Flow Matching: Learning the Transport Directly

Lecture 25 ▪ CSCI 4140/6960 ▪ Machine Learning & Optimization

Alex Gittens

Rensselaer Polytechnic Institute

Spring 2026

Course website: `cs.rpi.edu/.../mlandopt.html`

Learning Objectives

By the end of this lecture, you should be able to:

1. **Explain** the transport viewpoint of generation: how a time-dependent velocity field moves samples from p_0 to p_{data}
2. **State and interpret** the continuity equation as density-level mass conservation, and explain why it does not directly provide trainable pointwise labels
3. **Construct** the CFM training pair $(\mathbf{x}_t, \mathbf{u}_t^{\text{cond}})$ from sampled endpoints and a time t
4. **Explain** the regression principle $\mathbf{u}_t(\mathbf{x}) = \mathbb{E}[\dot{\mathbf{x}}_t \mid \mathbf{x}_t = \mathbf{x}]$ and why straight conditional paths can still induce curved sampled ODE trajectories
5. **State** the CFM theorem at a high level (same optimization target as ideal FM) and **describe** sampling as numerical ODE integration
6. **Compare** DDPM, DDIM, and flow matching in terms of training target and native sampling mechanism

Lecture Roadmap

From Denoising to Transport

The Marginal Velocity Field

Conditional Flow Matching

CFM Theorem and Sampling Implications

Comparison and Summary

Why Study Flow Matching?

Flow matching gives a clean transport-based view of generation.

- It makes the **deterministic ODE viewpoint** explicit: learn a velocity field and integrate it directly.
- Its training objective is a simple **regression on velocities** along chosen interpolation paths.
- It connects naturally to **optimal-transport intuition**, which is easier to reason about geometrically than score functions.
- Suitable path choices often produce trajectories that are **easier to integrate numerically**, so good samples can be obtained with fewer function evaluations.
- It is **closely related to continuous-time diffusion** — learning flow matching also clarifies diffusion rather than replacing it.
- It works in **latent space** just as well as pixel space — state-of-the-art models such as Stable Diffusion 3 and Flux apply flow matching to VAE latents, combining the efficiency of latent representations with the clean ODE sampler.

From Denoising to Transport

From Denoising to Transport

Two related training viewpoints

Diffusion-style training (DDPM/DDIM):

$$\min_{\theta} \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\|^2\right]$$

The network learns a *noise predictor* (equivalently, a score surrogate under standard parameterizations).

From Denoising to Transport

Two related training viewpoints

Diffusion-style training (DDPM/DDIM):

$$\min_{\theta} \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\|^2\right]$$

The network learns a *noise predictor* (equivalently, a score surrogate under standard parameterizations).

Deterministic or stochastic samplers are then **constructed from** this learned object.

From Denoising to Transport

Two related training viewpoints

Diffusion-style training (DDPM/DDIM):

$$\min_{\theta} \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\|^2\right]$$

The network learns a *noise predictor* (equivalently, a score surrogate under standard parameterizations).

Deterministic or stochastic samplers are then **constructed from** this learned object.

Flow matching training:

$$\min_{\theta} \mathbb{E}\left[\|\mathbf{u}_\theta(\mathbf{x}_t, t) - \mathbf{u}_t^{\mathrm{cond}}\|^2\right]$$

The network learns a *velocity field* $\mathbf{u}_\theta(\cdot, t)$ directly.

Sampling then **integrates that same field** as an ODE — the training target and the sampling object are the same function.

From Denoising to Transport

Two related training viewpoints

Diffusion-style training (DDPM/DDIM):

$$\min_{\theta} \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\|^2\right]$$

The network learns a *noise predictor* (equivalently, a score surrogate under standard parameterizations).

Deterministic or stochastic samplers are then **constructed from** this learned object.

Flow matching training:

$$\min_{\theta} \mathbb{E}\left[\|\mathbf{u}_\theta(\mathbf{x}_t, t) - \mathbf{u}_t^{\text{cond}}\|^2\right]$$

The network learns a *velocity field* $\mathbf{u}_\theta(\cdot, t)$ directly.

Sampling then **integrates that same field** as an ODE — the training target and the sampling object are the same function.

Flow matching changes which object is supervised: diffusion learns a noise/score predictor from which transport is derived; flow matching directly regresses the velocity field used in the ODE.

What Is Transport?

Fix a time-dependent velocity field $\mathbf{u}_t : \mathbb{R}^d \to \mathbb{R}^d$ and a distribution p_0.

1. Sample a random initial point $X_0 \sim p_0$.

What Is Transport?

Fix a time-dependent velocity field $\mathbf{u}_t : \mathbb{R}^d \to \mathbb{R}^d$ and a distribution p_0.

1. Sample a random initial point $X_0 \sim p_0$.
2. Flow it forward by solving the ODE

$$\frac{\mathrm{d}}{\mathrm{d}t} X_t = \mathbf{u}_t(X_t), \qquad t \in [0, 1].$$

What Is Transport?

Fix a time-dependent velocity field $\mathbf{u}_t : \mathbb{R}^d \to \mathbb{R}^d$ and a distribution p_0.

1. Sample a random initial point $X_0 \sim p_0$.
2. Flow it forward by solving the ODE

$$\frac{\mathrm{d}}{\mathrm{d}t} X_t = \mathbf{u}_t(X_t), \qquad t \in [0, 1].$$

3. Since X_0 is random, X_t is a *random variable* for every t. Let p_t denote the distribution of X_t.

What Is Transport?

Fix a time-dependent velocity field $\mathbf{u}_t : \mathbb{R}^d \to \mathbb{R}^d$ and a distribution p_0.

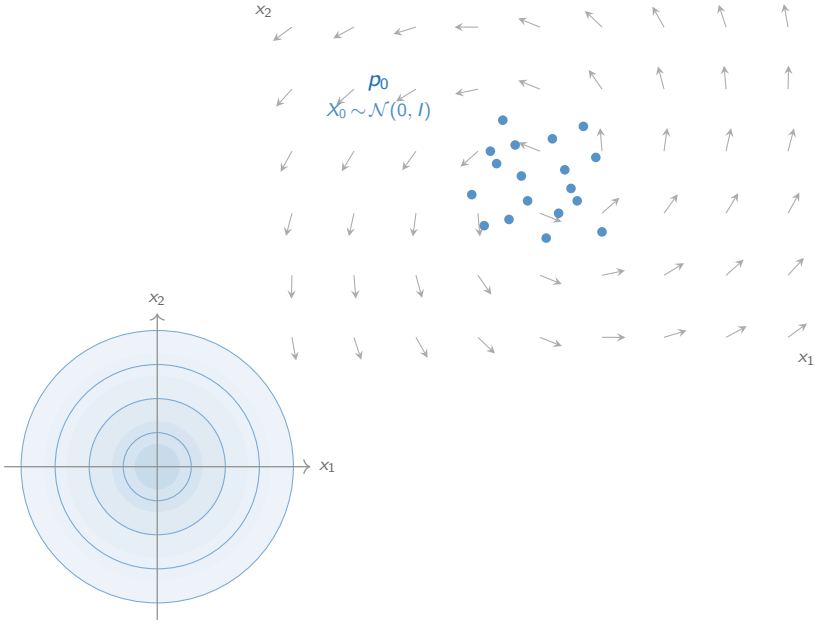
1. Sample a random initial point $X_0 \sim p_0$.
2. Flow it forward by solving the ODE

$$\frac{\mathrm{d}}{\mathrm{d}t} X_t = \mathbf{u}_t(X_t), \qquad t \in [0, 1].$$

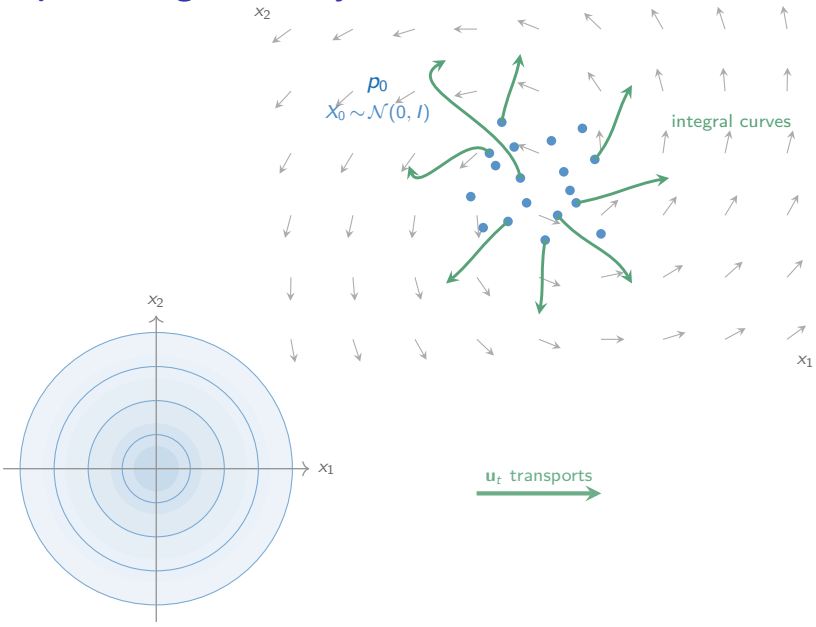
3. Since X_0 is random, X_t is a *random variable* for every t. Let p_t denote the distribution of X_t.
4. The field $\mathbf{u}_t$ therefore moves the whole distribution: $p_0 \xrightarrow{\mathbf{u}_t} p_1$.

We say $\mathbf{u}_t$ **transports** p_0 to p_1. Pointwise motion of samples induces motion of the whole probability mass.

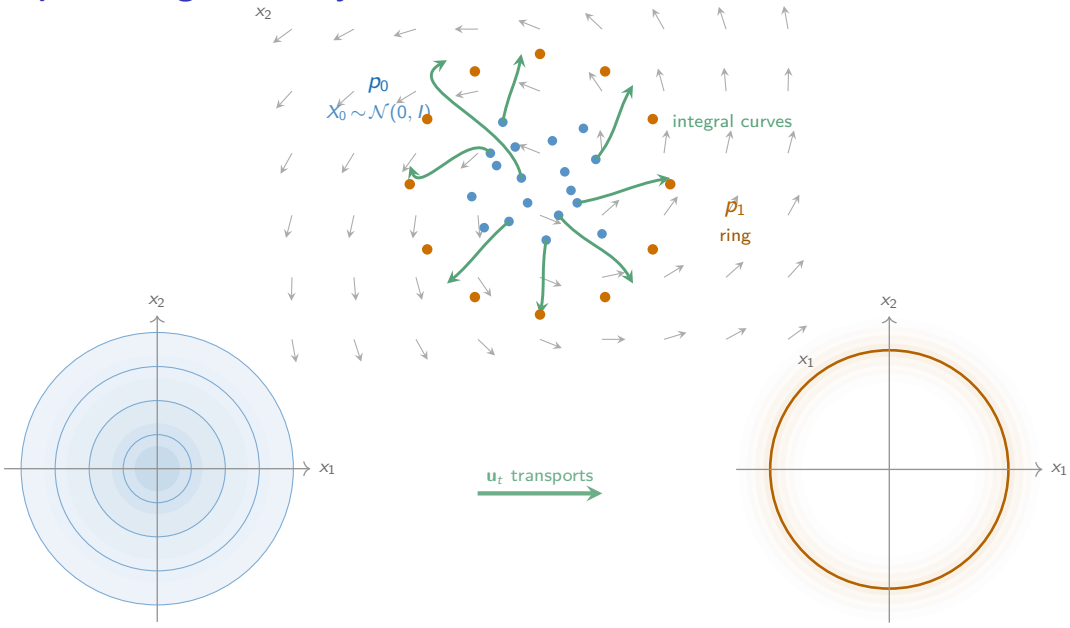
Transport Along a Velocity Field



Transport Along a Velocity Field



Transport Along a Velocity Field



Diffusion Also Defines Deterministic Transport

Convention change: flow matching indexes time so that $t = 0$ is noise and $t = 1$ is data — the reverse of the DDPM forward-process convention used in earlier lectures.

So far, diffusion was presented at discrete noise levels $t = 1, \dots, T$. We now pass to the **continuous-time viewpoint** that underlies ODE-based fast samplers and flow matching: consider a continuously indexed family of densities

$$(p_t)_{t \in [0,1]},$$

where p_0 is the noisiest distribution and $p_1 = p_{\text{data}}$. For each time t, define its score field by

$$s_t(\mathbf{x}) := \nabla_{\mathbf{x}} \log p_t(\mathbf{x}).$$

Diffusion Also Defines Deterministic Transport

Convention change: flow matching indexes time so that $t = 0$ is noise and $t = 1$ is data — the reverse of the DDPM forward-process convention used in earlier lectures.

So far, diffusion was presented at discrete noise levels $t = 1, \dots, T$. We now pass to the **continuous-time viewpoint** that underlies ODE-based fast samplers and flow matching: consider a continuously indexed family of densities

$$(p_t)_{t \in [0,1]},$$

where p_0 is the noisiest distribution and $p_1 = p_{\text{data}}$. For each time t, define its score field by

$$s_t(\mathbf{x}) := \nabla_{\mathbf{x}} \log p_t(\mathbf{x}).$$

Fact. There exists a deterministic velocity field $\mathbf{u}_t(\mathbf{x})$ such that if we solve

$$\frac{\mathrm{d}}{\mathrm{d}t} X_t = \mathbf{u}_t(X_t), \qquad X_0 \sim p_0,$$

then the random variable X_t has distribution p_t for every $t \in [0, 1]$. This ODE is called the **probability flow ODE**.

Diffusion Also Defines Deterministic Transport

Convention change: flow matching indexes time so that $t = 0$ is noise and $t = 1$ is data — the reverse of the DDPM forward-process convention used in earlier lectures.

So far, diffusion was presented at discrete noise levels $t = 1, \dots, T$. We now pass to the **continuous-time viewpoint** that underlies ODE-based fast samplers and flow matching: consider a continuously indexed family of densities

$$(p_t)_{t\in[0,1]},$$

where p_0 is the noisiest distribution and $p_1 = p_{\text{data}}$. For each time t, define its score field by

$$s_t(\mathbf{x}) := \nabla_{\mathbf{x}} \log p_t(\mathbf{x}).$$

Fact. There exists a deterministic velocity field $\mathbf{u}_t(\mathbf{x})$ such that if we solve

$$\frac{\mathrm{d}}{\mathrm{d}t} X_t = \mathbf{u}_t(X_t), \qquad X_0 \sim p_0,$$

then the random variable X_t has distribution p_t for every $t \in [0, 1]$. This ODE is called the **probability flow ODE**.

So the same evolving densities from diffusion can be viewed in two ways: as stochastic denoising, or as deterministic transport by a velocity field.

Diffusion Also Defines Deterministic Transport

Convention change: flow matching indexes time so that $t = 0$ is noise and $t = 1$ is data — the reverse of the DDPM forward-process convention used in earlier lectures.

So far, diffusion was presented at discrete noise levels $t = 1, \dots, T$. We now pass to the **continuous-time viewpoint** that underlies ODE-based fast samplers and flow matching: consider a continuously indexed family of densities

$$(p_t)_{t\in[0,1]},$$

where p_0 is the noisiest distribution and $p_1 = p_{\text{data}}$. For each time t, define its score field by

$$s_t(\mathbf{x}) := \nabla_{\mathbf{x}} \log p_t(\mathbf{x}).$$

Fact. There exists a deterministic velocity field $\mathbf{u}_t(\mathbf{x})$ such that if we solve

$$\frac{\mathrm{d}}{\mathrm{d}t} X_t = \mathbf{u}_t(X_t), \qquad X_0 \sim p_0,$$

then the random variable X_t has distribution p_t for every $t \in [0, 1]$. This ODE is called the **probability flow ODE**.

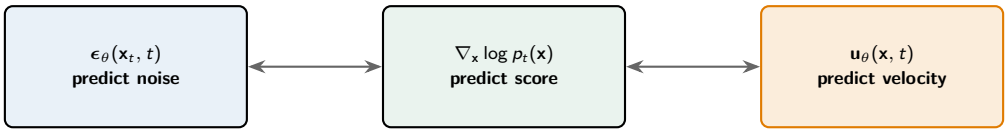
So the same evolving densities from diffusion can be viewed in two ways: as stochastic denoising, or as deterministic transport by a velocity field.

Why this matters for us. Flow matching takes the deterministic viewpoint as the main object: instead of learning a score field and then deriving transport, it learns the transport field directly.

Three Related Learned Objects in Generative Modeling

$-\frac{1}{\sigma_t}\epsilon_\theta \approx \nabla_{\mathbf{x}} \log p_t$

probability flow ODE

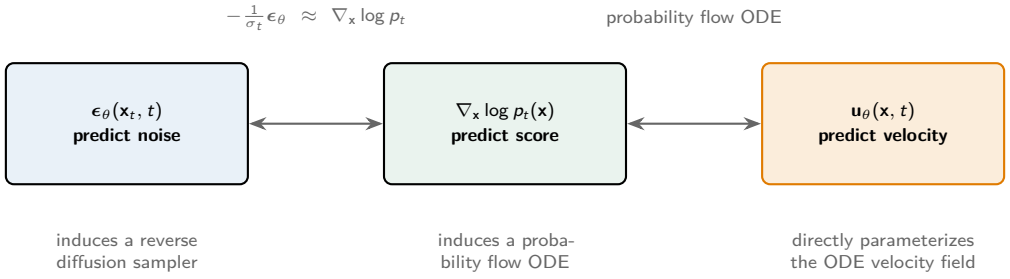


induces a reverse diffusion sampler

induces a probability flow ODE

directly parameterizes the ODE velocity field

Three Related Learned Objects in Generative Modeling



All three are related: the noise predictor is proportional to the score, and the score determines a probability flow ODE whose right-hand side is a velocity field. Flow matching directly regresses that velocity field.

The Marginal Velocity Field

The Central Object: a Velocity Field That Moves Samples

A flow-matching model aims to learn a time-dependent velocity field

$$\mathbf{u}_t(\mathbf{x}) \in \mathbb{R}^d$$

that tells us, for each location $\mathbf{x}$ at time t, the direction and speed of motion.

The Central Object: a Velocity Field That Moves Samples

A flow-matching model aims to learn a time-dependent velocity field

$$\mathbf{u}_t(\mathbf{x}) \in \mathbb{R}^d$$

that tells us, for each location $\mathbf{x}$ at time t, the direction and speed of motion.
Once such a field is known, we generate by solving the ODE

$$\frac{\mathrm{d}}{\mathrm{d}t} X_t = \mathbf{u}_t(X_t), \qquad X_0 \sim p_0 = \mathcal{N}(0, I), \qquad t \in [0, 1].$$

Ideal global object: $\mathbf{u}_t(\mathbf{x})$ is the field transporting $p_0 = \mathcal{N}(0, I)$ at $t = 0$ to p_{data} at $t = 1$.

The Central Object: a Velocity Field That Moves Samples

A flow-matching model aims to learn a time-dependent velocity field

$$\mathbf{u}_t(\mathbf{x}) \in \mathbb{R}^d$$

that tells us, for each location $\mathbf{x}$ at time t, the direction and speed of motion.

Once such a field is known, we generate by solving the ODE

$$\frac{\mathrm{d}}{\mathrm{d}t} X_t = \mathbf{u}_t(X_t), \qquad X_0 \sim p_0 = \mathcal{N}(0, I), \qquad t \in [0, 1].$$

Ideal global object: $\mathbf{u}_t(\mathbf{x})$ is the field transporting $p_0 = \mathcal{N}(0, I)$ at $t = 0$ to p_{data} at $t = 1$.

Does such a $\mathbf{u}_t$ exist? Yes — in fact, *infinitely many* velocity fields can transport p_0 to p_{data}. We construct one explicitly in Appendix B.

What Training Would Look Like (Ideally)

Suppose we had access to a valid $\mathbf{u}_t$ and could sample pairs $(\mathbf{x}_t, \mathbf{u}_t(\mathbf{x}_t))$.

What Training Would Look Like (Ideally)

Suppose we had access to a valid $\mathbf{u}_t$ and could sample pairs $(\mathbf{x}_t, \mathbf{u}_t(\mathbf{x}_t))$.

Then learning an approximation $\mathbf{u}_\theta(\mathbf{x}, t) \approx \mathbf{u}_t(\mathbf{x})$ is just **least-squares regression** minimizing the **flow-matching objective**:

$$\mathcal{L}_{\text{FM}}(\theta) \;=\; \mathbb{E}\Big[\big\|\mathbf{u}_\theta(\mathbf{x}_t, t) - \mathbf{u}_t(\mathbf{x}_t)\big\|^2\Big].$$

What Training Would Look Like (Ideally)

Suppose we had access to a valid $\mathbf{u}_t$ and could sample pairs $(\mathbf{x}_t, \mathbf{u}_t(\mathbf{x}_t))$.

Then learning an approximation $\mathbf{u}_\theta(\mathbf{x}, t) \approx \mathbf{u}_t(\mathbf{x})$ is just **least-squares regression** minimizing the **flow-matching objective**:

$$\mathcal{L}_{\text{FM}}(\theta) \;=\; \mathbb{E}\Big[\big\|\mathbf{u}_\theta(\mathbf{x}_t, t) - \mathbf{u}_t(\mathbf{x}_t)\big\|^2\Big].$$

Given a trained $\mathbf{u}_\theta$, generating a sample from p_{data} is straightforward:

1. Draw $X_0 \sim \mathcal{N}(0, I)$.
2. Numerically integrate $\frac{\mathrm{d}}{\mathrm{d}t} X_t = \mathbf{u}_\theta(X_t, t)$ from $t = 0$ to $t = 1$.
3. Return $X_1 \approx p_{\text{data}}$.

What Training Would Look Like (Ideally)

Suppose we had access to a valid $\mathbf{u}_t$ and could sample pairs $(\mathbf{x}_t, \mathbf{u}_t(\mathbf{x}_t))$.

Then learning an approximation $\mathbf{u}_\theta(\mathbf{x}, t) \approx \mathbf{u}_t(\mathbf{x})$ is just **least-squares regression** minimizing the **flow-matching objective**:

$$\mathcal{L}_{\text{FM}}(\theta) \;=\; \mathbb{E}\Big[\big\|\mathbf{u}_\theta(\mathbf{x}_t, t) - \mathbf{u}_t(\mathbf{x}_t)\big\|^2\Big].$$

Given a trained $\mathbf{u}_\theta$, generating a sample from p_{data} is straightforward:

1. Draw $X_0 \sim \mathcal{N}(0, I)$.
2. Numerically integrate $\frac{\mathrm{d}}{\mathrm{d}t} X_t = \mathbf{u}_\theta(X_t, t)$ from $t = 0$ to $t = 1$.
3. Return $X_1 \approx p_{\text{data}}$.

Problem: we do not have $\mathbf{u}_t$ to sample from.

Two Problems to Solve

We address two problems in turn:

1. **Construct a valid $\mathbf{u}_t$ in principle.**
   Use endpoint-conditioned interpolation paths to define a marginal velocity field that provably transports p_0 to p_{data}.

2. **Perform the regression without samples from $\mathbf{u}_t$.**
   Replace the intractable marginal target with computable *conditional* velocity labels, and show the two objectives share the same minimizer.

How to determine $\mathbf{u}_t$?

Let's tackle the first problem: finding a feasible $\mathbf{u}_t$. Of course, any valid velocity field $\mathbf{u}_t$ must be consistent with the evolution of the density p_t.

1. **Particle story:** each point follows $\frac{\mathrm{d}}{\mathrm{d}t} X_t = \mathbf{u}_t(X_t)$
2. **Region story:** probability mass flows into and out of every region
3. **Density story:** local conservation of mass gives a PDE for p_t

How to determine $\mathbf{u}_t$?

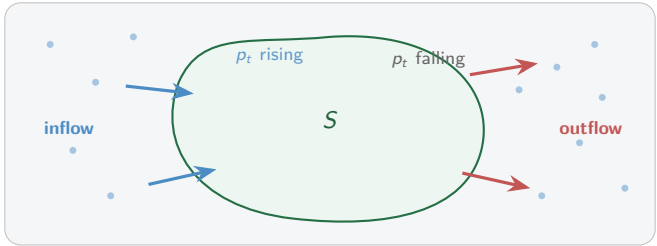
Let's tackle the first problem: finding a feasible $\mathbf{u}_t$. Of course, any valid velocity field $\mathbf{u}_t$ must be consistent with the evolution of the density p_t.

1. **Particle story:** each point follows $\frac{\mathrm{d}}{\mathrm{d}t} X_t = \mathbf{u}_t(X_t)$
2. **Region story:** probability mass flows into and out of every region
3. **Density story:** local conservation of mass gives a PDE for p_t

The density-level constraint is the **continuity equation**:

$$\boxed{\partial_t p_t(\mathbf{x}) + \nabla \cdot (p_t(\mathbf{x})\mathbf{u}_t(\mathbf{x})) = 0, \qquad p_0 = \mathcal{N}(0, I), \quad p_1 = p_{\mathrm{data}}.}$$

Physical Intuition for the Continuity Equation



- $p_t(\mathbf{x})\,\mathbf{u}_t(\mathbf{x})$ is the **probability flux**: how much mass passes through $\mathbf{x}$ per unit time
- $\nabla \cdot \big(p_t(\mathbf{x})\,\mathbf{u}_t(\mathbf{x})\big) = \sum_{i=1}^{d} \frac{\partial}{\partial x_i}\big[p_t(\mathbf{x})\,(u_t)_i(\mathbf{x})\big]$ measures net outflow per unit volume
- $\partial_t p_t(\mathbf{x})$ = rate of local density change

Restating:

$$\boxed{\partial_t p_t(\mathbf{x}) + \nabla \cdot \big(p_t(\mathbf{x})\,\mathbf{u}_t(\mathbf{x})\big) = 0}$$

In words: local density rises only when more mass flows in than out. Since total probability is always 1, mass is **transported**, not created or destroyed.

Why the Continuity Equation Is Not Enough

The continuity equation

$$\partial_t p_t + \nabla \cdot (p_t \mathbf{u}_t) = 0, \qquad p_0 = \mathcal{N}(0, I), \quad p_1 = p_{\text{data}}$$

is a **necessary and sufficient** condition for validity: $\mathbf{u}_t$ transports p_0 to p_{data} *if and only if* it satisfies this equation.

Why the Continuity Equation Is Not Enough

The continuity equation

$$\partial_t p_t + \nabla \cdot (p_t \mathbf{u}_t) = 0, \qquad p_0 = \mathcal{N}(0, I), \quad p_1 = p_{\text{data}}$$

is a **necessary and sufficient** condition for validity: $\mathbf{u}_t$ transports p_0 to p_{data} *if and only if* it satisfies this equation.

Yet it is **not sufficient as a training rule**:

- It is satisfied by **infinitely many** velocity fields (see Appendix B), so it does not pin down a unique regression target.
- It is a density-level PDE, not a pointwise supervised label — we cannot read off the value $\mathbf{u}_t(\mathbf{x})$ at a training point $(\mathbf{x}, t)$ without solving this PDE.
- The equation is **implicitly coupled**: p_t is not a given input — it is itself determined by $\mathbf{u}_t$ via the pushforward of p_0 along the flow. There is no direct way to solve for $\mathbf{u}_t$ without already knowing p_t, and you cannot know p_t without already having $\mathbf{u}_t$.

Why the Continuity Equation Is Not Enough

The continuity equation

$$\partial_t p_t + \nabla \cdot (p_t \mathbf{u}_t) = 0, \qquad p_0 = \mathcal{N}(0, I), \quad p_1 = p_{\text{data}}$$

is a **necessary and sufficient** condition for validity: $\mathbf{u}_t$ transports p_0 to p_{data} *if and only if* it satisfies this equation.

Yet it is **not sufficient as a training rule**:

- It is satisfied by **infinitely many** velocity fields (see Appendix B), so it does not pin down a unique regression target.
- It is a density-level PDE, not a pointwise supervised label — we cannot read off the value $\mathbf{u}_t(\mathbf{x})$ at a training point $(\mathbf{x}, t)$ without solving this PDE.
- The equation is **implicitly coupled**: p_t is not a given input — it is itself determined by $\mathbf{u}_t$ via the pushforward of p_0 along the flow. There is no direct way to solve for $\mathbf{u}_t$ without already knowing p_t, and you cannot know p_t without already having $\mathbf{u}_t$.

Our resolution: **construct** a $\mathbf{u}_t$ that satisfies the continuity equation and for which we can generate supervision data. That is the role of **conditional paths**.

Conditional Flow Matching

Step 1: Start with a Coupling

The transport problem asks for a joint law on $(\mathbf{x}_0, \mathbf{x}_1)$ whose marginals are p_0 and p_{data}. Such a joint law is called a **coupling**.

Step 1: Start with a Coupling

The transport problem asks for a joint law on $(\mathbf{x}_0, \mathbf{x}_1)$ whose marginals are p_0 and p_{data}. Such a joint law is called a **coupling**.

Definition. A **coupling** $\pi(\mathbf{x}_0, \mathbf{x}_1)$ is a joint distribution satisfying

$$(\mathbf{x}_0, \mathbf{x}_1) \sim \pi, \qquad \mathbf{x}_0 \sim p_0, \qquad \mathbf{x}_1 \sim p_{\text{data}}.$$

Step 1: Start with a Coupling

The transport problem asks for a joint law on $(\mathbf{x}_0, \mathbf{x}_1)$ whose marginals are p_0 and p_{data}. Such a joint law is called a **coupling**.

Definition. A **coupling** $\pi(\mathbf{x}_0, \mathbf{x}_1)$ is a joint distribution satisfying

$$(\mathbf{x}_0, \mathbf{x}_1) \sim \pi, \qquad \mathbf{x}_0 \sim p_0, \qquad \mathbf{x}_1 \sim p_{\text{data}}.$$

The default choice: the **independent coupling**

$$\pi \;=\; p_0 \otimes p_{\text{data}},$$

i.e., draw $\mathbf{x}_0$ and $\mathbf{x}_1$ independently. This is the default coupling in vanilla CFM, and it is also commonly used in straight-line rectified-flow setups. Other couplings (e.g., optimal transport) are possible but independence is the default.

Step 1: Start with a Coupling

The transport problem asks for a joint law on $(\mathbf{x}_0, \mathbf{x}_1)$ whose marginals are p_0 and p_{data}. Such a joint law is called a **coupling**.

Definition. A **coupling** $\pi(\mathbf{x}_0, \mathbf{x}_1)$ is a joint distribution satisfying

$$(\mathbf{x}_0, \mathbf{x}_1) \sim \pi, \qquad \mathbf{x}_0 \sim p_0, \qquad \mathbf{x}_1 \sim p_{\text{data}}.$$

The default choice: the **independent coupling**

$$\pi \;=\; p_0 \otimes p_{\text{data}},$$

i.e., draw $\mathbf{x}_0$ and $\mathbf{x}_1$ independently. This is the default coupling in vanilla CFM, and it is also commonly used in straight-line rectified-flow setups. Other couplings (e.g., optimal transport) are possible but independence is the default.

The coupling is the *only* place where knowledge of p_{data} enters. Everything that follows turns this pairing into a continuous-time process.

Step 2: A Coupling + Interpolation = A Stochastic Process

Given a coupling π, we pick a smooth interpolation rule that turns each pair $(\mathbf{x}_0, \mathbf{x}_1)$ into a path. The simplest rule is a straight line.

Step 2: A Coupling + Interpolation = A Stochastic Process

Given a coupling π, we pick a smooth interpolation rule that turns each pair $(\mathbf{x}_0, \mathbf{x}_1)$ into a path. The simplest rule is a straight line.

Interpolation rule (straight-line path):

$$\mathbf{x}_t \;=\; (1-t)\mathbf{x}_0 + t\mathbf{x}_1, \qquad (\mathbf{x}_0, \mathbf{x}_1) \sim \pi.$$

This defines a stochastic process $\{\mathbf{x}_t\}_{t\in[0,1]}$ with marginal density

$$p_t \;=\; \mathrm{Law}\big((1-t)\mathbf{x}_0 + t\mathbf{x}_1\big).$$

Step 2: A Coupling + Interpolation = A Stochastic Process

Given a coupling π, we pick a smooth interpolation rule that turns each pair $(\mathbf{x}_0, \mathbf{x}_1)$ into a path. The simplest rule is a straight line.

Interpolation rule (straight-line path):

$$\mathbf{x}_t \;=\; (1-t)\mathbf{x}_0 + t\mathbf{x}_1, \qquad (\mathbf{x}_0, \mathbf{x}_1) \sim \pi.$$

This defines a stochastic process $\{\mathbf{x}_t\}_{t\in[0,1]}$ with marginal density

$$p_t \;=\; \mathrm{Law}\big((1-t)\mathbf{x}_0 + t\mathbf{x}_1\big).$$

Marginals are correct by construction:

- At $t = 0$: $p_0 = \mathrm{Law}(\mathbf{x}_0) = \mathcal{N}(0, I)$. ✓
- At $t = 1$: $p_1 = \mathrm{Law}(\mathbf{x}_1) = p_{\mathsf{data}}$. ✓

Step 2: A Coupling + Interpolation = A Stochastic Process

Given a coupling π, we pick a smooth interpolation rule that turns each pair $(\mathbf{x}_0, \mathbf{x}_1)$ into a path. The simplest rule is a straight line.

Interpolation rule (straight-line path):

$$\mathbf{x}_t \;=\; (1-t)\mathbf{x}_0 + t\mathbf{x}_1, \qquad (\mathbf{x}_0, \mathbf{x}_1) \sim \pi.$$

This defines a stochastic process $\{\mathbf{x}_t\}_{t\in[0,1]}$ with marginal density

$$p_t \;=\; \mathrm{Law}\big((1-t)\mathbf{x}_0 + t\mathbf{x}_1\big).$$

Marginals are correct by construction:

- At $t = 0$: $p_0 = \mathrm{Law}(\mathbf{x}_0) = \mathcal{N}(0, I)$. ✓
- At $t = 1$: $p_1 = \mathrm{Law}(\mathbf{x}_1) = p_{\text{data}}$. ✓

The process $\{\mathbf{x}_t\}$ already transports p_0 to p_{data} — by definition of the coupling and interpolation rule. No continuity equation to solve; the marginals are correct automatically.

The *conditional* velocity is trivially computable: $\dot{\mathbf{x}}_t \,|\, \mathbf{x}_0, \mathbf{x}_1 = \mathbf{x}_1 - \mathbf{x}_0$. But $\{\mathbf{x}_t\}$ is a *stochastic* process — we need a *deterministic* velocity field $\mathbf{u}_t(\mathbf{x})$ to run the ODE at sampling time.

Step 3: Marginal Velocity as a Conditional Mean

We have a stochastic process with the right marginals. We want the deterministic field $\mathbf{u}_t(\mathbf{x})$ that drives the same marginal evolution.

Step 3: Marginal Velocity as a Conditional Mean

We have a stochastic process with the right marginals. We want the deterministic field $\mathbf{u}_t(\mathbf{x})$ that drives the same marginal evolution.

Answer (regression principle). Among all deterministic fields $f(\mathbf{x}, t)$, the one minimizing

$$\mathbb{E}\Big[\|f(\mathbf{x}_t, t) - \dot{\mathbf{x}}_t\|^2\Big]$$

is the conditional mean of $\dot{\mathbf{x}}_t$ given $\mathbf{x}_t$:

$$\mathbf{u}_t(\mathbf{x}) \;=\; \mathbb{E}[\dot{\mathbf{x}}_t \mid \mathbf{x}_t = \mathbf{x}] \;=\; \mathbb{E}[\mathbf{x}_1 - \mathbf{x}_0 \mid (1-t)\mathbf{x}_0 + t\mathbf{x}_1 = \mathbf{x}] \quad \text{(straight-line interpolant)}.$$

This is the standard least-squares fact, not a new assumption.

Step 3: Marginal Velocity as a Conditional Mean

We have a stochastic process with the right marginals. We want the deterministic field $\mathbf{u}_t(\mathbf{x})$ that drives the same marginal evolution.

Answer (regression principle). Among all deterministic fields $f(\mathbf{x}, t)$, the one minimizing

$$\mathbb{E}\Big[\|f(\mathbf{x}_t, t) - \dot{\mathbf{x}}_t\|^2\Big]$$

is the conditional mean of $\dot{\mathbf{x}}_t$ given $\mathbf{x}_t$:

$$\mathbf{u}_t(\mathbf{x}) \;=\; \mathbb{E}[\dot{\mathbf{x}}_t \mid \mathbf{x}_t = \mathbf{x}] \;=\; \mathbb{E}[\mathbf{x}_1 - \mathbf{x}_0 \mid (1-t)\mathbf{x}_0 + t\mathbf{x}_1 = \mathbf{x}] \quad \text{(straight-line interpolant)}.$$

This is the standard least-squares fact, not a new assumption.

Proposition (proved in Appendix C, slides C2–C3). The field $\mathbf{u}_t(\mathbf{x}) = \mathbb{E}[\dot{\mathbf{x}}_t \mid \mathbf{x}_t = \mathbf{x}]$ satisfies the continuity equation $\partial_t p_t + \nabla \cdot (p_t \mathbf{u}_t) = 0$, and the ODE $\dot{\mathbf{x}} = \mathbf{u}_t(\mathbf{x})$ with $X_0 \sim p_0$ produces marginals $\{p_t\}_{t\in[0,1]}$, reaching p_{data} at $t = 1$.

Step 3: Marginal Velocity as a Conditional Mean

We have a stochastic process with the right marginals. We want the deterministic field $\mathbf{u}_t(\mathbf{x})$ that drives the same marginal evolution.

Answer (regression principle). Among all deterministic fields $f(\mathbf{x}, t)$, the one minimizing

$$\mathbb{E}\Big[\|f(\mathbf{x}_t, t) - \dot{\mathbf{x}}_t\|^2\Big]$$

is the conditional mean of $\dot{\mathbf{x}}_t$ given $\mathbf{x}_t$:

$$\mathbf{u}_t(\mathbf{x}) \;=\; \mathbb{E}[\dot{\mathbf{x}}_t \mid \mathbf{x}_t = \mathbf{x}] \;=\; \mathbb{E}[\mathbf{x}_1 - \mathbf{x}_0 \mid (1-t)\mathbf{x}_0 + t\mathbf{x}_1 = \mathbf{x}] \quad \text{(straight-line interpolant)}.$$

This is the standard least-squares fact, not a new assumption.

Proposition (proved in Appendix C, slides C2–C3). The field $\mathbf{u}_t(\mathbf{x}) = \mathbb{E}[\dot{\mathbf{x}}_t \mid \mathbf{x}_t = \mathbf{x}]$ satisfies the continuity equation $\partial_t p_t + \nabla \cdot (p_t \mathbf{u}_t) = 0$, and the ODE $\dot{\mathbf{x}} = \mathbf{u}_t(\mathbf{x})$ with $X_0 \sim p_0$ produces marginals $\{p_t\}_{t\in[0,1]}$, reaching p_{data} at $t = 1$.

$\mathbf{u}_t$ is *not chosen* — it is *read off* from the stochastic process via the regression principle. Even if every conditional path is straight with constant velocity $\dot{\mathbf{x}}_t = \mathbf{x}_1 - \mathbf{x}_0$, the marginal field can still be state-dependent and time-dependent because it averages many different endpoint-conditioned directions using the posterior weights $p(\mathbf{x}_0, \mathbf{x}_1 \mid \mathbf{x}_t = \mathbf{x})$.

Step 4: Why We Can't Compute $\mathbf{u}_t$ Directly — and the Fix

Why $\mathbf{u}_t(\mathbf{x})$ is intractable. Evaluating

$$\mathbb{E}[\mathbf{x}_1 - \mathbf{x}_0 \mid (1-t)\mathbf{x}_0 + t\mathbf{x}_1 = \mathbf{x}]$$

requires integrating over all pairs $(\mathbf{x}_0, \mathbf{x}_1)$ consistent with $\mathbf{x}_t = \mathbf{x}$, weighted by their density — i.e., requires knowing $p_t(\mathbf{x})$. This same averaging is what makes the marginal field geometrically more complicated than the straight conditional labels: different endpoint pairs can contribute different directions at the same $(\mathbf{x}, t)$.

Step 4: Why We Can't Compute $\mathbf{u}_t$ Directly — and the Fix

Why $\mathbf{u}_t(\mathbf{x})$ is intractable. Evaluating

$$\mathbb{E}[\mathbf{x}_1 - \mathbf{x}_0 \mid (1-t)\mathbf{x}_0 + t\mathbf{x}_1 = \mathbf{x}]$$

requires integrating over all pairs $(\mathbf{x}_0, \mathbf{x}_1)$ consistent with $\mathbf{x}_t = \mathbf{x}$, weighted by their density — i.e., requires knowing $p_t(\mathbf{x})$. This same averaging is what makes the marginal field geometrically more complicated than the straight conditional labels: different endpoint pairs can contribute different directions at the same $(\mathbf{x}, t)$.

The fix — train directly on conditional labels. Replace the intractable target with the *per-sample* velocity:

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t,\,\mathbf{x}_0,\,\mathbf{x}_1}\left[\left\|\mathbf{u}_\theta(\mathbf{x}_t, t) - (\mathbf{x}_1 - \mathbf{x}_0)\right\|^2\right],$$

where $\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1$, with $\mathbf{x}_0 \sim p_0$, $\mathbf{x}_1 \sim p_{\text{data}}$, $t \sim \text{Uniform}[0,1]$. Every quantity is computable: sample a pair, form $\mathbf{x}_t$, compute $\mathbf{x}_1 - \mathbf{x}_0$, evaluate the network, backpropagate.

Step 4: Why We Can't Compute $\mathbf{u}_t$ Directly — and the Fix

Why $\mathbf{u}_t(\mathbf{x})$ is intractable. Evaluating

$$\mathbb{E}[\mathbf{x}_1 - \mathbf{x}_0 \mid (1-t)\mathbf{x}_0 + t\mathbf{x}_1 = \mathbf{x}]$$

requires integrating over all pairs $(\mathbf{x}_0, \mathbf{x}_1)$ consistent with $\mathbf{x}_t = \mathbf{x}$, weighted by their density — i.e., requires knowing $p_t(\mathbf{x})$. This same averaging is what makes the marginal field geometrically more complicated than the straight conditional labels: different endpoint pairs can contribute different directions at the same $(\mathbf{x}, t)$.

The fix — train directly on conditional labels. Replace the intractable target with the *per-sample* velocity:

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t,\,\mathbf{x}_0,\,\mathbf{x}_1}\left[\left\|\mathbf{u}_\theta(\mathbf{x}_t, t) - (\mathbf{x}_1 - \mathbf{x}_0)\right\|^2\right],$$

where $\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1$, with $\mathbf{x}_0 \sim p_0$, $\mathbf{x}_1 \sim p_{\text{data}}$, $t \sim \text{Uniform}[0,1]$. Every quantity is computable: sample a pair, form $\mathbf{x}_t$, compute $\mathbf{x}_1 - \mathbf{x}_0$, evaluate the network, backpropagate.

Training loop.

1. Sample $\mathbf{x}_0 \sim p_0$, $\mathbf{x}_1 \sim p_{\text{data}}$, $t \sim \text{Uniform}[0,1]$
2. Form $\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1$
3. Gradient step on $\|\mathbf{u}_\theta(\mathbf{x}_t, t) - (\mathbf{x}_1 - \mathbf{x}_0)\|^2$

Step 4: Why We Can't Compute $\mathbf{u}_t$ Directly — and the Fix

Why $\mathbf{u}_t(\mathbf{x})$ is intractable. Evaluating

$$\mathbb{E}[\mathbf{x}_1 - \mathbf{x}_0 \mid (1-t)\mathbf{x}_0 + t\mathbf{x}_1 = \mathbf{x}]$$

requires integrating over all pairs $(\mathbf{x}_0, \mathbf{x}_1)$ consistent with $\mathbf{x}_t = \mathbf{x}$, weighted by their density — i.e., requires knowing $p_t(\mathbf{x})$. This same averaging is what makes the marginal field geometrically more complicated than the straight conditional labels: different endpoint pairs can contribute different directions at the same $(\mathbf{x}, t)$.

The fix — train directly on conditional labels. Replace the intractable target with the *per-sample* velocity:

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t,\,\mathbf{x}_0,\,\mathbf{x}_1}\left[\left\|\mathbf{u}_\theta(\mathbf{x}_t, t) - (\mathbf{x}_1 - \mathbf{x}_0)\right\|^2\right],$$

where $\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1$, with $\mathbf{x}_0 \sim p_0$, $\mathbf{x}_1 \sim p_{\text{data}}$, $t \sim \text{Uniform}[0,1]$. Every quantity is computable: sample a pair, form $\mathbf{x}_t$, compute $\mathbf{x}_1 - \mathbf{x}_0$, evaluate the network, backpropagate.

Training loop.

1. Sample $\mathbf{x}_0 \sim p_0$, $\mathbf{x}_1 \sim p_{\text{data}}$, $t \sim \text{Uniform}[0,1]$
2. Form $\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1$
3. Gradient step on $\|\mathbf{u}_\theta(\mathbf{x}_t, t) - (\mathbf{x}_1 - \mathbf{x}_0)\|^2$

The CFM theorem (next section) proves $\mathcal{L}_{\text{CFM}}$ and the ideal loss $\mathcal{L}_{\text{FM}}$ share the same minimizer — so training on tractable labels converges to the right field.

From an Endpoint Pair to a Supervised Training Label

Let

$$\mathbf{x}_0 = (0, 0), \qquad \mathbf{x}_1 = (2, 1), \qquad t = 0.25.$$

$$\mathbf{x}_t = (1 - t)\mathbf{x}_0 + t\mathbf{x}_1 = (0.5, 0.25).$$

$$\mathbf{u}_t^{\text{cond}} = \mathbf{x}_1 - \mathbf{x}_0 = (2, 1).$$

From an Endpoint Pair to a Supervised Training Label

Let

$$\mathbf{x}_0 = (0, 0), \qquad \mathbf{x}_1 = (2, 1), \qquad t = 0.25.$$

For the straight conditional path,

$$\mathbf{x}_t = (1 - t)\mathbf{x}_0 + t\mathbf{x}_1 = (0.5, 0.25).$$

$$\mathbf{u}_t^{\text{cond}} = \mathbf{x}_1 - \mathbf{x}_0 = (2, 1).$$

From an Endpoint Pair to a Supervised Training Label

Let

$$\mathbf{x}_0 = (0, 0), \qquad \mathbf{x}_1 = (2, 1), \qquad t = 0.25.$$

For the straight conditional path,

$$\mathbf{x}_t = (1 - t)\mathbf{x}_0 + t\mathbf{x}_1 = (0.5, 0.25).$$

The target conditional velocity is

$$\mathbf{u}_t^{\text{cond}} = \mathbf{x}_1 - \mathbf{x}_0 = (2, 1).$$

From an Endpoint Pair to a Supervised Training Label

Let

$$\mathbf{x}_0 = (0, 0), \qquad \mathbf{x}_1 = (2, 1), \qquad t = 0.25.$$

For the straight conditional path,

$$\mathbf{x}_t = (1 - t)\mathbf{x}_0 + t\mathbf{x}_1 = (0.5, 0.25).$$

The target conditional velocity is

$$\mathbf{u}_t^{\text{cond}} = \mathbf{x}_1 - \mathbf{x}_0 = (2, 1).$$

So the network input is $(\mathbf{x}_t, t) = ((0.5, 0.25), 0.25)$, and the supervised target output is the velocity vector $(2, 1)$.

CFM Theorem and Sampling Implications

Setting Up the Comparison

Define two losses.

Setting Up the Comparison

Define two losses.

The ideal loss (intractable):

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}_{t,\,\mathbf{x}_t}\left[\|\mathbf{u}_\theta(\mathbf{x}_t, t) - \mathbf{u}_t(\mathbf{x}_t)\|^2\right].$$

Minimizing this drives $\mathbf{u}_\theta \to \mathbf{u}_t$ directly. But $\mathbf{u}_t(\mathbf{x}) = \mathbb{E}[\dot{\mathbf{x}}_t \mid \mathbf{x}_t = \mathbf{x}]$ is the marginal velocity field defined by the stochastic interpolant, which has no closed-form expression.

Setting Up the Comparison

Define two losses.

The ideal loss (intractable):

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}_{t,\,\mathbf{x}_t}\left[\|\mathbf{u}_\theta(\mathbf{x}_t, t) - \mathbf{u}_t(\mathbf{x}_t)\|^2\right].$$

Minimizing this drives $\mathbf{u}_\theta \to \mathbf{u}_t$ directly. But $\mathbf{u}_t(\mathbf{x}) = \mathbb{E}[\dot{\mathbf{x}}_t \mid \mathbf{x}_t = \mathbf{x}]$ is the marginal velocity field defined by the stochastic interpolant, which has no closed-form expression.

The tractable surrogate (what we actually train):

$$\mathcal{L}_{\mathrm{CFM}}(\theta) = \mathbb{E}_{t,\,\mathbf{x}_0,\,\mathbf{x}_1}\left[\|\mathbf{u}_\theta(\mathbf{x}_t, t) - \mathbf{u}_t^{\mathrm{cond}}(\mathbf{x}_t \mid \mathbf{x}_0, \mathbf{x}_1)\|^2\right].$$

Every quantity here is computable: we sample $(\mathbf{x}_0, \mathbf{x}_1, t)$ and evaluate the conditional path velocity.

Setting Up the Comparison

Define two losses.

The ideal loss (intractable):

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}_{t,\,\mathbf{x}_t}\left[\|\mathbf{u}_\theta(\mathbf{x}_t, t) - \mathbf{u}_t(\mathbf{x}_t)\|^2\right].$$

Minimizing this drives $\mathbf{u}_\theta \to \mathbf{u}_t$ directly. But $\mathbf{u}_t(\mathbf{x}) = \mathbb{E}[\dot{\mathbf{x}}_t \mid \mathbf{x}_t = \mathbf{x}]$ is the marginal velocity field defined by the stochastic interpolant, which has no closed-form expression.

The tractable surrogate (what we actually train):

$$\mathcal{L}_{\mathrm{CFM}}(\theta) = \mathbb{E}_{t,\,\mathbf{x}_0,\,\mathbf{x}_1}\left[\|\mathbf{u}_\theta(\mathbf{x}_t, t) - \mathbf{u}_t^{\mathrm{cond}}(\mathbf{x}_t \mid \mathbf{x}_0, \mathbf{x}_1)\|^2\right].$$

Every quantity here is computable: we sample $(\mathbf{x}_0, \mathbf{x}_1, t)$ and evaluate the conditional path velocity.

Question. Do $\mathcal{L}_{\mathrm{FM}}$ and $\mathcal{L}_{\mathrm{CFM}}$ have the same minimizer?
Stronger: do they have the same gradient in θ?

Least-Squares Minimizers Are Conditional Means

We already saw (Step 3) that $\mathbf{u}_t(\mathbf{x}) = \mathbb{E}[\dot{\mathbf{x}}_t \mid \mathbf{x}_t = \mathbf{x}]$ is the least-squares-optimal deterministic field consistent with the stochastic interpolant.

Least-Squares Minimizers Are Conditional Means

We already saw (Step 3) that $\mathbf{u}_t(\mathbf{x}) = \mathbb{E}[\dot{\mathbf{x}}_t \mid \mathbf{x}_t = \mathbf{x}]$ is the least-squares-optimal deterministic field consistent with the stochastic interpolant.

The CFM theorem makes this precise: among all θ-parameterized fields,

$$\arg\min_\theta \mathcal{L}_{\mathrm{CFM}}(\theta) \;=\; \arg\min_\theta \mathcal{L}_{\mathrm{FM}}(\theta).$$

Training on tractable conditional labels converges to the same minimizer as training on the ideal (intractable) marginal field.

The CFM Theorem

Assumptions. Let $\mathbf{x}_t(\mathbf{x}_0, \mathbf{x}_1)$ be any family of conditional paths that:

1. interpolates: $\mathbf{x}_0(\mathbf{x}_0, \mathbf{x}_1) = \mathbf{x}_0$ and $\mathbf{x}_1(\mathbf{x}_0, \mathbf{x}_1) = \mathbf{x}_1$,
2. is differentiable in t, and
3. induces the correct marginal: $\mathbf{x}_t \sim p_t$ when $(\mathbf{x}_0, \mathbf{x}_1) \sim p_0 \otimes p_{\text{data}}$.

The CFM Theorem

Assumptions. Let $\mathbf{x}_t(\mathbf{x}_0, \mathbf{x}_1)$ be any family of conditional paths that:

1. interpolates: $\mathbf{x}_0(\mathbf{x}_0, \mathbf{x}_1) = \mathbf{x}_0$ and $\mathbf{x}_1(\mathbf{x}_0, \mathbf{x}_1) = \mathbf{x}_1$,
2. is differentiable in t, and
3. induces the correct marginal: $\mathbf{x}_t \sim p_t$ when $(\mathbf{x}_0, \mathbf{x}_1) \sim p_0 \otimes p_{\text{data}}$.

Theorem (Lipman et al. 2022).

$$\nabla_\theta \mathcal{L}_{\text{CFM}}(\theta) \;=\; \nabla_\theta \mathcal{L}_{\text{FM}}(\theta) \quad \text{for all } \theta.$$

In particular, every local minimum of $\mathcal{L}_{\text{CFM}}$ is a local minimum of $\mathcal{L}_{\text{FM}}$.

The CFM Theorem

Assumptions. Let $\mathbf{x}_t(\mathbf{x}_0, \mathbf{x}_1)$ be any family of conditional paths that:

1. interpolates: $\mathbf{x}_0(\mathbf{x}_0, \mathbf{x}_1) = \mathbf{x}_0$ and $\mathbf{x}_1(\mathbf{x}_0, \mathbf{x}_1) = \mathbf{x}_1$,
2. is differentiable in t, and
3. induces the correct marginal: $\mathbf{x}_t \sim p_t$ when $(\mathbf{x}_0, \mathbf{x}_1) \sim p_0 \otimes p_{\text{data}}$.

Theorem (Lipman et al. 2022).

$$\nabla_\theta \mathcal{L}_{\text{CFM}}(\theta) \;=\; \nabla_\theta \mathcal{L}_{\text{FM}}(\theta) \quad \text{for all } \theta.$$

In particular, every local minimum of $\mathcal{L}_{\text{CFM}}$ is a local minimum of $\mathcal{L}_{\text{FM}}$.

Proof sketch. Expand $\mathcal{L}_{\text{FM}}$ by conditioning on $(\mathbf{x}_0, \mathbf{x}_1)$ via the tower property:

$$\mathcal{L}_{\text{FM}}(\theta) = \mathcal{L}_{\text{CFM}}(\theta) - \mathbb{E}_{t,\mathbf{x}_t}\Big[\|\mathbf{u}_t(\mathbf{x}_t)\|^2\Big] + C,$$

where C and the subtracted term are both independent of θ. Hence $\nabla_\theta \mathcal{L}_{\text{FM}} = \nabla_\theta \mathcal{L}_{\text{CFM}}$. □

Straight-Line Paths Satisfy the CFM Assumptions

We verify each assumption for the straight-line path $\mathbf{x}_t(\mathbf{x}_0, \mathbf{x}_1) = (1-t)\mathbf{x}_0 + t\mathbf{x}_1, \quad (\mathbf{x}_0, \mathbf{x}_1) \sim p_0 \otimes p_{\text{data}}$.

Straight-Line Paths Satisfy the CFM Assumptions

We verify each assumption for the straight-line path $\mathbf{x}_t(\mathbf{x}_0, \mathbf{x}_1) = (1-t)\mathbf{x}_0 + t\mathbf{x}_1, \quad (\mathbf{x}_0, \mathbf{x}_1) \sim p_0 \otimes p_{\text{data}}$.

Assumption 1 (interpolation). $\mathbf{x}_0(\mathbf{x}_0, \mathbf{x}_1) = \mathbf{x}_0$ ✓ and $\mathbf{x}_1(\mathbf{x}_0, \mathbf{x}_1) = \mathbf{x}_1$ ✓.

Straight-Line Paths Satisfy the CFM Assumptions

We verify each assumption for the straight-line path $\mathbf{x}_t(\mathbf{x}_0, \mathbf{x}_1) = (1-t)\mathbf{x}_0 + t\mathbf{x}_1, \quad (\mathbf{x}_0, \mathbf{x}_1) \sim p_0 \otimes p_{\text{data}}$.

Assumption 1 (interpolation). $\mathbf{x}_0(\mathbf{x}_0, \mathbf{x}_1) = \mathbf{x}_0$ ✓ and $\mathbf{x}_1(\mathbf{x}_0, \mathbf{x}_1) = \mathbf{x}_1$ ✓.

Assumption 2 (differentiability). $\dot{\mathbf{x}}_t = \mathbf{x}_1 - \mathbf{x}_0$, which is constant in t. Trivially smooth. ✓

Straight-Line Paths Satisfy the CFM Assumptions

We verify each assumption for the straight-line path $\mathbf{x}_t(\mathbf{x}_0, \mathbf{x}_1) = (1-t)\mathbf{x}_0 + t\mathbf{x}_1, \quad (\mathbf{x}_0, \mathbf{x}_1) \sim p_0 \otimes p_{\text{data}}$.

Assumption 1 (interpolation). $\mathbf{x}_0(\mathbf{x}_0, \mathbf{x}_1) = \mathbf{x}_0$ ✓ and $\mathbf{x}_1(\mathbf{x}_0, \mathbf{x}_1) = \mathbf{x}_1$ ✓.

Assumption 2 (differentiability). $\dot{\mathbf{x}}_t = \mathbf{x}_1 - \mathbf{x}_0$, which is constant in t. Trivially smooth. ✓

Assumption 3 (correct marginal). Define p_t to be the distribution of $(1-t)\mathbf{x}_0 + t\mathbf{x}_1$ when $\mathbf{x}_0 \sim p_0$ and $\mathbf{x}_1 \sim p_{\text{data}}$ are drawn independently. Concretely, p_t is the pushforward of $p_0 \otimes p_{\text{data}}$ under the linear map $(\mathbf{x}_0, \mathbf{x}_1) \mapsto (1-t)\mathbf{x}_0 + t\mathbf{x}_1$; it is a valid density for every $t \in [0,1]$. Then $\mathbf{x}_t \sim p_t$ holds *by definition* of p_t. The endpoint conditions in Assumption 1 guarantee $p_0 \to p_0$ and $p_1 \to p_{\text{data}}$. ✓

Straight-Line Paths Satisfy the CFM Assumptions

We verify each assumption for the straight-line path $\mathbf{x}_t(\mathbf{x}_0, \mathbf{x}_1) = (1-t)\mathbf{x}_0 + t\mathbf{x}_1, \quad (\mathbf{x}_0, \mathbf{x}_1) \sim p_0 \otimes p_{\text{data}}$.

Assumption 1 (interpolation). $\mathbf{x}_0(\mathbf{x}_0, \mathbf{x}_1) = \mathbf{x}_0$ ✓ and $\mathbf{x}_1(\mathbf{x}_0, \mathbf{x}_1) = \mathbf{x}_1$ ✓.

Assumption 2 (differentiability). $\dot{\mathbf{x}}_t = \mathbf{x}_1 - \mathbf{x}_0$, which is constant in t. Trivially smooth. ✓

Assumption 3 (correct marginal). Define p_t to be the distribution of $(1-t)\mathbf{x}_0 + t\mathbf{x}_1$ when $\mathbf{x}_0 \sim p_0$ and $\mathbf{x}_1 \sim p_{\text{data}}$ are drawn independently. Concretely, p_t is the pushforward of $p_0 \otimes p_{\text{data}}$ under the linear map $(\mathbf{x}_0, \mathbf{x}_1) \mapsto (1-t)\mathbf{x}_0 + t\mathbf{x}_1$; it is a valid density for every $t \in [0,1]$. Then $\mathbf{x}_t \sim p_t$ holds *by definition* of p_t. The endpoint conditions in Assumption 1 guarantee $p_0 \to p_0$ and $p_1 \to p_{\text{data}}$. ✓

All three assumptions hold. The CFM theorem therefore applies: training on straight-line conditional labels is gradient-equivalent to training directly on the marginal field.

Resolution: What the Learned Field Actually Is

The theorem closes the loop.

- Training on endpoint-conditioned labels is *not* a heuristic: it is gradient-equivalent to training directly on the marginal field.
- The network converges to $\mathbf{u}_t(\mathbf{x}) = \mathbb{E}[\mathbf{u}_t^{\text{cond}} \mid \mathbf{x}_t = \mathbf{x}]$ — the direction of net probability flow at $\mathbf{x}$ at time t.
- At sampling time there is no ambiguity: the learned field already averages over all conflicting paths in a least-squares-optimal way.

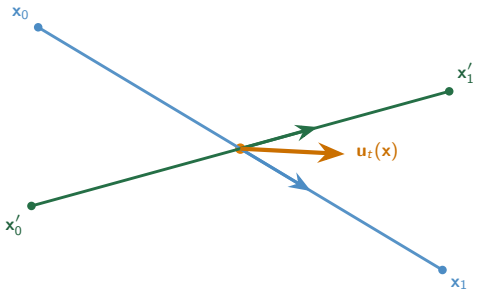
Resolution: What the Learned Field Actually Is

The theorem closes the loop.

- Training on endpoint-conditioned labels is *not* a heuristic: it is gradient-equivalent to training directly on the marginal field.
- The network converges to $\mathbf{u}_t(\mathbf{x}) = \mathbb{E}[\mathbf{u}_t^{\text{cond}} \mid \mathbf{x}_t = \mathbf{x}]$ — the direction of net probability flow at $\mathbf{x}$ at time t.
- At sampling time there is no ambiguity: the learned field already averages over all conflicting paths in a least-squares-optimal way.

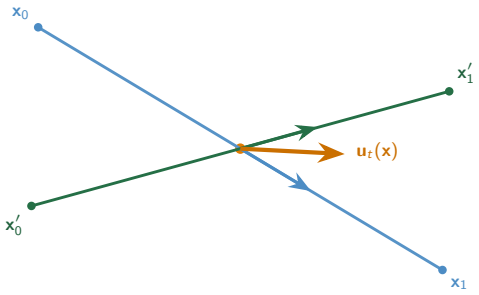
Why arbitrary paths matter. Any path family satisfying the three conditions gives the same theorem. Straight lines are the simplest choice; rectified flow is a particular straight-path-based variant rather than a synonym for vanilla CFM. Other choices — e.g. mini-batch optimal transport paths — can reduce path curvature, which can make ODE integration easier and may improve the speed-quality tradeoff empirically, but the theoretical guarantee is unchanged.

How Straight Conditional Paths Yield a Curved ODE



- Each endpoint pair $(\mathbf{x}_0, \mathbf{x}_1)$ defines a **straight conditional path** with constant velocity $\mathbf{x}_1 - \mathbf{x}_0$.
- At a fixed spacetime point $(\mathbf{x}, t)$, several such paths may pass through $\mathbf{x}$, but their directions need not agree.
- The marginal field $\mathbf{u}_t(\mathbf{x}) = \mathbb{E}[\mathbf{x}_1 - \mathbf{x}_0 \mid \mathbf{x}_t = \mathbf{x}]$ averages those directions with weights determined by how likely each pair is, given that it passes through $\mathbf{x}$ at time t.

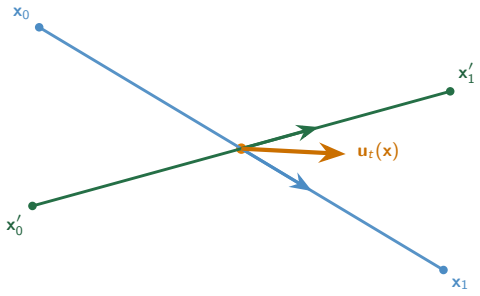
How Straight Conditional Paths Yield a Curved ODE



- Each endpoint pair $(\mathbf{x}_0, \mathbf{x}_1)$ defines a **straight conditional path** with constant velocity $\mathbf{x}_1 - \mathbf{x}_0$.
- At a fixed spacetime point $(\mathbf{x}, t)$, several such paths may pass through $\mathbf{x}$, but their directions need not agree.
- The marginal field $\mathbf{u}_t(\mathbf{x}) = \mathbb{E}[\mathbf{x}_1 - \mathbf{x}_0 \mid \mathbf{x}_t = \mathbf{x}]$ averages those directions with weights determined by how likely each pair is, given that it passes through $\mathbf{x}$ at time t.

Straight conditional labels $\neq$ *straight sampled trajectories*. Curvature appears after marginalization.

How Straight Conditional Paths Yield a Curved ODE



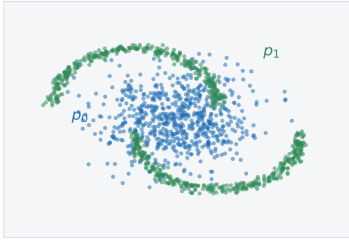
- Each endpoint pair $(\mathbf{x}_0, \mathbf{x}_1)$ defines a **straight conditional path** with constant velocity $\mathbf{x}_1 - \mathbf{x}_0$.
- At a fixed spacetime point $(\mathbf{x}, t)$, several such paths may pass through $\mathbf{x}$, but their directions need not agree.
- The marginal field $\mathbf{u}_t(\mathbf{x}) = \mathbb{E}[\mathbf{x}_1 - \mathbf{x}_0 \mid \mathbf{x}_t = \mathbf{x}]$ averages those directions with weights determined by how likely each pair is, given that it passes through $\mathbf{x}$ at time t.

Straight conditional labels $\neq$ *straight sampled trajectories*. Curvature appears after marginalization.

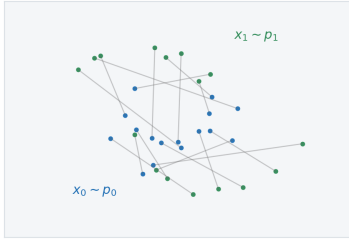
More curved integral curves are harder for ODE solvers to integrate accurately, typically requiring more function evaluations (more steps) to maintain sample quality.

Straight Conditional Paths vs. the Learned Marginal Field

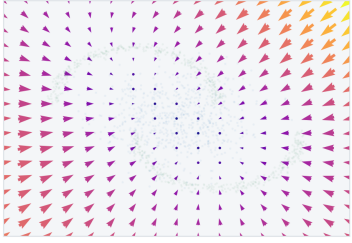
Distributions p_0 and p_1



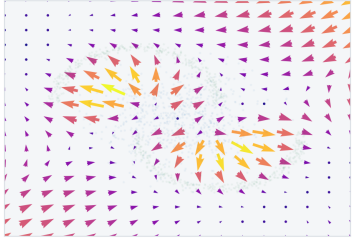
Independent coupling and straight conditional paths



Learned field u_t at $t = 0.10$

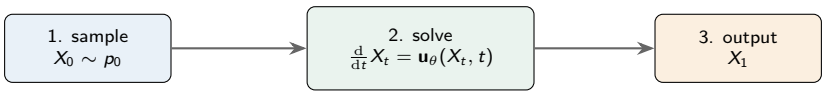


Learned field u_t at $t = 0.90$



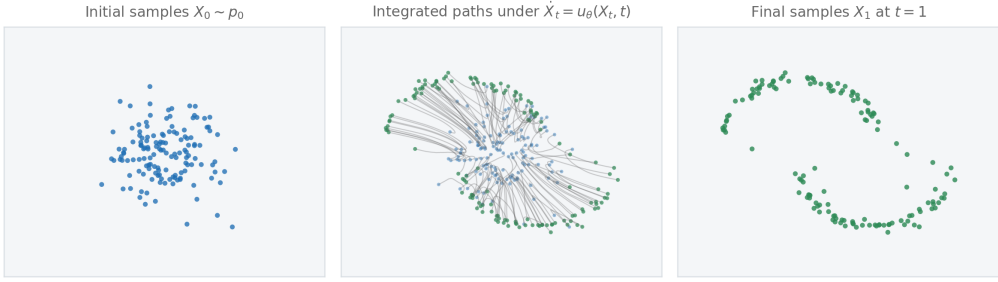
How Do We Generate a Sample?

Once training is done, we keep the learned velocity field $\mathbf{u}_\theta(\mathbf{x}, t)$ and use it directly.



The network provides the velocity at every point along the trajectory. Sampling is just following that learned direction field from noise to data.

Integrating the Learned Field on the Half-Moons Example



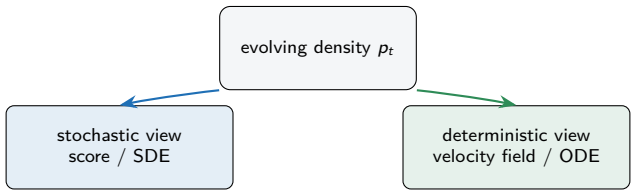
$$X_1 = X_0 + \int_0^1 \mathbf{u}_\theta(X_t, t)\,\mathrm{d}t.$$

In practice we solve this ODE numerically (e.g., Euler, Runge–Kutta, adaptive solvers) to move samples from the Gaussian base distribution to the learned data distribution.

Comparison and Summary

Flow Matching vs. Diffusion

Axis	DDPM	DDIM	Flow Matching
Direct training target	noise / score surrogate	same as DDPM	velocity field
Native sampler	reverse diffusion chain	deterministic denoising trajectory	ODE integration
Main caveat	many steps in basic form	still derived from denoiser training	path design and solver quality matter



Summary

Before: generation = reverse denoising **After:** generation = learned transport.

Three things that are new in this lecture:

1. **Stochastic interpolant.** Draw a coupling $(\mathbf{x}_0, \mathbf{x}_1)$ and connect them with a path. The path's marginals $\{p_t\}$ interpolate $p_0 \to p_{\text{data}}$ by construction.
2. **Regression principle.** The velocity field that transports those marginals is the conditional mean $\mathbf{u}_t(\mathbf{x}) = \mathbb{E}[\dot{\mathbf{x}}_t \mid \mathbf{x}_t = \mathbf{x}]$ — read off by least squares, not designed. Even straight conditional paths can average into a curved, time-dependent marginal field.
3. **CFM theorem + sampling.** We cannot compute $\mathbf{u}_t$ directly, but regressing on endpoint-conditioned labels $\mathbf{u}_t^{\text{cond}}$ is provably equivalent in optimization target; at inference we sample by numerically integrating $\dot{X}_t = \mathbf{u}_\theta(X_t, t)$ from Gaussian base points.

Appendix Overview

Appendix Overview

Appendix	What it proves
A: Continuity Equation	$\partial_t p_t + \nabla \cdot (p_t \mathbf{u}_t) = 0$ from mass conservation via the Jacobian of the flow map
B: Existence and Non-uniqueness of $\mathbf{u}_t$	reduces continuity equation to Poisson equation; constructs explicit $\mathbf{u}_t$; shows non-uniqueness of valid transport fields
C: CFM Theorem	$\nabla_\theta \mathcal{L}_{\mathrm{FM}} = \nabla_\theta \mathcal{L}_{\mathrm{CFM}}$ (tower property); why Gaussian base ensures regularity; why the learned ODE reaches p_{data} (DiPerna–Lions)
D: Guidance	classifier guidance and CFG in velocity space; guidance scale decoupled from training; principled alternatives

Appendix A: Continuity Equation

A1: Start from Mass Conservation

Let $S \subset \mathbb{R}^d$ be any measurable region, and suppose particles evolve by

$$\mathbf{x}_{t+h} = \mathbf{x}_t + h\,\mathbf{u}_t(\mathbf{x}_t) + o(h).$$

Define the first-order flow map $\Phi_h(\mathbf{x}) = \mathbf{x} + h\,\mathbf{u}_t(\mathbf{x})$, with Jacobian

$$J_{\Phi_h}(\mathbf{x}) = I + h\,J_{\mathbf{u}_t}(\mathbf{x}).$$

$$\int_{\Phi_h(S)} p_{t+h}(\mathbf{y})\,\mathrm{d}\mathbf{y} = \int_S p_t(\mathbf{x})\,\mathrm{d}\mathbf{x}.$$

$$\int_S p_{t+h}(\Phi_h(\mathbf{x}))\,|\det J_{\Phi_h}(\mathbf{x})|\,\mathrm{d}\mathbf{x} = \int_S p_t(\mathbf{x})\,\mathrm{d}\mathbf{x}.$$

A1: Start from Mass Conservation

Let $S \subset \mathbb{R}^d$ be any measurable region, and suppose particles evolve by

$$\mathbf{x}_{t+h} = \mathbf{x}_t + h\,\mathbf{u}_t(\mathbf{x}_t) + o(h).$$

Define the first-order flow map $\Phi_h(\mathbf{x}) = \mathbf{x} + h\,\mathbf{u}_t(\mathbf{x})$, with Jacobian

$$J_{\Phi_h}(\mathbf{x}) = I + h\,J_{\mathbf{u}_t}(\mathbf{x}).$$

If probability mass is conserved, then every moving region must satisfy

$$\int_{\Phi_h(S)} p_{t+h}(\mathbf{y})\,\mathrm{d}\mathbf{y} = \int_S p_t(\mathbf{x})\,\mathrm{d}\mathbf{x}.$$

$$\int_S p_{t+h}(\Phi_h(\mathbf{x}))\,|\det J_{\Phi_h}(\mathbf{x})|\,\mathrm{d}\mathbf{x} = \int_S p_t(\mathbf{x})\,\mathrm{d}\mathbf{x}.$$

A1: Start from Mass Conservation

Let $S \subset \mathbb{R}^d$ be any measurable region, and suppose particles evolve by

$$\mathbf{x}_{t+h} = \mathbf{x}_t + h\,\mathbf{u}_t(\mathbf{x}_t) + o(h).$$

Define the first-order flow map $\Phi_h(\mathbf{x}) = \mathbf{x} + h\,\mathbf{u}_t(\mathbf{x})$, with Jacobian

$$J_{\Phi_h}(\mathbf{x}) = I + h\,J_{\mathbf{u}_t}(\mathbf{x}).$$

If probability mass is conserved, then every moving region must satisfy

$$\int_{\Phi_h(S)} p_{t+h}(\mathbf{y})\,\mathrm{d}\mathbf{y} = \int_S p_t(\mathbf{x})\,\mathrm{d}\mathbf{x}.$$

Change variables $\mathbf{y} = \Phi_h(\mathbf{x})$, $\mathrm{d}\mathbf{y} = |\det J_{\Phi_h}(\mathbf{x})|\,\mathrm{d}\mathbf{x}$:

$$\int_S p_{t+h}(\Phi_h(\mathbf{x}))\,|\det J_{\Phi_h}(\mathbf{x})|\,\mathrm{d}\mathbf{x} = \int_S p_t(\mathbf{x})\,\mathrm{d}\mathbf{x}.$$

A2: First-Order Expansion Gives the PDE

Expand each factor to first order in h.

For the density, Taylor-expand in t and in $\mathbf{x}$:

$$p_{t+h}(\Phi_h(\mathbf{x})) \approx p_t(\mathbf{x}) + h\,\partial_t p_t(\mathbf{x}) + h\,\nabla p_t(\mathbf{x}) \cdot \mathbf{u}_t(\mathbf{x}).$$

For the Jacobian determinant, use $\det(I + h\,A) \approx 1 + h^\top(A)$ for small h:

$$|\det J_{\Phi_h}(\mathbf{x})| = \det(I + h\,J_{\mathbf{u}_t}(\mathbf{x})) \approx 1 + h^\top(J_{\mathbf{u}_t}(\mathbf{x})) = 1 + h\,\nabla \cdot \mathbf{u}_t(\mathbf{x}).$$

$$p_t(\mathbf{x}) + h\Big[\partial_t p_t(\mathbf{x}) + \nabla p_t(\mathbf{x}) \cdot \mathbf{u}_t(\mathbf{x}) + p_t(\mathbf{x})\,\nabla \cdot \mathbf{u}_t(\mathbf{x})\Big].$$

$$\partial_t p_t(\mathbf{x}) + \nabla \cdot (p_t(\mathbf{x})\mathbf{u}_t(\mathbf{x})) = 0.$$

This is the density-level consistency law for deterministic transport.

A2: First-Order Expansion Gives the PDE

Expand each factor to first order in h.

For the density, Taylor-expand in t and in $\mathbf{x}$:

$$p_{t+h}(\Phi_h(\mathbf{x})) \approx p_t(\mathbf{x}) + h\,\partial_t p_t(\mathbf{x}) + h\,\nabla p_t(\mathbf{x}) \cdot \mathbf{u}_t(\mathbf{x}).$$

For the Jacobian determinant, use $\det(I + hA) \approx 1 + h^\top(A)$ for small h:

$$|\det J_{\Phi_h}(\mathbf{x})| = \det(I + h\,J_{\mathbf{u}_t}(\mathbf{x})) \approx 1 + h^\top(J_{\mathbf{u}_t}(\mathbf{x})) = 1 + h\,\nabla \cdot \mathbf{u}_t(\mathbf{x}).$$

Multiplying and keeping only first-order terms, the integrand on the left becomes

$$p_t(\mathbf{x}) + h\Big[\partial_t p_t(\mathbf{x}) + \nabla p_t(\mathbf{x}) \cdot \mathbf{u}_t(\mathbf{x}) + p_t(\mathbf{x})\,\nabla \cdot \mathbf{u}_t(\mathbf{x})\Big].$$

$$\partial_t p_t(\mathbf{x}) + \nabla \cdot (p_t(\mathbf{x})\mathbf{u}_t(\mathbf{x})) = 0.$$

This is the density-level consistency law for deterministic transport.

A2: First-Order Expansion Gives the PDE

Expand each factor to first order in h.

For the density, Taylor-expand in t and in $\mathbf{x}$:

$$p_{t+h}(\Phi_h(\mathbf{x})) \approx p_t(\mathbf{x}) + h\,\partial_t p_t(\mathbf{x}) + h\,\nabla p_t(\mathbf{x}) \cdot \mathbf{u}_t(\mathbf{x}).$$

For the Jacobian determinant, use $\det(I + h\,A) \approx 1 + h^{\top}(A)$ for small h:

$$|\det J_{\Phi_h}(\mathbf{x})| = \det(I + h\,J_{\mathbf{u}_t}(\mathbf{x})) \approx 1 + h^{\top}(J_{\mathbf{u}_t}(\mathbf{x})) = 1 + h\,\nabla \cdot \mathbf{u}_t(\mathbf{x}).$$

Multiplying and keeping only first-order terms, the integrand on the left becomes

$$p_t(\mathbf{x}) + h\Big[\partial_t p_t(\mathbf{x}) + \nabla p_t(\mathbf{x}) \cdot \mathbf{u}_t(\mathbf{x}) + p_t(\mathbf{x})\,\nabla \cdot \mathbf{u}_t(\mathbf{x})\Big].$$

Setting equal to $p_t(\mathbf{x})$ and recognising the product rule $\nabla \cdot (p_t \mathbf{u}_t) = \nabla p_t \cdot \mathbf{u}_t + p_t\,\nabla \cdot \mathbf{u}_t$, the $O(h)$ coefficient must vanish for every S:

$$\partial_t p_t(\mathbf{x}) + \nabla \cdot (p_t(\mathbf{x})\mathbf{u}_t(\mathbf{x})) = 0.$$

This is the density-level consistency law for deterministic transport.

Appendix B: Existence and Non-uniqueness of Velocity Fields

B1: Setting Up the Problem

We want to find *some* velocity field $\mathbf{u}_t$ that transports p_0 to p_{data}. By the continuity equation (Appendix A), $\mathbf{u}_t$ must satisfy

$$\partial_t p_t + \nabla \cdot (p_t \, \mathbf{u}_t) = 0.$$

B1: Setting Up the Problem

We want to find *some* velocity field $\mathbf{u}_t$ that transports p_0 to p_{data}. By the continuity equation (Appendix A), $\mathbf{u}_t$ must satisfy

$$\partial_t p_t + \nabla \cdot (p_t \, \mathbf{u}_t) = 0.$$

Choose a simple interpolating density. Define

$$p_t(\mathbf{x}) = (1 - t)\, p_0(\mathbf{x}) + t\, p_1(\mathbf{x}), \qquad t \in [0, 1],$$

where $p_1 = p_{\text{data}}$. This is a valid probability density for every t (a convex combination of two densities).

B1: Setting Up the Problem

We want to find *some* velocity field $\mathbf{u}_t$ that transports p_0 to p_{data}. By the continuity equation (Appendix A), $\mathbf{u}_t$ must satisfy

$$\partial_t p_t + \nabla \cdot (p_t \, \mathbf{u}_t) = 0.$$

Choose a simple interpolating density. Define

$$p_t(\mathbf{x}) = (1 - t)\, p_0(\mathbf{x}) + t\, p_1(\mathbf{x}), \qquad t \in [0, 1],$$

where $p_1 = p_{\text{data}}$. This is a valid probability density for every t (a convex combination of two densities). Differentiating with respect to t:

$$\partial_t p_t = p_1 - p_0.$$

The continuity equation becomes

$$\nabla \cdot (p_t \, \mathbf{u}_t) = p_0 - p_1.$$

This is a PDE for $\mathbf{u}_t$ at each fixed t.

B2: Reduction to a Poisson Equation

Introduce the **flux field** $\mathbf{v}_t = p_t\,\mathbf{u}_t$. The equation to solve is

$$\nabla \cdot \mathbf{v}_t = p_0 - p_1.$$

B2: Reduction to a Poisson Equation

Introduce the **flux field** $\mathbf{v}_t = p_t\,\mathbf{u}_t$. The equation to solve is

$$\nabla \cdot \mathbf{v}_t = p_0 - p_1.$$

Ansatz: try $\mathbf{v}_t = \nabla\phi_t$ for a scalar potential ϕ_t. Then

$$\nabla \cdot (\nabla\phi_t) = \Delta\phi_t = p_0 - p_1.$$

This is a **Poisson equation** — the same PDE you have seen in electrostatics.

B2: Reduction to a Poisson Equation

Introduce the **flux field** $\mathbf{v}_t = p_t\,\mathbf{u}_t$. The equation to solve is

$$\nabla \cdot \mathbf{v}_t = p_0 - p_1.$$

Ansatz: try $\mathbf{v}_t = \nabla\phi_t$ for a scalar potential ϕ_t. Then

$$\nabla \cdot (\nabla\phi_t) = \Delta\phi_t = p_0 - p_1.$$

This is a **Poisson equation** — the same PDE you have seen in electrostatics.

Is it solvable? A necessary condition (the compatibility condition) is that the right-hand side integrates to zero:

$$\int_{\mathbb{R}^d} (p_0 - p_1)\,\mathrm{d}\mathbf{x} = 1 - 1 = 0.\checkmark$$

So $\Delta\phi_t = p_0 - p_1$ has a solution ϕ_t.

B2: Reduction to a Poisson Equation

Introduce the **flux field** $\mathbf{v}_t = p_t\,\mathbf{u}_t$. The equation to solve is

$$\nabla \cdot \mathbf{v}_t = p_0 - p_1.$$

Ansatz: try $\mathbf{v}_t = \nabla \phi_t$ for a scalar potential ϕ_t. Then

$$\nabla \cdot (\nabla \phi_t) = \Delta \phi_t = p_0 - p_1.$$

This is a **Poisson equation** — the same PDE you have seen in electrostatics.

Is it solvable? A necessary condition (the compatibility condition) is that the right-hand side integrates to zero:

$$\int_{\mathbb{R}^d} (p_0 - p_1)\,\mathrm{d}\mathbf{x} = 1 - 1 = 0. \checkmark$$

So $\Delta \phi_t = p_0 - p_1$ has a solution ϕ_t.

Recover the velocity field (wherever $p_t > 0$):

$$\mathbf{u}_t(\mathbf{x}) = \frac{\nabla \phi_t(\mathbf{x})}{p_t(\mathbf{x})}.$$

A valid velocity field exists.

B3: Non-Uniqueness and Its Consequence

The solution ϕ_t to $\Delta\phi_t = p_0 - p_1$ is not unique: adding any harmonic function (satisfying $\Delta\psi = 0$) gives another solution.

B3: Non-Uniqueness and Its Consequence

The solution ϕ_t to $\Delta\phi_t = p_0 - p_1$ is not unique: adding any harmonic function (satisfying $\Delta\psi = 0$) gives another solution.

More generally, if $\mathbf{v}_t = \nabla\phi_t$ solves $\nabla \cdot \mathbf{v}_t = p_0 - p_1$, then so does

$$\mathbf{v}_t' = \nabla\phi_t + \mathbf{w}_t$$

for *any* divergence-free field $\mathbf{w}_t$ (i.e., $\nabla \cdot \mathbf{w}_t = 0$). In $\mathbb{R}^3$, curl fields $\mathbf{w}_t = \nabla \times \mathbf{A}$ are one family of examples.

B3: Non-Uniqueness and Its Consequence

The solution ϕ_t to $\Delta\phi_t = p_0 - p_1$ is not unique: adding any harmonic function (satisfying $\Delta\psi = 0$) gives another solution.

More generally, if $\mathbf{v}_t = \nabla\phi_t$ solves $\nabla \cdot \mathbf{v}_t = p_0 - p_1$, then so does

$$\mathbf{v}_t' = \nabla\phi_t + \mathbf{w}_t$$

for *any* divergence-free field $\mathbf{w}_t$ (i.e., $\nabla \cdot \mathbf{w}_t = 0$). In $\mathbb{R}^3$, curl fields $\mathbf{w}_t = \nabla \times \mathbf{A}$ are one family of examples.

Conclusion. Infinitely many velocity fields transport p_0 to p_{data} along the linear interpolant. There is no single canonical choice.

This non-uniqueness is precisely why a naive regression target $\dot{\mathbf{x}}_t$ is ill-posed: many valid fields agree on p_t but disagree pointwise. Flow matching resolves this by *conditioning on endpoint pairs*, selecting a tractable representative.

Appendix C: CFM Theorem Proof and Regularity

C1: Regularity Assumptions and a Practical Caveat

The proof requires that for a.e. $t \in [0, 1]$:

- p_t is a well-defined probability density (absolutely continuous w.r.t. Lebesgue measure),
- $\mathbf{u}_t$ is square-integrable under p_t: $\mathbb{E}_{\mathbf{x} \sim p_t}[\|\mathbf{u}_t(\mathbf{x})\|^2] < \infty$,
- $\mathbf{u}_t^{\text{cond}}$ is square-integrable under the joint distribution of $(\mathbf{x}_t, \mathbf{x}_0, \mathbf{x}_1)$.

C1: Regularity Assumptions and a Practical Caveat

The proof requires that for a.e. $t \in [0,1]$:

- p_t is a well-defined probability density (absolutely continuous w.r.t. Lebesgue measure),
- $\mathbf{u}_t$ is square-integrable under p_t: $\mathbb{E}_{\mathbf{x}\sim p_t}[\|\mathbf{u}_t(\mathbf{x})\|^2] < \infty$,
- $\mathbf{u}_t^{\text{cond}}$ is square-integrable under the joint distribution of $(\mathbf{x}_t, \mathbf{x}_0, \mathbf{x}_1)$.

Do these hold for image data? Image distributions are believed to concentrate near a low-dimensional manifold, so p_{data} may have zero density off that manifold — violating absolute continuity.

Rectified flow saves us. For $t \in (0,1)$, the interpolant $\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1$ has a Gaussian component $(1-t)\mathbf{x}_0$ with $\mathbf{x}_0 \sim \mathcal{N}(0, I)$. This spreads mass over all of $\mathbb{R}^d$, making p_t *absolutely continuous with full support* for every $t \in (0,1)$, regardless of how degenerate p_{data} is.

C1: Regularity Assumptions and a Practical Caveat

The proof requires that for a.e. $t \in [0,1]$:

- p_t is a well-defined probability density (absolutely continuous w.r.t. Lebesgue measure),
- $\mathbf{u}_t$ is square-integrable under p_t: $\mathbb{E}_{\mathbf{x}\sim p_t}[\|\mathbf{u}_t(\mathbf{x})\|^2] < \infty$,
- $\mathbf{u}_t^{\text{cond}}$ is square-integrable under the joint distribution of $(\mathbf{x}_t, \mathbf{x}_0, \mathbf{x}_1)$.

Do these hold for image data? Image distributions are believed to concentrate near a low-dimensional manifold, so p_{data} may have zero density off that manifold — violating absolute continuity.

Rectified flow saves us. For $t \in (0,1)$, the interpolant $\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1$ has a Gaussian component $(1-t)\mathbf{x}_0$ with $\mathbf{x}_0 \sim \mathcal{N}(0, I)$. This spreads mass over all of $\mathbb{R}^d$, making p_t *absolutely continuous with full support* for every $t \in (0,1)$, regardless of how degenerate p_{data} is.

The Gaussian base distribution is not merely a convenience: it is what makes the regularity assumptions of the CFM theorem actually hold in practice, even for manifold-supported data.

C2: The Stochastic Process Satisfies the Continuity Equation

Claim. Let $p_t = \mathrm{Law}(\mathbf{x}_t)$ be the marginals of the stochastic interpolant, and define $\mathbf{u}_t(\mathbf{x}) = \mathbb{E}[\dot{\mathbf{x}}_t \mid \mathbf{x}_t = \mathbf{x}]$. Then

$$\partial_t p_t + \nabla \cdot (p_t \mathbf{u}_t) = 0.$$

Proof. For any smooth test function $\phi : \mathbb{R}^d \to \mathbb{R}$, differentiate $\mathbb{E}[\phi(\mathbf{x}_t)]$ in t:

$$\frac{d}{dt}\mathbb{E}[\phi(\mathbf{x}_t)] = \mathbb{E}[\nabla\phi(\mathbf{x}_t) \cdot \dot{\mathbf{x}}_t]\,.$$

Apply the tower property, conditioning on $\mathbf{x}_t$:

$$\mathbb{E}[\nabla\phi(\mathbf{x}_t) \cdot \dot{\mathbf{x}}_t] = \mathbb{E}[\nabla\phi(\mathbf{x}_t) \cdot \mathbb{E}[\dot{\mathbf{x}}_t \mid \mathbf{x}_t]] = \mathbb{E}[\nabla\phi(\mathbf{x}_t) \cdot \mathbf{u}_t(\mathbf{x}_t)]\,.$$

Writing both sides in terms of p_t and integrating by parts:

$$\int \phi\, \partial_t p_t \,\mathrm{d}\mathbf{x} = \int \nabla\phi \cdot \mathbf{u}_t\, p_t \,\mathrm{d}\mathbf{x} = -\int \phi\, \nabla \cdot (p_t \mathbf{u}_t)\,\mathrm{d}\mathbf{x}.$$

Since ϕ is arbitrary, $\partial_t p_t + \nabla \cdot (p_t \mathbf{u}_t) = 0$. □

What this says. The stochastic process and its conditional-mean velocity field are consistent: $\mathbf{u}_t$ transports mass at the correct rate to reproduce the marginal path $\{p_t\}$. **What this does not say.** Whether the *deterministic ODE* $\dot{\mathbf{x}} = \mathbf{u}_t(\mathbf{x})$ produces those same marginals — that requires a separate argument (see C3).

C3: The ODE Produces the Same Marginals

Step 1: The ODE also satisfies the continuity equation. Let q_t be the marginal path induced by the ODE $\dot{\mathbf{x}} = \mathbf{u}_t(\mathbf{x})$, $\mathbf{x}_0 \sim p_0$. By the classical push-forward identity (Liouville's theorem / method of characteristics), any smooth ODE $\dot{\mathbf{x}} = v_t(\mathbf{x})$ satisfies

$$\partial_t q_t + \nabla \cdot (q_t \, \mathbf{u}_t) = 0.$$

So q_t satisfies the *same* continuity equation as p_t, with the *same* $\mathbf{u}_t$.

Step 2: The initial conditions agree. Both the stochastic process and the ODE start from $p_0 = \mathcal{N}(0, I)$, so $q_0 = p_0$.

Step 3: Uniqueness forces $q_t = p_t$.

Theorem (DiPerna–Lions 1989; Ambrosio 2004)

If $\mathbf{u}_t \in L^\infty([0,1] \times \mathbb{R}^d)$ with $\nabla_{\mathbf{x}} \mathbf{u}_t \in L^1([0,1]; L^1(\mathbb{R}^d))$, then the continuity equation has at most one weakly continuous solution μ_t for a given initial condition μ_0.

For the Gaussian-regularized interpolant, $\mathbf{u}_t$ satisfies these conditions for $t \in (0,1)$ (the Gaussian component of p_t gives full-support density and finite moments). Hence $q_t = p_t$ for all t.

The ODE $\dot{\mathbf{x}} = \mathbf{u}_t(\mathbf{x})$, $\mathbf{x}_0 \sim p_0$, transports p_0 to $p_1 = p_{\text{data}}$.
Training $\mathbf{u}_\theta \approx \mathbf{u}_t$ is enough: the learned ODE will do the same.

C4: Write Out Both Objectives

Recall the two losses:

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}_{t,\,\mathbf{x}_t \sim p_t}\left[\left\|\mathbf{u}_\theta(\mathbf{x}_t, t) - \mathbf{u}_t(\mathbf{x}_t)\right\|^2\right],$$

$$\mathcal{L}_{\mathrm{CFM}}(\theta) = \mathbb{E}_{t,\,\mathbf{x}_0,\mathbf{x}_1}\left[\left\|\mathbf{u}_\theta(\mathbf{x}_t, t) - \mathbf{u}_t^{\mathrm{cond}}(\mathbf{x}_t \mid \mathbf{x}_0, \mathbf{x}_1)\right\|^2\right],$$

where in $\mathcal{L}_{\mathrm{CFM}}$ the intermediate point is $\mathbf{x}_t = \mathbf{x}_t(\mathbf{x}_0, \mathbf{x}_1)$.

C4: Write Out Both Objectives

Recall the two losses:

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}_{t,\,\mathbf{x}_t \sim p_t}\left[\left\|\mathbf{u}_\theta(\mathbf{x}_t, t) - \mathbf{u}_t(\mathbf{x}_t)\right\|^2\right],$$

$$\mathcal{L}_{\mathrm{CFM}}(\theta) = \mathbb{E}_{t,\,\mathbf{x}_0,\mathbf{x}_1}\left[\left\|\mathbf{u}_\theta(\mathbf{x}_t, t) - \mathbf{u}_t^{\mathrm{cond}}(\mathbf{x}_t \mid \mathbf{x}_0, \mathbf{x}_1)\right\|^2\right],$$

where in $\mathcal{L}_{\mathrm{CFM}}$ the intermediate point is $\mathbf{x}_t = \mathbf{x}_t(\mathbf{x}_0, \mathbf{x}_1)$.

Strategy. Expand each squared norm and show the two objectives differ only by terms that do not depend on θ. Expand $\|\boldsymbol{a} - \boldsymbol{b}\|^2 = \|\boldsymbol{a}\|^2 - 2\boldsymbol{a}\cdot\boldsymbol{b} + \|\boldsymbol{b}\|^2$ with $\boldsymbol{a} = \mathbf{u}_\theta(\mathbf{x}_t, t)$:

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}\left[\left\|\mathbf{u}_\theta\right\|^2\right] - 2\,\mathbb{E}[\mathbf{u}_\theta(\mathbf{x}_t, t)\cdot\mathbf{u}_t(\mathbf{x}_t)] + \underbrace{\mathbb{E}\left[\left\|\mathbf{u}_t(\mathbf{x}_t)\right\|^2\right]}_{=:\,C_1,\ \theta\text{-free}},$$

$$\mathcal{L}_{\mathrm{CFM}}(\theta) = \mathbb{E}\left[\left\|\mathbf{u}_\theta\right\|^2\right] - 2\,\mathbb{E}\left[\mathbf{u}_\theta(\mathbf{x}_t, t)\cdot\mathbf{u}_t^{\mathrm{cond}}\right] + \underbrace{\mathbb{E}\left[\left\|\mathbf{u}_t^{\mathrm{cond}}\right\|^2\right]}_{=:\,C_2,\ \theta\text{-free}}.$$

C5: The Tower Property Equates the Cross Terms

It remains to show the two cross terms are equal:

$$\mathbb{E}[\mathbf{u}_\theta(\mathbf{x}_t, t) \cdot \mathbf{u}_t(\mathbf{x}_t)] \;=\; \mathbb{E}\Big[\mathbf{u}_\theta(\mathbf{x}_t, t) \cdot \mathbf{u}_t^{\text{cond}}(\mathbf{x}_t \mid \mathbf{x}_0, \mathbf{x}_1)\Big] \,.$$

C5: The Tower Property Equates the Cross Terms

It remains to show the two cross terms are equal:

$$\mathbb{E}[\mathbf{u}_\theta(\mathbf{x}_t, t) \cdot \mathbf{u}_t(\mathbf{x}_t)] \;=\; \mathbb{E}\Big[\mathbf{u}_\theta(\mathbf{x}_t, t) \cdot \mathbf{u}_t^{\text{cond}}(\mathbf{x}_t \mid \mathbf{x}_0, \mathbf{x}_1)\Big] \,.$$

Take the right-hand side and apply the tower property, conditioning on $\mathbf{x}_t$:

$$\mathbb{E}\Big[\mathbf{u}_\theta(\mathbf{x}_t, t) \cdot \mathbf{u}_t^{\text{cond}}\Big] = \mathbb{E}_{\mathbf{x}_t}\Big[\mathbf{u}_\theta(\mathbf{x}_t, t) \cdot \mathbb{E}\Big[\mathbf{u}_t^{\text{cond}}(\mathbf{x}_t \mid \mathbf{x}_0, \mathbf{x}_1) \;\Big|\; \mathbf{x}_t\Big]\Big] \,.$$

The key step: by *definition* of the marginal velocity field,

$$\mathbb{E}\Big[\mathbf{u}_t^{\text{cond}}(\mathbf{x}_t \mid \mathbf{x}_0, \mathbf{x}_1) \;\Big|\; \mathbf{x}_t = \mathbf{x}\Big] = \mathbf{u}_t(\mathbf{x}).$$

(This is exactly the least-squares conditional mean from the previous section.)

C5: The Tower Property Equates the Cross Terms

It remains to show the two cross terms are equal:

$$\mathbb{E}[\mathbf{u}_\theta(\mathbf{x}_t, t) \cdot \mathbf{u}_t(\mathbf{x}_t)] \;=\; \mathbb{E}\Big[\mathbf{u}_\theta(\mathbf{x}_t, t) \cdot \mathbf{u}_t^{\text{cond}}(\mathbf{x}_t \mid \mathbf{x}_0, \mathbf{x}_1)\Big] \,.$$

Take the right-hand side and apply the tower property, conditioning on $\mathbf{x}_t$:

$$\mathbb{E}\Big[\mathbf{u}_\theta(\mathbf{x}_t, t) \cdot \mathbf{u}_t^{\text{cond}}\Big] = \mathbb{E}_{\mathbf{x}_t}\Big[\mathbf{u}_\theta(\mathbf{x}_t, t) \cdot \mathbb{E}\Big[\mathbf{u}_t^{\text{cond}}(\mathbf{x}_t \mid \mathbf{x}_0, \mathbf{x}_1) \;\Big|\; \mathbf{x}_t\Big]\Big] \,.$$

The key step: by *definition* of the marginal velocity field,

$$\mathbb{E}\Big[\mathbf{u}_t^{\text{cond}}(\mathbf{x}_t \mid \mathbf{x}_0, \mathbf{x}_1) \;\Big|\; \mathbf{x}_t = \mathbf{x}\Big] = \mathbf{u}_t(\mathbf{x}).$$

(This is exactly the least-squares conditional mean from the previous section.)
Substituting:

$$\mathbb{E}\Big[\mathbf{u}_\theta(\mathbf{x}_t, t) \cdot \mathbf{u}_t^{\text{cond}}\Big] = \mathbb{E}_{\mathbf{x}_t}[\mathbf{u}_\theta(\mathbf{x}_t, t) \cdot \mathbf{u}_t(\mathbf{x}_t)] = \mathbb{E}[\mathbf{u}_\theta(\mathbf{x}_t, t) \cdot \mathbf{u}_t(\mathbf{x}_t)] \,. \checkmark$$

C6: Conclude

From C4 and C5:

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \mathcal{L}_{\mathrm{CFM}}(\theta) + C_1 - C_2,$$

where $C_1 = \mathbb{E}[\|\mathbf{u}_t(\mathbf{x}_t)\|^2]$ and $C_2 = \mathbb{E}[\|\mathbf{u}_t^{\mathrm{cond}}\|^2]$ are both independent of θ.

C6: Conclude

From C4 and C5:

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \mathcal{L}_{\mathrm{CFM}}(\theta) + C_1 - C_2,$$

where $C_1 = \mathbb{E}[\|\mathbf{u}_t(\mathbf{x}_t)\|^2]$ and $C_2 = \mathbb{E}[\|\mathbf{u}_t^{\mathrm{cond}}\|^2]$ are both independent of θ.

Differentiating:

$$\nabla_\theta \mathcal{L}_{\mathrm{FM}}(\theta) = \nabla_\theta \mathcal{L}_{\mathrm{CFM}}(\theta) \quad \text{for all } \theta.$$

Every critical point (and in particular every local minimum) of $\mathcal{L}_{\mathrm{CFM}}$ is a critical point of $\mathcal{L}_{\mathrm{FM}}$, and vice versa. □

C6: Conclude

From C4 and C5:

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \mathcal{L}_{\mathrm{CFM}}(\theta) + C_1 - C_2,$$

where $C_1 = \mathbb{E}[\|\mathbf{u}_t(\mathbf{x}_t)\|^2]$ and $C_2 = \mathbb{E}[\|\mathbf{u}_t^{\mathrm{cond}}\|^2]$ are both independent of θ.

Differentiating:

$$\nabla_\theta \mathcal{L}_{\mathrm{FM}}(\theta) = \nabla_\theta \mathcal{L}_{\mathrm{CFM}}(\theta) \quad \text{for all } \theta.$$

Every critical point (and in particular every local minimum) of $\mathcal{L}_{\mathrm{CFM}}$ is a critical point of $\mathcal{L}_{\mathrm{FM}}$, and vice versa. □

What the proof actually uses.

- The tower property (law of total expectation) — no special structure of the paths.
- The definition $\mathbf{u}_t(\mathbf{x}) = \mathbb{E}[\mathbf{u}_t^{\mathrm{cond}} \mid \mathbf{x}_t = \mathbf{x}]$ — this is what makes conditioning on endpoint pairs the *right* surrogate.
- The regularity in C1 — needed for the tower property to apply and for $C_1, C_2 < \infty$.

Appendix D: Guidance in Velocity Space

D1: The Guidance Problem

After training a velocity network $\mathbf{u}_\theta(\mathbf{x}_t, t)$ (or conditionally, $\mathbf{u}_\theta(\mathbf{x}_t, t, c)$), we want to steer sampling toward a target condition c — a class label, text embedding, or any side information not necessarily used during training.

The key structural fact. Guidance is a *sampling-time* modification of the ODE velocity field. The training objective $\mathcal{L}_{\text{CFM}}$ determines the network's weights; it plays no role in how the learned field is used at inference. Replacing $\mathbf{u}_t$ with a guided field $\tilde{\mathbf{u}}_t$ changes which density the ODE transports toward, by the continuity equation.

The regression training scale (average $\|\mathbf{x}_1 - \mathbf{x}_0\|^2$) and the scale of $\log p(c \mid \mathbf{x}_t)$ are quantities in completely different spaces and have no intrinsic relationship to each other. Any magnitude mismatch is absorbed by a scalar guidance strength s tuned at inference.

D2: Classifier Guidance

Setup. Train a time-aware classifier $p_\phi(c \mid \mathbf{x}_t)$ separately, using interpolated samples $\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1$ at each $t \in [0,1]$.

Guided ODE. At sampling time, perturb the velocity field:

$$\tilde{\mathbf{u}}_t(\mathbf{x}) \;=\; \mathbf{u}_t(\mathbf{x}) \;+\; s\,\nabla_{\mathbf{x}} \log p_\phi(c \mid \mathbf{x}_t).$$

Why the scales are commensurable. Both $\mathbf{u}_t(\mathbf{x})$ and $\nabla_{\mathbf{x}} \log p_\phi(c \mid \mathbf{x}_t)$ are vectors in $\mathbb{R}^d$ — the same ambient space as the data. The guidance strength s is a free inference hyperparameter (no correct value; tuned empirically).

Practical considerations.

- Requires a separately trained classifier on *interpolated* (not clean) samples, since the classifier must be well-calibrated at every t.
- Quality of guidance degrades if the classifier is poorly calibrated at intermediate t.
- Introduced for diffusion models by Dhariwal & Nichol (2021); directly applicable to flow matching with no modification.

D3: Classifier-Free Guidance (CFG)

Setup. Train a single network $\mathbf{u}_\theta(\mathbf{x}_t, t, c)$ with *condition dropout*: during training, replace c with a null token $\varnothing$ independently with probability p_{drop} (typically 10–20%). The training loss is the standard $\mathcal{L}_{\text{CFM}}$ — no modification.

Guided ODE. At sampling time, extrapolate in velocity space:

$$\tilde{\mathbf{u}}_t(\mathbf{x}) \;=\; \mathbf{u}_\theta(\mathbf{x}_t, t, \varnothing) \;+\; s\,\big(\mathbf{u}_\theta(\mathbf{x}_t, t, c) - \mathbf{u}_\theta(\mathbf{x}_t, t, \varnothing)\big).$$

The difference $\mathbf{u}_\theta(\cdot, c) - \mathbf{u}_\theta(\cdot, \varnothing)$ estimates the velocity perturbation attributable to the condition; s amplifies it.

$s = 0$ fully unconditional
$s = 1$ conditional only (no extrapolation)
$s > 1$ amplify conditioning (trades sample diversity for fidelity to c)

CFG is the standard approach in deployed flow-matching models (Stable Diffusion 3, Flux, etc. — Esser et al. 2024). Introduced for diffusion by Ho & Salimans (2022); adapted to flow matching by Lipman et al. (2022).

D4: Score Space vs. Velocity Space — and Extensions

Why guidance for flow matching is simpler than for diffusion.

In score-based diffusion (DDPM/DDIM), the drift in the reverse SDE is the *score* $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$, which has units $(\Delta \mathbf{x})^{-1}$ tied to the noise schedule. Adding a classifier score $s\, \nabla_{\mathbf{x}} \log p(c \mid \mathbf{x}_t)$ requires careful calibration because both terms live in score space and the noise schedule sets the natural scale.

In flow matching, $\mathbf{u}_t(\mathbf{x})$ is in *data displacement space* ($\mathbb{R}^d$, units $\Delta \mathbf{x}/\Delta t$). The classifier gradient $\nabla_{\mathbf{x}} \log p_\phi \in \mathbb{R}^d$ lives in the same space. The guidance scale s has no intrinsically correct value; it is always treated as an inference hyperparameter regardless of the training procedure.

More principled extensions.

- **Stochastic guidance / Langevin correction** (Pokle et al. 2023; Dou & Song 2024): augment the ODE with a stochastic term $\propto \nabla_{\mathbf{x}} \log p_\phi(c \mid \mathbf{x}_t)\, \mathrm{d}W_t$, giving a principled posterior-sampling interpretation at the cost of stochastic trajectories.
- **Coupling-based conditioning**: encode c directly into the coupling $\pi^{(c)}$ (e.g., pair noise with class-specific data), so the conditional structure is baked into the training paths rather than appended at inference.

CFG dominates in practice for its simplicity. Langevin correction offers a theoretically grounded alternative for inverse problems and likelihood-based guidance.