Diffusion models and Optimal Transport

Gabriel Peyré

December 4, 2024

Abstract

This note recaps the main ingredient of the very neat proof of Hugo Lavenant and Filippo Santambrogio [1] that in general, diffusion models do not define optimal transport mapping. We keep the derivation informal, the goal being to insist on the proof techniques, which is not fully constructive but relies on the nice idea of differentiating through time a family of diffusion mappings to reach a contradiction.

Generative models aim to build a transportation map T between a reference distribution α (typically an isotropic Gaussian) and the data distribution β. We denote by $T_\sharp\alpha$ the push-forward of α by T (if α is made of Dirac masses at x_i, then $T_\sharp\alpha$ is composed of Dirac masses at $T(x_i)$). The goal is thus to find T such that $T_\sharp\alpha = \beta$. It is easy to see that such a map always exists for any β, but finding an explicit constructive method for T is surprisingly non-trivial. Two standard approaches are optimal transport and integrating backward the advection field associated with diffusion. Two other classes of methods are flow matching (but for a Gaussian latent space, it is the same as the diffusion model up to a time reparameterization, so the same reasoning applies) and Dacorogna-Moser construction where $\beta_t = (1-t)\alpha + t\beta$ (which is problematic if the densities do not have the same support).

1 Optimal Transport

Optimal transport finds T by solving the Monge problem:

$$\min_T \left\{ \int \|T(x) - x\|^2 \,\mathrm{d}\alpha(x) : T_\sharp\alpha = \beta \right\}. \tag{1}$$

Brenier's celebrated theorem from 1991 states that this map exists, is unique, and can be written as the gradient of a convex function, $T = \nabla\varphi$. Conservation of mass, $T_\sharp\alpha = \beta$, equivalently means that φ solves the Monge-Ampère equation:

$$\det(\partial^2\varphi(x)) = \frac{\alpha(x)}{\beta(\nabla\varphi(x))}, \tag{2}$$

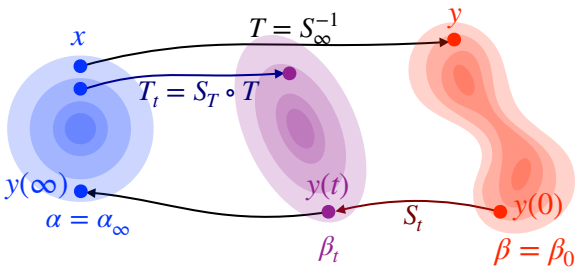
where the right-hand side involves the ratio of the densities of the measures.

2 Inverse Flow Map

Diffusion models consider an interpolation β_t between $\beta_0 = \beta$ and $\beta_\infty = \alpha = \mathcal{N}(0, \mathrm{Id})$, defined by solving:

$$\frac{\partial \beta_t}{\partial t} = \mathrm{div}(y\beta_t) + \Delta \beta_t, \tag{3}$$

(note that use y to denote the space variable since the evolution is in the reverse direction, from the data β to the latter variable α) which converges to $\beta_\infty = \alpha$. The solution depends linearly on β_0 through a Gaussian convolution.



Rewriting these equations in divergence form as

$$\frac{\partial \beta_t}{\partial t} + \mathrm{div}[v\beta_t] = 0 \quad \text{where} \quad v(y) := -y - \nabla \log(\beta_t)(y)$$

shows that, if one has already computed β_t, then this evolution can be obtained by evolving particles according to the vector field v. A map S_t between $\beta_0 = \beta$ and β_t is thus obtained by integrating the ODE defined by the score function $\nabla \log(\beta_t)$

$$\dot{y}(t) = v(y(t)) \tag{4}$$

The map S_t is the flow map

$$S_t : y(0) \mapsto y(t), \quad \text{where } y(t) \text{ solves (4)} \tag{5}$$

One has $S_0 = \mathrm{Id}$ and $(S_t)_\sharp \beta_0 = \beta_t$. Thus, in the limit $t \to \infty$, $(S_\infty)_\sharp \beta = \alpha$, so $T := S_\infty^{-1}$ is a valid transport, i.e., $T_\sharp \alpha = \beta$. We call this T the *inverse flow map*, which represents a deterministic diffusion model.

As a side note, the ML community is particularly excited about this idea because $\nabla \log(\beta_t)$ can be efficiently approximated from samples using denoising-based optimization. Numerically, the inverse flow map is approximated by considering a large, finite t and computing $T_t = S_t^{-1}$ by integrating the ODE. In practice, a stochastic ODE is often used, making the map non-deterministic, though this requires adjusting the bias term x in the ODE to account for extra diffusion.

3 Flow matching

Another related construction is flow matching. It is more general because α is not required to be an isotropic Gaussian, and it contains OT as a special case. With respect to diffusion models, it considers $t \in [0,1]$ in place of $t \in [0,+\infty]$. It assumes the knowledge of a coupling π between α and β. The simplest choice is the "independent" coupling $\pi = \alpha \otimes \beta$. More complex couplings, such as an OT coupling, could also be considered, but they are computationally too expensive. Using π, the interpolation between $\alpha = \alpha_0$ (at $t = 0$) and $\beta = \alpha_1$ (at $t = 1$) is obtained by push-forwarding using a linear interpolation:

$$\alpha_t := (P_t)_\sharp \pi, \quad \text{where } P_t(x,y) := (1-t)x + ty. \tag{6}$$

If $\pi = \alpha \otimes \beta$ and $\alpha = \frac{1}{n}\sum_i \delta_{x_i}$, $\beta = \frac{1}{m}\sum_j \delta_{y_j}$, then α_t consists of $n \times m$ Dirac masses traveling in straight lines:

$$\alpha_t = \frac{1}{nm}\sum_{i,j} \delta_{(1-t)x_i + ty_j}. \tag{7}$$

If $\pi = (\mathrm{Id}, T)_\sharp \alpha$ is a Brenier-type coupling, then $\alpha_t = ((1-t)\mathrm{Id} + tT)_\sharp \alpha$ is the so-called McCann OT interpolation. This interpolation is not directly useful for sampling from β, but it can be used to define a flow field v_t so that the Eulerian advection equation holds:

$$\frac{\partial \alpha_t}{\partial t} + \mathrm{div}(\alpha_t v_t) = 0. \tag{8}$$

A valid v_t can be found by solving the regression problem:

$$\min_{(v_t)_t} \int \|v_t((1-t)x + ty) - (y-x)\|^2 \,\mathrm{d}\pi(x,y). \tag{9}$$

Intuitively, this means that $v_t(x)$ at some point x should be the average velocity of all trajectories passing through x

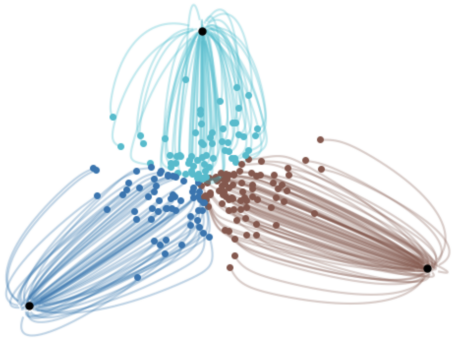
$$v_t(z) = \mathbb{E}_{(x,y)\sim\pi}\big[y - x \,\big|\, z = (1-t)x + ty\big].$$

Numerically, $v_t(x)$ can be parameterized by a neural network (e.g., a U-Net for vision tasks) and estimated using stochastic gradient descent on the objective in (9). If α is an isotropic, after a change of variable $t \in [0,+\infty[\to 1 - e^{e^t} \in [0,1[$ then v_t is the score $\nabla \log(\alpha_t)$ and score matching interpolation is equivalent to computing the inverse flow map. In the following, we thus only consider the inverse flow map.

4 Inverse Flow Maps Are Not Optimal Transport

It is natural to wonder whether the inverse flow map $T = S_\infty^{-1}$ is the solution to (1). In dimension 1, S_t defines a diffeomorphism, so S_∞ is monotonic and so

is $T = S_\infty^{-1}$. Thus, it is the gradient of a convex function, making it optimal by Brenier's theorem. If β is Gaussian, up to a rotation of the space to make the covariance diagonal, the diffusion map is defined by a monotonic map along each axis and is also an optimal transport. A last example is the radial case, so that β_t is also radial, and only needs to study the radial evolution, and in this case, it is also an optimal transport. The figure below shows the diffusion map for the case where β is supported in three Dirac. In this case, the mapping defines a tesselation of the space (a Voronoi tesselation for OT) and the tesselation of diffusion is most likely different from the one of OT.



Lavenant and Santambrogio showed by contradiction that, in general, the inverse flow map is not the optimal transport. They construct a β close to the isotropic Gaussian α, but instead of proving the conjecture is false for β, they show there exists some $t \geq 0$ such that the inverse flow map T_t from α to β_t is not an optimal transport. They actually shows that T_t is not an optimal transform for t which can be arbitrary close to 0 (but not necessarily 0 ...).

We denote S_t the flow map from $\beta_0 = \beta$ to β_t. If the conjecture is true, the inverse flow map T_t from α to β_t is an optimal transport for all t. By the composition rule of flow maps, this map is:

$$T_t := S_t \circ S_\infty^{-1} = S_t \circ T, \tag{10}$$

and $(T_t)_\sharp \alpha = \beta_t$. The goal is to show that T_t being an optimal transport for all t leads to a contradiction if β is well-chosen (specifically, very close to α with specific second- and fourth-order log-density derivatives at 0).

By Brenier's theorem, T_t being an optimal transport implies that it is the gradient of a convex function, which is equivalent to:

$$\partial T_t(x) \text{ is a positive symmetric matrix for all } x \text{ and } t. \tag{11}$$

Combining:

- Differentiating (11) with respect to t,
- Differentiating the flow ODE (4) with respect to x,

and then evaluating the obtained equation at $t = 0$, Hugo and Filippo showed by explicit computation that this leads to:

$$[\partial^2 \log(\beta)(T(x))] \times [\partial T(x)] \text{ is symmetric.} \tag{12}$$

Using the fundamental property:

$$A, B \text{ symmetric and } AB \text{ symmetric } \iff AB = BA,$$

it follows that (12) implies:

$$\forall y = T(x), \quad G(y) := \partial^2 \log(\beta)(y) \text{ commutes with } H(y) := \partial S_\infty(y). \tag{13}$$

To reach a contradiction, assume $G(y)$ and $H(y)$ commute for all y. Since $\beta = T_\sharp \alpha$, $\alpha = S_\sharp \beta$, and T and S are inverse optimal transport maps, we denote $S = \nabla \psi$ with ψ convex. The Monge-Ampère equation (2) implies:

$$\begin{aligned} G(y) &= \partial^2 \log(\beta)(y) = \partial^2 \left[\log\det(\partial^2 \psi) - \frac{1}{2}\|\nabla \psi\|^2 \right](y), \\ H(y) &= \partial S_\infty(y) = \partial^2 \psi(y). \end{aligned}$$

To make β close to α, consider:

$$\psi(y) = \frac{\|y\|^2}{2} + \varepsilon h(y),$$

for small ε. Expanding in Taylor series, after some computation:

$$\begin{aligned} G(y) &= -\mathrm{Id} + \varepsilon\{\partial^2[\Delta h] + \partial^2[y \cdot \Delta h]\} + o(\varepsilon), \\ H(y) &= \mathrm{Id} + \varepsilon \partial^2 h(y). \end{aligned}$$

Focusing on $y = 0$, the goal is to reach a contradiction by crafting h such that $\partial^2 h(0)$ and $\partial^2[\Delta h](0)$ do not commute. Around 0, h must be at least a polynomial of degree 4. An example in dimension 2 is:

$$h(y) = y_1 y_2 + y_1^4,$$

yielding:

$$\partial^2 h(0) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \partial^2[\Delta h](0) \propto \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

References

[1] H. Lavenant and F. Santambrogio, *The flow map of the Fokker–Planck equation does not provide optimal transport*, SIAM Journal on Mathematical Analysis, 2020.