

A Dynamic View on Distribution Morphing

Lecture 3 — release notes

Romain Lopez

Monday 28 September 2026

These notes record what was derived in the third lecture, together with the background it assumed. §1 fixes the vocabulary for motion — velocity fields, flows, probability paths — and proves that a flow solves the continuity equation. §2 is flow matching: the target we cannot evaluate, the conditional construction that replaces it, the marginalisation trick, and the objective that makes the whole thing a least-squares regression. §3 returns to optimal transport for the quadratic cost, states Benamou–Brenier with its optimiser, and asks what flow matching gave up by declining to solve it.

Contents

1	The language of dynamics	2
1.1	Velocity fields, flows, and probability paths	2
1.2	The continuity equation	3
2	Flow matching	4
2.1	Flow models and the training problem	5
2.2	Conditional probability paths	5
2.3	Marginal paths and the marginal vector field	7
2.4	The marginalisation trick	8
2.5	The conditional flow matching objective	9
2.6	The training algorithm	11
3	Dynamic optimal transport	11
3.1	The static problem, recalled	12
3.2	Benamou–Brenier	12
3.3	What flow matching gave up	13

1 The language of dynamics

Everything in the course so far has been *static*. A coupling, or a transport map, says where mass ends up; neither says a word about the route it took. This section introduces the objects needed to describe the route, and proves the one result that ties them together: a family of distributions transported by a flow satisfies a conservation law, the continuity equation.

1.1 Velocity fields, flows, and probability paths

Definition 1.1 (Velocity field). A *velocity field* is a map

$$u: \begin{cases} [0, 1] \times \mathbb{R}^d & \longrightarrow \mathbb{R}^d \\ (t, x) & \longmapsto u_t(x), \end{cases}$$

assigning to each time and each location the velocity of whatever sits there.

Definition 1.2 (Flow). The *flow* of a velocity field u_t is the family of maps $\psi_t: \mathbb{R}^d \rightarrow \mathbb{R}^d$ obtained by following that velocity from a given starting point $x \in \mathbb{R}^d$:

$$\begin{cases} \frac{d}{dt} \psi_t(x) &= u_t(\psi_t(x)), & t \in [0, 1], \\ \psi_0(x) &= x. \end{cases} \quad (1)$$

Thus $\psi_t(x)$ is the position at time t of the particle that started at x at time 0 and has been carried along by u ever since.

Definition 1.2 writes down an ordinary differential equation and calls its solution “the” flow, which presumes there is exactly one. That is what the next theorem delivers: if u is smooth enough, the flow exists, is unique, and is defined for the whole of $[0, 1]$.

Theorem 1.3 (Picard–Lindelöf). *Let u be a velocity field that is continuous in t and Lipschitz in x , uniformly in t :*

$$\exists L < \infty : \quad \|u_t(x) - u_t(y)\| \leq L \|x - y\| \quad \forall t \in [0, 1], \quad \forall x, y \in \mathbb{R}^d.$$

Then

$$\forall x \in \mathbb{R}^d : \quad \exists! \text{ solution } t \mapsto \psi_t(x) \text{ of (1), defined on all of } [0, 1].$$

Not proved here; it is the standard existence-and-uniqueness theorem for ordinary differential equations, and any text on the subject has it. The Lipschitz constant being uniform in t is what rules out escape to infinity in finite time, so the solution survives the whole interval rather than only a neighbourhood of $t = 0$.

Remark 1.4 (Why the hypothesis is not decorative). Without a Lipschitz bound the ODE can branch, and “the” flow is not a thing. The standard example is $\dot{x} = 2\sqrt{|x|}$ started at $x(0) = 0$, which is solved both by $x \equiv 0$ and by $x = t^2$: two flows, same field, same starting point.

Definition 1.5 (Probability path). Let u_t be a velocity field with flow ψ_t , and let ρ_0 be a probability measure on $\mathbb{R}^d$. The associated *probability path* is the family of laws

$$\rho_t := (\psi_t)_\# \rho_0, \quad t \in [0, 1]; \quad (2)$$

that is, ρ_t is the law of $\psi_t(X_0)$ for $X_0 \sim \rho_0$.

Recall the *pushforward* in its integrated form, which is the only property of it used below: for T measurable and g bounded measurable,

$$\int g \, d(T_\# \mu) = \int g(T(x)) \, d\mu(x). \quad (3)$$

1.2 The continuity equation

Definition 1.6 (Continuity equation). Let $(\rho_t)_{t \in [0,1]}$ be probability measures on $\mathbb{R}^d$ with densities, written ρ_t as well, and let u_t be a velocity field. The pair satisfies the *continuity equation* if

$$\partial_t \rho_t + \operatorname{div}(\rho_t u_t) = 0. \quad (4)$$

This is conservation of mass. The density at a point changes only because u_t carries mass in or out; nothing is created, nothing destroyed.

Proposition 1.7 (Flows solve the continuity equation). *Let u_t be a velocity field with flow ψ_t , let ρ_0 be a probability density on $\mathbb{R}^d$, and let $\rho_t := (\psi_t)_\# \rho_0$ be the associated probability path. Then (ρ_t, u_t) solves the continuity equation.*

Proof. Fix a test function $\varphi \in C_c^\infty(\mathbb{R}^d)$, that is, a map $\varphi: \mathbb{R}^d \rightarrow \mathbb{R}$ that is C^∞ with compact support. Since ρ_t has a density, $\int f \, d\rho_t = \int_{\mathbb{R}^d} f(x) \rho_t(x) \, dx$, and we start from the left-hand side of (4), integrated against φ .

(i) **Pull the time derivative out.** The integrand $\varphi \rho_t$ vanishes outside the compact set $\operatorname{supp} \varphi$, so the derivative passes through the integral and

$$\int_{\mathbb{R}^d} \varphi \partial_t \rho_t \, dx = \frac{d}{dt} \int_{\mathbb{R}^d} \varphi \rho_t \, dx = \frac{d}{dt} \int \varphi \, d\rho_t. \quad (5)$$

(ii) **Compute that derivative by following particles.** By (3),

$$\int \varphi \, d\rho_t = \int \varphi(\psi_t(x)) \, d\rho_0(x).$$

For each fixed x the integrand is differentiable in t , and its derivative is bounded uniformly in x : $\nabla \varphi$ is bounded, and u is bounded on the compact set swept out by the flow over $[0, 1]$. So we may differentiate under the integral sign, and the chain rule together with the ODE (1) gives

$$\frac{d}{dt} \int \varphi \, d\rho_t = \int \left\langle \nabla \varphi(\psi_t(x)), \frac{d}{dt} \psi_t(x) \right\rangle d\rho_0(x) = \int \left\langle \nabla \varphi(\psi_t(x)), u_t(\psi_t(x)) \right\rangle d\rho_0(x).$$

The integrand is now one fixed function of position, namely $y \mapsto \langle \nabla \varphi(y), u_t(y) \rangle$, evaluated at $\psi_t(x)$. Applying (3) a second time — reading it from right to left this time — converts the integral against ρ_0 into one against ρ_t :

$$\frac{d}{dt} \int \varphi d\rho_t = \int \langle \nabla \varphi, u_t \rangle d\rho_t = \int_{\mathbb{R}^d} \langle \nabla \varphi, u_t \rangle \rho_t dx. \quad (6)$$

(iii) Put the derivative back on $\rho_t u_t$. Since ρ_t is a scalar, $\langle \nabla \varphi, u_t \rangle \rho_t = \langle \nabla \varphi, \rho_t u_t \rangle$, and the product rule, applied one coordinate at a time, gives

$$\operatorname{div}(\varphi \rho_t u_t) = \langle \nabla \varphi, \rho_t u_t \rangle + \varphi \operatorname{div}(\rho_t u_t). \quad (7)$$

Now integrate (7) over $\mathbb{R}^d$. A compactly supported divergence integrates to zero — by the fundamental theorem of calculus in each variable separately, the endpoints being 0 — and $\varphi \rho_t u_t$ vanishes outside $\operatorname{supp} \varphi$, so the left-hand side drops out, leaving

$$\int_{\mathbb{R}^d} \langle \nabla \varphi, u_t \rangle \rho_t dx = \int_{\mathbb{R}^d} \langle \nabla \varphi, \rho_t u_t \rangle dx = - \int_{\mathbb{R}^d} \varphi \operatorname{div}(\rho_t u_t) dx. \quad (8)$$

Chaining (5), (6) and (8), we have shown that for every $\varphi \in C_c^\infty(\mathbb{R}^d)$ and every t ,

$$\int_{\mathbb{R}^d} \varphi \partial_t \rho_t dx = - \int_{\mathbb{R}^d} \varphi \operatorname{div}(\rho_t u_t) dx, \quad \text{equivalently} \quad \int_{\mathbb{R}^d} \varphi \left(\partial_t \rho_t + \operatorname{div}(\rho_t u_t) \right) dx = 0. \quad (9)$$

A pair (ρ_t, u_t) satisfying (9) for every compactly supported test function is called a *weak solution* of the continuity equation. $\square$

Remark 1.8 (Weak solutions). The proof delivered (9), not (4): the continuity equation holds against every test function, rather than pointwise. A weak solution is enough under regularity assumptions: if $\partial_t \rho_t + \operatorname{div}(\rho_t u_t)$ is continuous, then (9) forces it to vanish identically, since a continuous function that integrates to zero against every test function is zero.

The converse of Proposition 1.7 also holds, and we shall need it twice: in §2.4 and in §3.2. It is not proved in this course.

Theorem 1.9 (Uniqueness for the continuity equation). *Let u be a velocity field, Lipschitz in x uniformly in t , and let ρ_0 be a probability measure on $\mathbb{R}^d$. Then (4) has at most one solution with initial datum ρ_0 — necessarily the probability path (2).*

A complete treatment is in [Santambrogio, 2015]; the general theory of continuity equations is developed in [Ambrosio et al., 2008], and the low-regularity case is [DiPerna & Lions, 1989].

2 Flow matching

We now put a neural network where the velocity field goes. The obstacle is immediate: to fit a field by regression we need a target, and every field we know to carry noise to data is defined through objects we cannot evaluate. Flow matching constructs a usable target by conditioning on a single data point, where everything is explicit, and then averaging.

Throughout this section the two distributions have fixed roles: the source is noise, $p_{\text{noise}} = \mathcal{N}(0, I_d)$, and the target is the data, p_{data} . Noise and data are sampled independently of each other, as in the original construction [Lipman et al., 2023]; OT-CFM [Tong et al., 2024] extends the framework to non-independent couplings.

2.1 Flow models and the training problem

Definition 2.1 (Flow model). A *flow model* is a velocity field $u_t^\theta(x)$ parameterised by a neural network with weights θ , together with a source distribution $p_{\text{noise}} = \mathcal{N}(0, I_d)$. It generates a sample by drawing a starting point and following the field:

$$X_0 \sim p_{\text{noise}}, \quad \frac{d}{dt}X_t = u_t^\theta(X_t), \quad \text{output } X_1.$$

The network parameterises the *field*, not the flow. To produce X_1 one must integrate the ODE, in practice with a numerical scheme; Algorithm 1 uses the Euler method, and any better integrator substitutes.

Algorithm 1 Sampling from a flow model (Euler)

Input: trained field u^θ , step count N

- 1: draw $x \sim p_{\text{noise}}$; $h \leftarrow 1/N$; $t \leftarrow 0$
 - 2: **for** $\ell = 1, \dots, N$ **do**
 - 3: $x \leftarrow x + h u_t^\theta(x)$; $t \leftarrow t + h$
 - 4: **end for**
 - 5: **return** x
-

What we require of training is exactly one thing: that $X_1 \sim p_{\text{data}}$. By Proposition 1.7 the flow of *any* regular field carries p_{noise} along *some* path; we need the path that ends at the data.

The natural attack is regression. Suppose we possessed a *target*: a field u_t already known to carry p_{noise} to p_{data} . Then we could fit

$$\min_{\theta} \mathbb{E}_{t, X_t} \|u_t^\theta(X_t) - u_t(X_t)\|^2$$

and inherit its guarantee. The obstacle is that we have no such target we can evaluate: every construction of one runs through either the Brenier map of p_{data} or the density of p_{data} , and we have neither. We have samples. The question of this section is therefore:

how do you regress onto a function you cannot evaluate?

2.2 Conditional probability paths

Carrying p_{noise} to all of p_{data} is hard because p_{data} is only known through samples. Carrying p_{noise} to a *single, known* data point z is not hard at all — we can simply decide how to do it.

Definition 2.2 (Conditional probability path). A *conditional probability path* is a family of probability measures $p_t(\cdot \mid z)$ on $\mathbb{R}^d$, indexed by $t \in [0, 1]$ and by a data point $z \in \mathbb{R}^d$, with

$$p_0(\cdot \mid z) = p_{\text{noise}} \quad \text{and} \quad p_1(\cdot \mid z) = \delta_z \quad \text{for every } z.$$

Conditioned on z , the path interpolates the noise to that one point: it starts at p_{noise} , ignores every other data point, and collapses onto z at time one. Read backwards, it is a data point progressively drowned in noise. One choice dominates practice, and it is the obvious one.

Example 2.3 (The linear path). Take the Gaussian conditional path

$$p_t(\cdot \mid z) := \mathcal{N}(tz, (1-t)^2 I_d), \quad (10)$$

whose endpoints are right by inspection: at $t = 0$ it is $\mathcal{N}(0, I_d) = p_{\text{noise}}$, and at $t = 1$ the variance vanishes, so the Gaussian degenerates to δ_z . It is realised by pushing the noise toward z along a straight line: if

$$X_t := tz + (1-t)X_0, \quad X_0 \sim p_{\text{noise}} = \mathcal{N}(0, I_d), \quad (11)$$

then $X_t \sim p_t(\cdot \mid z)$.

Proof. X_t is an affine image of the Gaussian vector X_0 , hence Gaussian, so it is determined by its first two moments. Its mean is $\mathbb{E}[X_t] = tz + (1-t)\mathbb{E}[X_0] = tz$, and since tz is deterministic its covariance is

$$\text{Cov}(X_t) = \text{Cov}((1-t)X_0) = (1-t)^2 \text{Cov}(X_0) = (1-t)^2 I_d,$$

which is (10). □

Remark 2.4 (Noise schedulers). The literature writes the conditional path as $\mathcal{N}(\alpha_t z, \beta_t^2 I_d)$ for *noise schedulers* (α_t, β_t) with $\alpha_0 = \beta_1 = 0$ and $\alpha_1 = \beta_0 = 1$; (10) is the linear choice $\alpha_t = t$, $\beta_t = 1-t$, and it is all we need. General schedulers are the point of contact between flow matching and diffusion models, for which see [Holderrieth & Erives, 2026]. Nothing below depends on the Gaussian as such: what is needed is a path whose conditional field is available in closed form.

Having decided the path, we need the field that produces it. Reading (11) as a map of the starting point,

$$\psi_t(x_0 \mid z) := tz + (1-t)x_0, \quad (12)$$

gives a flow whose probability path is the one we want. All that is left is to read off the field it follows.

Proposition 2.5 (The conditional field of the linear path). *For $0 \leq t < 1$, the velocity field*

$$u_t(x \mid z) := \frac{z - x}{1-t} \quad (13)$$

generates the conditional path (10), in the sense that the pair $(p_t(\cdot \mid z), u_t(\cdot \mid z))$ solves the continuity equation (4) for every z .

Proof. The defining equation. A flow and its field are tied by (1), which for $\psi_t(\cdot \mid z)$ reads

$$\frac{d}{dt} \psi_t(x_0 \mid z) = u_t(\psi_t(x_0 \mid z) \mid z) \quad \text{for every } x_0 \in \mathbb{R}^d. \quad (14)$$

Substitute the flow. Putting (12) into both sides of (14), the left-hand side is $\frac{d}{dt}(tz + (1-t)x_0) = z - x_0$, so

$$u_t(tz + (1-t)x_0 \mid z) = z - x_0. \quad (15)$$

Change variables. Equation (15) gives the field at the point a particle started from x_0 has reached; a field must be indexed by position instead. For $t < 1$ the map $x_0 \mapsto tz + (1-t)x_0$ is invertible,

$$x = tz + (1-t)x_0 \iff x_0 = \frac{x - tz}{1-t},$$

and substituting this for x_0 on both sides of (15),

$$u_t(x \mid z) = z - \frac{x - tz}{1-t} = \frac{(1-t)z - x + tz}{1-t} = \frac{z - x}{1-t},$$

which is (13).

Finally, the flow of this field started at p_{noise} is (12), whose pushforward is $p_t(\cdot \mid z)$, so Proposition 1.7 applies and the pair solves the continuity equation. $\square$

Remark 2.6 (What the particles are doing). Along the flow, (13) takes the constant value $z - x_0$: each particle moves in a straight line, at constant speed, from where it started to z . The $1 - t$ in the denominator is not a singularity in the motion, only in the coordinates — as $t \rightarrow 1$ the surviving particles are already at z , and the field has to be large because the remaining distance is being covered in the remaining time. Note also that (13) depends on z , which we know, and on nothing we do not.

2.3 Marginal paths and the marginal vector field

A conditional path regenerates one known data point; it is not a generative model. Averaging over the data gives the object we want.

Definition 2.7 (Marginal probability path). The *marginal probability path* of a conditional path is its mixture over the data,

$$p_t := \int_{\mathbb{R}^d} p_t(\cdot \mid z) dp_{\text{data}}(z), \quad t \in [0, 1]; \quad (16)$$

equivalently, p_t is the law of $X_t \sim p_t(\cdot \mid Z)$ for $Z \sim p_{\text{data}}$.

Proposition 2.8 (The marginal path interpolates noise and data). $p_0 = p_{\text{noise}}$ and $p_1 = p_{\text{data}}$.

Proof. At $t = 0$ the conditional is p_{noise} whatever z is, so the mixture (16) is $p_{\text{noise}} \int dp_{\text{data}}(z) = p_{\text{noise}}$. At $t = 1$ the conditional is δ_z , so for any measurable $A \subseteq \mathbb{R}^d$,

$$p_1(A) = \int_{\mathbb{R}^d} \delta_z(A) dp_{\text{data}}(z) = \int_{\mathbb{R}^d} \mathbf{1}\{z \in A\} dp_{\text{data}}(z) = p_{\text{data}}(A). \quad \square$$

Remark 2.9 (We can sample p_t , and that is all). Drawing $Z \sim p_{\text{data}}$ and then $X_t \sim p_t(\cdot \mid Z)$

produces a sample of p_t ; for the linear path this is just

$$X_t = t Z + (1 - t) X_0, \quad Z \sim p_{\text{data}}, \quad X_0 \sim p_{\text{noise}} \text{ independent.}$$

The *density* p_t , by contrast, is an integral against p_{data} and is out of reach. This is not a temporary embarrassment; it is the permanent situation, and the method is built around it. Every step below must be checked against it: if a quantity needs $p_t(x)$ as a number, it is not implementable.

The field is averaged too — but not in the same way. Velocities do not add like masses; the right average weights each conditional velocity by how much that conditional path contributes to being at x at time t .

Definition 2.10 (Marginal vector field). The *marginal vector field* is

$$u_t(x) := \int_{\mathbb{R}^d} u_t(x | z) \underbrace{\frac{p_t(x | z) p_{\text{data}}(z)}{p_t(x)}}_{\text{posterior over } z \text{ given } x} dz, \quad (17)$$

densities written for measures.

Remark 2.11 (It is a conditional expectation). The weight in (17) is Bayes' rule: it is the posterior density of Z given $X_t = x$, since the joint density of (X_t, Z) is $p_t(x | z)p_{\text{data}}(z)$ and its x -marginal is $p_t(x)$ by (16). So for $Z \sim p_{\text{data}}$ and $X_t \sim p_t(\cdot | Z)$,

$$u_t(x) = \mathbb{E}[u_t(X_t | Z) | X_t = x] : \quad (18)$$

the average velocity demanded at x by the conditional paths passing through x . This reading is what makes §2.5 work.

Note what (17) costs: the quotient contains $p_t(x)$, which Remark 2.9 says we cannot compute. The definition is honest and, for now, useless.

2.4 The marginalisation trick

We now have a candidate target. The claim is that it is the right one.

Theorem 2.12 (The marginalisation trick). *Suppose that for every z the conditional field $u_t(\cdot | z)$ generates the conditional path $p_t(\cdot | z)$. Then the marginal field (17) generates the marginal path: the pair (p_t, u_t) is a weak solution of the continuity equation.*

Proof. Fix a test function $\varphi \in C_c^\infty(\mathbb{R}^d)$, and start as in Proposition 1.7, from the continuity

equation for (p_t, u_t) tested against φ :

$$\begin{aligned}
\int_{\mathbb{R}^d} \varphi \partial_t p_t \, dx &\stackrel{(i)}{=} \frac{d}{dt} \int \varphi \, dp_t \\
&\stackrel{(ii)}{=} \frac{d}{dt} \int_{\mathbb{R}^d} \left(\int \varphi \, dp_t(\cdot | z) \right) dp_{\text{data}}(z) \\
&\stackrel{(iii)}{=} \int_{\mathbb{R}^d} \frac{d}{dt} \left(\int \varphi \, dp_t(\cdot | z) \right) dp_{\text{data}}(z) \\
&\stackrel{(iv)}{=} \int_{\mathbb{R}^d} \left(\int \langle \nabla \varphi, u_t(\cdot | z) \rangle \, dp_t(\cdot | z) \right) dp_{\text{data}}(z) \\
&\stackrel{(v)}{=} \int_{\mathbb{R}^d} \left\langle \nabla \varphi(x), \int_{\mathbb{R}^d} u_t(x | z) \frac{p_t(x | z) p_{\text{data}}(z)}{p_t(x)} \, dz \right\rangle p_t(x) \, dx \\
&\stackrel{(vi)}{=} \int_{\mathbb{R}^d} \langle \nabla \varphi, u_t \rangle p_t \, dx \stackrel{(vii)}{=} - \int_{\mathbb{R}^d} \varphi \operatorname{div}(p_t u_t) \, dx.
\end{aligned}$$

Steps (i) and (vii) are steps (i) and (iii) of the proof of Proposition 1.7, which used nothing but the compact support of φ . Between them, (ii) is Definition 2.7, which says exactly that p_t is the mixture of the conditionals; (iii) exchanges the time derivative with the integral over z ; (iv) is the continuity equation for the conditional pair, which holds for every z by hypothesis — this is the only step that uses anything; (v) exchanges the order of integration and multiplies and divides by $p_t(x)$; and (vi) recognises Definition 2.10. The only analytic point is the exchange in (iii), which holds under standard domination hypotheses, satisfied by the Gaussian path.

As φ was arbitrary, the pair (p_t, u_t) satisfies (9). $\square$

Theorem 2.12 is a statement about an equation; what we want is a statement about a sampler, and Theorem 1.9 bridges them.

Corollary 2.13 (The marginal field is a generative model). *Let $X_0 \sim p_{\text{noise}}$ and let X_t solve $\frac{d}{dt} X_t = u_t(X_t)$. Then $X_t \sim p_t$ for every t , and in particular*

$$X_1 \sim p_{\text{data}}.$$

Proof. The flow of u_t started at p_{noise} pushes it forward to a weak solution of the continuity equation, by Proposition 1.7; the marginal path p_t is a weak solution with the same initial datum, by Theorem 2.12 and Proposition 2.8. By Theorem 1.9 they coincide, so $X_t \sim p_t$; at $t = 1$ that law is p_{data} . $\square$

Remark 2.14 (Half an answer). The question of §2.1 asked for a target field known to carry p_{noise} to p_{data} . Corollary 2.13 produces one. But it is defined by (17), an integral against p_{data} divided by $p_t(x)$, so we still cannot evaluate it at a single point. We have a target we cannot see. The next section shows that we can regress onto it anyway.

2.5 The conditional flow matching objective

Write down both regressions: the one we want, and the one we can run.

Definition 2.15 (Flow matching losses). For a flow model u^θ ,

$$L_{\text{FM}}(\theta) := \mathbb{E}_{t \sim \mathcal{U}([0,1]), X_t \sim p_t} \|u_t^\theta(X_t) - u_t(X_t)\|^2, \quad (19)$$

$$L_{\text{CFM}}(\theta) := \mathbb{E}_{t \sim \mathcal{U}([0,1]), Z \sim p_{\text{data}}, X_t \sim p_t(\cdot | Z)} \|u_t^\theta(X_t) - u_t(X_t | Z)\|^2. \quad (20)$$

The first is the regression we want and cannot run: its target (17) is unavailable. The second replaces that target by the conditional field, which Proposition 2.5 hands us in closed form, and every quantity in it is samplable. The two look like different problems — sample by sample their targets do not even agree.

Theorem 2.16 (The two losses share their gradients). *Assume p_{noise} and p_{data} have finite second moments. Then*

$$L_{\text{CFM}}(\theta) = L_{\text{FM}}(\theta) + c \quad \text{for every } \theta,$$

where c does not depend on θ . In particular $\nabla_\theta L_{\text{FM}} = \nabla_\theta L_{\text{CFM}}$, and both are minimised exactly at the marginal field u_t .

Proof. (i) **Expand.** From (19),

$$\begin{aligned} L_{\text{FM}}(\theta) &= \mathbb{E}_{t, X_t \sim p_t} \|u_t^\theta(X_t)\|^2 - 2 \mathbb{E}_{t, X_t \sim p_t} \langle u_t^\theta(X_t), u_t(X_t) \rangle \\ &\quad + \mathbb{E}_{t, X_t \sim p_t} \|u_t(X_t)\|^2, \end{aligned}$$

and the third term carries no θ .

(ii) **The first term costs nothing.** By Definition 2.7, drawing $Z \sim p_{\text{data}}$ and then $X_t \sim p_t(\cdot | Z)$ produces $X_t \sim p_t$, so

$$\mathbb{E}_{t, X_t \sim p_t} \|u_t^\theta(X_t)\|^2 = \mathbb{E}_{t, Z \sim p_{\text{data}}, X_t \sim p_t(\cdot | Z)} \|u_t^\theta(X_t)\|^2.$$

The integrand does not mention the target at all, so this is a change of sampling scheme and nothing more.

(iii) **The cross term.** Here the marginal field is traded for the conditional one:

$$\begin{aligned} \mathbb{E}_{t, X_t \sim p_t} \langle u_t^\theta(X_t), u_t(X_t) \rangle &\stackrel{(a)}{=} \int_0^1 \int \langle u_t^\theta(x), u_t(x) \rangle p_t(x) \, dx \, dt \\ &\stackrel{(b)}{=} \int_0^1 \int \left\langle u_t^\theta(x), \int u_t(x | z) \frac{p_t(x | z) p_{\text{data}}(z)}{p_t(x)} \, dz \right\rangle p_t(x) \, dx \, dt \\ &\stackrel{(c)}{=} \int_0^1 \int \int \langle u_t^\theta(x), u_t(x | z) \rangle p_t(x | z) p_{\text{data}}(z) \, dz \, dx \, dt \\ &\stackrel{(d)}{=} \mathbb{E}_{t, Z \sim p_{\text{data}}, X_t \sim p_t(\cdot | Z)} \langle u_t^\theta(X_t), u_t(X_t | Z) \rangle. \end{aligned}$$

Step (a) writes the expectation as an integral; (b) is Definition 2.10; (c) cancels the $p_t(x)$ in the denominator against the one outside, and uses linearity of the inner product to move $u_t^\theta(x)$ inside; (d) reads the triple integral back as an expectation. The left end mentions the marginal field; the right end does not. That cancellation is the whole theorem.

(iv) **Complete the square.** Substituting (ii) and (iii) into (i), all three surviving expectations under the conditional scheme combine:

$$\begin{aligned} L_{\text{FM}}(\theta) &= \mathbb{E}_{t, Z, X_t} \left[\|u_t^\theta(X_t)\|^2 - 2\langle u_t^\theta(X_t), u_t(X_t | Z) \rangle \right] + \mathbb{E}_{t, X_t \sim p_t} \|u_t(X_t)\|^2 \\ &= \underbrace{\mathbb{E}_{t, Z, X_t} \|u_t^\theta(X_t) - u_t(X_t | Z)\|^2}_{L_{\text{CFM}}(\theta)} - \underbrace{\mathbb{E}_{t, Z, X_t} \|u_t(X_t | Z)\|^2 + \mathbb{E}_{t, X_t \sim p_t} \|u_t(X_t)\|^2}_{:= -c}, \end{aligned}$$

where we added and subtracted $\|u_t(X_t | Z)\|^2$, and c involves no θ . Finally $L_{\text{FM}} \geq 0$ vanishes precisely when $u_t^\theta(X_t) = u_t(X_t)$ almost surely, so both losses are minimised exactly at the marginal field. $\square$

2.6 The training algorithm

One simplification before implementing. Along the conditional path, $x = tz + (1 - t)x_0$, so

$$u_t(x | z) = \frac{z - x}{1 - t} = \frac{z - tz - (1 - t)x_0}{1 - t} = \frac{(1 - t)(z - x_0)}{1 - t} = z - x_0, \quad (21)$$

which is Remark 2.6 again: the regression target is just the displacement from the noise sample to the data sample.

Algorithm 2 Flow matching training

Input: samples from p_{data} , noise $p_{\text{noise}} = \mathcal{N}(0, I_d)$, field u^θ

- 1: **for** each gradient step **do**
 - 2: draw $t \sim \mathcal{U}([0, 1])$, $z \sim p_{\text{data}}$, $x_0 \sim p_{\text{noise}}$
 - 3: $x \leftarrow tz + (1 - t)x_0$
 - 4: gradient step on $\nabla_\theta \|u_t^\theta(x) - (z - x_0)\|^2$
 - 5: **end for**
-

Remark 2.17 (Simulation-free). Two features carry the method. First, training touches only closed-form quantities — no ODE is ever solved during training; the integrator appears only at sampling time, in Algorithm 1. Contrast what training would cost if the loss required integrating the ODE: a full solve inside every gradient step, backpropagated through, which is what earlier continuous normalising-flow work did [Chen et al., 2018]. Second, the objective is a pure least-squares regression, so everything known about training supervised models transfers unchanged.

3 Dynamic optimal transport

One loop left to close. We have been building fields that carry one distribution to another *without ever asking what they cost*. Optimal transport does ask, and the two questions meet in the Benamou–Brenier formula.

Throughout this section μ and ν have finite second moments, μ is absolutely continuous, and every path and field is as smooth as the argument requires.

3.1 The static problem, recalled

A *coupling* of $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ is a probability measure π on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals μ and ν ; the set of them is $\Pi(\mu, \nu)$, and the squared 2-Wasserstein distance is

$$\mathcal{W}_2^2(\mu, \nu) := \inf_{\pi \in \Pi(\mu, \nu)} \int |x - y|^2 d\pi(x, y).$$

A transport map T with $T_\# \mu = \nu$ induces the *deterministic coupling*

$$\pi_T := (\text{id}, T)_\# \mu, \quad (22)$$

carried by the graph of T , whose cost is $\int |x - T(x)|^2 d\mu$ by (3). And by Brenier's theorem [Brenier, 1991], the optimal coupling is unique and is π_T for a single map $T = \nabla \Phi$ with Φ convex — the *Brenier map* T .

3.2 Benamou–Brenier

Theorem 3.1 (Benamou–Brenier).

$$\mathcal{W}_2^2(\mu, \nu) = \inf_{(\rho_t, u_t)} \int_0^1 \int_{\mathbb{R}^d} |u_t(x)|^2 d\rho_t(x) dt, \quad (23)$$

the infimum being over all pairs (ρ_t, u_t) solving the continuity equation (4) with $\rho_0 = \mu$ and $\rho_1 = \nu$. The infimum is attained, at a pair built from the Brenier map T alone: writing

$$T_t := (1 - t) \text{id} + tT \quad (\text{injective for } t < 1), \quad (24)$$

the optimal path and the optimal field are

$$\rho_t^\star = (T_t)_\# \mu, \quad u_t^\star = (T - \text{id}) \circ T_t^{-1}. \quad (25)$$

The formula is due to [Benamou & Brenier, 2000]; see [Santambrogio, 2015] for a complete treatment.

Proof. The two inequalities are proved separately: no path beats $\mathcal{W}_2^2$, and one path attains it.

Part 1: every path costs at least $\mathcal{W}_2^2$. Let (ρ_t, u_t) be a solution and ψ_t the flow of u . By Proposition 1.7 the family $(\psi_t)_\# \rho_0$ solves the same continuity equation with the same initial datum, so by Theorem 1.9 the two coincide: $(\psi_t)_\# \rho_0$ is ρ_t . That step is not a formality; without it we would know only that the flow couples μ with $\psi_1_\# \mu$, and nothing would identify that endpoint with ν . With it, $\psi_1_\# \mu = \rho_1 = \nu$.

So ψ_1 is a transport map, and (22) gives a coupling $\pi := (\text{id}, \psi_1)_\# \mu \in \Pi(\mu, \nu)$. Bounding the infimum by this particular coupling, then writing the displacement as the integral of the velocity along the trajectory,

$$\mathcal{W}_2^2(\mu, \nu) \leq \int |\psi_1(x) - \psi_0(x)|^2 d\mu(x) = \int \left| \int_0^1 u_t(\psi_t(x)) dt \right|^2 d\mu(x) \leq \int_0^1 \int |u_t|^2 d\rho_t dt,$$

the last step by Jensen applied to the time integral, and the final change of variables by (3). Taking the infimum over paths gives “ $\leq$ ”.

Part 2: one path attains it. Take T_t, ρ_t^*, u_t^* as in (24)–(25). Three checks.

(i) *The field is well defined.* We want u_t^* with

$$u_t^*(T_t(x)) = T(x) - x, \quad (26)$$

which assigns a velocity to a *location*, so it is legitimate only if T_t is injective — otherwise two particles meet at one point demanding different velocities. Put $\Phi_t(x) := (1-t)\frac{1}{2}|x|^2 + t\Phi(x)$, so that $T_t = \nabla\Phi_t$. For $t < 1$, Φ_t is *strictly* convex, being the sum of the strictly convex $(1-t)|x|^2/2$ and the convex $t\Phi$; the gradient of a strictly convex function is injective, so T_t is, and T_t^{-1} in (25) makes sense. At the single time $t = 1$ injectivity may fail, and one time contributes nothing to the time integral.

(ii) *The pair solves the continuity equation.* Differentiating T_t gives $\frac{d}{dt}T_t(x) = T(x) - x$, which by (26) is $u_t^*(T_t(x))$; and $T_0 = \text{id}$. So $t \mapsto T_t(x)$ is the flow of u^* in the sense of (1), and $\rho_t^* = (T_t)_\# \mu$ is its pushforward, so Proposition 1.7 applies. The endpoints are right too: $\rho_0^* = \mu$ and $\rho_1^* = T_\# \mu = \nu$.

(iii) *Its action is $\mathcal{W}_2^2$.* By (26) the integrand $|u_t^*|^2$ pulls back along T_t to $|T(x) - x|^2$, which does not depend on t , so

$$\int_0^1 \int |u_t^*|^2 d\rho_t^* dt = \int_0^1 \int |T(x) - x|^2 d\mu(x) dt = \int |T(x) - x|^2 d\mu(x) = \mathcal{W}_2^2(\mu, \nu),$$

the last equality because T is the optimal map. So the infimum is attained. $\square$

Remark 3.2 (One object, two descriptions). Static optimal transport answers *where does the mass go*; dynamic optimal transport answers *along which path*; and for the quadratic cost the second answer is determined by the first — every symbol in (25) is built from T and nothing else. Since $T_t(x) = x + t(T(x) - x)$, the path ρ_t^* is McCann’s *displacement interpolation* [McCann, 1997]: every particle walks in a straight line, at constant speed, from x to its Brenier image $T(x)$, and u_t^* is just that constant velocity $T(x) - x$, read off at the particle’s current position rather than its starting one.

Remark 3.3 (The same T_t is a Wasserstein barycenter). For the quadratic cost, ρ_t^* is also the *Wasserstein barycenter* of μ and ν with weights $(1-t, t)$:

$$\rho_t^* = \arg \min_{\rho \in \mathcal{P}(\mathbb{R}^d)} (1-t) \mathcal{W}_2^2(\mu, \rho) + t \mathcal{W}_2^2(\rho, \nu). \quad (27)$$

Not proved here; see [Agueh & Carlier, 2011].

Two measures is a special case, and a misleadingly easy one. For three or more the barycenter of [Agueh & Carlier, 2011] is a genuinely different object: it has no closed form in terms of pairwise maps, and computing it is a problem in its own right.

3.3 What flow matching gave up

Remark 3.4 (The pairing is what was discarded). Compare (25) with Proposition 2.5. Both describe particles moving in straight lines at constant speed. The difference is the *pairing*: displacement interpolation sends x to $T(x)$, the Brenier image; flow matching sends x_0 to z , where $x_0 \sim p_{\text{noise}}$ and $z \sim p_{\text{data}}$ were drawn *independently*. Conditionally on the pairing,

both are optimal; the pairing itself is what flow matching discards. There are two reasons to discard it.

(i) *Scale*. Computing the optimal pairing means solving optimal transport, and at the size of a generative model — millions of images — even an $\mathcal{O}(nm)$ -per-iteration Sinkhorn solver is out of reach.

(ii) *Modelling*. Optimality is always relative to a cost, and we do not know the right one: what does the quadratic cost mean between two images, compared pixel by pixel?

Flow matching asks only that the marginals be right, which is why Theorem 2.12 was a statement about endpoints and never about cost.

This is a live question rather than a closed one. OT-CFM [Tong et al., 2024] puts an optimal coupling back inside the minibatch, buying straighter paths and cheaper sampling at the price of a small transport solve per batch.

References

- [Agueh & Carlier, 2011] M. Agueh and G. Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.
- [Ambrosio et al., 2008] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics ETH Zürich, Birkhäuser, 2nd edition, 2008.
- [Benamou & Brenier, 2000] J.-D. Benamou and Y. Brenier. A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.
- [Brenier, 1991] Y. Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4):375–417, 1991.
- [Chen et al., 2018] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, 2018. <https://arxiv.org/abs/1806.07366>.
- [DiPerna & Lions, 1989] R. J. DiPerna and P.-L. Lions. Ordinary differential equations, transport theory and Sobolev spaces. *Inventiones Mathematicae*, 98(3):511–547, 1989.
- [Holderrieth & Erives, 2026] P. Holderrieth and E. Erives. *An Introduction to Flow Matching and Diffusion Models*. MIT 6.S184 lecture notes, 2026. <https://diffusion.csail.mit.edu/>.
- [Lipman et al., 2023] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. <https://arxiv.org/abs/2210.02747>.

- [McCann, 1997] R. J. McCann. A convexity principle for interacting gases. *Advances in Mathematics*, 128(1):153–179, 1997.
- [Peyré & Cuturi, 2019] G. Peyré and M. Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019. <https://arxiv.org/abs/1803.00567>.
- [Santambrogio, 2015] F. Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Progress in Nonlinear Differential Equations and Their Applications 87, Birkhäuser, 2015.
- [Tong et al., 2024] A. Tong, K. Fatras, N. Malkin, G. Hugué, Y. Zhang, J. Rector-Brooks, G. Wolf, and Y. Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. <https://arxiv.org/abs/2302.00482>.
- [Villani, 2009] C. Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften 338, Springer, 2009.