35 Flow Matching Models

Flow matching models are a class of probabilistic generative models. Consider the continuous evolution of a probability distribution over time from a simple prior distribution (e.g., a standard Gaussian distribution) to a complex distribution (e.g., the data distribution). The path of evolution is called the probability path or probability flow. The probability path is determined by a time-dependent velocity field and the prior distribution. The entire process is described by an ordinary differential equation (ODE) based on the velocity field.

The goal of flow matching learning is to fit the velocity field using training data. The key idea is not to directly fit the intractable velocity field itself, but to construct an equivalent learning objective that fits a tractable conditional velocity field, thereby enabling a continuous evolution from a standard Gaussian distribution to the data distribution.

In the data generation process, a sample is first randomly sampled from a standard Gaussian distribution. Then, using the learned velocity field, the ODE is solved through a numerical integration method, gradually computing the evolution trajectory of the sample, ultimately obtaining a sample from the true data distribution.

The flow matching method was proposed and developed by Lipman et al. in 2022.

Section 35.1 outlines the basic principle of flow matching models. Section 35.2 describes the learning and generation algorithms for flow matching, including theoretical derivations. Section 35.3 explains the relationship between flow matching models and diffusion models.

35.1 Basic Principle

In physics, ”flow” is often used to describe the continuous evolution of matter over time under deterministic dynamics, such as the flow and deformation of a fluid. ”Flow” in flow matching borrows this concept to characterize the continuous evolution of probability distributions.

The evolution of a probability distribution over time can be intuitively understood as the "flow" of probability mass in the sample space. At the initial time $t = 0$, the probability mass is distributed as a simple prior distribution $p_0(\boldsymbol{x})$; over time, the probability mass flows and deforms continuously and smoothly, becoming the distribution $p_t(\boldsymbol{x})$ at time t; at the final time $t = 1$, the probability mass precisely evolves into the complex data distribution $p_1(\boldsymbol{x})$.

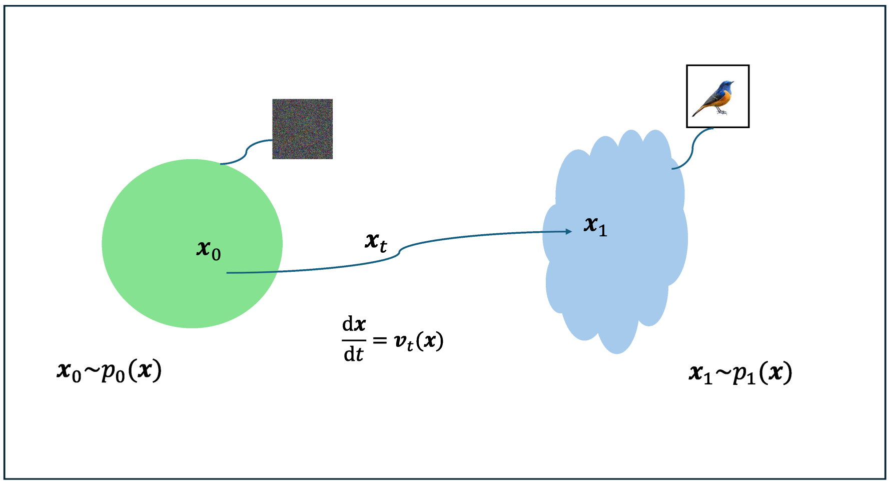

图 35.1: Flow matching: deterministic continuous evolution from prior distribution to data distribution

Figure 35.1 illustrates the basic principle of the flow matching problem. Assume the prior distribution $p_0(\boldsymbol{x})$ is a standard Gaussian distribution, with its samples being random noise images, while the data distribution $p_1(\boldsymbol{x})$ is the target distribution to be learned, with its samples being real images. The flow matching method formalizes the learning and sampling problems as modeling the aforementioned probability flow: by learning from image data, it captures the trends of distribution evolution over time and generates new images accordingly.

The evolution of the probability flow is determined by a time-dependent velocity field $\boldsymbol{v}_t(\boldsymbol{x})$ and the prior distribution. The velocity field represents the direction and rate of evolution of a sample in space over time. The solution to the ODE based on the velocity field is the evolution trajectory of the sample in space.

$$\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = \boldsymbol{v}_t(\boldsymbol{x})$$

The core technique of flow matching lies in constructing learning of a conditional velocity field equivalent to direct learning of the velocity field, thereby learning the desired velocity field. In practice, a neural network $\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x}, t)$ is used to represent this velocity field. After learning, the velocity field $\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x}, t)$ serves as the model for data generation. Given a sample from the prior distribution at the initial time, solving the ODE defined by this velocity field using numerical integration methods gradually evolves the sample to the final time, thus obtaining a sample following the data distribution.

Flow matching models are a class of flow-based models, with forms similar to normalizing flows and continuous normalizing flows (CNF), as they all achieve generative modeling by characterizing the deterministic continuous evolution of probability distributions. In practical applications, flow matching models typically demonstrate higher sampling efficiency and better generation quality.

Flow matching models are fundamentally different from diffusion models, the former based on deterministic dynamics and the latter on stochastic dynamics; however, there is also a close relationship between them. In terms of probability flow modeling, a correspondence can be established between the two. The stochastic diffusion process in diffusion models can be represented as a deterministic continuous evolution process through the probability flow ODE, and flow matching models can precisely model the deterministic continuous evolution of the probability flow. Flow matching models generally have higher sampling efficiency than diffusion models. Diffusion models typically define the data at time 0 and the Gaussian noise at time T; flow matching models typically define the Gaussian noise at time 0 and the data at time 1.

35.2 Flow Matching Problem

This section describes concepts such as probability path, velocity field, and the continuity equation, and defines the flow matching learning problem.

35.2.1 Probability Paths and Velocity Fields

Consider a sequence of probability distributions defined on the continuous time interval $[0, 1]$. At the initial time $t = 0$, the probability density function of the distribution is $p_0(\boldsymbol{x})$ (abbreviated as p_0); at the final time $t = 1$, the density function is $p_1(\boldsymbol{x})$ (abbreviated as p_1); at any intermediate time $t \in (0, 1)$, the density function is $p_t(\boldsymbol{x})$ (abbreviated as p_t). Here, $\boldsymbol{x} \in \mathbb{R}^d$ represents a sample in the sample space, also a location in the sample space, and d is the dimension of the sample space.

The probability distribution evolves gradually over time, starting from the prior distribution $p_0(\boldsymbol{x})$ at time $t = 0$: at time t, it becomes the intermediate distribution $p_t(\boldsymbol{x})$, and at time $t = 1$, it becomes the data distribution $p_1(\boldsymbol{x})$. This probability distribution evolution process $\{p_t(\boldsymbol{x})\}_{t\in[0,1]}$ is called the probability path or probability flow. Sometimes, for convenience, only the density function $p_t(\boldsymbol{x})$ at time t is used to represent the entire probability path. It is usually assumed that the prior distribution $p_0(\boldsymbol{x})$ is a standard Gaussian distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$, and the data distribution $p_1(\boldsymbol{x}) = q(\boldsymbol{x})$ is the target distribution to be learned.

Let $\boldsymbol{x} \in \mathbb{R}^d$ represent the position of a sample at time t, satisfying $\boldsymbol{x} \sim p_t(\cdot)$. The evolution of the sample $\boldsymbol{x}$ over time is described by an ordinary differential equation (ODE):

$$\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = \boldsymbol{v}(\boldsymbol{x}, t) \tag{35.1}$$

where $\boldsymbol{v}(\boldsymbol{x}, t) : \mathbb{R}^d \times [0, 1] \to \mathbb{R}^d$ is a vector field that indicates the direction and rate of evolution of a sample at the given time t and position $\boldsymbol{x}$. $\boldsymbol{v}(\boldsymbol{x}, t)$ is called the velocity field, abbreviated as $\boldsymbol{v}_t(\boldsymbol{x})$. Assume $\boldsymbol{x}_0$ is the initial position.

If the velocity field $\boldsymbol{v}(\boldsymbol{x}, t)$ is locally Lipschitz continuous in the spatial variable $\boldsymbol{x}$, continuous in the time variable t, and satisfies a linear growth condition, then for any initial position $\boldsymbol{x}_0$, the ODE (35.1) has a unique solution on $t \in [0, 1]$.

Since the velocity field is time-dependent, i.e., there is a velocity field for each time t, when the initial position $\boldsymbol{x}_0$ and the velocity field $\boldsymbol{v}_t(\boldsymbol{x})$ are determined, the positions of the sample at all times, i.e., its evolution trajectory $\boldsymbol{x}_0 \to \cdots \to \boldsymbol{x} \to \cdots \to \boldsymbol{x}_1$, is uniquely determined by this ODE. Its solution can be written in integral form

$$\boldsymbol{x} = \boldsymbol{x}_0 + \int_0^t \boldsymbol{v}(\boldsymbol{x}_\tau, \tau)\, d\tau \tag{35.2}$$

The ODE can also be solved using numerical integration methods. So it can be said that the solution of the ODE based on the velocity field represents the evolution trajectory of the sample in space.

In this case, the velocity field $\boldsymbol{v}(\boldsymbol{x}, t)$ is said to induce the probability path $p_t(\boldsymbol{x})$, expressed in the following form.

$$\boldsymbol{x}_0 \sim p_0(\cdot), \frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = \boldsymbol{v}_t(\boldsymbol{x}) \Rightarrow \boldsymbol{x} \sim p_t(\cdot) \tag{35.3}$$

It should be noted that the velocity field that can induce the same probability path is not necessarily unique.

There exists a function $\boldsymbol{\psi}$ between the initial position $\boldsymbol{x}_0$ of the ODE and the solution $\boldsymbol{x}$ at each time t, called the flow.

$$\boldsymbol{x} = \boldsymbol{\psi}(\boldsymbol{x}_0, t) \tag{35.4}$$

The ODE can also be written as

$$\begin{aligned} \frac{\mathrm{d}\boldsymbol{\psi}(\boldsymbol{x}_0, t)}{\mathrm{d}t} &= \boldsymbol{v}(\boldsymbol{\psi}(\boldsymbol{x}_0, t), t) \\ \boldsymbol{\psi}(\boldsymbol{x}_0, 0) &= \boldsymbol{x}_0 \end{aligned} \tag{35.5}$$

Here, $\boldsymbol{\psi} : \mathbb{R}^d \times [0, 1] \to \mathbb{R}^d$ represents the evolution trajectory of a sample from $\boldsymbol{x}_0$ to $\boldsymbol{x}$. $\boldsymbol{\psi}(\boldsymbol{x}_0, t)$ is abbreviated as $\boldsymbol{\psi}_t(\boldsymbol{x}_0)$.

It is usually assumed that the flow $\boldsymbol{\psi}_t(\boldsymbol{x}_0)$ is a diffeomorphism, i.e., under appropriate smoothness conditions, it is a smooth, invertible transformation with a smooth inverse. Flow is also a concept used in continuous normalizing flows.

It can be seen that the ODE is defined by the velocity field $\boldsymbol{v}_t(\boldsymbol{x})$, and its solution $\boldsymbol{x}$ can be expressed as the flow $\boldsymbol{\psi}_t(\boldsymbol{x}_0)$. The velocity field $\boldsymbol{v}_t(\boldsymbol{x})$ together with the prior distribution $p_0(\boldsymbol{x})$ determines a probability path. If we can find a "correct" velocity field $\boldsymbol{v}_t(\boldsymbol{x})$, then by solving the above ODE, samples from the prior distribution $p_0(\boldsymbol{x})$ can be evolved into samples of the data distribution $p_1(\boldsymbol{x})$.

35.2.2 Continuity Equation

At each time t, the probability mass equals 1, and is constant.

$$\int p_t(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 1$$

This means that the probability path must satisfy this constraint during evolution.

In physics, the motion of matter must obey the law of conservation of mass, whose mathematical expression is the continuity equation. Correspondingly, the evolution of a probability path must also obey the conservation law and has its corresponding continuity equation. Specifically, the relationship between the probability density $p_t(\boldsymbol{x})$ and the velocity field $\boldsymbol{v}_t(\boldsymbol{x})$ is described by the probabilistic continuity equation:

$$\frac{\partial p_t(\boldsymbol{x})}{\partial t} + \nabla_{\boldsymbol{x}} \cdot (p_t(\boldsymbol{x})\boldsymbol{v}_t(\boldsymbol{x})) = 0 \tag{35.6}$$

where $\nabla_{\boldsymbol{x}}\cdot$ is the divergence operator.

In the continuity equation, $\frac{\partial p_t(\boldsymbol{x})}{\partial t}$ represents the rate of change of the probability density at position $\boldsymbol{x}$ over time, and $\nabla_{\boldsymbol{x}} \cdot (p_t(\boldsymbol{x})\boldsymbol{v}_t(\boldsymbol{x}))$ represents the probability density flowing into or out of that position from the surroundings. When the former is greater than 0 and the latter is less than 0, probability mass flows in from the surroundings (converges); when the former is less than 0 and the latter is greater than 0, probability mass flows out to the surroundings (diverges).

The following theorem holds regarding the continuity equation.

Theorem 35.1 (Continuity Equation) For a probability path $p_t(\boldsymbol{x})$ induced by a velocity field $\boldsymbol{v}_t(\boldsymbol{x})$, if the velocity field $\boldsymbol{v}_t(\boldsymbol{x})$ satisfies certain smoothness conditions (locally Lipschitz continuous in the spatial variable $\boldsymbol{x}$, continuous in the time variable t, and satisfies a linear growth condition), then the probability path $p_t(\boldsymbol{x})$ satisfies the following continuity equation:

$$\frac{\partial}{\partial t} p_t(\boldsymbol{x}) = -\nabla_{\boldsymbol{x}} \cdot \big(p_t(\boldsymbol{x})\boldsymbol{v}_t(\boldsymbol{x})\big) \tag{35.7}$$

where $\nabla_{\boldsymbol{x}}\cdot$ is the divergence operator. Conversely, if a probability path $p_t(\boldsymbol{x})$ satisfies this equation and the initial condition $p_0(\boldsymbol{x})$, and the velocity field $\boldsymbol{v}_t(\boldsymbol{x})$ in the equation satisfies the aforementioned smoothness conditions, then the probability path $p_t(\boldsymbol{x})$ is induced by the velocity field $\boldsymbol{v}_t(\boldsymbol{x})$.

35.2.3 Flow Matching

Assume we have a training dataset whose samples $\boldsymbol{x}$ come from an unknown data distribution $q(\boldsymbol{x})$. The goal is to learn this distribution. Assume there exists a probability path evolving from the prior distribution $p_0(\boldsymbol{x})$ (a standard Gaussian distribution) to the data distribution $p_1(\boldsymbol{x}) = q(\boldsymbol{x})$, which generated the samples in the training set. Flow matching aims to learn a neural network $\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x}, t)$ that can fit the velocity field $\boldsymbol{v}_t(\boldsymbol{x})$ that determines this probability path (i.e., the aforementioned correct velocity field). The loss function is defined as

$$L_{\mathrm{FM}}(\boldsymbol{\theta}) = \mathbb{E}_{t,\boldsymbol{x}}||\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x}, t) - \boldsymbol{v}_t(\boldsymbol{x})||_2^2, \quad t \sim U[0,1], \boldsymbol{x} \sim p_t(\cdot) \tag{35.8}$$

where the neural network $\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x}, t) : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ represents the velocity vector at position $\boldsymbol{x}$ and time t, $\boldsymbol{\theta}$ are the parameters of the neural network, $\boldsymbol{v}_t(\boldsymbol{x}) \in \mathbb{R}^d$ is the target velocity field, $||\cdot||$ is the L_2 norm, and $U[0,1]$ is the uniform distribution on the interval $[0,1]$.

After learning the neural network velocity field, samples from the data distribution can be generated starting from a standard Gaussian distribution. The problem here is that the loss function $L_{\mathrm{FM}}(\boldsymbol{\theta})$ cannot be directly computed because it depends on the true unknown velocity field $\boldsymbol{v}_t(\boldsymbol{x})$ and the intermediate distribution $p_t(\boldsymbol{x})$, which is difficult to sample directly. The flow matching method solves this problem by implementing an equivalent learning objective for a conditional velocity field.

35.3 Flow Matching Method

This section elaborates on the flow matching algorithm and its theoretical foundation. First, a conditional probability path and a conditional velocity field are defined. Then, the relationship between the target probability path and target velocity field (i.e., the marginal probability path and marginal velocity field) to be learned is derived. Next, it is proven that learning the target marginal velocity field is equivalent to learning the conditional velocity field. Finally, when the conditional probability path is assumed to be Gaussian, the learning algorithm for flow matching is derived.

35.3.1 Conditional Probability Paths and Conditional Velocity Fields

First, construct the conditional velocity field and conditional probability path. Define a sequence of conditional probability distributions on the continuous time interval $[0,1]$, called a conditional probability path. Given a sample (or location) $\boldsymbol{x}_1$, the conditional probability distribution is $p_t(\boldsymbol{x}|\boldsymbol{x}_1)$, where $\boldsymbol{x}$ represents a location in the sample space at time $t \in (0,1)$. Assume the initial conditional probability distribution and final conditional probability distribution are respectively

$$p_0(\boldsymbol{x}|\boldsymbol{x}_1) = \pi(\boldsymbol{x}), \quad p_1(\boldsymbol{x}|\boldsymbol{x}_1) = \delta_{\boldsymbol{x}_1}(\boldsymbol{x}) \tag{35.9}$$

where $\pi(\boldsymbol{x})$ is an arbitrary probability distribution, usually a standard Gaussian distribution; $\delta_{\boldsymbol{x}_1}(\boldsymbol{x})$ denotes the Dirac δ distribution satisfying

$$\delta_{\boldsymbol{x}'}(\boldsymbol{x}) = \begin{cases} +\infty, & \boldsymbol{x} = \boldsymbol{x}', \\ 0, & \boldsymbol{x} \neq \boldsymbol{x}', \end{cases} \quad \text{and} \quad \int_{\mathbb{R}^d} \delta_{\boldsymbol{x}'}(\boldsymbol{x})\, \mathrm{d}^d\boldsymbol{x} = 1.$$

Correspondingly, define the conditional velocity field. Given a sample $\boldsymbol{x}_1$, the conditional velocity field is $\boldsymbol{v}_t(\boldsymbol{x}|\boldsymbol{x}_1)$, where $\boldsymbol{x}$ represents the location of the sample at time $t \in (0,1)$.

Thus, the ordinary differential equation (ODE) describing the evolution of the sample at each time t can be defined as

$$\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = \boldsymbol{v}_t(\boldsymbol{x}|\boldsymbol{x}_1) \tag{35.10}$$

Given the initial sample position $\boldsymbol{x}_0$ and the conditional velocity field $\boldsymbol{v}_t(\boldsymbol{x}|\boldsymbol{x}_1)$, the conditional probability path $p_t(\boldsymbol{x}|\boldsymbol{x}_1)$ is determined. That is, the conditional velocity field induces the conditional probability path.

$$\boldsymbol{x}_0 \sim \pi(\cdot), \boldsymbol{x}_1 \sim q(\cdot), \frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = \boldsymbol{v}_t(\boldsymbol{x}|\boldsymbol{x}_1) \Rightarrow \boldsymbol{x} \sim p_t(\cdot|\boldsymbol{x}_1) \tag{35.11}$$

According to Theorem 35.1, the conditional probability path satisfies the corresponding continuity equation.

$$\frac{\partial}{\partial t} p_t(\boldsymbol{x}|\boldsymbol{x}_1) + \nabla_{\boldsymbol{x}} \cdot (p_t(\boldsymbol{x}|\boldsymbol{x}_1)\boldsymbol{v}_t(\boldsymbol{x}|\boldsymbol{x}_1)) = 0 \tag{35.12}$$

35.3.2 Marginal Probability Paths and Marginal Velocity Fields

Next, it is derived that the marginal probability path corresponding to the conditional probability path is exactly the target probability path, and the marginal velocity field corresponding to the conditional velocity field is exactly the target velocity field.

Define the marginal probability path as the expectation of the conditional probability path $p_t(\boldsymbol{x}|\boldsymbol{x}_1)$ with respect to the data distribution $q(\boldsymbol{x}_1)$.

$$p_t(\boldsymbol{x}) = \int p_t(\boldsymbol{x}|\boldsymbol{x}_1)q(\boldsymbol{x}_1)\mathrm{d}\boldsymbol{x}_1 \tag{35.13}$$

where $q(\boldsymbol{x}_1)$ is the data distribution. It is easy to deduce that the initial marginal distribution and final marginal distribution satisfy respectively

$$p_0(\boldsymbol{x}) = \pi(\boldsymbol{x}), \quad p_1(\boldsymbol{x}) = q(\boldsymbol{x}) \tag{35.14}$$

where $\boldsymbol{x}$ represents a location in the sample space.

Define the marginal velocity field as the expectation of the conditional velocity field $\boldsymbol{v}_t(\boldsymbol{x}|\boldsymbol{x}_1)$ with respect to the posterior probability distribution $p_t(\boldsymbol{x}_1|\boldsymbol{x})$. According to Bayes' theorem, we have

$$\boldsymbol{v}_t(\boldsymbol{x}) = \int \boldsymbol{v}_t(\boldsymbol{x}|\boldsymbol{x}_1)\frac{p_t(\boldsymbol{x}|\boldsymbol{x}_1)q(\boldsymbol{x}_1)}{p_t(\boldsymbol{x})}\mathrm{d}\boldsymbol{x}_1 \tag{35.15}$$

Theorem 35.2 below guarantees that the above marginal velocity field induces the marginal probability path, i.e., the target velocity field and probability path.

Theorem 35.2 If the conditional velocity field $\boldsymbol{v}_t(\boldsymbol{x}|\boldsymbol{x}_1)$ induces the conditional probability path $p_t(\boldsymbol{x}|\boldsymbol{x}_1)$, then the corresponding marginal velocity field $\boldsymbol{v}_t(\boldsymbol{x})$ induces the corresponding marginal probability path $p_t(\boldsymbol{x})$, i.e., the marginal velocity field and marginal probability path satisfy the continuity equation (35.6).

$$\boldsymbol{x}_0 \sim \pi(\cdot), \frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = \boldsymbol{v}_t(\boldsymbol{x}) \Rightarrow \boldsymbol{x} \sim p_t(\cdot) \tag{35.16}$$

Proof

$$
\begin{aligned}
\frac{\partial p_t(\boldsymbol{x})}{\partial t} &= \frac{\partial}{\partial t}\int p_t(\boldsymbol{x}|\boldsymbol{x}_1)q(\boldsymbol{x}_1)\mathrm{d}\boldsymbol{x}_1 \\
&= \int \frac{\partial}{\partial t} p_t(\boldsymbol{x}|\boldsymbol{x}_1)q(\boldsymbol{x}_1)\mathrm{d}\boldsymbol{x}_1 \\
&= \int -\nabla_{\boldsymbol{x}} \cdot \left(p_t(\boldsymbol{x}|\boldsymbol{x}_1)\boldsymbol{v}_t(\boldsymbol{x}|\boldsymbol{x}_1)\right) q(\boldsymbol{x}_1)\mathrm{d}\boldsymbol{x}_1 \\
&= -\nabla_{\boldsymbol{x}} \cdot \int p_t(\boldsymbol{x}|\boldsymbol{x}_1)\boldsymbol{v}_t(\boldsymbol{x}|\boldsymbol{x}_1)q(\boldsymbol{x}_1)\mathrm{d}\boldsymbol{x}_1 \\
&= -\nabla_{\boldsymbol{x}} \cdot \left(p_t(\boldsymbol{x}) \int \boldsymbol{v}_t(\boldsymbol{x}|\boldsymbol{x}_1)\frac{p_t(\boldsymbol{x}|\boldsymbol{x}_1)q(\boldsymbol{x}_1)}{p_t(\boldsymbol{x})}\mathrm{d}\boldsymbol{x}_1\right) \\
&= -\nabla_{\boldsymbol{x}} \cdot \left(p_t(\boldsymbol{x})\boldsymbol{v}_t(\boldsymbol{x})\right)
\end{aligned}
$$

The first step uses the definition of the marginal probability path, the second step uses the Leibniz integral rule, the third step uses the continuity equation, the fourth step uses the Leibniz integral rule, the fifth step uses properties of integrals, and the sixth step uses the definition of the marginal velocity field. It is assumed here that the relevant functions satisfy integrability and smoothness conditions such that integration and differentiation are interchangeable.

■

35.3.3 Objective Function for Flow Matching

Consider learning a neural network $\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x}, t)$ that can fit the conditional velocity field $\boldsymbol{v}_t(\boldsymbol{x}|\boldsymbol{x}_1)$ which determines the conditional probability path. The loss function is defined as

$$
\begin{aligned}
L_{\mathrm{CFM}}(\boldsymbol{\theta}) &= \mathbb{E}_{t,\boldsymbol{x}_1,\boldsymbol{x}}||\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x}, t) - \boldsymbol{v}_t(\boldsymbol{x}|\boldsymbol{x}_1)||_2^2 \\
& t \sim U[0,1], \boldsymbol{x}_1 \sim q(\cdot), \boldsymbol{x} \sim p_t(\cdot|\boldsymbol{x}_1)
\end{aligned} \tag{35.17}
$$

where the neural network $\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x}, t) : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ represents the velocity vector at time t and position $\boldsymbol{x}$, $\boldsymbol{\theta}$ are the parameters of the neural network, $\boldsymbol{v}_t(\boldsymbol{x}|\boldsymbol{x}_1) \in \mathbb{R}^d$ is the target conditional velocity field, $||\cdot||$ is the L_2 norm, and $U[0,1]$ is the uniform distribution on the interval $[0,1]$.

It can be proven that learning the conditional velocity field is equivalent to learning the target velocity field (marginal velocity field). The following theorem holds. Therefore, by learning a conditional velocity field, we can obtain the target (marginal) velocity field. This is the core idea of the flow matching method.

Theorem 35.3 Minimizing the objective function L_{FM} of flow matching is equivalent to minimizing the objective function L_{CFM} of conditional flow matching.

$$\min_{\boldsymbol{\theta}} L_{\mathrm{FM}}(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} L_{\mathrm{CFM}}(\boldsymbol{\theta}) \tag{35.18}$$

Proof

The loss function L_{FM} can be written as

$$\begin{aligned} L_{\mathrm{FM}}(\boldsymbol{\theta}) &= \mathbb{E}_{t,\boldsymbol{x}}||\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x},t) - \boldsymbol{v}_t(\boldsymbol{x})||_2^2 \\ &= \mathbb{E}_{t,\boldsymbol{x}}\left[||\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x},t)||_2^2 - 2\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x},t)^{\mathrm{T}}\boldsymbol{v}_t(\boldsymbol{x}) + ||\boldsymbol{v}_t||_2^2\right] \\ &= \mathbb{E}_{t,\boldsymbol{x}}\left[||\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x},t)||_2^2\right] - 2\mathbb{E}_{t,\boldsymbol{x}}\left[\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x},t)^{\mathrm{T}}\boldsymbol{v}_t(\boldsymbol{x})\right] + \mathbb{E}_{t,\boldsymbol{x}}\left[||\boldsymbol{v}_t(\boldsymbol{x})||_2^2\right] \\ &= \mathbb{E}_{t,\boldsymbol{x},\boldsymbol{x}_1}\left[||\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x},t)||_2^2\right] - 2\mathbb{E}_{t,\boldsymbol{x}}\left[\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x},t)^{\mathrm{T}}\boldsymbol{v}_t(\boldsymbol{x})\right] + C' \end{aligned} \tag{35.19}$$

where the third step uses properties of mathematical expectation; the first term in the fourth step further takes expectation with respect to the data distribution $q(\boldsymbol{x}_1)$, and the third term is a constant C' independent of the parameter $\boldsymbol{\theta}$.

Next, we derive the second term.

$$\begin{aligned} \mathbb{E}_{t,\boldsymbol{x}}\left[\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x},t)^{\mathrm{T}}\boldsymbol{v}_t(\boldsymbol{x})\right] &= \int_0^1 \int p_t(\boldsymbol{x})\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x},t)^{\mathrm{T}}\boldsymbol{v}_t(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}t \\ &= \int_0^1 \int p_t(\boldsymbol{x})\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x},t)^{\mathrm{T}}\left[\int \boldsymbol{v}_t(\boldsymbol{x}|\boldsymbol{x}_1)\frac{p_t(\boldsymbol{x}|\boldsymbol{x}_1)q(\boldsymbol{x}_1)}{p_t(\boldsymbol{x})}\mathrm{d}\boldsymbol{x}_1\right]\mathrm{d}\boldsymbol{x}\mathrm{d}t \\ &= \int_0^1 \int\int \boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x},t)^{\mathrm{T}}\boldsymbol{v}_t(\boldsymbol{x}|\boldsymbol{x}_1)p_t(\boldsymbol{x}|\boldsymbol{x}_1)q(\boldsymbol{x}_1)\mathrm{d}\boldsymbol{x}_1\mathrm{d}\boldsymbol{x}\mathrm{d}t \\ &= \mathbb{E}_{t,\boldsymbol{x},\boldsymbol{x}_1}\left[\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x},t)^{\mathrm{T}}\boldsymbol{v}_t(\boldsymbol{x}|\boldsymbol{x}_1)\right] \end{aligned} \tag{35.20}$$

where the second step uses the definition of the marginal velocity field, the third step uses properties of integrals, and the fourth step uses the definition of mathematical expectation.

The loss function L_{CFM} can be written as

$$\begin{aligned} L_{\mathrm{CFM}}(\boldsymbol{\theta}) &= \mathbb{E}_{t,\boldsymbol{x},\boldsymbol{x}_1}||\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x},t) - \boldsymbol{v}_t(\boldsymbol{x}|\boldsymbol{x}_1)||_2^2 \\ &= \mathbb{E}_{t,\boldsymbol{x},\boldsymbol{x}_1}\left[||\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x},t)||_2^2 - 2\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x},t)^{\mathrm{T}}\boldsymbol{v}_t(\boldsymbol{x}|\boldsymbol{x}_1) + ||\boldsymbol{v}_t(\boldsymbol{x}|\boldsymbol{x}_1)||_2^2\right] \\ &= \mathbb{E}_{t,\boldsymbol{x},\boldsymbol{x}_1}\left[||\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x},t)||_2^2\right] - 2\mathbb{E}_{t,\boldsymbol{x},\boldsymbol{x}_1}\left[\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x},t)^{\mathrm{T}}\boldsymbol{v}_t(\boldsymbol{x}|\boldsymbol{x}_1)\right] + \mathbb{E}_{t,\boldsymbol{x},\boldsymbol{x}_1}\left[||\boldsymbol{v}_t(\boldsymbol{x}|\boldsymbol{x}_1)||_2^2\right] \\ &= \mathbb{E}_{t,\boldsymbol{x},\boldsymbol{x}_1}\left[||\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x},t)||_2^2\right] - 2\mathbb{E}_{t,\boldsymbol{x},\boldsymbol{x}_1}\left[\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x},t)^{\mathrm{T}}\boldsymbol{v}_t(\boldsymbol{x}|\boldsymbol{x}_1)\right] + (C' - C) \end{aligned} \tag{35.21}$$

where the third step uses the definition of mathematical expectation, and in the fourth step, the third term is a constant $(C' - C)$ independent of $\boldsymbol{\theta}$.

Comparing equations (35.19) and (35.21), and using equation (35.20), we obtain

$$L_{\mathrm{FM}}(\boldsymbol{\theta}) = L_{\mathrm{CFM}}(\boldsymbol{\theta}) + C$$

where C is a constant term. ■

The advantage of indirectly optimizing the objective function L_{CFM} is that there is no need to know or compute the marginal velocity field $\boldsymbol{v}_t(\boldsymbol{x})$ and the marginal probability path $p_t(\boldsymbol{x})$; one only needs to sample from the simple conditional distribution $p_t(\boldsymbol{x}|\boldsymbol{x}_1)$ and use the known, analytically expressible conditional velocity field $\boldsymbol{v}_t(\boldsymbol{x}_t|\boldsymbol{x}_1)$. This addresses the difficulty of directly optimizing the objective function L_{FM}.

35.3.4 Flow Matching Algorithm

With the general method for flow matching learning, the remaining problem is how to define a specific conditional probability path. It is usually assumed to be a Gaussian conditional probability path, leading to a commonly used algorithm.

For a Gaussian conditional probability path, the conditional probability distribution at each time t is defined as

$$p_t(\boldsymbol{x}|\boldsymbol{x}_1) = \mathcal{N}(\alpha_t\boldsymbol{x}_1, \beta_t^2\boldsymbol{I}).$$

Equivalently, introduce $\boldsymbol{x}_0 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, independent of $\boldsymbol{x}_1$, and define

$$\boldsymbol{x} = \alpha_t\boldsymbol{x}_1 + \beta_t\boldsymbol{x}_0.$$

where α_t and β_t are parameters.

Set $\alpha_0 = \beta_1 = 0$ and $\alpha_1 = \beta_0 = 1$, i.e., the conditional distribution satisfies

$$p_0(\boldsymbol{x}|\boldsymbol{x}_1) = \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \quad p_1(\boldsymbol{x}|\boldsymbol{x}_1) = \delta_{\boldsymbol{x}_1}(\boldsymbol{x})$$

Then the corresponding marginal distribution satisfies

$$p_0(\boldsymbol{x}) = \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \quad p_1(\boldsymbol{x}) = q(\boldsymbol{x})$$

Thus, the conditional velocity field can induce a Gaussian conditional probability path

$$\begin{gathered}\boldsymbol{x}_0 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \boldsymbol{x}_1 \sim q(\cdot), \frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = \boldsymbol{v}_t(\boldsymbol{x}|\boldsymbol{x}_1) \Rightarrow \boldsymbol{x} \sim p_t(\cdot|\boldsymbol{x}_1) \\ \boldsymbol{x} = \alpha_t \boldsymbol{x}_1 + \beta_t \boldsymbol{x}_0\end{gathered} \tag{35.22}$$

The conditional velocity field is

$$\boldsymbol{v}_t(\boldsymbol{x}|\boldsymbol{x}_1) = \Big(\dot{\alpha}_t - \frac{\dot{\beta}_t}{\beta_t}\alpha_t\Big)\boldsymbol{x}_1 + \frac{\dot{\beta}_t}{\beta_t}\boldsymbol{x} \tag{35.23}$$

where $\dot{\alpha}_t$ denotes the derivative of α_t with respect to time t, and $\dot{\beta}_t$ denotes the derivative of β_t with respect to time t. The proof is left as an exercise.

When $\boldsymbol{x}_0$ is determined, it can be deduced that

$$\boldsymbol{v}_t(\boldsymbol{x}|\boldsymbol{x}_1) = \dot{\alpha}_t \boldsymbol{x}_1 + \dot{\beta}_t \boldsymbol{x}_0$$

Consider a specific case: the conditional probability path is a linear interpolation of $\boldsymbol{x}_1$ and $\boldsymbol{x}_0$. That is,

$$\alpha_t = t, \beta_t = 1 - t$$

$$\boldsymbol{x} = t\boldsymbol{x}_1 + (1 - t)\boldsymbol{x}_0$$

Thus we have

$$\boldsymbol{v}_t(\boldsymbol{x}|\boldsymbol{x}_1) = \boldsymbol{x}_1 - \boldsymbol{x}_0$$

The objective function becomes very simple:

$$\begin{aligned} L_{\mathrm{CFM}}(\boldsymbol{\theta}) &= \mathbb{E}_{t,\boldsymbol{x},\boldsymbol{x}_1} ||\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x}, t) - (\boldsymbol{x}_1 - \boldsymbol{x}_0)||_2^2 \\ &= \mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{x}_1} ||\boldsymbol{u}_{\boldsymbol{\theta}}(t\boldsymbol{x}_1 + (1-t)\boldsymbol{x}_0, t) - (\boldsymbol{x}_1 - \boldsymbol{x}_0)||_2^2 \end{aligned} \tag{35.24}$$

The linear interpolation path is, in the sense of optimal transport, a path of minimal energy that continuously transports the prior distribution to the data distribution. This property is not only optimal in optimal transport theory but also leads to more stable training processes and higher sampling efficiency in practice, making flow matching widely adopted in practical applications.

Based on the training data, optimizing this objective function allows learning a neural network consistent with the marginal velocity field. Algorithm 35.1 presents the learning algorithm for flow matching in the linear interpolation case.

Algorithm 35.1 Flow matching (linear interpolation) − learning algorithm

Input: Training dataset $\mathcal{T}$.

Output: Velocity field neural network $\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x}, t)$.

(1) Initialize neural network parameters $\boldsymbol{\theta}$.

(2) Repeat the following until convergence:

(2-1) Sample a data sample $\boldsymbol{x}_1 \sim q(\cdot)$ from the training dataset $\mathcal{T}$;

(2-2) Randomly sample time $t \sim U[0, 1]$;

(2-3) Randomly sample initial noise $\boldsymbol{x}_0 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$;

(2-4) Construct the intermediate state sample (linear interpolation path):

$$\boldsymbol{x} = t\boldsymbol{x}_1 + (1 - t)\boldsymbol{x}_0$$

(2-5) Compute the loss function gradient and update the neural network parameters:

$$\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \left\| \boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x}, t) - (\boldsymbol{x}_1 - \boldsymbol{x}_0) \right\|_2^2$$

(3) Output the learned velocity field neural network $\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x}, t)$. ■

Algorithm 35.2 presents the generation algorithm for flow matching in the linear interpolation case. This algorithm uses the Euler method to solve the corresponding ODE, where the velocity field is represented by the learned neural network. In principle, higher-precision numerical integration methods for ODEs can also be used.

Algorithm 35.2 Flow matching (linear interpolation – generation algorithm)

Input: Velocity field neural network $\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x}, t)$.

Output: Generated sample $\boldsymbol{x}_1$.

Hyperparameters: Number of time steps T, step size $\Delta t = 1/T$.

(1) Randomly sample initial noise $\boldsymbol{x}_0 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$.

(2) For $(k = 0, 1, \cdots, T-1)$\{

(2-1) Set current time $t_k = k\Delta t$;

(2-2) Update the sample according to the velocity field:

$$\boldsymbol{x}_{t_{k+1}} = \boldsymbol{x}_{t_k} + \Delta t\, \boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_k}, t_k)$$

\}

(3) Output the generated sample $\boldsymbol{x}_1 = \boldsymbol{x}_{t_T}$. ■

35.4 Flow Matching Models and Diffusion Models

This section first points out that flow matching is a more general framework applicable to diffusion processes, and then compares flow matching models and diffusion models.

35.4.1 Flow Matching Models for Diffusion Processes

Stochastic Differential Equation (SDE)

Diffusion processes describe the stochastic evolution of samples over continuous time using stochastic differential equations (SDEs), including a forward process (adding noise) and a corresponding reverse process (denoising).

The forward process is predefined, and its SDE form is:

$$\mathrm{d}\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{x}, t)\mathrm{d}t + g(t)\mathrm{d}\boldsymbol{w} \tag{35.25}$$

where $\boldsymbol{f}(\boldsymbol{x}, t)$ is the drift coefficient, $g(t)$ is the diffusion coefficient, and $\boldsymbol{w}$ is a standard Wiener process. This process starts from a sample of the data distribution $p_0(\boldsymbol{x})$ and, by gradually adding random noise, ultimately obtains a sample from the prior distribution (e.g., a standard Gaussian distribution $p_1(\boldsymbol{x})$).

The goal is to learn the reverse process. This process starts from a sample of the prior distribution $p_1(\boldsymbol{x})$ and, by gradually removing random noise, ultimately obtains a sample from the data distribution $p_0(\boldsymbol{x})$. According to stochastic process theory, the form of the reverse SDE is:

$$\mathrm{d}\boldsymbol{x} = \left[\boldsymbol{f}(\boldsymbol{x}, t) - g^2(t)\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x})\right] \mathrm{d}t + g(t)\mathrm{d}\bar{\boldsymbol{w}} \tag{35.26}$$

where $\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x})$ is the score function or score field, and $\bar{\boldsymbol{w}}$ is a standard Wiener process in reverse time. The learning objective becomes estimating the score field at each time t.

Fokker-Planck Equation

In the forward or reverse process of a diffusion process, there exists a probability distribution at each time, whose density function is $p_t(\boldsymbol{x})$. Together, they form a forward or reverse probability flow or probability path. The probability path satisfies a partial differential equation called the Fokker-Planck equation or Kolmogorov forward equation.

Theorem 35.4 (Fokker-Planck Equation) For the forward process (35.25) of a diffusion process, if its drift coefficient $\boldsymbol{f}(\boldsymbol{x}, t)$ and diffusion coefficient $g(t)$ satisfy certain smoothness conditions ($\boldsymbol{f}(\boldsymbol{x}, t)$ is locally Lipschitz continuous in the spatial variable $\boldsymbol{x}$, continuous in the time variable t, and satisfies linear growth; and $g(t)$ is continuous and bounded in the time variable t), then its probability path $p_t(\boldsymbol{x})$ satisfies the following Fokker-Planck equation:

$$\frac{\partial}{\partial t} p_t(\boldsymbol{x}) = -\nabla_{\boldsymbol{x}} \cdot \big(p_t(\boldsymbol{x})\boldsymbol{f}(\boldsymbol{x}, t)\big) + \frac{g^2(t)}{2}\nabla^2_{\boldsymbol{x}} p_t(\boldsymbol{x}) \tag{35.27}$$

where $\nabla_{\boldsymbol{x}}\cdot$ is the divergence operator, and $\nabla^2_{\boldsymbol{x}}$ is the Laplacian operator. Conversely, if a probability path $p_t(\boldsymbol{x})$ satisfies this equation and the initial condition $p_0(\boldsymbol{x})$, then it represents the probability path of the diffusion process (35.25).

The Fokker-Planck equation (35.27) is the specific form of the law of conservation of probability mass (continuity equation) in a diffusion process.

Ordinary Differential Equation (ODE)

The probability path or probability flow of the above diffusion process can be described by a deterministic ODE.

Corollary 35.1 For the forward process (35.25) and reverse process (35.26) of a diffusion process, there exists a velocity field $\boldsymbol{v}_t(\boldsymbol{x})$ as follows that induces the probability path represented by the probability density function $p_t(\boldsymbol{x})$.

$$\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = \boldsymbol{v}_t(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x}, t) - \frac{g^2(t)}{2}\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}) \tag{35.28}$$

Proof:

For the forward process, starting from the Fokker-Planck equation (35.27):

$$\begin{aligned}
\frac{\partial}{\partial t}p_t(\boldsymbol{x}) &= -\nabla_{\boldsymbol{x}} \cdot \big(p_t(\boldsymbol{x})\boldsymbol{f}(\boldsymbol{x}, t)\big) + \frac{g^2(t)}{2}\nabla^2_{\boldsymbol{x}}p_t(\boldsymbol{x}) \\
&= -\nabla_{\boldsymbol{x}} \cdot \big(p_t(\boldsymbol{x})\boldsymbol{f}(\boldsymbol{x}, t)\big) + \frac{g^2(t)}{2}\nabla_{\boldsymbol{x}} \cdot \big(p_t(\boldsymbol{x})\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x})\big) \\
&= -\nabla_{\boldsymbol{x}} \cdot \left[p_t(\boldsymbol{x}) \left(\boldsymbol{f}(\boldsymbol{x}, t) - \frac{g^2(t)}{2}\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x})\right)\right].
\end{aligned}$$

Let $\boldsymbol{v}_t(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x}, t) - \frac{g^2(t)}{2}\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x})$. The above equation can be written in the continuity equation form:

$$\frac{\partial}{\partial t}p_t(\boldsymbol{x}) = -\nabla_{\boldsymbol{x}} \cdot \big(p_t(\boldsymbol{x})\boldsymbol{v}_t(\boldsymbol{x})\big)$$

The reverse process is the time reversal of the forward process given its probability flow. Reversing time from $t = 1$ to $t = 0$, the probability flow ODE still takes the form (35.28). ■

From Corollary 35.1, it can be seen that after a diffusion model learns the score field $\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x})$, it can directly use the ODE (35.28) for deterministic generation, obtaining samples from the data distribution, without the need of utilizing the reverse SDE for stochastic sampling.

For example, in the DDPM model, the coefficients of the forward SDE are $f(\boldsymbol{x}, t) = -\frac{\beta(t)}{2}\boldsymbol{x}$, $g(t) = \sqrt{\beta(t)}$. According to Corollary 35.1, the corresponding probability flow ODE is:

$$\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = -\frac{\beta(t)}{2}\left(\boldsymbol{x} + \nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x})\right) \tag{35.29}$$

表 35.1: Comparison of diffusion models and flow matching models

	Flow matching models	Diffusion models (SDE)
Approach	deterministic, probability distribution evolution	stochastic, forward noising + reverse denoising of samples
Equation	ordinary differential equation (ODE)	stochastic differential equation (SDE)
Learning objective	velocity field $\boldsymbol{v}_t(\boldsymbol{x})$	score field $\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x})$ or noise
Constraint	continuity equation	Fokker-Planck equation
Sampling efficiency	can use fewer steps	requires more steps; also has deterministic form

Data can be generated directly using this probability flow. Accelerated sampling algorithms like DDIM are essentially based on this idea.

35.4.2 Comparison Between Flow Matching Models and Diffusion Models

Flow matching models and diffusion models both belong to the class of probabilistic generative models, but there are fundamental differences in their basic principles. Flow matching models characterize the deterministic evolution of probability distributions through deterministic dynamics, a process described by an ordinary differential equation (ODE). This process is constrained by the continuity equation. It learns the velocity field $\boldsymbol{v}_t(\boldsymbol{x})$ to evolve the prior distribution into the target data distribution. Since the evolution process is deterministic, sampling can usually be computed in fewer steps, offering high sampling efficiency.

In contrast, diffusion models characterize the stochastic evolution of data samples through stochastic dynamics: the forward process gradually adds noise to data, transforming the data distribution into a prior distribution, described by an SDE; the reverse process gradually denoises data, transforming the prior distribution back into the data distribution, also described by an SDE. Both processes are constrained by the Fokker-Planck equation. They learn the score field $\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x})$. Because the reverse process is stochastic, sampling typically requires more steps, although it also has an equivalent deterministic evolution form (the probability flow ODE).

Furthermore, in diffusion models, the reverse process also has a velocity field determined by the score function. Therefore, score matching can be seen as an indirect way of learning the velocity field within the framework of stochastic dynamics; this velocity field can be linked to the framework of flow matching models.

In summary, flow matching models and diffusion models each have their own characteristics in terms of approach, equations, learning objectives, constraints, and sampling efficiency, collectively illustrating the underlying dynamics and capabilities of probabilistic generative models (Table 35.1).

35.5 Summary

1. Basic concepts of flow matching models A sequence of probability distributions $\{p_t(\boldsymbol{x})\}_{t\in[0,1]}$ defined on the continuous time interval $t \in [0, 1]$ is called a probability path or probability flow. The probability flow starts from a prior distribution (e.g., a standard Gaussian distribution) p_0 at time $t = 0$, evolves continuously over time, becomes an intermediate distribution p_t at time t, and finally becomes the true data distribution p_1 at time $t = 1$.

The evolution of a sample $\boldsymbol{x}$ over time can be described by an ordinary differential equation (ODE):

$$\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = \boldsymbol{v}(\boldsymbol{x}, t)$$

where $\boldsymbol{v}(\boldsymbol{x}, t)$ is a time-dependent velocity field, indicating the direction and rate of sample evolution.

Under certain smoothness conditions, the velocity field $\boldsymbol{v}_t(\boldsymbol{x})$ together with the initial distribution p_0 uniquely determines a probability path $p_t(\boldsymbol{x})$. The velocity field is said to induce this probability path.

2. Continuity equation The continuity equation describes the conservation of mass during the evolution of a probability distribution:

$$\frac{\partial p_t(\boldsymbol{x})}{\partial t} + \nabla_{\boldsymbol{x}} \cdot \big(p_t(\boldsymbol{x})\boldsymbol{v}_t(\boldsymbol{x})\big) = 0$$

Under certain smoothness conditions, this equation is a necessary and sufficient condition for the velocity field $\boldsymbol{v}_t(\boldsymbol{x})$ to induce the probability path $p_t(\boldsymbol{x})$.

3. Core idea of the flow matching method The flow matching method constructs conditional velocity fields and conditional probability paths. It can be proven that the resulting marginal velocity field induces the target marginal probability path, and learning the target marginal velocity field is equivalent to learning the conditional velocity field.

3-1. Conditional probability path and conditional velocity field Given a data sample $\boldsymbol{x}_1$, define a conditional velocity field and conditional probability path, assuming the conditional velocity field induces the conditional probability path:

$$\boldsymbol{x}_0 \sim \pi(\cdot), \boldsymbol{x}_1 \sim q(\cdot), \frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = \boldsymbol{v}_t(\boldsymbol{x}|\boldsymbol{x}_1) \Rightarrow \boldsymbol{x} \sim p_t(\cdot|\boldsymbol{x}_1)$$

Set the initial and final conditional distributions as

$$p_0(\boldsymbol{x}|\boldsymbol{x}_1) = \pi(\boldsymbol{x}) \qquad p_1(\boldsymbol{x}|\boldsymbol{x}_1) = \delta_{\boldsymbol{x}_1}(\boldsymbol{x})$$

where $\delta_{\boldsymbol{x}_1}$ denotes the Dirac distribution centered at $\boldsymbol{x}_1$.

Define the marginal distribution

$$p_t(\boldsymbol{x}) = \int p_t(\boldsymbol{x}|\boldsymbol{x}_1)\ q(\boldsymbol{x}_1)\,\mathrm{d}\boldsymbol{x}_1$$

and the marginal velocity field

$$\boldsymbol{v}_t(\boldsymbol{x}) = \int \boldsymbol{v}_t(\boldsymbol{x}|\boldsymbol{x}_1)\frac{p_t(\boldsymbol{x}|\boldsymbol{x}_1)q(\boldsymbol{x}_1)}{p_t(\boldsymbol{x})}\,\mathrm{d}\boldsymbol{x}_1$$

Under these conditions, the marginal velocity field induces the marginal probability path:

$$\boldsymbol{x}_0 \sim \pi(\cdot), \frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = \boldsymbol{v}_t(\boldsymbol{x}) \Rightarrow \boldsymbol{x} \sim p_t(\cdot)$$

3-2. Learning objective for conditional flow matching Use a neural network $\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x}, t)$ to fit the conditional velocity field. The learning objective is

$$L_{\mathrm{CFM}}(\boldsymbol{\theta}) = \mathbb{E}_{t,\boldsymbol{x},\boldsymbol{x}_1} \left\| \boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x}, t) - \boldsymbol{v}_t(\boldsymbol{x}|\boldsymbol{x}_1) \right\|_2^2$$

It can be proven that

$$\min_{\boldsymbol{\theta}} L_{\mathrm{FM}}(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} L_{\mathrm{CFM}}(\boldsymbol{\theta})$$

where $L_{\mathrm{FM}}(\boldsymbol{\theta})$ is the objective when learning the marginal velocity field.

3-3. Gaussian conditional probability path A common choice is the Gaussian conditional probability path:

$$\boldsymbol{x}_0 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \qquad \boldsymbol{x}_1 \sim q(\cdot), \qquad \boldsymbol{x} = \alpha_t \boldsymbol{x}_1 + \beta_t \boldsymbol{x}_0$$

where $\alpha_0 = 0, \beta_0 = 1$, $\alpha_1 = 1, \beta_1 = 0$.

Its conditional velocity field is

$$\boldsymbol{v}_t(\boldsymbol{x}|\boldsymbol{x}_1) = \left(\dot{\alpha}_t - \frac{\dot{\beta}_t}{\beta_t}\alpha_t\right)\boldsymbol{x}_1 + \frac{\dot{\beta}_t}{\beta_t}\boldsymbol{x}$$

4. Flow matching algorithm with linear interpolation In the Gaussian conditional path, a common choice is $\alpha_t = t, \beta_t = 1 - t$, i.e., the linear interpolation path

$$\boldsymbol{x} = t\boldsymbol{x}_1 + (1 - t)\boldsymbol{x}_0$$

The optimization objective becomes

$$\mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{x}_1} \left\| \boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x}, t) - (\boldsymbol{x}_1 - \boldsymbol{x}_0) \right\|_2^2$$

Sampling is performed by numerical integration

$$\boldsymbol{x}_{t_{k+1}} = \boldsymbol{x}_{t_k} + \Delta t\, \boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x}_{t_k}, t_k)$$

to compute the generated sample.

5. Diffusion models and the probability flow ODE Consider the forward diffusion process

$$\mathrm{d}\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{x}, t)\,\mathrm{d}t + g(t)\,\mathrm{d}\boldsymbol{w}_t$$

Its probability distribution satisfies the Fokker-Planck equation:

$$\frac{\partial}{\partial t} p_t(\boldsymbol{x}) = -\nabla_{\boldsymbol{x}} \cdot \big(p_t(\boldsymbol{x})\boldsymbol{f}(\boldsymbol{x}, t)\big) + \frac{g^2(t)}{2} \nabla^2_{\boldsymbol{x}} p_t(\boldsymbol{x})$$

After learning the score field $\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x})$, the probability flow ODE

$$\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = \boldsymbol{f}(\boldsymbol{x}, t) - \frac{1}{2} g^2(t) \nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x})$$

represents the deterministic evolution process of its probability distribution.

Further Reading

The original work on flow matching models can be found in [1], and a significant subsequent improvement includes Rectified Flow [2]. Methodologically, flow matching models are closely related to existing generative model techniques, including Normalizing Flows [3], Continuous Normalizing Flows (CNF) [4], SDE-based diffusion models and their corresponding probability flow ODE [5], and Denoising Diffusion Implicit Models (DDIM) [6]. For review and introductory work on flow matching models, see [7–8].

Exercises

35.1 Write down the ODE for a flow matching model, and define the velocity field and flow, explaining their roles in the flow matching algorithm.

35.2 Suppose the flow $\boldsymbol{\psi}_t(\boldsymbol{x}_0)$ is a diffeomorphism described by an ODE. If two different initial positions $\boldsymbol{x}_0$ and $\boldsymbol{x}_0'$ evolve to the same position $\boldsymbol{x}$ at the same time t, which property of the ODE does this violate?

35.3 In the flow matching algorithm, assuming the path is a Gaussian conditional path, propose a design for a non-linear path and discuss its potential advantages and disadvantages.

35.4 Let the velocity field be a constant vector $\boldsymbol{v}_t(\boldsymbol{x}) = \boldsymbol{a}$, and the prior distribution be $p_0(\boldsymbol{x}) \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. (1) Find the analytical expression for the distribution $p_t(\boldsymbol{x})$ at time t; (2) Verify that this distribution satisfies the continuity equation; (3) Draw schematic diagrams of the distribution at times $t = 0, 0.5, 1.0$ (for one and two dimensions).

35.5 Flow and probability flow are different concepts. List their differences.

35.6 Prove the computational formula (35.23) for the Gaussian conditional velocity field.

35.7 The general form of the stochastic differential equation (SDE) for a diffusion process is

$$\mathrm{d}\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{x}, t)\mathrm{d}t + \boldsymbol{G}(\boldsymbol{x}, t)\mathrm{d}\boldsymbol{w}$$

where $\boldsymbol{f}(\boldsymbol{x}, t)$ is the drift coefficient, $\boldsymbol{G}(\boldsymbol{x}, t)$ is the diffusion coefficient, and $\mathrm{d}\boldsymbol{w}$ is a standard Wiener process. The drift coefficient is a vector, and the diffusion coefficient is a matrix. Derive the corresponding general form of the Fokker-Planck equation.

$$\frac{\partial p_t(\boldsymbol{x})}{\partial t} = -\nabla \cdot (\boldsymbol{f}(\boldsymbol{x}, t)p_t(\boldsymbol{x})) + \frac{1}{2}\nabla \cdot \nabla \cdot \big(\boldsymbol{G}(\boldsymbol{x}, t)\boldsymbol{G}^\top(\boldsymbol{x}, t)p_t(\boldsymbol{x})\big)$$

35.8 The stochastic differential equation form of Langevin dynamics is

$$\mathrm{d}\boldsymbol{x} = -\frac{1}{\gamma}\nabla_{\boldsymbol{x}} U(\boldsymbol{x})\mathrm{d}t + \sqrt{\frac{2k_B T}{\gamma}}\mathrm{d}\boldsymbol{w} \tag{35.30}$$

where γ is the friction coefficient, $U(\boldsymbol{x})$ is the potential energy function, $\nabla_{\boldsymbol{x}} U(\boldsymbol{x})$ represents the conservative force, k_B is Boltzmann's constant, T is the temperature, and $\mathrm{d}\boldsymbol{w}$ is a Wiener process. Write down the corresponding Fokker-Planck equation.

References

[1] LIPMAN Y, CHEN R T, BEN-HAMU H, et al. Flow matching for generative modeling[C]//International Conference on Learning Representations, 2023.

[2] LIU X, GONG C, et al. Flow straight and fast: Learning to generate and transfer data with rectified flow[C]//International Conference on Learning Representations, 2023.

[3] KOBYZEV I, PRINCE S J, BRUBAKER M A. Normalizing flows: An introduction and review of current methods[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 43(11): 3964–3979.

[4] CHEN R T, RUBANOVA Y, BETTENCOURT J, et al. Neural ordinary differential equations[C]//Advances in Neural Information Processing Systems, 2018.

[5] SONG Y, SOHL-DICKSTEIN J, KINGMA D P, et al. Score-based generative modeling through stochastic differential equations[C]//International Conference on Learning Representations, 2021.

[6] SONG J, MENG C, ERMON S. Denoising diffusion implicit models[C]//International Conference on Learning Representations, 2021.

[7] LIPMAN Y, HAVASI M, HOLDERRIETH P, et al. Flow matching guide and code[J]. arXiv preprint arXiv:2412.06264, 2024.

[8] HOLDERRIETH P, ERIVES E. An introduction to flow matching and diffusion models[J]. arXiv preprint arXiv:2506.02070, 2025.