A Gentle Introduction to

Flow Matching

August 21, 2025

1 Prerequisites

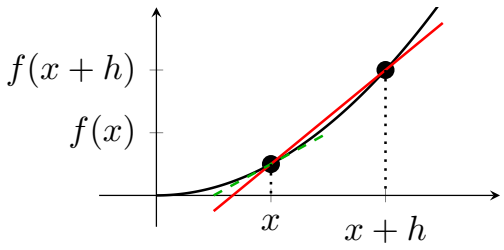
1.1 Vector Calculus

Explaining vector calculus lies beyond the scope of this document, but we will nevertheless cover the fundamental concepts in a single page. We won't provide deeper explanations and intuitions here, but it's crucial to understand these concepts well as they form the foundation not only for this work but for much of mathematics and nature.

1.1.1 Differential Calculus

For $f(x)$ a function of one variable, $\frac{df}{dx}$ tells us how rapidly $f(x)$ varies. This is the **ordinary derivative**:

$$\frac{df}{dx} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} \tag{1}$$



1.1.2 Gradient

For a function with multiple variables, we can define the gradient as the collection of partial derivatives with respect to each variable. Consider a function $f(x, y, z)$:

$$\frac{\partial f}{\partial x}, \quad \frac{\partial f}{\partial y}, \quad \frac{\partial f}{\partial z} \tag{2}$$

We introduce the **del operator** ∇ (nabla):

$$\nabla = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right) \tag{3}$$

We can now express the gradient with the proper symbol:

$$\nabla f = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z} \right) \tag{4}$$

1.1.3 Vector Fields

A vector field assigns a vector to each point in space. We can write a vector field $\mathbf{F}$ as:

$$\mathbf{F}(x, y, z) = (F_x(x, y, z), F_y(x, y, z), F_z(x, y, z)) \tag{5}$$

where each component F_x, F_y, and F_z is a scalar function of position.

1.1.4 Divergence

The divergence measures how much a vector field spreads out from a point. For a vector field $\mathbf{F} = (F_x, F_y, F_z)$:

$$\nabla \cdot \mathbf{F} = \frac{\partial F_x}{\partial x} + \frac{\partial F_y}{\partial y} + \frac{\partial F_z}{\partial z} \tag{6}$$

1.1.5 Curl

The curl measures how much a vector field rotates around a point. For a vector field $\mathbf{F} = (F_x, F_y, F_z)$:

$$\nabla \times \mathbf{F} = \left(\frac{\partial F_z}{\partial y} - \frac{\partial F_y}{\partial z}, \frac{\partial F_x}{\partial z} - \frac{\partial F_z}{\partial x}, \frac{\partial F_y}{\partial x} - \frac{\partial F_x}{\partial y} \right) \tag{7}$$

Insight

> For a deeper understanding of divergence and curl, see `https://math.libretexts.org/Bookshelves/Calculus/Calculus_(OpenStax)/16:_Vector_Calculus/16.05:_Divergence_and_Curl`.

1.2 Probability Basics

1.2.1 Marginalization

Let's start with a concrete example using real U.S. population data. We'll use age and height distributions from the CDC National Health and Nutrition Examination Survey

(NHANES) 2015-2018[1]:

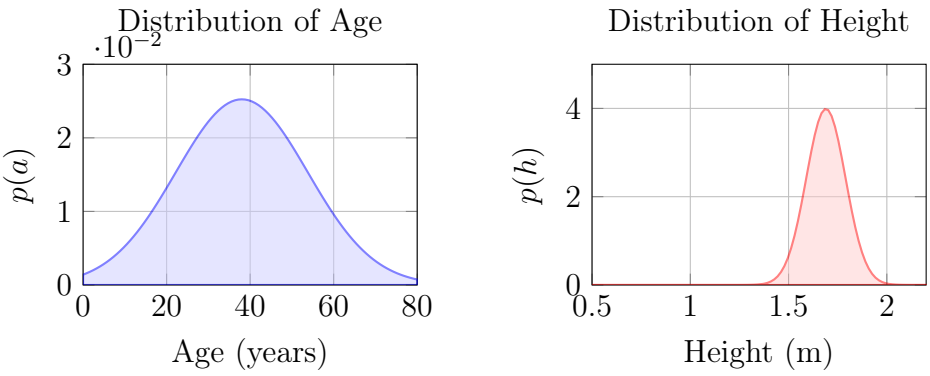


Figure 1: Marginal distributions from U.S. population data. Left: Age distribution with median around 38 years (U.S. Census 2020). Right: Height distribution centered around 1.69 meters (average of male 1.76m and female 1.62m from NHANES 2015-2018).

Now, if we tried to sample from these distributions *independently*, we could get nonsensical results - like a 2-year-old who is 1.8 meters tall (when CDC growth charts show 2-year-olds average 0.87m), or a 30-year-old who is 0.8 meters tall (when adults average 1.69m). This tells us something important: these variables are **related**!

When we sample a specific age, say age $= 2$ (shown as a point on the age distribution below), the height distribution must change accordingly:

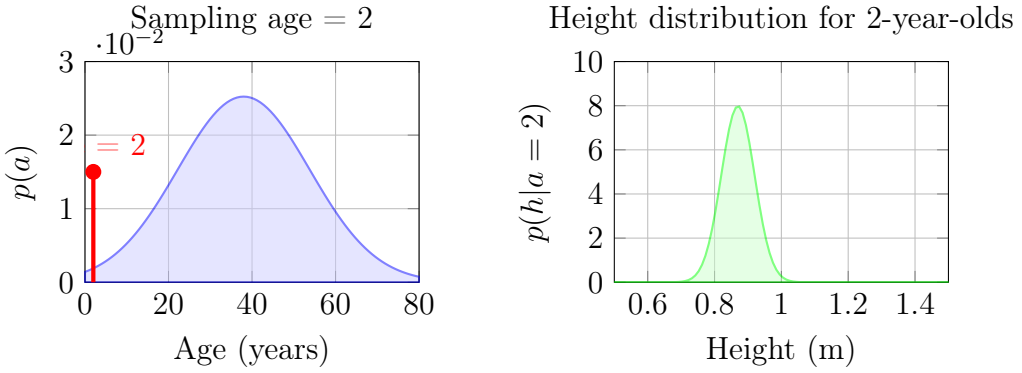


Figure 2: When we condition on age $= 2$ (left), the height distribution changes dramatically (right), now centered around 0.87m with much smaller variance. This matches CDC growth charts which show the 50th percentile height for 2-year-olds is 87cm.

This relationship is captured by the joint distribution $p(a, h)$, which encodes how age and height vary together:

The mathematical relationship between these distributions is given by marginalization:

$$p(h) = \int p(h|a)p(a)\, da \tag{8}$$

[1]Fryar CD, et al. Anthropometric reference data for children and adults: United States, 2015–2018. National Center for Health Statistics. Vital Health Stat 3(46). 2021.

This equation tells us that to get the overall height distribution, we integrate the conditional height distribution for each age, weighted by how common that age is in the population.

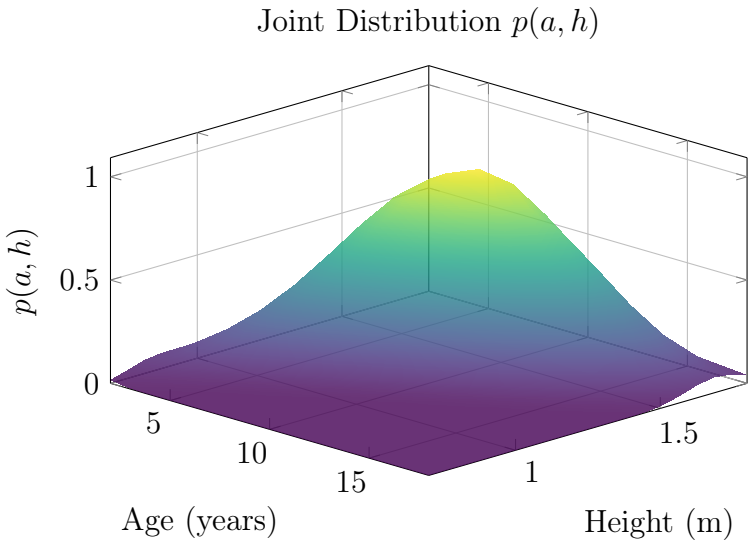


Figure 3: The joint distribution $p(a, h)$ captures the relationship between age and height. Based on CDC data, the peak shifts from 0.75m at age 1 to 0.87m at age 2, reaching 1.69m for adults. Notice how height increases with age following the CDC growth curves.

This principle of marginalization - integrating conditional distributions weighted by a prior distribution - is exactly what we use in flow matching to construct marginal probability paths from conditional ones.

1.3 Continuity Equation

The continuity equation is one of the fundamental equations in physics:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{u}) = \sigma \tag{9}$$

where ρ is density, $\mathbf{u}$ is velocity field, and σ is a source term.

We'll pick this equation apart piece by piece until it becomes intuitive. For our purposes, let's work with a more concrete example, the flow of water, where:

- p = water density (scalar field)
- $\mathbf{v}$ = water velocity (vector field)
- S = source/sink rate (constant)

Our continuity equation becomes:

$$\frac{\partial p}{\partial t} + \nabla \cdot (p\mathbf{v}) = S \tag{10}$$

Let's visualize an example of these fields:

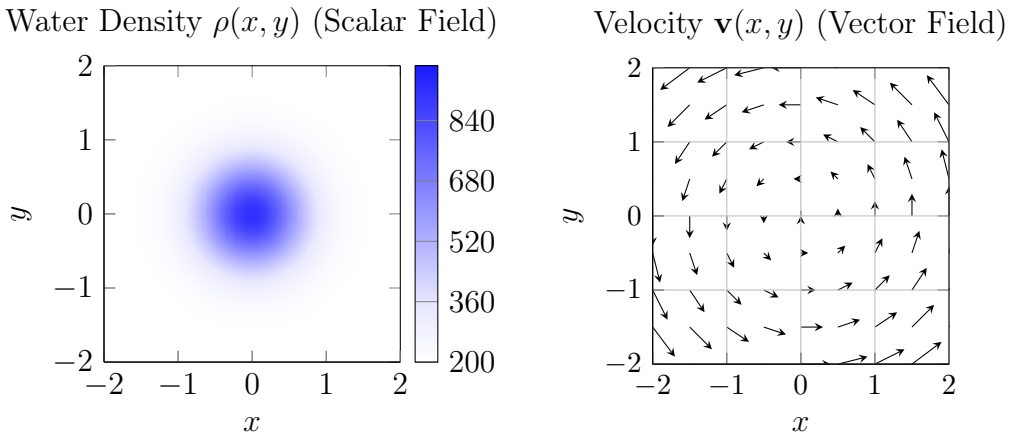


Figure 4: Left: Density is a scalar field - each point in space has a single value. Right: Velocity is a vector field - each point has both magnitude and direction.

Now, let's rearrange our equation to see how density changes over time:

$$\frac{\partial p}{\partial t} = S - \nabla \cdot (p\mathbf{v}) \tag{11}$$

First, consider the term $p\mathbf{v}$. Since p is a scalar field and $\mathbf{v}$ is a vector field, their product $p\mathbf{v}$ creates a new vector field where each vector from $\mathbf{v}$ is scaled by the local density p. This gives us a "density-weighted velocity field": the flow is stronger where there's more density.

Now let's analyze $\nabla \cdot (p\mathbf{v})$, the divergence of this density-weighted flow. Divergence can be thought of as a "generalized derivative": it's the sum of partial derivatives:

$$\nabla \cdot (p\mathbf{v}) = \frac{\partial (pv_x)}{\partial x} + \frac{\partial (pv_y)}{\partial y} \tag{12}$$

But intuitively, divergence measures the concentration or spread of mass in our field. It tells us whether the flow lines are converging (negative divergence) or diverging (positive divergence) at each point. Think of it as measuring "how much is flowing in versus flowing out" at each location.

So we can conclude that $\nabla \cdot (p\mathbf{v})$ tells us the net flow in versus out of our density-weighted velocity field. Essentially, how much water is flowing away from or toward each point.

Now consider the left side of our equation: $\frac{\partial p}{\partial t}$. This is a partial derivative asking "how does density p change with a small change in time?" It measures the rate of change of density at each point:

$$\frac{\partial p}{\partial t} = \lim_{\Delta t \to 0} \frac{p(x, y, t + \Delta t) - p(x, y, t)}{\Delta t} \tag{13}$$

Putting it all together, our equation:

$$\frac{\partial p}{\partial t} = S - \nabla \cdot (p\mathbf{v}) \tag{14}$$

is saying: the change in density over time equals a constant source/sink rate minus the concentration/spread of flow. If more is flowing out than in at a point, the density there decreases. If more is flowing in than out, the density increases. The source term S adds or removes density uniformly.

Let's analyze S more intuitively. Assume for a moment that $\nabla \cdot (p\mathbf{v}) = 0$. This means our field has no net flow in or out: all arrows flowing in equal those flowing out. There are no concentrations or spreads in the flow. This leaves us with:

$$\frac{\partial p}{\partial t} = S \tag{15}$$

Now we can see what S really does:

- If $S = 0$: The change in density over time is zero. Nothing changes.
- If $S > 0$: We have a true source! Even when divergence is zero (no flow concentration), the density still increases over time. Particles are entering the system from "nowhere". This is why it's called a source term. In our water example, imagine a pipe injecting water uniformly throughout the space.
- If $S < 0$: We have a true sink. The density decreases over time even though the flow field itself has no sinks. Particles are leaving the system. Think of water evaporating uniformly throughout the space.

The key idea: S represents changes in density that aren't explained by the flow itself. When density changes but the flow is perfectly balanced (divergence $= 0$), we know there must be an external source or sink. This is why S is called the source term: it accounts for particles entering or leaving the system independently of the flow.

Now, what if we set $S = 0$? This means we have no external sources or sinks. All density changes come purely from the flow:

$$\frac{\partial p}{\partial t} = -\nabla \cdot (p\mathbf{v}) \tag{16}$$

Notice the negative sign! This is crucial:

- If $\nabla \cdot (p\mathbf{v}) > 0$ (positive divergence, flow spreading out), then $\frac{\partial p}{\partial t} < 0$: density decreases
- If $\nabla \cdot (p\mathbf{v}) < 0$ (negative divergence, flow converging), then $\frac{\partial p}{\partial t} > 0$: density increases

Insight

> The negative sign in front of divergence often confuses people, but it makes perfect physical sense.
>
> When flow diverges from a point (positive divergence), particles are leaving that region, so density decreases (negative change). When flow converges to a point

(negative divergence), particles are accumulating there, so density increases (positive change).

In our water example: if water flows out of a region, the water density there drops. If water flows into a region, the density there rises. The negative sign ensures this relationship is correct.

This relationship between divergence and density change is fundamental to conservation laws in physics: mass cannot be created or destroyed, only moved around by the flow.

Exercise

Consider the continuity equation with a positive sign instead of negative:

$$\frac{\partial \rho}{\partial t} = +\nabla \cdot (\rho \mathbf{u}) \tag{17}$$

What physical phenomena could this equation be describing? What would it mean for density to increase when there's positive divergence (flow spreading out)?

This thought exercise is valuable to understand that equations are not godly commands but descriptions of nature! We're modeling the world around us, and we shouldn't take equations as absolute truths. We should question and play with them to understand what other phenomena they could represent.

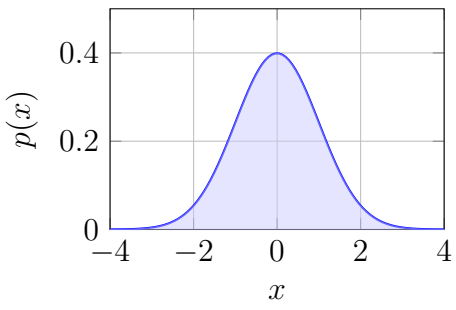
2 Introduction

Let us consider two distributions in $\mathbb{R}$:

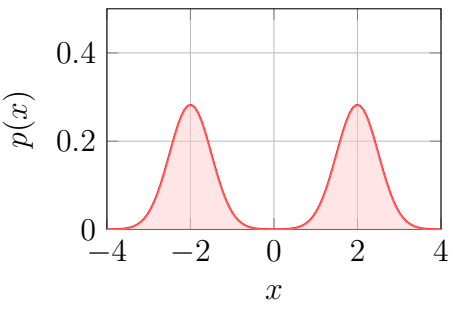
- q: A centered Gaussian distribution at the origin
- p: A mixture of Gaussians separated by distance a from the origin

These will serve as our toy distributions to illustrate the flow matching framework.

Distribution q: Centered Gaussian



Distribution p: Mixture of Gaussians



We define π_t to be a time-dependent probability distribution that interpolates between q and p, with boundary conditions:

$$\pi_0 = q \quad (\text{at } t = 0) \tag{18}$$

$$\pi_1 = p \quad (\text{at } t = 1) \tag{19}$$

Flow matching is fundamentally about learning this time-dependent distribution π_t that smoothly transforms from the source distribution q to the target distribution p over the time interval $t \in [0, 1]$.

To find this family of distributions, we need to define two key concepts:

- **Conditional Probability Path**: A path from the source distribution to a *single point* in the target distribution. This describes how probability mass flows from the entire source distribution to concentrate at a specific target location.
- **Marginal Probability Path**: A path from the source distribution to the target distribution as a whole. This is what we ultimately want - the evolution of the entire distribution over time.

Let's visualize the difference between these two concepts. Consider 10 sample points drawn from the source distribution q:

(Visualization: We'll show the difference between conditional and marginal probability paths here - 10 points flowing from source to either a single target point or distributed across the target)

Figure 5: Comparison of probability paths. Left: Conditional path where all points from q converge to a single target point. Right: Marginal path where points from q distribute across the target distribution p.

The key insight is that the marginal probability path can be constructed by integrating over all possible conditional probability paths, weighted by the target distribution p.

2.1 Conditional Probability Path

Let's define the mathematics of the conditional probability path. For this path, we have:

$$\pi_0 = q \quad (\text{we start at the source distribution}) \tag{20}$$

$$\pi_1 = \delta(x - z) \quad (\text{Dirac delta at point } z) \tag{21}$$

For our specific example where q is a Gaussian, we can define the conditional probability path as:

$$\pi_t(x|z) = \mathcal{N}(\alpha_t z, \beta_t I) \tag{22}$$

where:

$$\alpha_t = t \tag{23}$$

$$\beta_t = 1 - t \tag{24}$$

and z is the target point we're converging towards.

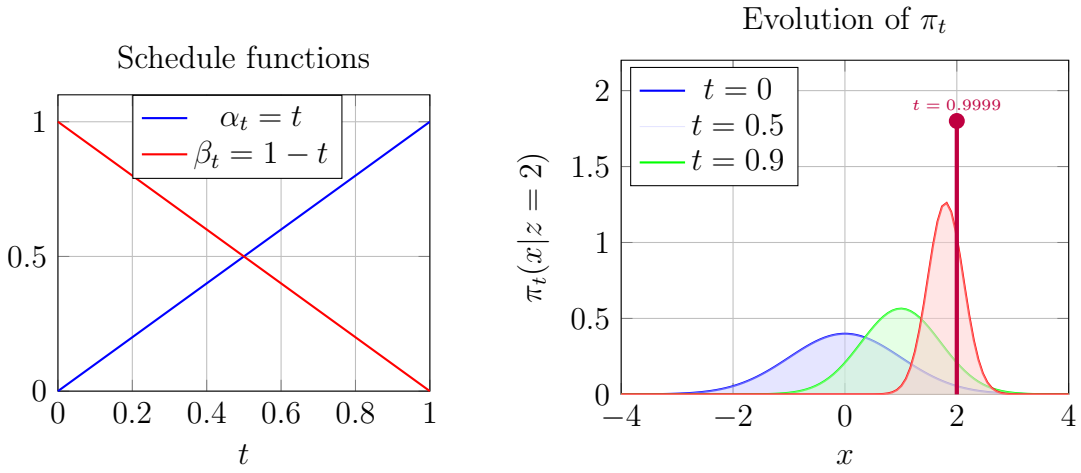


Figure 6: Left: Evolution of α_t and β_t over time. Right: The conditional distribution $\pi_t(x|z = 2)$ at different times, showing convergence from $\mathcal{N}(0, 1)$ at $t = 0$ to a point at $z = 2$ as $t \to 1$.

Notice that:

- At $t = 0$: $\pi_0(x|z) = \mathcal{N}(0, I)$ - the original Gaussian distribution
- At $t = 1$: $\pi_1(x|z) = \mathcal{N}(z, 0)$ - a point mass at z (Dirac delta)

This is our conditional probability path! It describes how probability mass flows from the entire source distribution to concentrate at a single target point z.

2.2 Marginal Probability Path

Now, instead of flowing to a single point, we want to define the path to the entire target distribution. The key insight is that we can construct this by considering conditional probability paths to *all* possible target points z, weighted by their probability in the target distribution!

For the marginal probability path, we have:

$$\pi_0 = q \quad \text{(we start at the source distribution)} \tag{25}$$
$$\pi_1 = p \quad \text{(we end at the target distribution)} \tag{26}$$

The marginal distribution at time t is obtained by integrating over all possible conditional paths:

$$p_t(x) = \int \pi_t(x|z)p(z)\,dz \tag{27}$$

Insight

> Remember the age-height example from the prerequisites? There, we had a joint distribution and used marginalization to extract individual distributions.
>
> Here, we're doing the reverse: we have many conditional paths (each going to a specific point z), and we combine them through marginalization to build the overall flow.
>
> Each particle knows only its destination. But when we consider all particles together, each weighted by how likely its destination is in $p(z)$, the collective motion reconstructs the entire target distribution p.
>
> It's the same $\int \cdot p(z)\, dz$ operation, just applied to build rather than decompose.

This elegant formulation allows us to interpolate between the source distribution q and the target distribution p!

2.3 The Velocity Field Approach

Now here's the crucial insight: we don't actually want to compute these distributions π_t directly. This would be impossible!

Imagine the source distribution q as simple noise and the target distribution p as the distribution of all natural images. The intermediate distributions π_t would be extraordinarily complex and intractable. We don't even know what the distribution of natural images looks like explicitly! Therefore, any π_t would be as impossible to compute as p_{images} itself.

But here's the key realization: we don't actually care about the shape of the distributions themselves. What we really care about is how to move individual particles from one distribution to the other. We want to know: given a particle at position x at time t, which direction should it move and how fast?

So here's our recipe: we want a velocity field $\mathbf{u}_t$ such that:

$$X_0 \sim q \quad \text{(sample from initial distribution)} \tag{28}$$

$$\frac{dX_t}{dt} = \mathbf{u}_t(X_t) \quad \text{(follow the velocity field)} \tag{29}$$

$$X_1 \sim p \quad \text{(end up at target distribution)} \tag{30}$$

Insight

> This is why flow matching takes a different approach: instead of computing the probability distributions, we focus on the velocity field that transports particles from source to target.
>
> Think of it this way: if you're directing a crowd from a stadium to various exits, you don't need to know the exact crowd density at every point at every moment. You just need to tell each person which way to walk when they reach a certain

> location. The velocity field provides these local instructions.
>
> It's all about the journey. We don't stop to smell the flowers.

Just as we had conditional and marginal probability paths, we can define conditional and marginal velocity fields.

2.3.1 Conditional Velocity Field

Let's start with the conditional velocity field. When we know the target point z (where $z \sim p$, sampled from the target distribution), we can define a velocity field $\mathbf{u}_t(x|z)$ for $t \in [0, 1]$, where $x, z \in \mathbb{R}^d$ (in our examples, just $\mathbb{R}$).

This velocity field tells particles how to move from the source to the specific target point z:

$$\frac{d}{dt}X_t = \mathbf{u}_t(X_t|z) \tag{31}$$

Insight

> We're defining the law of motion for each particle explicitly. We literally specify how a particle moves through space over time, defining its position at every moment.
>
> For example, with a linear trajectory: $X_t = (1-t)X_0 + tz$. The velocity field is just the derivative of this position: $\mathbf{u}_t(X|z) = z - X_0$. Nothing global is being solved here. It's just local, particle-by-particle dynamics.

This is an ordinary differential equation (ODE) that describes the trajectory of a particle starting from the source distribution and ending at point z.

For our Gaussian example, we can derive the exact conditional velocity field:

$$\mathbf{u}_t(x|z) = \left(\dot{\alpha}_t - \frac{\dot{\beta}_t}{\beta_t}\alpha_t\right) z + \frac{\dot{\beta}_t}{\beta_t}x \tag{32}$$

where $\dot{\alpha}_t = \frac{d\alpha_t}{dt}$ and $\dot{\beta}_t = \frac{d\beta_t}{dt}$. The derivation is quite straightforward and can be found in Annex A.

In our example, we defined $\alpha_t = t$ and $\beta_t = 1 - t$. Therefore $\dot{\alpha}_t = 1$ and $\dot{\beta}_t = -1$, giving us:

$$\mathbf{u}_t(x|z) = \left(1 - \frac{-1}{1-t}t\right) z + \frac{-1}{1-t}x = \frac{z-x}{1-t} \tag{33}$$

This is the velocity field that transforms a Gaussian into a point! Let's visualize how this conditional velocity field evolves over time:

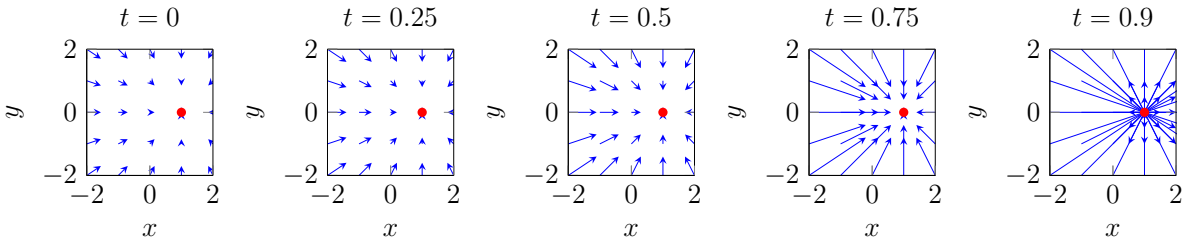


Figure 7: Evolution of the conditional velocity field $\mathbf{u}_t(x|z)$ for $z = (1, 0)$ (red dot). As t approaches 1, the velocity field becomes stronger, pulling all points toward the target.

Insight

Think about what we've just created: a vector field that collapses an entire Gaussian distribution down to a single point. Imagine you're a particle somewhere in that Gaussian cloud. You look around and see arrows telling you which way to go. You follow them, and so does every other particle around you. Everyone takes different paths, but somehow you all end up at exactly the same spot.

Notice how the field gets stronger over time. Early on, particles drift slowly. Near t=1, the velocities become huge because every particle needs to reach z at exactly t=1. The math forces this: as the denominator (1-t) approaches zero, the velocity blows up.

2.3.2 Marginal Velocity Field

Now we want something more ambitious: a velocity field that transforms one distribution into another distribution, not just to a single point. We need a field that works for all possible endpoints simultaneously, creating a distribution of trajectories rather than a single trajectory.

Insight

It's critical we stop here and think for a second.

With the conditional probability path, we had an explicit trajectory for a single particle. We could say exactly where the particle is at any time t: position X_t. To get the velocity, we just take the derivative: $v_t = \frac{dX_t}{dt}$. This is basic physics.

But the marginal path isn't an explicit trajectory. It's an entire evolving distribution $\pi_t(x)$ describing where many particles are. The derivative now exists at the distribution level: $\frac{\partial \pi_t}{\partial t}$. How do we relate this density evolution to a velocity field?

The answer is the continuity equation. The velocity field $\mathbf{u}_t(x)$ describes the aggregate motion of particles passing through point x:

$$\frac{\partial \pi_t}{\partial t} + \nabla \cdot (\pi_t \mathbf{u}_t) = 0 \tag{34}$$

> The conditional shows how a single particle moves explicitly, and we relate position to velocity through the derivative. The marginal shows how a collection of particles evolves, now related to velocity through the continuity equation.
>
> The clear, fundamental difference is that one is taking the derivative of a position function while the other is taking the derivative of a distribution. If this is still unclear, I have provided a detailed definition and explanation in the Annex for what a "derivative" means in terms of distributions.

The marginal velocity field can be derived from the conditional velocity fields through the following equation:

$$\mathbf{u}_t(x) = \int \mathbf{u}_t(x|z) \frac{\pi_t(x|z)p(z)}{\pi_t(x)} \, dz \tag{35}$$

where:

- $\mathbf{u}_t(x|z)$ is the conditional velocity field (which we know)
- $\pi_t(x|z)$ is the conditional probability path
- $p(z)$ is the target distribution
- $\pi_t(x)$ is the marginal distribution at time t

This shows that the marginal velocity field is a weighted average of conditional velocity fields, where the weights are the posterior probabilities of reaching different targets z given the current position x.

Insight

> This might seem like a significant jump in complexity, but two important points:
>
> 1) The fundamental concept is the same: we're deriving the mathematical equation for the velocity field that transports particles. The actual derivation is in the Annex.
>
> 2) Here's a spoiler: we won't actually use this exact equation for flow matching training! It's too expensive to compute in practice. The clever trick is that flow matching sidesteps this entirely and learns the field without ever calculating this integral.

This marginal velocity field satisfies:

$$X_0 \sim q \quad \text{(initial distribution)} \tag{36}$$

$$\frac{dX_t}{dt} = \mathbf{u}_t(X_t) \quad \text{(follow the marginal velocity field)} \tag{37}$$

$$X_t \sim \pi_t \quad \text{(particles follow the evolving distribution)} \tag{38}$$

The derivation of this equation is detailed in Annex A.4.

Insight

> This is significant! We now have a mathematical description of how to transform one distribution into another using a velocity field.
>
> The marginal velocity field $\mathbf{u}_t(x)$ tells us exactly how to move particles from any source distribution q to any target distribution p. Each particle follows the field, and collectively they reshape from one distribution to the other.

3 Training

3.1 Flow Matching Loss

Now we have a target for our neural network: the velocity field $\mathbf{u}_t(x)$. Our goal is to train a neural network $\mathbf{u}_t^\theta(x)$ with parameters θ such that:

$$\mathbf{u}_t^\theta(x) \approx \mathbf{u}_t(x) \tag{39}$$

With this in mind, we can define the mean squared error between the predicted velocity field and the real velocity field. The **Flow Matching loss** is:

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{t,z,x_t}\left[\|\mathbf{u}_t^\theta(x_t) - \mathbf{u}_t(x_t)\|^2\right] \tag{40}$$

where the expectation is taken over time $t \sim \mathcal{U}[0,1]$, data points $z \sim p_{\text{data}}$, and positions x_t sampled from the marginal distribution $\pi_t(x)$.

Insight

> t: Remember we defined our process to go from 0 to 1. We need a specific point in time to compute the velocity field for training.
>
> z: We sample from our data distribution p_{data} - in our example that's the mixture of Gaussians, in practice it's an image from the dataset.
>
> x_t: We sample from the conditional probability path $\pi_t(x|z)$ at time t given the data point z.
>
> Looking at our marginal velocity field $\mathbf{u}_t(x) = \int \mathbf{u}_t(x|z)\frac{\pi_t(x|z)p(z)}{\pi_t(x)}\,dz$, we need all three: time t, data z, and position x_t to compute the conditional velocity field that goes into this equation.

The mean squared error is our minimizer: it's never negative, and equals zero if and only if $\mathbf{u}_t^\theta(x_t) = \mathbf{u}_t(x_t)$ exactly.

Unfortunately, this loss is not tractable! The marginal velocity field $\mathbf{u}_t(x)$ requires computing:

$$\mathbf{u}_t(x) = \int \mathbf{u}_t(x|z)\frac{\pi_t(x|z)p(z)}{\pi_t(x)}\,dz \tag{41}$$

This integral is high-dimensional (when z represents images, it could be millions of dimensions) and therefore computationally intractable. We can't compute $\pi_t(x)$ explicitly, nor can we evaluate this integral over all possible data points z.

But don't despair, the story doesn't end here.

3.2 Conditional Flow Matching Loss

Let's analyze what happens when we try to use the conditional velocity field instead of the marginal. The **Conditional Flow Matching loss** would be:

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t,z,x_t}\left[\|\mathbf{u}_t^\theta(x_t) - \mathbf{u}_t(x_t|z)\|^2\right] \tag{42}$$

Once again, the expectation is taken over time $t \sim \mathcal{U}[0,1]$, data points $z \sim p_{\text{data}}$, and positions x_t.

Insight

> At the risk of being pedantic, let's quickly check what this is.
>
> Remember that our conditional velocity field is the time derivative of a single trajectory function. For each fixed z, we have a deterministic path X_t that flows from the source distribution to the point z, and $\mathbf{u}_t(x|z) = \frac{dX_t}{dt}$ is just its velocity.
>
> So unlike the marginal velocity field (which requires that complex integral), the conditional velocity field is something we can compute directly from our trajectory definition.

3.3 The Equivalence of Losses

We now come across a fascinating result! As it turns out, the marginal and conditional losses are related:

$$\mathcal{L}_{\text{FM}}(\theta) = \mathcal{L}_{\text{CFM}}(\theta) + C \tag{43}$$

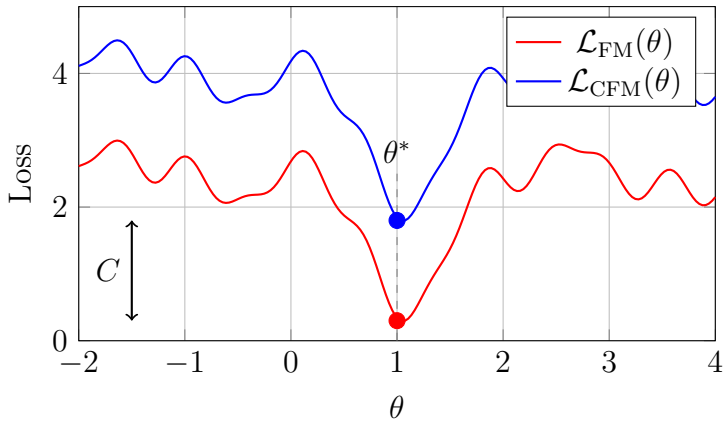
where C is a constant independent of θ.

Exercise

> Before we move on to analyze how this relationship is achieved and what it means, try to predict the consequences of this result.
>
> We are saying that the marginal loss is exactly the same as the conditional loss plus a constant. What does this imply?

The proof of this relationship is provided in Annex A.5.

Let's visualize what this relationship means:



Notice the key implications:

- Since they're the same function separated only by a constant, their minima occur at the same θ^*. Minimizing one is equivalent to minimizing the other!

  For a minimizer θ^* of $\mathcal{L}_{\text{CFM}}$, we have $\mathbf{u}_t^{\theta^*} = \mathbf{u}_t$. This means that if we find the minimizer for the conditional flow matching loss, we also get the minimizer for the marginal velocity field!

- $\mathcal{L}_{\text{CFM}}$ will never reach zero at the minimum. Since $\mathcal{L}_{\text{CFM}}(\theta) = \mathcal{L}_{\text{FM}}(\theta) + C$, even at the minimum we still have the constant C.

- The gradients are identical: $\nabla_\theta \mathcal{L}_{\text{CFM}}(\theta) = \nabla_\theta \mathcal{L}_{\text{FM}}(\theta)$. Trivially, the gradient of a constant is zero, so gradient-based optimization behaves identically for both losses.

Insight

> For our purposes, we don't care about the actual values of each function - we're simply trying to find the minimum! This is why we can use this result to create the actual training objective.

3.4 The Flow Matching Algorithm

Now that we have an objective, we can define the training algorithm. We have a dataset (in our example, the mixture of Gaussians distribution) and our neural network $\mathbf{u}_t^\theta$.

Flow Matching Training Algorithm

Initialize: Neural network u_t^θ with parameters θ
Dataset: p_{data} (e.g., mixture of Gaussians)

for each training iteration do:
  # Sample mini-batch
  $z \sim p_{\text{data}}$ # Sample from dataset
  $t \sim \mathcal{U}(0,1)$ # Sample random time
  $x \sim p_t(x|z)$ # Sample from conditional path
  # Compute loss
  $\mathcal{L}(\theta) = \|u_t^\theta(x) - u_t(x|z)\|^2$
  # Update parameters
  $\theta \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}(\theta)$
end for

3.4.1 Example: Training with Gaussian Initial Distribution

Let's make this concrete with our toy example from the Introduction: the initial distribution is a standard Gaussian $q = \mathcal{N}(0, I)$ and the target is our mixture of Gaussians p with modes at ± 2 on the x-axis. Since we chose a Gaussian initial distribution, we have a closed-form expression for the conditional velocity field:

$$\mathbf{u}_t(x|z) = \frac{z - (1-\beta_t)x}{\beta_t} \tag{44}$$

where we use our previously defined schedule $\beta_t = 1 - t$. Let's walk through one training iteration:

Example: One Training Step

1. **Sample from data:** We sample z from our mixture of Gaussians p (the target distribution from our Introduction). Let's say we get $z = (2.1, 0)$ from the right mode at $+2$.

2. **Sample time:** We draw $t \sim \mathcal{U}(0,1)$. Let's say $t = 0.7$.

3. **Sample along path:** With $\beta_t = 1 - 0.7 = 0.3$, we sample from:

$$p_t(x|z) = \mathcal{N}((1-\beta_t)z, \beta_t^2 I) = \mathcal{N}(0.7 \cdot 2.1, 0.09) = \mathcal{N}(1.47, 0.09) \tag{45}$$

   So we sample x from a Gaussian centered at 1.47 with variance 0.09. Let's say we get $x = 1.52$.

4. **Compute true velocity:** Using our closed-form expression:

$$\mathbf{u}_t(x|z) = \frac{z - (1-\beta_t)x}{\beta_t} = \frac{2.1 - 0.7 \cdot 1.52}{0.3} = \frac{2.1 - 1.064}{0.3} = \frac{1.036}{0.3} = 3.45 \tag{46}$$

5. **Compute neural network prediction:** We pass $(t = 0.7, x = 1.52)$ through our neural network to get $\mathbf{u}_t^\theta(x) = 3.2$ (initially this will be random).

6. **Compute loss and update:**

$$\mathcal{L} = |3.2 - 3.45|^2 = 0.0625 \tag{47}$$

   We then backpropagate this loss to update θ.

The beauty is that even though the marginal velocity field $\mathbf{u}_t(x)$ is complex (it needs to split the Gaussian into two modes at ± 2 as introduced in our toy example), we only ever train on the simpler conditional velocity field $\mathbf{u}_t(x|z)$, which just points from the current position toward the specific target z.

4 Integration and Inference

Having trained our neural network $\mathbf{u}_t^\theta$ to approximate the velocity field, we now turn to the question of generating samples. The process is elegantly straightforward: we begin with a point x_0 sampled from our source distribution q (typically a standard Gaussian), and we follow the learned velocity field to transform this simple noise into a sample from our target distribution p.

4.1 Euler Method

The simplest approach is to use the Euler method for numerical integration. Given:

- Initial condition: $x_0 \sim q$ (sampled from source distribution)
- Learned velocity field: $\mathbf{u}_t^\theta$
- Number of steps: N

Euler Integration for Flow Matching

Input: Initial point x_0, velocity field $\mathbf{u}_t^\theta$, number of steps N

$t \leftarrow 0$ # Start at time 0
$h \leftarrow 1/N$ # Step size
$X_t \leftarrow x_0$ # Initialize position
for $i = 0, 1, \ldots, N-1$ do:
    $X_{t+h} \leftarrow X_t + h \cdot \mathbf{u}_t^\theta(X_t)$ # Euler step
    $t \leftarrow t + h$ # Update time
end for
Return: X_1 # Final position at $t = 1$

The Euler method approximates the continuous ODE $\frac{dX_t}{dt} = \mathbf{u}_t(X_t)$ by taking small discrete steps. At each time step, we:

1. Evaluate the velocity field at the current position
2. Move in that direction by a small amount proportional to the step size h
3. Update the time and repeat

For better accuracy, one can also store and return the entire trajectory $\{X_0, X_h, X_{2h}, \ldots, X_1\}$, which shows how the sample evolves from the source to the target distribution.

4.2 Inference

Let's see this in action with our toy example. We sample points from the source distribution $q = \mathcal{N}(0, 1)$ and follow the learned velocity field to generate samples from our target mixture of Gaussians:

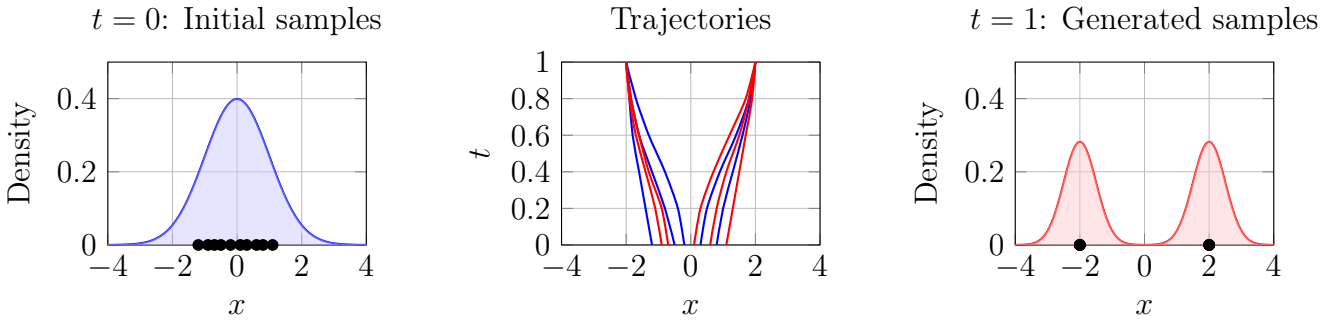


Figure 8: Inference with flow matching. Left: We sample 10 points from the source Gaussian. Middle: Each sample follows a trajectory through the learned velocity field. Right: The samples arrive at the target distribution, correctly split between the two modes.
The learned velocity field automatically routes each initial sample to the appropriate mode. Notice how samples starting near zero can end up at either mode—the neural network has learned the correct probability split!

A Annex

A.1 Derivation of Gaussian Conditional Velocity Field

For the Gaussian probability path, we define the position at time t as:

$$X_t = \alpha_t z + \beta_t \epsilon \tag{48}$$

where z is the target point and $\epsilon \sim \mathcal{N}(0, I)$ is Gaussian noise.

To find the velocity field, we take the time derivative:

$$\frac{dX_t}{dt} = \dot{\alpha}_t z + \dot{\beta}_t \epsilon \tag{49}$$

Now we need to express ϵ in terms of the current position X_t. From the original equation:

$$\epsilon = \frac{X_t - \alpha_t z}{\beta_t} \tag{50}$$

Substituting this back into the velocity equation:

$$\frac{dX_t}{dt} = \dot{\alpha}_t z + \dot{\beta}_t \left(\frac{X_t - \alpha_t z}{\beta_t} \right) \tag{51}$$

$$= \dot{\alpha}_t z + \frac{\dot{\beta}_t}{\beta_t} X_t - \frac{\dot{\beta}_t}{\beta_t} \alpha_t z \tag{52}$$

$$= \left(\dot{\alpha}_t - \frac{\dot{\beta}_t}{\beta_t} \alpha_t \right) z + \frac{\dot{\beta}_t}{\beta_t} X_t \tag{53}$$

Therefore, the conditional velocity field is:

$$\mathbf{u}_t(x|z) = \left(\dot{\alpha}_t - \frac{\dot{\beta}_t}{\beta_t} \alpha_t \right) z + \frac{\dot{\beta}_t}{\beta_t} x \tag{54}$$

A.2 Derivatives: Functions vs Probability Measures

You might be familiar with the definition of a derivative as:

$$\lim_{h \to 0} \frac{f(x+h) - f(x)}{h} \tag{55}$$

This is true! More concretely, this is "partially" true. This definition holds up, but only in a space where subtraction and addition of scalars makes sense. This "ordinary derivative" works specifically in Banach spaces (like functions $f(t, \cdot)$ in L^2).

But if $f(t) = \pi_t$ is a probability measure, we run into problems:

- You can't subtract measures pointwise like numbers. For instance, $\pi_{t+h}(x) - \pi_t(x)$ may not even exist if there are no densities.
- Even when densities exist, they may not be smooth enough to make the pointwise limit well-defined.

The naive derivative is too restrictive here. So we avoid subtraction and generalize differentiation to capture the rate of change of density where the "ordinary derivative" might not exist. This leads us to the continuity equation:

$$\frac{\partial \pi_t}{\partial t} + \nabla \cdot (\pi_t \mathbf{u}_t) = 0 \tag{56}$$

This weak formulation captures how distributions evolve without requiring pointwise differentiability.

If this is not satisfactory enough, look at the following Annex section to get a formal definition.

A.3 Weak Derivatives and Wasserstein Gradient Flows

[To be added: Formal mathematical definition of derivatives in distribution spaces]

A.4 Derivation of Marginal Velocity Field

We start with two key equations: the marginal probability path and the continuity equation that any probability path must satisfy:

$$\pi_t(x) = \int \pi_t(x|z)p(z)\,dz \tag{57}$$

$$\frac{\partial \pi_t}{\partial t} + \nabla \cdot (\pi_t \mathbf{u}_t) = 0 \tag{58}$$

Let's take the time derivative of the marginal probability path:

$$\frac{\partial \pi_t(x)}{\partial t} = \frac{\partial}{\partial t} \int \pi_t(x|z)p(z)\,dz \tag{59}$$

Since $p(z)$ doesn't depend on time, we can move the derivative inside the integral:

$$\frac{\partial \pi_t(x)}{\partial t} = \int \frac{\partial \pi_t(x|z)}{\partial t} p(z)\,dz \tag{60}$$

Now, each conditional path $\pi_t(x|z)$ satisfies its own continuity equation with conditional velocity field $\mathbf{u}_t(x|z)$:

$$\frac{\partial \pi_t(x|z)}{\partial t} = -\nabla \cdot (\pi_t(x|z)\mathbf{u}_t(x|z)) \tag{61}$$

Substituting this into our integral:

$$\frac{\partial \pi_t(x)}{\partial t} = \int -\nabla \cdot (\pi_t(x|z)\mathbf{u}_t(x|z))p(z)\,dz \tag{62}$$

$$= -\int \nabla \cdot (\pi_t(x|z)\mathbf{u}_t(x|z))p(z)\,dz \tag{63}$$

We can pull the divergence operator outside the integral (since it acts on x, not z):

$$\frac{\partial \pi_t(x)}{\partial t} = -\nabla \cdot \left(\int \pi_t(x|z)\mathbf{u}_t(x|z)p(z)\,dz \right) \tag{64}$$

Now we can factor out $\pi_t(x)$ by multiplying and dividing:

$$\frac{\partial \pi_t(x)}{\partial t} = -\nabla \cdot \left(\pi_t(x) \int \mathbf{u}_t(x|z) \frac{\pi_t(x|z)p(z)}{\pi_t(x)} \, dz \right) \tag{65}$$

This has the same form as the continuity equation! Comparing with:

$$\frac{\partial \pi_t}{\partial t} = -\nabla \cdot (\pi_t \mathbf{u}_t) \tag{66}$$

We can identify the marginal velocity field as:

$$\mathbf{u}_t(x) = \int \mathbf{u}_t(x|z) \frac{\pi_t(x|z)p(z)}{\pi_t(x)} \, dz \tag{67}$$

A.5 Equivalence of Flow Matching Losses

We will prove that the marginal Flow Matching loss and the Conditional Flow Matching loss differ only by a constant independent of θ:

$$\mathcal{L}_{\text{FM}}(\theta) = \mathcal{L}_{\text{CFM}}(\theta) + C \tag{68}$$

[Proof to be added]

B References

References

[1] Fryar CD, Carroll MD, Gu Q, Afful J, Ogden CL. *Anthropometric reference data for children and adults: United States, 2015–2018*. National Center for Health Statistics. Vital Health Stat 3(46). 2021.

[2] Centers for Disease Control and Prevention. *CDC Growth Charts: United States*. National Center for Health Statistics. 2000. Available at: `https://www.cdc.gov/growthcharts/`

[3] U.S. Census Bureau. *2020 Census Demographic and Housing Characteristics*. Available at: `https://www.census.gov/data/`