right) \rho_t \mathrm{d}x\mathrm{d}t, \tag{1}$$

subject to the constraint that the pair also satisfies the *continuity equation*

$$\partial_t \rho_t + \nabla \cdot (v_t \rho_t) = 0, \tag{2}$$

and that both $\rho_0 = \mu$ and $\rho_1 = \nu$ (i.e. that the initial and final endpoints respectively agree). The scalar ε will control the amount of regularization provided by the Fisher information term $\|\nabla \log \rho_t\|^2$, and corresponds to entropic regularization of OT [10].

Entropic regularization will play an important role in our approach. In particular, since in general we do not have access to the true data distribution, this type of regularization introduces inherent stochasticity to the learned flows, which may heuristically be beneficial during training and inference. In addition, entropic regularization will simplify our numerical method by allowing us to approximate a particular function transformation with Monte-Carlo integration. Notice that when $\varepsilon = 0$, the variational problem (1) selects, among all pairs (ρ_t, v_t), the one that minimizes the kinetic energy of the vector field.

It is from the continuity equation (2) that a CNF is defined, by the ODE

$$X_t' = v_t(X_t(x)) \quad \text{s.t.} \quad X_0(x) = x; \quad x \in \mathbb{R}^d \tag{3}$$

where x is a particle drawn from the data. The continuity equation can be interpreted as the equation ruling the evolution of a family of particles initially drawn from the data measure μ, and where each particle follows the path defined by the solution operator of (3), flowing the particles to the Normal distribution.

Main contributions

In practice, directly optimizing the variational problem (1) may not be feasible in high dimensions, and optimizing the CNF generated by (3) introduces its own difficulties. In this paper we instead use theoretical results from OT theory to provide implicit formulas for the optimal flow (ρ_t, v_t) solving (1). The flow so defined will only depend on two scalar functions φ and ψ, called *Entropy-Kantorovich potentials*. The CNF will be defined through the following vector field

$$v_t(x) = \nabla \frac{1}{2}(\varphi_t(x) - \psi_t(x)), \tag{4}$$

where $\varphi_t := \varepsilon \log \mathcal{H}_{t\varepsilon}[e^{\varphi/\varepsilon}]$ and $\psi_t := \varepsilon \log \mathcal{H}_{(1-t)\varepsilon}[e^{\psi/\varepsilon}]$, with $\mathcal{H}$ the heat kernel (see Section 2.4; $\mathcal{H}$ is also known as Gaussian averaging). Additionally, the density-flow will be defined by

$$\log \rho_t(x) = (\varphi_t(x) + \psi_t(x))/\varepsilon. \tag{5}$$

We will show how the Entropy-Kantorovich potentials can be found by optimizing a static *dual problem* of (1); the optimization procedure itself will not require the numerical solution of an ODE.

To summarize, our main contributions are the following:

- We introduce a novel framework for constructing CNFs from potential functions, based on ideas from entropy-regularized OT. The framework is theoretically well motivated, and interprets the CNF so defined as a curve in the 2-Wasserstein space of probability measures, connecting the data to the Normal distribution.
- Our method is computationally efficient: training is sample-based and mesh-free, and only requires optimizing two time-invariant Entropy-Kantorovich potential functions φ and ψ. The method completely avoids solving an ODE during training.
- Once the potential functions have been trained, it is straightforward to perform density estimation and generative modeling through a CNF recovered from the learned potential functions. The CNF is evaluated easily, and can be applied in higher dimensions through the use of Monte-Carlo integration.

1.1 Related work

The tools provided by the literature on OT [11, 41] are the cornerstone of our methodology. Connections between particle-based methods in numerical analysis and OT first appeared in the seminal work of [3], and is referred to as the "Benamou-Brenier" formulation in the modern literature. This provides a *dynamic* perspective on OT, which can be generalized to entropic optimal transport (EOT) [15, 16, 25]. These connections will be made more explicit in Section 2.

Meanwhile, the literature on flow-based methods in deep learning is rich with applications of OT theory, particularly in the context of normalizing flows; examples include [30, 33, 34, 39, 43]. A relevant connection to our work appears in [13], where the authors exploit the Benamou-Brenier formulation by adding the relevant kinetic energy term to the objective function. Following work in [29] cast the vector field as a gradient potential. However, in both these works the authors directly solve the ODE generated by the time-dependent vector field during training, which is in contrast to our approach.

In the context of Wasserstein Generative Adversarial Networks (GANs), different approaches have been taken to learn Kantorovich potentials using neural networks; see for example [2, 19], and the entropy-regularized case in [26].

2 CNFs from entropic optimal transport

Notation: $\mathcal{P}(\mathbb{R}^d)$ denotes the set of probability distributions over $\mathbb{R}^d$, with $\mathcal{P}_p(\mathbb{R}^d)$ being the set of probabilities with $p \geq 1$ finite moments. The density of the standard Normal distribution in $\mathbb{R}^d$ is denoted $p_{\mathcal{N}}$ and the data measure is denoted by $p_{\mathcal{D}}$. In the OT framework below, we identify $p_{\mathcal{N}}$ with the target measure ν, while $p_{\mathcal{D}}$ is the source measure μ. In principle, we could let the target measure be any closed-form density function, but in this work we always take ν to be the standard Gaussian measure, as done in the normalizing flow literature. We sometimes refer to $p_{\mathcal{N}}$ (ν) as the Normal distribution, despite it being a measure.

The 2-Wasserstein distance between $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ is defined as

$$W_2(\mu,\nu) := \left(\min_{\gamma \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \frac{1}{2} \|x-y\|_2^2 \, d\gamma(x,y) \right)^{1/2}, \tag{6}$$

where $\Pi(\mu,\nu)$ denotes the set of probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals equal to μ and ν,

$$\Pi(\mu,\nu) := \left\{ \gamma \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d) \;\middle|\; \gamma(A \times \mathbb{R}^d) = \mu(A), \gamma(\mathbb{R}^d \times A) = \nu(A) \right\}.$$

An element in $\Pi(\mu,\nu)$ is called a coupling or a transport plan, and $\gamma^{\text{opt}} \in \Pi(\mu,\nu)$ realising the minimum in (6) is called the optimal transport plan. For a map $T : \mathbb{R}^d \to \mathbb{R}^d$, $T_\sharp\mu$ denotes the *pushforward* of μ with T, i.e. $\mu(T^{-1}(A)) = T_\sharp\mu(A)$, for any $A \subseteq \mathbb{R}^d$ Borel measurable[1].

[1] We refer the reader to Appendix A for background on optimal transport maps.

The space $\mathcal{P}_2(\mathbb{R}^d)$ endowed with W_2 is a complete and separable metric space, denoted by $\mathbb{W}_2(\mathbb{R}^d)$. We can also show that $\mathbb{W}_2(\mathbb{R}^d)$ is a geodesic space, i.e. any two points in $\mathbb{W}_2(\mathbb{R}^d)$ can be connected by a geodesic[2] (see Appendix B for details).

Finally, we define the function space $\mathrm{L}^{\exp}_{\varepsilon}(\mathbb{R}^d; \mathrm{d}x) := \{f : \mathbb{R}^d \to \mathbb{R} \mid \int_{\mathbb{R}^d} \exp(f/\varepsilon)\mathrm{d}x < +\infty\}$ and the (c, ε)-transformation of $f \in \mathrm{L}^{\exp}_{\varepsilon}(\mathbb{R}^d; \mathrm{d}x)$:

$$f^{(c,\varepsilon)}(y; \nu) := \varepsilon \log \left(\frac{1}{(2\pi\varepsilon)^{\frac{d}{2}}} \int_{\mathbb{R}^d} e^{\frac{f(x) - \frac{1}{2}\|x-y\|^2}{\varepsilon}} \mathrm{d}x \right) - \varepsilon \log(\nu(y)), \tag{7}$$

and analogously for $g \in \mathrm{L}^{\exp}_{\varepsilon}(\mathbb{R}^d; \mathrm{d}y)$ (with ν swapped for μ). This transformation is critical to our method. It arises throughout many branches of math and physics under various names: (i) the Hopf-Cole transformation of f, and is the solution operator of a partial differential equation from stochastic control theory; (ii) the softmax operation of f convolved with a Normal distribution of variance ε; (iii) additionally the (c, ε)-transform is a smoothed version of the quadratic c-transform arising in OT theory and convex analysis. The (c, ε)-transform lends itself to evaluation in high dimensions or in mesh-free environments by Monte-Carlo integration, whereas the c-transform is difficult to compute in these scenarios. Both the (c, ε)-transform and the space $\mathrm{L}^{\exp}_{\varepsilon}(\mathbb{R}^d; \mathrm{d}x)$ are intimately tied to energy-based models (see Equation (13) below).

2.1 Transport maps and normalizing flows

Here we recall some basic facts about normalizing flows and transport maps, with a clearer exposition in Appendix A. Under some regularity assumptions on μ, ν, the optimal transport map T is 1-Lipschitz [6, 7] and then, $T_\sharp\mu = \nu$ can be equivalently written by the change of variables formula $\mu(x) = \nu(T(x))|\det \mathrm{J}_T(x)|$, where J_T is the Jacobian of T. Therefore, when $\mu = p_{\mathcal{D}}$ is the data density and $\nu = p_{\mathcal{N}}$ is the standard Normal density, T is a normalizing flow.

Normalizing flows are designed to maximize the log-likelihood of the data under a transformation T. The normalizing flow so defined is not necessarily an optimal transport map. Since the data density is not known analytically, the difficulty of evaluating the log-likelihood of a sample x under $p_{\mathcal{D}}$ is pushed onto evaluating the log-likelihood of $T(x)$ under $p_{\mathcal{N}}$:

$$\log p_{\mathcal{D}}(x) = \log p_{\mathcal{N}}(T(x)) + \log |\det \mathrm{J}_T(x)|. \tag{8}$$

In the normalizing flow literature, T is a composition of 'simple' analytic functions, so that the log-determinant of the Jacobian can be computed tractably. For example, in CNFs, where the transport map is defined as the solution operator of the ODE (3), the Jacobian log-determinant is evaluated by integrating the divergence of the vector field v_t along the solution path [18]. Once a family of maps with tractable Jacobian determinants have been constructed, and given data $x_i \sim p_{\mathcal{D}}$, the objective of normalizing flows is to simply maximize the log-likelihood of the data, $\max_T \sum_i \log p_{\mathcal{N}}(T(x_i)) + \log |\det \mathrm{J}_T(x_i)|$.

2.2 Geodesics flows in the 2-Wasserstein space $\mathbb{W}_2(\mathbb{R}^d)$

In this section, we briefly discuss how to construct a CNF which is a *constant-speed geodesic* between $p_{\mathcal{D}}$ and $p_{\mathcal{N}}$ in $\mathbb{W}_2(\mathbb{R}^d)$, deferring technical details to Appendix B.

Suppose there exists an optimal transport map T between $p_{\mathcal{D}}$ and $p_{\mathcal{N}}$, and consider the convex combination between the identity map Id and T, $\pi_t(x) = (1 - t)x + tT(x)$. This can be viewed as a interpolant between our two measures of interest. In fact, the continuous deformation $\rho_t = (\pi_t)_\sharp p_{\mathcal{D}}$, for $t \in [0, 1]$, is a constant-speed geodesic between the data and the Normal distribution.

By the celebrated Brenier's Theorem [5], the optimal transport map T can be expressed as the gradient of a scalar-valued potential $\phi : \mathbb{R}^d \to \mathbb{R}$, $T(x) = \nabla(\frac{1}{2}\|x\|^2 - \phi(x))$, then $\rho_t = (\pi_t)_\sharp p_{\mathcal{D}}$ reads

$$\rho_t = (\mathrm{Id} + t\nabla\phi(x))_\sharp p_{\mathcal{D}}. \tag{9}$$

In ODE terms, the velocity field defining the ODE (3) of this continuous normalizing flow is given by $v_t(x) = T(x) - x = \nabla\phi(x)$, and is *time-invariant*, depending only on the point x, and is hence

[2] A geodesic is a curve that minimizes the "length" between two end-points.

constant-speed. The function ϕ is called a *Kantorovich potential* and is related to the dual problem of (6). One can verify that ρ_t solves the continuity equation $\partial_t \rho_t + \nabla \cdot (\nabla \phi \rho_t) = 0$ from which the continuous normalizing flow is read off. In the context of normalizing flows, the continuity equation dictates the evolution of data moving from $p_{\mathcal{D}}$ to $p_{\mathcal{N}}$, if the paths were to truly take the optimal trajectory.

2.3 Entropy-regularized 2-Wasserstein distance

While the approach of Section 2.2 is elegant, it is difficult to optimize the Kantorovich potential directly ϕ in a mesh-free environment, or in high-dimensions. We instead turn to the entropy-regularized optimal transport problem, which as we shall see, lends itself to a computationally tractable method to determine the potential functions.

The entropic regularization of the W_2 distance with regularization parameter $\varepsilon > 0$ [10] is defined by

$$W_\varepsilon^2(p_{\mathcal{D}}, p_{\mathcal{N}}) = \min_{\gamma \in \Pi(p_{\mathcal{D}}, p_{\mathcal{N}})} \int_{\mathbb{R}^d \times \mathbb{R}^d} \frac{1}{2}\|x-y\|^2 d\gamma(x,y) + \varepsilon \mathrm{H}(\gamma). \tag{10}$$

Here H is the entropy of a probability measure $\gamma \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d)$, defined by $\mathrm{H}(\gamma) = \int_{\mathbb{R}^d \times \mathbb{R}^d} \gamma(x) \log(\gamma(x,y))\, \mathrm{d}x\mathrm{d}y$ if γ is a density and $\mathrm{H}(\mu) = +\infty$ otherwise. By strong convexity (10) always admits a *unique* minimizer [11, 14, 25]. Entropic regularization has the effect of 'diffusing' or 'fuzzing' the transport plans.

An equivalent formulation of (10), the so-called dual or Entropy-Kantorovich formulation of (10) allows us to obtain the distance between $p_{\mathcal{D}}$ and $p_{\mathcal{N}}$ by maximizing over pairs of Entropy-Kantorovich potentials (ψ, φ) rather than minimizing over measures γ [11, 12, 25],

$$W_\varepsilon^2(p_{\mathcal{D}}, p_{\mathcal{N}}) = \sup\left\{D_\varepsilon(\varphi,\psi) : \varphi \in \mathrm{L}_\varepsilon^{\exp}(\mathbb{R}^d; \mathrm{d}x), \psi \in \mathrm{L}_\varepsilon^{\exp}(\mathbb{R}^d; \mathrm{d}y)\right\} + \varepsilon, \tag{11}$$

where $D_\varepsilon : \mathrm{L}_\varepsilon^{\exp}(\mathbb{R}^d; \mathrm{d}x) \otimes \mathrm{L}_\varepsilon^{\exp}(\mathbb{R}^d; \mathrm{d}y) \to \mathbb{R}$ is the dual functional

$$\begin{aligned} D_\varepsilon(\varphi,\psi) = &\int_{\mathbb{R}^d} \varphi(x)\, \mathrm{d}p_{\mathcal{D}}(x) + \int_{\mathbb{R}^d} \psi(y)\, \mathrm{d}p_{\mathcal{N}}(y) \\ &- \varepsilon \int_{\mathbb{R}^d \times \mathbb{R}^d} \exp\left\{\frac{\varphi(x) + \psi(y) - \frac{1}{2}\|x-y\|^2}{\varepsilon}\right\} \mathrm{d}x\mathrm{d}y. \end{aligned} \tag{12}$$

Note that the functional D_ε is strictly concave in each variable and, under mild hypotheses, one can show the existence of maximizers in (11) which are unique up to additive constants [12]. Useful characterizations of the primal (10) and dual problem (12) are given by the following theorem.

Theorem 1 (Proposition 2.11, [12]) *Let $\varepsilon > 0$ be a positive number, $\Omega \subset \mathbb{R}^d$ be a compact set, $p_{\mathcal{D}}, p_{\mathcal{N}} \in \mathcal{P}(\Omega)$. Then given $\varphi \in \mathrm{L}_\varepsilon^{\exp}(\mathbb{R}^d; \mathrm{d}x)$ and $\psi \in \mathrm{L}_\varepsilon^{\exp}(\mathbb{R}^d; \mathrm{d}y)$, the following are equivalent:*

1. *(Maximizers) φ and ψ are maximizing potentials for* (11)*;*
2. *(Maximality condition) $\varphi^{(c,\varepsilon)} = \psi$, $\psi^{(c,\varepsilon)} = \varphi$ and, moreover, $\varphi, \psi \in L^\infty(\Omega)$.*
3. *(Primal problem) $\gamma_\varepsilon^{opt} = \exp\left((\varphi(x) + \psi(y) - \frac{1}{2}\|x-y\|^2)/\varepsilon\right) \in \Pi(p_{\mathcal{D}}, p_{\mathcal{N}})$;*
4. *(Duality attainment) $W_\varepsilon^2(p_{\mathcal{D}}, p_{\mathcal{N}}) = D_\varepsilon(\varphi,\psi) + \varepsilon$.*

Moreover, the optimal coupling γ_ε^{opt} is also the (unique) minimizer for the problem (10).

When ψ and φ are optimal, we may read off the data and normal log-densities from the Entropy-Kantorovich potential functions:

$$\begin{cases} \varphi(x) + \psi^{(c,\varepsilon)}(x) = \varepsilon \log p_{\mathcal{D}}(x) + C_1 \\ \psi(y) + \varphi^{(c,\varepsilon)}(y) = \varepsilon \log p_{\mathcal{N}}(y) + C_2 \end{cases} \tag{13}$$

for some normalizing constants C_1 and C_2. In other words, the potential functions parameterize the data and Normal distributions as energy-based models.

2.4 A bridge between CNFs and potentials: the dynamic formulation

We are now in a position to bridge entropic optimal transport with continuous normalizing flows. The variational problem (10) can be expressed in a *dynamic* form [17, 25]

$$\frac{\varepsilon}{2}\left(\mathrm{H}(p_{\mathcal{D}}) + \mathrm{H}(p_{\mathcal{N}})\right) + \inf_{(\rho_t^\varepsilon, v_t^\varepsilon)} \int_0^1 \int_{\mathbb{R}^d} \left(\frac{\|v_t^\varepsilon\|^2}{2} + \frac{\varepsilon^2}{8} \|\nabla \log \rho_t^\varepsilon\|^2 \right) \rho_t^\varepsilon \mathrm{d}x \mathrm{d}t, \tag{14}$$

with the constraint that $(\rho_t^\varepsilon, v_t^\varepsilon)$ solves the continuity equation $\partial_t \rho_t^\varepsilon + \nabla \cdot (v_t^\varepsilon \rho_t^\varepsilon) = 0$, and that $\rho_0^\varepsilon = p_{\mathcal{D}}$, $\rho_1^\varepsilon = p_{\mathcal{N}}$. The time-dependent density ρ_t^ε is a curve between the data and Normal distributions in the 2-Wasserstein space parameterized by $t \in [0, 1]$. Once v_t^ε is known, this time dependent vector field defines a CNF via the ODE (3).

Equation (14) also has an associated dual problem (equivalent to (12)), where again instead of minimizing over pairs $(\rho_t^\varepsilon, v_t^\varepsilon)$, optimization takes place across the following two functionals:

$$\begin{cases} J(\varphi) = \varepsilon H(p_{\mathcal{D}}) + \sup\limits_{\varphi} \displaystyle\int_{\mathbb{R}^d} \varphi \, \mathrm{d}p_{\mathcal{N}} + \int_{\mathbb{R}^d} \varphi^{(c,\varepsilon)} \, \mathrm{d}p_{\mathcal{D}}, \\ I(\psi) = \varepsilon H(p_{\mathcal{N}}) + \sup\limits_{\psi} \displaystyle\int_{\mathbb{R}^d} \psi \, \mathrm{d}p_{\mathcal{D}} + \int_{\mathbb{R}^d} \psi^{(c,\varepsilon)} \, \mathrm{d}p_{\mathcal{N}}, \end{cases} \tag{15}$$

$\varphi^{(c,\varepsilon)}$ and $\psi^{(c,\varepsilon)}$ are, respectively the (c, ε)-transforms of φ and ψ defined in (7). We refer the reader to [16] for a derivation of this result. In practice, it is through (15) that we will build our numerical method: we will solve for φ and ψ, after which the CNF will be recovered.

Recovering the flow and density: We first define the convolution operator

$$\mathcal{H}_s[f](y) := \frac{1}{(2\pi s)^{\frac{d}{2}}} \int_{\mathbb{R}^d} f(x) \exp\left(-\frac{1}{2s}\|x - y\|^2\right) \mathrm{d}x, \tag{16}$$

which smooths the operand with the Normal distribution of variance s; this is sometimes called the heat kernel. Note the similarities with the (c, ε)-transform. Let (φ, ψ) be the optimal Entropy-Kantorovich potentials in (15), and define $\varphi_t := \varepsilon \log \mathcal{H}_{t\varepsilon}[e^{\varphi/\varepsilon}]$ and $\psi_t := \varepsilon \log \mathcal{H}_{(1-t)\varepsilon}[e^{\psi/\varepsilon}]$. Then the *entropic-displacement interpolation* $\rho_t^\varepsilon : [0, 1] \to \mathcal{P}_2(\mathbb{R}^d)$ between the probability densities $p_{\mathcal{D}}$ and $p_{\mathcal{N}}$ and the corresponding velocity field v_t^ε are given by [25]

$$\rho_t^\varepsilon(x) = \exp\left((\varphi_t(x) + \psi_t(x))/\varepsilon\right), \text{ and} \tag{17}$$

$$v_t^\varepsilon(x) = \nabla\left(\varphi_t(x) - \psi_t(x)\right)/2, \tag{18}$$

The entropic interpolant ρ_t^ε given by (17) is the regularized analogue to the constant speed geodesic defined in Section 2.2. Moreover, as $\varepsilon \to 0$, $\rho_t^\varepsilon \to \rho_t$, the 2-Wasserstein geodesic between $p_{\mathcal{D}}$ and $p_{\mathcal{N}}$ (9) introduced in Section 2.2 (see e.g. [25]). In Appendix C, we illustrate the smoothing effect of entropic regularization on the 2-Wasserstein geodesic between two Gaussian distributions, where a closed-form solution is known.

We emphasize that once φ and ψ are known, we can completely defined a continuous normalizing flow between the data distribution and the standard Normal via equations (18) and the ODE (3).

3 Numerics

We now have the necessary tools to build CNFs by solving for Entropy-Kantorovich potentials. We parameterize the pair of Entropy-Kantorovich potentials (φ, ψ) as neural networks $(\varphi_\theta, \psi_\omega)$ with respective parameters θ and ω. We solve the dual problem (11) by maximizing the pair of functionals (15) over batches sampled from $p_{\mathcal{D}}$ and $p_{\mathcal{N}}$. The complete pseudo-code of our training procedure is outlined in Algorithm 1.

Alternating between optimizing φ and ψ In practice, we take alternating gradient ascent steps on the functional J in φ and the functional I in ψ. This alternating approach is motivated by the following.

Proposition 1 (Lemma 2.6 in [12]) *The dual function $D_\varepsilon : \mathrm{L}^{\exp}_{\varepsilon}(\mathbb{R}^d; \mathrm{d}x) \times \mathrm{L}^{\exp}_{\varepsilon}(\mathbb{R}^d; \mathrm{d}y) \to \mathbb{R}$ defined as in (12) is concave in each one of the variables. Moreover,*

$$D_\varepsilon(\varphi, \varphi^{(c,\varepsilon)}) \geq D_\varepsilon(\varphi, \psi),\ \forall\, \varphi \in \mathrm{L}^{\exp}_{\varepsilon}(\mathbb{R}^d; \mathrm{d}x),$$

$$D_\varepsilon(\varphi, \varphi^{(c,\varepsilon)}) = D_\varepsilon(\varphi, \psi) \text{ if and only if } \psi = \varphi^{(c,\varepsilon)}.$$

In particular we can say that $\varphi^{(c,\varepsilon)} \in \operatorname{argmax}\{D_\varepsilon(\varphi, \psi) \ :\ \psi \in \mathrm{L}^{\exp}_{\varepsilon}(\mathbb{R}^d; \mathrm{d}x)\}$. Clearly, an analogous results holds by exchanging the roles of φ and ψ.

In other words, we can create an increasing sequence of objective values by alternating between placing only φ and only ψ (with their respective (c, ε)-transforms in place of the other potential function) in the arguments of D_ε. We note that this sequence of objective function values is increasing only up to error induced by mini-batch sampling. An analysis of this error is outside the scope of this paper.

Then, at optimum, Theorem 1 tells us that the optimal potentials in (11) satisfy $\varphi^{(c,\varepsilon)} = \psi$ and $\psi^{(c,\varepsilon)} = \varphi$. Moreover, $D_\varepsilon(\varphi, \psi)$ is bounded above by the entropy-regularized functional $W_\varepsilon(p_{\mathcal{D}}, p_{\mathcal{N}})$, since $D_\varepsilon(\varphi, \psi) \leq W_\varepsilon(p_{\mathcal{D}}, p_{\mathcal{N}}) - \varepsilon$, $\forall\, \varphi, \psi$ (see also Lemma 2.10 in [12]).

Fast approximate (c, ε)-transform The (c, ε)-transformation building the above sequences can be approximated efficiently via Monte-Carlo (MC) integration with N samples $x_i \sim \mathcal{N}(y, \varepsilon)$:

$$(\varphi_\theta)^{(c,\varepsilon)}(y) = \varepsilon \log\left(\frac{1}{(2\pi\varepsilon)^{\frac{d}{2}}} \int_{\mathbb{R}^d} e^{\frac{\varphi_\theta(x) - \frac{1}{2}\|x-y\|^2}{\varepsilon}} \mathrm{d}x\right) \approx \varepsilon \log\left(\frac{1}{N} \sum_{i=1}^{N} e^{\varphi_\theta(x_i)/\varepsilon}\right). \tag{19}$$

Monte-Carlo integration is well known to be close the true integral point-wise with an error of $\mathcal{O}(N^{-1/2})$ in the number of samples (for fixed dimension d) [32]. We can safely omit the second term in (7), as we are only interested in the argmax of the objective function, and not the optimal function value. We will also use MC integration for a fast evaluation of the heat kernel $\mathcal{H}$.

Constructing the CNF and the velocity field v_t Finally, upon optimizing for φ_θ and ψ_ω, the optimal vector field generating the CNF is given by (18). The CNF framework [18] allows us to both estimate probability density and generate samples. For a given $x_i \in \mathcal{D}$, the log-likelihood of the data point is computed via (8), where the transformation is provided by solving (3). Generation is done by sampling $z_i \sim \mathcal{N}(0, 1)$ and running (3) backwards in time. Note that because we use MC integration for the heat kernel, computation of v_t is mesh-free, quick, and scales easily to high dimensions.

Speeding optimization by reinforcing the (c, ε)-transform In practice we have found optimization is helped by reinforcing the constraint that $\varphi^{(c,\varepsilon)} = \psi$ and $\psi^{(c,\varepsilon)} = \varphi$. To do so, we re-define

Algorithm 1 Dual ascent of potential functions, parameterized by neural networks

```
Input: Target dataset $\mathcal{D} \subseteq \mathbb{R}^d$, $\varepsilon > 0$; $N,\ B,\ k_{\max} \in \mathbb{N}$; $k = 0$ and step-size $\eta > 0$
Initialize networks $\varphi_{(0)}, \psi_{(0)}$
while $k < k_{\max}$ do
    for $x_B \in \mathcal{D}$ do
        Sample $z_B \sim \mathcal{N}(0, I_d)$                                ▷ Sampling from $p_{\mathcal{N}}$
        Compute $\varphi^{(c,\varepsilon)}_{(k)}$ and $\psi^{(c,\varepsilon)}_{(k)}$ with MC integration, using $N$ samples
        Update $\varphi_{(k)}$ using a stochastic optimizer over data $(x_B, z_B)$, with $\psi_{(k)}$ fixed:
            $\varphi_{(k+1)} \leftarrow \varphi_{(k)} + \eta \nabla \tilde{J}(\varphi_{(k)})$
        Update $\psi_{(k)}$ using a stochastic optimizer over data $(x_B, z_B)$, with $\varphi_{(k+1)}$ fixed:
            $\psi_{(k+1)} \leftarrow \psi_{(k)} + \eta \nabla \tilde{I}(\psi_{(k)})$
    end for
    $k \leftarrow k + 1$
end while
```

the objective dual function (12) with an extra auxiliary variable

$$D_\varepsilon(\varphi, \tilde{\varphi}, \psi) = \int_{\mathbb{R}^d} \varphi dp_{\mathcal{D}} + \int_{\mathbb{R}^d} \tilde{\varphi} dp_{\mathcal{N}} - \varepsilon \int_{\mathbb{R}^d \times \mathbb{R}^d} \exp\left\{ \frac{\varphi + \psi - \frac{1}{2}\|x - y\|^2}{\varepsilon} \right\} \mathrm{d}x\mathrm{d}y.$$

Optimization then alternates over the twin functionals $\tilde{J}$ and $\tilde{I}$, which are motivated by Proposition 1 and equation (15)

$$\tilde{J}(\varphi_\theta) := D_\varepsilon(\varphi_\theta, \varphi_\theta^{(c,\varepsilon)}, \psi_\omega) + \alpha \|\varphi_\theta^{(c,\varepsilon)} - \psi_\omega\|^2, \tag{20}$$

$$\tilde{I}(\psi_\omega) := D_\varepsilon(\psi_\omega, \psi_\omega^{(c,\varepsilon)}, \varphi_\theta) + \alpha \|\psi_\omega^{(c,\varepsilon)} - \varphi_\theta\|^2. \tag{21}$$

We have incorporated an additional L_2-regularization term with strength $\alpha > 0$ for extra reinforcement of the optimality conditions over mini-batches.

Examples We consider several low-dimensional distributions commonly used in the normalizing flow literature, some of which are discontinuous (e.g. checkerboard). For these experiments, we parameterize the two Entropy-Kantorovich potential functions $(\varphi_\theta, \psi_\omega)$ using four fully connected linear layers with ReLU activations, with hidden dimension 64. The hyper-parameters for the experiments are provided in Appendix D and, apart from the total number of iterations, are the same for each dataset. Indeed, we observed that some of the distributions were more difficult to model than others, and needed more time to optimize over the function space.

In Figure 2, we present the ground-truth log-densities, our estimated log-densities, and generated samples flowing from a standard Normal distribution to the target. The added blur in our estimated log-densities highlights the effect of the entropic interpolation (we trained with $\varepsilon = 1$), though the generated samples seem largely unaffected, and are well-concentrated.

4 Discussion and future work

We have presented a novel framework for density estimation and generative modelling with CNFs, based on well-establish results from entropy-regularized optimal transport. Rather than solving a dynamic problem, we exploit a dual formulation that easily takes advantage of the function-approximation abilities of neural networks. This allows us to define the estimated densities and their normalizing flows in (near) closed form. We studied toy problems, but the method we have presented readily extends to higher-dimensions, which we leave for future work.

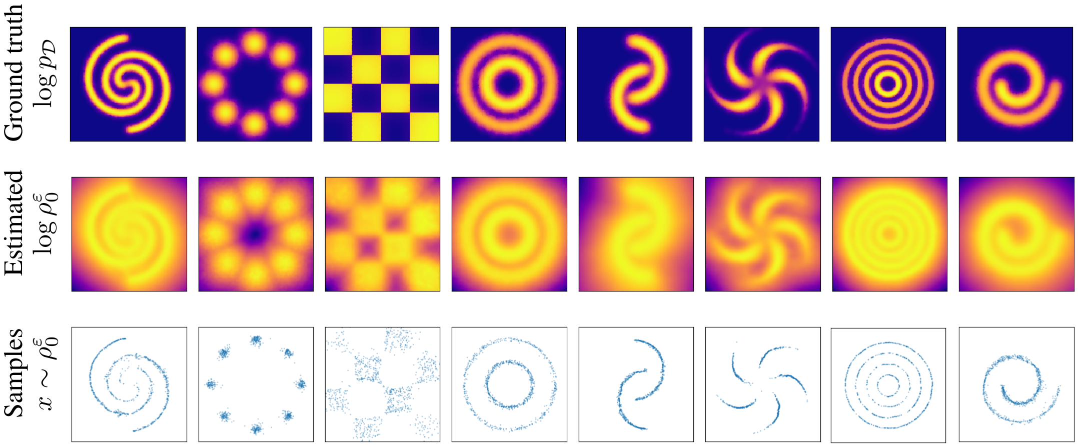

Figure 2: Estimated densities and generated samples using Entropy-Kantorovic potentials, on 2D examples. (Top row) Ground-truth log-densities; (Middle row) Our approximated log-density ρ_0^ε; (Bottom row) Generated samples flowing from standard Normal.

Broader Impact

This is a theoretical work and in the opinion of the authors, a discussion of broader impact is not applicable.

References

[1] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein Generative Adversarial Networks. In *Proceedings of the 34nd International Conference on Machine Learning, ICML 2017, Sydney, Australia, 7-9 August, 2017*, 2017.

[3] Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

[4] Rianne van den Berg, Leonard Hasenclever, Jakub M Tomczak, and Max Welling. Sylvester normalizing flows for variational inference. *arXiv preprint arXiv:1803.05649*, 2018.

[5] Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.

[6] Luis A Caffarelli. Monotonicity properties of optimal transportation and the FKG and related inequalities. *Communications in Mathematical Physics*, 214(3):547–563, 2000.

[7] Luis A Caffarelli. Erratum: Monotonocity of optimal transportation and the FKG and related inequalities (communication in mathematical physics (2000) 214 (547-563)). *Communications in Mathematical Physics*, 225(2):449–450, 2002.

[8] Tian Qi Chen, Jens Behrmann, David K Duvenaud, and Jörn-Henrik Jacobsen. Residual flows for invertible generative modeling. In *Advances in Neural Information Processing Systems*, pages 9913–9923, 2019.

[9] Tian Qi Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, pages 6571–6583, 2018.

[10] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pages 2292–2300, 2013.

[11] Marco Cuturi and Gabriel Peyré. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

[12] Simone Di Marino and Augusto Gerolin. An Optimal Transport approach for the Schrödinger bridge problem and convergence of Sinkhorn algorithm. *ArXiv: 1911.06850*, 2019.

[13] Chris Finlay, Jörn-Henrik Jacobsen, Levon Nurbekyan, and Adam M Oberman. How to train your neural ODE: the world of Jacobian and kinetic regularization. *International Conference on Machine Learning*, 2020.

[14] Joel Franklin and Jens Lorenz. On the scaling of multidimensional matrices. *Linear Algebra and its applications*, 114:717–735, 1989.

[15] Ivan Gentil, Christian Léonard, and Luigia Ripani. About the analogy between optimal transport and minimal entropy. In *Annales de la Faculté des Sciences de Toulouse. Mathématiques*, volume 3, pages 569–600, 2017.

[16] N. Gigli and L. Tamanini. Second order differentiation formula on $RCD^*(K, N)$ spaces. *J. Eur. Math. Soc. (JEMS)*, 2018.

[17] Nicola Gigli and Luca Tamanini. Benamou-Brenier and duality formulas for the entropic cost on $RCD^*(K, N)$ spaces. *Probab. Theory Related Fields*, 2018.

[18] Will Grathwohl, Ricky T. Q. Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. FFJORD: Free-form continuous dynamics for scalable reversible generative models. *International Conference on Learning Representations*, 2019.

[19] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems*, pages 5767–5777, 2017.

[20] Eldad Haber and Lars Ruthotto. Stable architectures for deep neural networks. *Inverse Problems*, 34(1):014004, 2017.

[21] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, pages 10215–10224, 2018.

[22] Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems*, pages 4743–4751, 2016.

[23] AV Kolesnikov. On sobolev regularity of mass transport and transportation inequalities. *Theory of Probability & Its Applications*, 57(2):243–264, 2013.

[24] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. A tutorial on energy-based learning. 2006.

[25] Christian Léonard. A survey of the Schrödinger problem and some of its connections with optimal transport. *Discrete & Continuous Dynamical Systems-A*, 34(4):1533–1574, 2014.

[26] Anton Mallasto, Jes Frellsen, Wouter Boomsma, and Aasa Feragen. (q, p)-Wasserstein GANs: Comparing ground metrics for Wasserstein GANs. *arXiv preprint arXiv:1902.03642*, 2019.

[27] Stefano Massaroli, Michael Poli, Michelangelo Bin, Jinkyoo Park, Atsushi Yamashita, and Hajime Asama. Stable neural flows. *arXiv preprint arXiv:2003.08063*, 2020.

[28] Robert J McCann. A convexity principle for interacting gases. *Advances in mathematics*, 128(1):153–179, 1997.

[29] 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.

[30] Michele Pavon, Esteban G Tabak, and Giulio Trigila. The data-driven Schroedinger bridge. *arXiv preprint arXiv:1806.01364*, 2018.

[31] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning*, pages 1530–1538, 2015.

[32] Christian Robert and George Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

[33] Lars Ruthotto and Eldad Haber. Deep neural networks motivated by partial differential equations. *Journal of Mathematical Imaging and Vision*, pages 1–13, 2019.

[34] Lars Ruthotto, Stanley J. Osher, Wuchen Li, Levon Nurbekyan, and Samy Wu Fung. A machine learning framework for solving high-dimensional mean field game and mean field control problems. *Proceedings of the National Academy of Sciences*, 117(17):9183–9193, 2020.

[35] Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

[36] Esteban G Tabak and Cristina V Turner. A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics*, 66(2):145–164, 2013.

[37] Esteban G Tabak, Eric Vanden-Eijnden, et al. Density estimation by dual ascent of the log-likelihood. *Communications in Mathematical Sciences*, 8(1):217–233, 2010.

[38] Jakub M Tomczak and Max Welling. Improving variational Auto-Encoders using Householder flow. *arXiv preprint arXiv:1611.09630*, 2016.

[39] Giulio Trigila and Esteban G Tabak. Data-driven optimal transport. *Communications on Pure and Applied Mathematics*, 69(4):613–648, 2016.

[40] Aaron van den Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George van den Driessche, Edward Lockhart, Luis Cobo, Florian Stimberg, Norman Casagrande, Dominik Grewe, Seb Noury, Sander Dieleman, Erich Elsen, Nal Kalchbrenner, Heiga Zen, Alex Graves, Helen King, Tom Walters, Dan Belov, and Demis Hassabis. Parallel WaveNet: Fast high-fidelity speech synthesis. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3918–3926, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[41] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[42] Hanshu Yan, Jiawei Du, Vincent Tan, and Jiashi Feng. On robustness of neural ordinary differential equations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[43] Linfeng Zhang, E Weinan, and Lei Wang. Monge-Ampère flow for generative modeling. *arXiv preprint arXiv:1809.10188*, 2018.

Suplementary material: Learning normalizing flows from Entropy-Kantorovich potentials

A Some concepts from Optimal Transport Theory

We briefly introduce the Monge problem in $\mathbb{R}^d$ for the distance square cost function and highlight the relation with the Kantorovich relaxation. For more detail, see e.g. [35, 41]. First, let us recall the definition of push-forward of a measure.

The push-forward of a measure

Let $T : \mathbb{R}^d \to \mathbb{R}^d$ be a Borel function and μ, ν be probability measures in $\mathbb{R}^d$. The push-forward measure $T_\sharp\mu \in \mathcal{P}(\mathbb{R}^d)$ is defined by

$$T_\sharp\mu(A) := \mu(T^{-1}(A)) = \mu\left(\left\{x \in \mathbb{R}^d : T(x) \in A\right\}\right) \text{ for any Borel measurable set } A \subset \mathbb{R}^d. \quad (1)$$

Equivalently, one can write $T_\sharp\mu$ in integral terms

$$\int_{\mathbb{R}^d} h(y) dT_\sharp\mu(y) = \int_{\mathbb{R}^d} h(T(x)) d\mu(x), \quad \forall\, h : \mathbb{R}^d \to \mathbb{R} \text{ Borel.} \quad (2)$$

In particular, if we assume T additionally differentiable, $T_\sharp\mu = \nu$ can be simply written as classical change of variables formula

$$\mu(x) = \nu(T(x))|\det \mathrm{J}_T(x)|. \quad (3)$$

Then, by applying the $\log$ in both sides in (3) one has

$$\log \mu(x) = \log \nu(T(x)) + \log |\det \mathrm{J}_T(x)|,$$

where J_T denotes the Jacobian of a map T.

Monge problem and its Kantorovich relaxation

The Monge problem seeks to find the best optimal transport map *transporting* μ and ν, i.e. $T_\sharp\mu = \nu$, that minimizes the total work

$$\inf_{T_\sharp\mu=\nu} \int_{\mathbb{R}^d} \frac{1}{2}\|x - T(x)\|^2 d\mu(x) = \inf\left\{\int_{\mathbb{R}^d} \frac{1}{2}\|x - T(x)\|^2 d\mu(x) : T : \mathbb{R}^d \to \mathbb{R}^d \text{ Borel and } T_\sharp\mu = \nu\right\}. \quad (4)$$

In general, the problem (4) does not always admit a minimizer. The class of functions $\mathcal{T}(\mu, \nu) = \{T : \mathbb{R}^d \to \mathbb{R}^d : T_\sharp\mu = \nu \text{ and } T \text{ Borel }\}$ can even be empty. It is enough to take, for example $\mu = \delta_{x_1}$ and $\nu = \frac{1}{2}\delta_{y_1} + \frac{1}{2}\delta_{y_2}$.

The Kantorovich *relaxation* instead

$$\min_{\gamma\in\Pi(\mu,\nu)} \mathcal{C}(\gamma) := \min_{\gamma\in\Pi(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \frac{1}{2}\|x - y\|^2 d\gamma(x, y), \quad (5)$$

admits a minimizer, since the set $\Pi(\mu, \nu)$ is compact and the cost function $\mathcal{C}(\gamma)$ is lower semi-continuous in the weak*-topology (convergence in law). Notice that the set of transport maps $\mathcal{T}(\mu, \nu)$ can be identified with a subset of $\Pi(\mu, \nu)$ by writing for every $T \in \mathcal{T}(\mu, \nu)$, $\gamma_T = (\mathrm{Id}, T)_\sharp\mu \in \Pi(\mu, \nu)$. Then,

$$\min_{\gamma\in\Pi(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \frac{1}{2}\|x - y\|^2 d\gamma(x, y) \le \inf_{T_\sharp\mu=\nu} \int_{\mathbb{R}^d} \frac{1}{2}\|x - T(x)\|^2 d\mu(x).$$

Under some hypothesis on μ and ν, one can also show that the equality holds in the above equation. In other words, the solution of (5) is of *Monge-type*, $\gamma_T = (\mathrm{Id}, T)_\sharp\mu$. This is precisely the statement of Brenier's Theorem.

Theorem 2 (Brenier) *Let μ and ν be Borel probability measures on $\mathbb{R}^n$, $c(x,y) = \frac{1}{2}\|x-y\|^2$ be a cost function and suppose μ has a density with respect to Lebesgue. Then the optimal plan γ solving (5) is supported on the graph of a map $T : \mathbb{R}^d \to \mathbb{R}^n$ satisfying $T_\sharp \mu = \nu$ (i.e. $T \in \mathcal{T}(\mu,\nu)$), i.e. $\gamma = (\mathrm{Id}, T)_\sharp \mu$. Moreover, this map is unique and there exists a convex function u such that $T(x) = \nabla u(x)$.*

As a consequence, the Monge problem (4) admits a unique minimizer.

A natural question is to enquire when the optimal map T in (4) is differentiable, allowing us to write the condition $T_\sharp \mu = \nu$ as in (3). One theoretical and insightful result due to Caffarelli guarantees the regularity of the potentials u. Assume that μ has compact support and ν has finite second moments. Then, at least when $\mu(x) = \exp(-W(x) - |x|^2)\mathrm{d}y$ and $\nu(y) = \exp(V(y) - |y|^2)\mathrm{d}x$ with V, W convex, the map $T = \nabla u$ is 1-Lipschitz and $\nu = T_\sharp \mu$ [6, 7, 23].

B Absolutely continuous curves and geodesics in $\mathbb{W}_2$

Let $\rho(t)$ be a curve in $\mathcal{P}(\mathbb{R}^d)$, i.e. $\rho : [0,1] \to \mathcal{P}(\mathbb{R}^d)$, the metric derivative of $\rho(t)$ denoted by $|\dot\rho|(t)$ is defined by

$$|\dot\rho|(t) = \lim_{h\to 0^+} \frac{W_2(\rho(t+h), \rho(t))}{h} \qquad \text{provided the limit exists.}$$

The following theorems guarantee the existence of the metric derivative for Lipschitz curves in $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ and relate absolutely continuous curves in $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ with solutions of the continuity equation. We refer to [1] for the prove and further details.

Theorem 3 *Suppose that $\rho : [0,1] \to \mathcal{P}(\mathbb{R}^d)$ is Lipschitz continuous, i.e. for all $s,t \in [0,1]$, $W_2(\rho_s, \rho_t) \le L|t-s|$, for $L > 0$. Then the metric derivative $|\dot\rho|(t)$ exists for almost every $t \in [0,1]$. Moreover, for all $t < s$*

$$W_2(\rho(t), \rho(s)) \le \int_t^s |\dot\rho|(a) da.$$

Definition 1 *A curve $\rho : [0,1] \to \mathcal{P}(\mathbb{R}^d)$ is said to be absolutely continuous if there exists a function f such that*

$$W_2(\rho(t), \rho(s)) \le \int_t^s f(a) da, \quad \forall\, s < t.$$

The next theorems relates the continuity equation and an ODE flows constructed in this paper. We refer to [35] for the proofs and in-depth discussion of the results.

Theorem 4 *Let $(\rho_t)_{t\in[0,1]}$ be an absolutely continuous curve in $(\mathcal{P}_2(\mathbb{R}^d), W_2)$. Then, there exists a vector field $v_t \in L^2(\rho_t, \mathbb{R}^d)$ such that the continuity equation $\partial_t \rho_t + \nabla \cdot (v_t \rho_t) = 0$ is satisfied in the weak sense and, for almost every $t \in [0,1]$, $|v_t|_{L^2(\rho_t)} \le |\dot\rho|(t)$. Moreover, the converse also holds: if $(\rho_t)_{t\in[0,1]}$ is a curve in $(\mathcal{P}_2(\mathbb{R}^d), W_2)$, $v_t \in L^2(\mathbb{R}^d, \rho_t)$ such that $\int_0^1 \int_\Omega |v_t|^2 \rho_t \mathrm{d}x\mathrm{d}t < +\infty$ solving $\partial_t \rho_t + \nabla \cdot (v_t \rho_t) = 0$, then ρ_t is absolutely continuous in W_2 and for almost every $t \in [0,1]$, $|\dot\rho|(t) \le |v_t|_{L^2(\rho_t)}$.*

Definition 2 *A curve $\rho : [0,1] \to X$ is said to be a geodesic between μ and $\nu \in X$ if it minimizes the length among all curves such that $\rho(0) = \mu$ and $\rho(1) = \nu$.*

Let us denote by $\mathrm{L}(\rho)$ the length of a curve $\rho : [0,1] \to X$,

$$\mathrm{L}(\rho) := \sup\left\{ \sum_{k=0}^{n-1} d(\rho(t_k), \rho(t_{k+1})) \;:\; n \ge 1,\, 0 = t_0 < t_1 < \cdots < t_n = 1 \right\}.$$

A space (X, d) is said to be a *geodesic space* if it holds

$$d(\mu,\nu) = \min\{\mathrm{L}(\rho) \;:\; \rho \text{ is absolutely continuous, } \rho(0) = \mu, \rho(1) = \nu\},$$

i.e. there exist geodesics between arbitrary points.

Proposition 2 **($(\mathcal{P}_p(\Omega), W_2)$ is a geodesic space)** *Let $\Omega \subset \mathbb{R}^d$ be convex, $\mu, \nu \in \mathcal{P}_p(\Omega)$ and $\gamma \in \Pi(\mu, \nu)$ an optimal transport plan for the cost $c(x, y) = |x - y|^p$, $p \geq 1$. Define the curve $\pi_t : \Omega \times \Omega \to \Omega$ through $\pi_t(x, y) = (1 - t)x + ty$. Then the curve $\rho_t = (\pi_t)_\sharp\gamma$ is a constant speed geodesic in $(\mathcal{P}_p(\Omega), W_p)$ from μ to ν. In particular, when an optimal transport plan $\gamma = \gamma_T$ is concentrated in a map T, the curve $\rho_t = ((1 - t)\mathrm{Id} + tT)_\sharp\mu$.*

Proposition 3 *Let μ, ν be two densities in $\mathcal{P}_p(\Omega), p \geq 2$, $\rho_t = (\pi_t)_\#\gamma$ be the geodesic connecting μ to ν introduced in Proposition 2 and $T_t(x) = (1 - t)x + tT(x)$, where T is the optimal transport map between μ to ν. Then the velocity field $v_t(y) = (T - \mathrm{Id})(T_t^{-1}(y))$ is well defined on* $\mathrm{spt}(\rho_t)$ *for each $t \in]0, 1[$ and satisfies*

$$\partial_t\rho_t + \nabla \cdot (\rho_t v_t) = 0, \quad \|v_t\|_{L^p(\rho_t)} = |\dot{\rho}|(t) = W_p(\mu, \nu).$$

C Dynamical formulation of the Entropy-regularized Optimal Transport

The variational problem (10) can be alternatively writen in the *dynamic* formulation [15, 17, 25]

$$W_\varepsilon^2(p_\mathcal{D}, p_\mathcal{N}) = \min_{(\rho_t^\varepsilon, w_t^\varepsilon)} \int_0^1 \int_{\mathbb{R}^d} \frac{\|w_t^\varepsilon\|^2}{2} d\rho_t^\varepsilon dt + \frac{\varepsilon}{2}\left(\mathrm{H}(p_\mathcal{D}) + \mathrm{H}(p_\mathcal{N})\right), \tag{6}$$

$$= \sup_{(\varphi_t^\varepsilon, \psi_t^\varepsilon)} \int_{\mathbb{R}^d} (\varphi_1^\varepsilon - \psi_1^\varepsilon) dp_\mathcal{D} + \int_{\mathbb{R}^d} (\varphi_0^\varepsilon - \psi_0^\varepsilon) dp_\mathcal{N} + \frac{\varepsilon}{2}\left(\mathrm{H}(p_\mathcal{D}) + \mathrm{H}(p_\mathcal{N})\right), \tag{7}$$

where the minimum must be understood as taken among all couples $(\rho_t^\varepsilon, w_t^\varepsilon)$ solving the backward and forward Fokker-Planck equations

$$-\partial_t\rho_t^\varepsilon + \nabla \cdot (\nabla\varphi_t^\varepsilon\rho_t^\varepsilon) = \frac{\varepsilon}{2}\Delta\rho_t^\varepsilon, \quad \text{and} \quad \partial_t\rho_t^\varepsilon + \nabla \cdot (\nabla\psi_t^\varepsilon\rho_t^\varepsilon) = \frac{\varepsilon}{2}\Delta\rho_t^\varepsilon,$$

for $t \in [0, 1]$ such that $\rho_0^\epsilon = p_\mathcal{D}, \rho_1^\varepsilon = p_\mathcal{N}$; while the supremum is taking over the couple $(\varphi_t^\varepsilon, \psi_t^\varepsilon)$ solving the Hamilton-Jacobi-Bellman equations

$$\partial_t\varphi_t^\varepsilon = \frac{\|\nabla\varphi_t^\varepsilon\|^2}{2} + \frac{\varepsilon}{2}\Delta\varphi_t^\varepsilon, \quad \text{and} \quad -\partial_t\psi_t^\varepsilon = \frac{\|\nabla\psi_t^\varepsilon\|^2}{2} + \frac{\varepsilon}{2}\Delta\psi_t^\varepsilon.$$

The optimal vector field w_t^ε is given by the Entropy-Kantorovich potentials $w_t^\varepsilon = \nabla(\varphi_t^\varepsilon - \psi_t^\varepsilon)/2$, which corresponds to the regularized constant speed geodesic in the 2-Wasserstein space.

By writing $w_t^\varepsilon = v_t^\varepsilon - \varepsilon\nabla\log(\rho_t^\varepsilon)$, the variational problem (6) corresponds to eq (1)

$$\frac{\varepsilon}{2}\left(\mathrm{H}(p_\mathcal{D}) + \mathrm{H}(p_\mathcal{N})\right) + \inf_{(\rho_t, v_t^\varepsilon)} \int_0^1 \int_{\mathbb{R}^d} \left(\frac{\|v_t^\varepsilon\|^2}{2} + \frac{\varepsilon^2}{8}\|\nabla\log\rho_t\|^2\right)\rho_t\, dx\, dt, \tag{8}$$

where $(\rho_t^\varepsilon, v_t^\varepsilon)$ is such that $\rho_0^\varepsilon = p_\mathcal{D}$, $\rho_1 = p_\mathcal{N}$ and solves the continuity equation

In the following, we give a formal computation explaining the optimal conditions obtained via the above primal-dual relation.

Characterization (6) via primal-dual problems

Let us assume that φ_t and ψ_t solves the respective HJB equations and define $\alpha_t = (\varphi_t - \psi_t)/2$. Let us compute

$$\left.\frac{d}{ds}\right|_{s=t} \int_{\mathbb{R}^d} \mu_s^\varepsilon\alpha_s dx = \int_{\mathbb{R}^d} \left.\frac{d}{ds}\right|_{s=t} \mu_s^\varepsilon\alpha_s dx + \int_{\mathbb{R}^d} \mu_s^\varepsilon \left.\frac{d}{ds}\right|_{s=t} \alpha_s dx =: (\mathrm{I}) + (\mathrm{II}). \tag{9}$$

Since φ_t and ψ_t solves the respective HJB equations, we have

$$\begin{aligned}(\mathrm{II}) = \int_{\mathbb{R}^d} \mu_s^\varepsilon \left.\frac{d}{ds}\right|_{s=t} \alpha_s dx &= \int_{\mathbb{R}^d} \left(-\frac{\|\nabla\psi_t\|^2}{4} - \frac{\|\nabla\varphi_t\|^2}{4} - \frac{\varepsilon}{4}\Delta(\psi_t + \varphi_t)\right)\mu_s^\varepsilon dx \\ &= \int_{\mathbb{R}^d} \left(-\frac{\|\nabla\psi_t\|^2}{4} - \frac{\|\nabla\varphi_t\|^2}{4} + \frac{1}{4}\langle\nabla(\psi_t + \varphi_t), \varepsilon\nabla\log(\mu_s^\varepsilon)\rangle\right)\mu_s^\varepsilon dx \\ &\leq \int_{\mathbb{R}^d} \left(-\frac{\|\nabla\psi_t\|^2}{4} - \frac{\|\nabla\varphi_t\|^2}{4} + \frac{1}{8}\|\nabla(\psi_t + \varphi_t)\|^2 + \frac{\varepsilon^2}{8}\|\nabla\log(\mu_s^\varepsilon)\|^2\right)\mu_s^\varepsilon dx\end{aligned}$$

The second line follows from the first by applying integration by parts to pass a gradient onto the measure μ^ε, then multiplying and dividing by μ^ε, and then using $(\nabla\mu^\varepsilon)/\mu^\varepsilon = \nabla \log \mu^\varepsilon$. The last line is with equality if and only if $\varepsilon\nabla\log(\mu_t^\varepsilon) = \nabla(\psi_t + \varphi_t)$ almost everywhere.

Now, if (μ_t^ε, v_t) solves the continuity equation then

$$\text{(I)} = \int_{\mathbb{R}^d} \frac{d}{ds}\bigg|_{s=t} \mu_s^\varepsilon \alpha_s dx = -\int_{R^d} \alpha_s \nabla\cdot(v_t\mu_t^\varepsilon)dx = \int_{\mathbb{R}^d} \langle \nabla(\psi_t - \varphi_t)/2, v_t\rangle \mu_t^\varepsilon dx \tag{10}$$

$$\leq \int_{\mathbb{R}^d} \frac{1}{2}\|\nabla(\psi_t - \varphi_t)/2\|^2 + \frac{1}{2}\|v_t\|^2\mu_t^\varepsilon dx, \tag{11}$$

with equality if and only if $v_t = \nabla(\psi_t - \varphi_t)/2$. Finally, integrating (9) over time one has

$$\frac{1}{2}\int_{\mathbb{R}^d}(\psi_1 - \varphi_1)d\rho_1 + \frac{1}{2}\int_{\mathbb{R}^d}(\psi_0 - \varphi_0)d\rho_0 \leq \int_0^1\int_{\mathbb{R}^d} \frac{\|v_t\|^2}{2} + \frac{\varepsilon}{8}\|\nabla\log\mu_t^\varepsilon\|^2 dxdt. \tag{12}$$

Since all the computations above are arbitrary we have that

$$\sup_{(\psi_t,\varphi_t)} \left\{ \frac{1}{2}\int_{\mathbb{R}^d}(\psi_1 - \varphi_1)d\rho_1 + \frac{1}{2}\int_{\mathbb{R}^d}(\psi_0 - \varphi_0)d\rho_0 : \begin{array}{r} \partial_t\varphi_t = \dfrac{\|\nabla\varphi_t\|^2}{2} + \dfrac{\varepsilon}{2}\Delta\varphi_t \\ -\partial_t\psi_t = \dfrac{\|\nabla\psi_t\|^2}{2} + \dfrac{\varepsilon}{2}\Delta\psi_t \end{array} \right\} \leq$$

$$\leq \inf_{(\mu_t^\varepsilon, v_t)} \left\{ \int_0^1\int_{\mathbb{R}^d} \frac{\|v_t\|^2}{2} + \frac{\varepsilon}{8}\|\nabla\log\mu_t^\varepsilon\|^2 dxdt \ : \begin{array}{r} \partial_t\mu_t^\varepsilon + \nabla\cdot(v_t\mu_t^\varepsilon) = 0 \\ \mu_0^\varepsilon = p_{\mathcal{D}}, \mu_1^{\varepsilon^2} = p_{\mathcal{N}} \end{array} \right\}$$

The equality is reached when $v_t = \nabla(\psi_t - \varphi_t)/2$ and μ_t^ε is the entropic interpolation

$$\mu_t^\varepsilon := \mathcal{H}_{t\varepsilon}(e^{\varphi^\varepsilon})\,\mathcal{H}_{(1-t)\varepsilon}(e^{\psi^\varepsilon}).$$

In particular, at the optimal (μ_t^ε, v_t) we have that

$$\frac{1}{2}\int_{\mathbb{R}^d}(\psi_1 - \varphi_1)p_{\mathcal{N}}\mathrm{d}y + \frac{1}{2}\int_{\mathbb{R}^d}(\psi_0 - \varphi_0)p_{\mathcal{D}}\mathrm{d}x = \int_0^1\int_{\mathbb{R}^d} \frac{\|v_t\|^2}{2} + \frac{\varepsilon^2}{8}\|\nabla\log\mu_t^\varepsilon\|^2\mu_t^\varepsilon \mathrm{d}x\mathrm{d}t.$$

Closed-form solutions for d-dimensional Gaussians

We illustrate in the following example and accompanied Figure 1 the smoothing effect of the regularization on 2-Wasserstein geodesics for two Gaussian distributions. In the example, we notice that the initial distribution has a degenerate covariance structure that is maintained in the 2-Wasserstein case. When incorporating regularization, we see a smoothed out distribution that more closely resembles the target throughout the flow.

Example 1 (Comparing geodesics and entropic interpolation for Gaussian distributions) *Consider two two multivariate Gaussian distributions $\rho_0 = \mathcal{N}(m_0, \Sigma_0)$ and $\rho_1 = \mathcal{N}(m_1, \Sigma_1)$. The geodesics under the Wasserstein metric is given by $\rho_t = \mathcal{N}(m_t, \Sigma_t)$ [28] with $m_t = (1-t)m_0 + tm_1$ and*

$$\Sigma_t = (1-t)^2\Sigma_0 + t^2\Sigma_1 + t(1-t)[(\Sigma_0\Sigma_1)^{1/2} + (\Sigma_1\Sigma_0)^{1/2}].$$

The entropic interpolation is $\rho_t^\varepsilon = (m_t, \Sigma_t^\varepsilon)$ $t \in [0,1]$, where m_t is the same as before, and

$$\Sigma_t^\varepsilon = (1-t)^2\Sigma_0 + t^2\Sigma_1 + t(1-t)\left[\left(\frac{\varepsilon^2}{16}I + \Sigma_0\Sigma_1\right)^{1/2} + \left(\frac{\varepsilon^2}{16}I + \Sigma_1\Sigma_0\right)^{1/2}\right].$$

Notice that the covariance structures are the same up to a function of ε that appears in the mixing term. Clearly $\Sigma_t^\varepsilon \to \Sigma_t$ when $\varepsilon \to 0$.

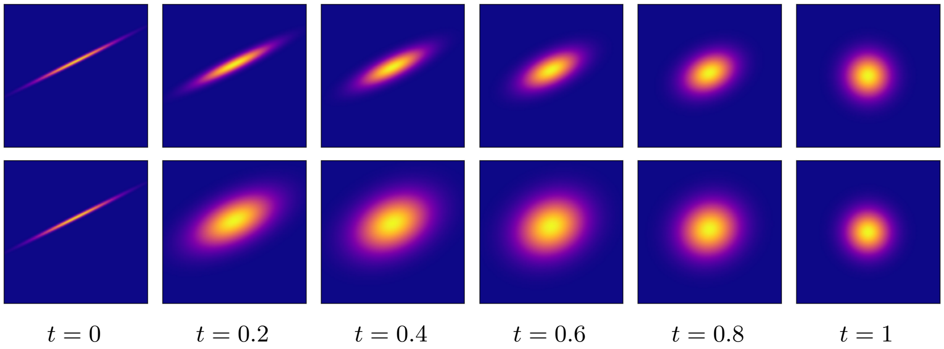

Figure 1: Flow between two Normal distributions: the source distribution has a degenerate covariance structure and the target is the standard Normal distribution. (Top) W_2 geodesics (Bottom) The entropy-regularized interpolation.

D Algorithm details and hyperparameters

For all the experiments, we use four fully connected linear layers with ReLU activations. The hidden dimension of the layers was 64.

The 2D datasets considered are: ‘Checkerboard’, ‘Swissroll’, ‘Rings’ (four concentric rings), ‘Moons’, ‘Circles’ (two concentric rings), ‘2spirals’, ‘Pinwheel’, and ‘8gaussians’.

For *training*, the following hyperparameters are constant across all datasets:

- Batch-size for the sampled data (and sampling from the Normal distribution) was 1000
- Number of samples for the Monte-Carlo (MC) integration was 100
- Learning rate for stochastic gradient descent was 10^{-3}
- L_2 penalty term to enforce the (c, ε)-transformation was 10^{-5}

For all but the ‘Rings’ dataset, the number of iterations was 20000 — for ‘Rings’, we needed to use 40000 iterations.

Finally, for *generating* samples, we used a batch-size of 1000, 200 MC samples, and used the default Dormand-Prince Runge-Kutta 4(5) adaptive solver (`dopri5`) ODE integrator from the `torchdiffeq` Python package [9].