Flow Matching Notes

Jonathan Ma
johnma@udel.edu

March 6, 2025

Contents

Preface

Notes for https://arxiv.org/pdf/2412.06264, sections 3, 4. I have handwritten notes for section 1, 2, but I haven't included them here. There should be code in a repo I will link here with this document included: https://github.com/jma02/flow-matching.

1 Flow Models

Flow matching (FM) is very similar to the ideas of DDPM, DDIM we have been discussing in our past meetings, in fact, we have been told by our other paper (https://arxiv.org/abs/2406.08929) that DDIM is a special case of FM. Note that Lipman's guide is agnostic of knowledge of either DDPM, DDIM, but I will try to connect the ideas of FM, DDPM, DDIM as best as I can. Lipman argues three motivators for why FM is advantageous as a generative learning paradigm.

1. Flows are a simple continuous time Markov process, being deterministic, and having "compact parametrization via velocities," which roughly means that flow models can transform any source distribution p into any target distribution q given that the two have densities.

2. Flows can be sampled efficiently by approximating the solutions of ODEs, compared to the harder-to-simulate SDE we've discussed in DDPM/DDIM.

3. The deterministic nature of flows facilitates more unbiased model likelihood estimation, while more general stochastic processes require working with lower bounds. (?)

1.1 Preliminary Knowledge

Consider data in the d-dimensional Euclidean space, $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ with the standard Euclidean inner product and norm. We will consider random variables (RVs) $X \in \mathbb{R}^d$. We will omit the integration interval when integrating over the whole space, that is $\int \equiv \int_{\mathbb{R}^d}$. We will refer to the PDF p_{X_t} of the random variable X_t as simply p_t. A common PDF in generative modeling is the d-dimensional isotropic Gaussian:

$$N(x \mid \mu, \sigma^2 I) = (2\pi\sigma^2)^{-d/2} \exp\left(-\frac{\|x-\mu\|_2^2}{2\sigma^2}\right),$$

where $\mu \in \mathbb{R}^d, \sigma \in \mathbb{R}_{>0}$, standing for the mean and standard deviation of the distribution, respectively. The following are particularly important for our discussion. Recall from probability that the conditional expectation $\mathbf{E}\left[X \mid Y\right]$ is the best approximating function $g_*(Y)$ to X in the least-squares sense:

$$\begin{aligned} g_* &:= \underset{g:\mathbb{R}^d \to \mathbb{R}^d}{\operatorname{argmin}} \mathbf{E}\left[\|X - g(Y)\|^2\right] = \underset{g:\mathbb{R}^d \to \mathbb{R}^d}{\operatorname{argmin}} \int \|x - g(y)\|^2 p_{X,Y}(x,y)\, dx\, dy \\ &= \underset{g:\mathbb{R}^d \to \mathbb{R}^d}{\operatorname{argmin}} \int \left(\int \|x - g(y)\|^2 p_{X|Y}(x \mid y)\, dx \right) p_Y(y)\, dy. \end{aligned} \tag{*}$$

For $y \in \mathbb{R}^d$ such that $p_Y(y) > 0$, the conditional expectation function is therefore

$$\mathbf{E}\left[X \mid Y = y\right] := g_*(y) = \int x p_{X|Y}(x \mid y)\, dx,$$

where the second equality follows from taking the minimizer of the inner integral in $(*)$ for $Y = y$. Composing g_* with Y, we get

$$\mathbf{E}\left[X \mid Y\right] := g_*(Y),$$

which is a random variable in $\mathbb{R}^d$. Note that $\mathbf{E}\left[X \mid Y = y\right]$ is a function $\mathbb{R}^d \to \mathbb{R}^d$, while $\mathbf{E}\left[X \mid Y\right]$ is a random variable assuming values in $\mathbb{R}^d$. Also, recall the tower property:

$$\mathbf{E}\left[\mathbf{E}\left[X \mid Y\right]\right] = \mathbf{E}\left[X\right],$$

which will be helpful later.

1.2 Diffeomorphisms, Push-Forward Maps

Denote by $C^r(\mathbb{R}^m, \mathbb{R}^n)$, the collection of functions $f : \mathbb{R}^m \to \mathbb{R}^n$ with continuous partial derivatives of order r. We will write $C^r(\mathbb{R}^n) : C^r(\mathbb{R}^m, \mathbb{R})$ to be the space of such functionals. An important class of functions are the C^r diffeomorphisms. These are invertible functions $\psi \in C^r(\mathbb{R}^n, \mathbb{R}^n)$ with $\psi^{-1} \in C^r(\mathbb{R}^n, \mathbb{R}^n)$.

Now, given a RV $X \sim p_X$, (with density p_X), let us consider a RV $Y = \psi(X)$, where $\psi : \mathbb{R}^d \to \mathbb{R}^d$ is a C^1 diffeomorphism. The PDF of Y, denoted p_Y is called the *push-forward* of p_X. Then the PDF p_Y can be computed via a change of variables:

$$\mathbf{E}\left[f(Y)\right] = \mathbf{E}\left[f(\psi(X))\right] = \int f(\psi(x))p_X(x)\,dx = \int f(y)p_X(\psi^{-1}(y))|\det \partial_y \psi^{-1}(y)|\,dy,$$

where the third equality is due to the change of variables $x = \psi^{-1}(y)$, and $\partial_y(\psi^{-1}(y))$ denotes the Jacobian matrix of first partial derivatives of $\psi^{-1}(y)$. Therefore, we conclude that the PDF of Y is

$$p_Y(y) = p_X(\psi^{-1}(y))|\det \partial_y \psi^{-1}(y)|.$$

We will denote the push-forward operator with the symbol $\sharp$, that is

$$[\psi_\sharp p_X](y) := p_X(\psi^{-1}(y))|\det \partial_y \psi^{-1}(y)|. \tag{Push-Forward}$$

1.3 Flows as Generative Models

The goal of generative modeling is to transform samples $X_0 = x_0$ from a source distribution p into samples $X_1 = x_1$ from a target distribution q. Let's start building the tools neceasary to address this problem by means of a flow mapping ψ_t.

Definition 1.1. A C^r flow is a time dependent mapping $\psi : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ implementing $\psi : (t,x) \mapsto \psi_t(x)$. Also, $\psi \in C^r([0,1] \times \mathbb{R}^d, \mathbb{R}^d)$ function, such that the function $\psi_t(x)$ is a C^r diffeomorphism in x, for all $t \in [0,1]$.

Definition 1.2. A flow model is a continuous-time Markov process $(X_t)_{0 \le t \le 1}$ defined by applying a flow ψ_t to the RV X_0 :

$$X_t = \psi_t(X_0), \quad t \in [0,1], \text{ where } X_0 \sim p.$$

To see why X_t is Markov, note that for any choice of $0 \le t < s \le 1$, we have

$$X_s = \psi_s(X_0) = \psi_s(\psi_t^{-1}(\psi_t(X_0))) = \psi_{s|t}(X_t),$$

where we have defined $\psi_{s|t} := \psi_s \circ \psi_t^{-1}$, which we claim is also a diffeomorphism. (It shouldn't be hard to prove this – Jonathan.) Therefore, $X_s = \psi_{s|t}(X_t)$ implies that states later than X_t depend only on X_t. Thus, (X_t) is Markov. In fact, for flow models, this dependence is deterministic. In summary, the goal of generative flow modeling is to find a flow ψ_t such that $X_1 = \psi_1(X_0) \sim q$.

1.3.1 Equivalence Between Flows and Velocity Fields

A C^r flow ψ can be defined in terms of a $C^r([0,1] \times \mathbb{R}^d, \mathbb{R}^d)$ velocity field $u : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ implementing $u : (t,x) \mapsto u_t(x)$ via the following ODE:

$$\frac{d}{dt}\psi_t(x) = u_t(\psi_t(x)), \quad \psi_0(x) = x. \tag{Flow-ODE}$$

Theorem 1.3 (Flow local existence and uniqueness)**.** *If u is $C^r([0,1] \times \mathbb{R}^d, \mathbb{R}^d), r \geq 1$ (in particular, locally Lipschitz), then the ODE described above has a unique solution which is a $C^r(\Omega, \mathbb{R}^d)$ diffeomorphism $\psi_t(x)$ defined over an open set Ω, which is a super-set of $\{0\} \times \mathbb{R}^d$.*

The proof of this is due to several authors which Lipman cites. Note that this theorem guarantees only the local existence and uniqueness of a C^r flow moving each point $x \in \mathbb{R}^d$ by $\psi_t(x)$ during a potentially limited amount of time $t \in [0, t_x)$. To guarantee a solution until $t = 1$, for all $x \in \mathbb{R}^d$, one must place additional assumptions beyond local Lipschitzness. For instance, one caould consider global Lipschitzness, guaranteed by bounded first derivatives in the C^1 case. We will later rely on a different condition, namely integrability, to guarantee the existence of the flow almost everywhere and until $t = 1$.

So far, we have shown a velocity field uniquely defines a flow. Conversely, given a C^1 flow psi_t, one can extract its defining velocity field $u_t(x)$ for arbitrary $x \in \mathbb{R}^d$, by considering the equation $\frac{d}{dt}\psi_t(x') = u_t(\psi_t(x'))$, and using the fact that ψ_t is an diffeomorphism, thus, invertible for every $t \in [0,1]$, and let $x' = \psi_t^{-1}(x)$. Thus, the unique velocity field u_t determining ψ_t is

$$u_t(x) = \dot{\psi}_t(\psi_t^{-1}(x)),$$

where $\dot{\psi}_t := \frac{d}{dt}\psi_t$. In conclusion, we have shown the equivalence between C^r flows ψ_t and C^r velocity fields u_t.

1.3.2 Computing Target Samples From Source Samples

Computing a target sample X_1, or in general, any sample X_t, entails approximating the solution of the ODE in Eq. (Flow-ODE). The simplest of these approximation schemes is the Euler method:

$$X_{t+h} = X_t + hu_t(X_t),$$

where $h = n^{-1} > 0$ is a step size hyper-parameter, with $n \in \mathbb{Z}^+$. To draw a sample X_1 from the target distribution, apply the Euler method starting at some $X_0 \sim p$, to produce the sequence $X_h, X_{2h}, \ldots, X_1$. The Euler method coincides with the first-order Taylor expansion of X_t :

$$X_{t+h} = X_t + h\dot{X}_t + o(h) = X_t + hu_t(X_t) + o(h),$$

showing that the Euler method accumulates $o(h)$ error per step. The code in the repo linked in the preface uses the midpoint method, which is a second-order method, which often outperforms the Euler method in practice.

1.4 Probability Paths and the Continuity Equation

We call a time-dependent probability $(p_t)_{0 \leq t \leq 1}$ a probability path. For our purposes, an important probability path is the marginal PDF of a flow model $X_t = \psi_t(X_0)$ at time t:

$$X_t \sim p_t.$$

For each time $t \in [0,1]$, these marginal PDFs are obtained via the push-forward formula, Eq. (Push-Forward):

$$p_t(x) = [\psi_{t\sharp p}](x).$$

Definition 1.4. Given some arbitrary probability path p_t, we say u_t generates p_t if $X_t = \psi_t(X_0) \sim p_t$, for all $t \in [0, 1)$.

In this way, we establish a close relationship between velocity fields, their flows, and the generated probability paths. We use the time interval $[0, 1)$ to allow dealing with target distributions q with compact support where the velocity is not defined precisely at $t = 1$.

To verify that a velocity field u_t generates a probability path p_t, one can verify if the pair (u_t, p_t) satisfies a PDE known as the continuity equation:

$$\frac{d}{dt}p_t(x) + \operatorname{div}(p_t u_t)(x) = 0. \tag{Continuity-Eq}$$

The following theorem, a rephrased version of the mass conservation formula, states that a solution u_t to the continuity equation generates the probability path p_t.

Theorem 1.5 (Mass Conservation)**.** *Let (p_t) be a probability path, and u_t a locally Lipschitz integrable vector field. Then TFAE:*

1. *Equation* (Continuity-Eq) *holds, for $t \in [0, 1)$.*
2. *The vector field u_t generates p_t.*

Local Lipschitzness assumes that there exists a local neighborhood over which $u_t(x)$ is Lipschitz, for all (t, x). Assuming u is integrable means that

$$\int_0^1 \int \|u_t(x)\| p_t(x)\, dx\, dt < \infty.$$

Specifically, integrating a solution to the flow ODE Eq. (Flow-ODE) across times $[0, t]$ leads to the integral equation

$$\psi_t(x) = x + \int_0^t u_s(\psi_s(x))\, ds.$$

Hence,

$$\begin{aligned}
\mathbf{E}\left[\|X_t\|\right] &= \int \|\psi_t(x)\| p(x)\, dx \\
&= \int \left\| x + \int_0^t u_s(\psi_s(x))\, ds \right\| p(x)\, dx \\
&\leq \mathbf{E}\left[\|X_0\|\right] + \int_0^1 \int \|u_s(x)\| p_t(x)\, dt \\
&< \infty,
\end{aligned}$$

where we've used the triangle inequality in the second to last inequality, and last inequality holds by the integrebiliy condition, and $\mathbf{E}\left[\|X_0\|\right] < \infty$. In sum, integrability allows assuming that X_t has bounded expected norm, if X_0 also does.

1.4.1 Continuity Equation

We can rewrite the continuity equation in integral form using the divergence theorem. For some domain D, and some smooth vector field $u : \mathbb{R}^d \to \mathbb{R}^d$, accumulating the divergences of u inside D equals the flux leaving D by orthogonally crossing its boundary ∂D :

$$\int_D \operatorname{div}(u)(x)\, dx = \int_{\partial D} \langle u(y), n(y) \rangle\, ds_y,$$

where $n(y)$ is a unit norm field pointing outward to the domain's boundary ∂D, and ds_y is the boundary's area element. To apply these insights to the continuity equation, let us integrate Eq. (Continuity-Eq) over a small domain $D \subset \mathbb{R}^d$, and apply the divergence theorem to obtain

$$\frac{d}{dt} = \int_D p_t(x)\,dx = -\int_D \operatorname{div}(p_t u_t)(x)\,dx = -\int_{\partial D} \langle p_t(y)u_t(y), n(y)\rangle\,ds_y.$$

This equation expresses the rate of change of total probability mass in the volume D as the negative probability flux leaving the domain.

1.5 Instantaneous Change of Variables

An important benefit of using flows as generative models is that they allow tractable computation of exact likelihoods $\log p_1(x)$ for all $x \in \mathbb{R}^d$. This is a consequence of the continuity equation, called instantaneous change of variables:

$$\frac{d}{dt}\log p_t(\psi_t(x)) = -\operatorname{div}(u_t)(\psi_t(x)). \tag{ICoV}$$

This is the ODE governing change in log-likelihood, $\log p_t(\psi_t(x))$, along a sampling trajectory $\psi_t(x)$, defined by the flow ODE Eq. (Flow-ODE). To derive Eq. (ICoV), differentiate $\log p_t(\psi_t(x))$ with respect to time, and apply the continuity equation Eq. (Continuity-Eq) and the flow ODE Eq. (Flow-ODE). Integrating Eq. (ICoV) from $t = 0$ to $t = 1$, and rearranging yields

$$\log p_1(\psi_1(x)) = \log p_0(\psi_0(x)) - \int_0^1 \operatorname{div}(u_t)(\psi_t(x))\,dt. \tag{**}$$

In practice, computing $\operatorname{div}(u_t)$, which is the trace of the Jacobian $\partial_x u_t(x) \in \mathbb{R}^{d\times d}$ is increasingly challenging as d grows. Because of this, previous work employ unbiased estimators, such as Hutchinson's trace estimator:

$$\operatorname{div}(u_t)(x) = \operatorname{tr}\partial_x u_t(x) = \mathbf{E}_Z \operatorname{tr}(Z^T \partial_x u_t(x) Z),$$

where $Z \in \mathbb{R}^{d\times d}$ is any random variable with $\mathbf{E}[Z] = 0$, and $\operatorname{Cov}(Z, Z) = I$. Plugging the equation above into $(**)$ and switching the order of integration and expectation, we obtain the following unbiased log-likelihood estimator:

$$\log p_1(\psi_1(x)) = \log p_0(\psi_0(x)) - \mathbf{E}_Z \int_0^1 \operatorname{tr}(Z^T \partial_x u_t(\psi_t(x)) Z)\,dt.$$

For a fixed sample Z, this above equation can be done with a single backard pass via a vector-Jacobian product (JVP).

In summary, computing an unbiased estimate of $\log p_1(x)$ entails simulating the ODE

$$\frac{d}{dt}\begin{bmatrix} f(t) \\ g(t) \end{bmatrix} = \begin{bmatrix} u_t(f(t)) \\ -\operatorname{tr}(Z^T \partial_x u_t(f(t)) Z \end{bmatrix}, \quad \begin{bmatrix} f(1) \\ g(1) \end{bmatrix} = \begin{bmatrix} x \\ 0 \end{bmatrix}, \tag{Log-Flow-ODE}$$

backwards in time, from $t = 1$, to $t = 0$, and setting

$$\widehat{\log p_1}(x) = \log p_0(f(0)) - g(0).$$

1.6 Training Flow Models With Simulation

The work we've done with instantaneous change of variables, and the resulting ODE system allows training a flow model by maximizing the log-likelihood of training data. Specifically, let u_t^θ be a velocity field with learnable parameters $\theta \in \mathbb{R}^p$, and consider the problem of learning θ such that

$$p_1^\theta \approx q.$$

We can pursue this goal by minimizing the KL-divergence of p_1^θ and q:

$$\mathcal{L}(\theta) = D_{KL}(q, p_q^\theta) = -\mathbf{E}\left[Y \sim q\right] \log {p_1}^\theta(Y) + O(1),$$

where p_1^θ is the distribution of $X_1 = \psi_1^\theta(X_0)$, ψ_t^θ is defined by u_t^θ, and we can obtain an unbiased estimate of $\log p_1^\theta(Y)$ via a solution to Eq. (Log-Flow-ODE). Computing this loss, as well as its gradients, requires precise ODE simulations during training, whereas only errorless solutions constitute unbiased gradients. We will now present Flow Matching (FM) as a simulation-free framework for training flow generative models, without the need of solving ODEs during training.

2 Flow Matching

We claim Flow Matching (FM) is a scalable approach for training a flow model, defined by a learnable velocity u_t^θ, and solving the Flow Matching Problem:

Find u_t^θ generating p_t, with $p_0 = p$, and $p_1 = q$.

2.1 Data

Let source samples be a RV $X_0 \sim p$, and target samples be a RV $X_1 \sim q$. Typically, source samples follow an easy-to-sample, known distribution, and target samples are given in terms of a finite-sized dataset. Target samples might constitute images, videos, audio segments, or other types of high-dimensional, richly structured data. Source and target samples can be independent, or they can originate from a general joint distribution known as coupling:

$$(X_0, X_1) \sim \pi_{0,1}(X_0, X_1), \qquad \text{(Coupling)}$$

where, if no coupling is known, the source-target samples are following the independent coupling $\pi_{0,1}(X_0, X_1) = p(X_0)q(X_1)$.

Example 2.1. The case of producing images from random Gaussian noise vectors is an example of independent source-target distribution.

Example 2.2. The production of high-resulution images from their low resolution versions, or producing colorized versions from their gray-scale counterparts, are examples of dependent coupling.

2.2 Building Probability Paths

Flow Matching drastically simplifies designing probability paths (p_t), along with corresponding velocity fields (u_t) by adopting a conditional strategy. As a first example, consider conditioning the design of p_t on a single target example, $X_1 = x_1$, yielding the conditional probability path

$p_{t|1}(x \mid x_1)$. Then we may construct the overall marginal probability path (p_t) by aggregating such conditional probability paths $p_{t|1}$:

$$p_t(x) = \int p_{t|1}(x \mid x_1) q(x_1)\, dx_1. \tag{Marginal-Path}$$

To solve the Flow Matching problem, we would like p_t to satisfy the following boundary conditions:

$$p_0 = p, \quad p_1 = q, \tag{Boundary-Cond}$$

that is, the marginal probability path p_t interpolates from the source distribution p at time $t = 0$ to the target distribution q at time $t = 1$. These boundary conditions can be enforced by requiring the conditional probability paths to satisfy

$$p_{0|1}(x \mid x_1) = \pi_{0|1}(x \mid x_1) \text{ and } p_{1|1}(x \mid x_1) = \delta_{x_1}(x),$$

where the conditional coupling $\pi_{0|1}(x_0, x_1)/q(x_1)$, and δ_{x_1} is the delta measure centered at x_1. For independent coupling $\pi_{0,1}(x_0, x_1) = p(x_0)q(x_1)$, the first constraint above reduces to $p_{0|1}(x \mid x_1) = p(x)$. Because the delta measure does not have a density, the second constraint should be read as $\int p_{t|1} f(y)\, dy \to f(x)$ as $t \to 1$, for continuous functions f. The boundary conditions can be verified by plugging Eq. (Boundary-Cond) into Eq. (Marginal-Path).

A popular example of a conditional probability path satisfying the conditions in Eq. (Boundary-Cond) is

$$p_t(x) = \int p_{t|1}(x \mid x_1)\, dx_1, \text{ where } p_{t|1}(x \mid x_1) = N(x \mid tx_1, (1-t)^2 I, \tag{Conditional-OT}$$

which is called the conditional optimal transport:

$$N(\cdot \mid tx_1, (1-t)^2 I) \to \delta_{x_1}(\cdot), \text{ as } t \to 1.$$

2.3 Deriving Generating Velocity Fields

Given a marginal probability path (p_t), we now build a velocity field u_t generating p_t. The generating velocity field u_t is an average of multiple conditional velocity fields $u_t(x \mid x_1)$, satisfying

$$u_t(\cdot \mid x_1) \text{ generates } p_{t|1}(\cdot \mid x_1).$$

Then the marginal velocity field $u_t(x)$ is given by averaging the conditional velocity fields $u_t(x \mid x_1)$ across target examples:

$$u_t(x) = \int u_t(x \mid x_1) p_{1|t}(x_1 \mid x)\, dx_1. \tag{Marginal-VF}$$

To express the equation above using known terms, using Bayes' rule,

$$p_{1|t}(x_1 \mid x) = \frac{p_{t|1}(x \mid x_1) q(x_1)}{p_t(x)},$$

defined for all x with $p_t(x) > 0$. Equation (Marginal-VF) can be interpreted as the weighted average of the conditional velocities $u_t(x \mid x_1)$ with weights $p_{1|t}(x_1 \mid x)$ representing the posterior probability of target samples x_1 given the current sample x. Another interpretation of Eq. (Marginal-VF) can be given with conditional expectations. Namely, if X_t is any RV such that $X_t \sim p_{t|1}(\cdot \mid X_1)$, or equivalently, the joint distribution of (X_t, X_1) has density $p_{t,1}(x, x_1) = p_{t|1}(x \mid x_1)q(x_1)$, then LOTUS to write Eq. (Marginal-VF) as a conditional expectation, we obtain

$$u_t(x) = \mathbf{E}\left[u_t(X_t \mid X_1 \mid X_t = x\right], \tag{Marginal-CE}$$

which yields the useful interpretation of $u_t(x)$ as the least squares approximation to $u_t(X_t \mid X_1)$ given $X_t = x$. Note that X_t in *Eq.* (Marginal-CE) is in general a different RV than X_t defined by the final flow model, but they share the same marginal probability $p_t(x)$.

2.4 General Conditioning and the Marginalization Trick

To justify the constructions above, we need to show that the marginal velocity field u_t from Eq. (Marginal-VF) and Eq. (Marginal-CE) generates the marginal probability path (p_t) under mild assumptions. The mathematical tool we use to prove this is the mass conservation theorem (Theorem 1.5). To proceed, consider a more general setting, that is, consider conditioning any arbitrary RV $Z \in \mathbb{R}^m$, with PDF p_Zm yielding the marginal probability path

$$p_t(x) = \int p_{t|Z}(x \mid z)p_Z(z)\,dz,$$

which in turn is generated by the marginal velocity field

$$u_t(x) = \int u_t(x \mid z)p_{Z|t}(z \mid x)\,dz = \mathbf{E}\left[u_t(X_t \mid Z) \mid X_t = x\right],$$

where $u_t(\cdot \mid z)$ generates $p_{t|Z}(\cdot \mid z)$, and $p_{Z|t}(z \mid x) = \frac{p_{t|Z}(x|z)p_Z(z)}{p_t(x)}$ follows from Bayes' rule given $p_t(x) > 0$, and $X_t \sim p_{t|Z}(\cdot \mid Z)$. Naturally, we can recover the constructions in previous sections by setting $Z = X_1$. Before we prove the main result, we need some regularity assumptions, encapsulated as follows.

Assumption 1. Assume $p_{t|Z}(x \mid z)$ is $C^1([0,1) \times \mathbb{R}^d)$ and $u_t(x \mid z)$ is $C^1([0,1) \times \mathbb{R}^d, \mathbb{R}^d)$ as a function of (t, x). Furthermore, suppose p_Z has bounded support, that is, $p_Z(x) = 0$ outside some bounded set in $\mathbb{R}^m$. Finally, assume $p_t(x) > 0$, for all $x \in \mathbb{R}^d$, and $t \in [0, 1)$.

These are mild assumptions. One can show that $p_t(x) > 0$ by finding a condition z such that $p_Z(z) > 0$, and $p_{t|Z}(\cdot \mid z) > 0$. In practice, one can satisfy this by considering $(1 - (1-t)\varepsilon)p_{t|Z} + (1-t)\varepsilon N(0, I)$ for an arbitrarily small $\varepsilon > 0$. One example of $p_{t|Z}(\cdot \mid z)$ satisfying this assumption is the path in Eq. (Conditional-OT) where $Z = X_1$. We are now ready to state the main result.

Theorem 2.3. *Under the previous assumption, if $u_t(x \mid z)$ is conditionally integrable, and generates the conditional probability path $p_t(\cdot \mid z)$, then the marginal velocity field u_t generates the marginal probability path p_t, for all $t \in [0, 1)$.*

In the theorem above, conditionally integrable refers to a conditional version of the integrability condition from Theorem 1.5, namely:

$$\int_0^1 \iint \|u_t(x \mid z)\| p_{t|Z}p_Z(x)\,dz\,dx\,dt < \infty.$$

I need to save time, so I will not repeat any proofs from this point forward. Again, the paper we are following is https://arxiv.org/pdf/2412.06264.

2.5 Flow Matching Loss

After having established that the target velocity field u_t generates the prescribed probability path p_t from p to q, the missing ingredient is a tractable loss function to learn a velocity field model u_t^θ as close as possible to the target u_t. One major roadblock towards stating this loss function directly is that computing the target u_t is unfeasible, as it requires marginalizing over the entire training set. Fortunately, a family of loss functions known as Bregman divergences provide unbiased gradients to learn $u_t^\theta(x)$ in terms of conditional velocities $u_t(x \mid z)$ alone.

Bregman divergences measure dissimilarity between two vectors $u, v \in \mathbb{R}^d$ as

$$D(u,v) := \Phi(u) - [\Phi(v) + \langle u - v, \nabla\Phi(v)\rangle],$$

where $\Phi : \mathbb{R}^d \to \mathbb{R}$ is a strictly convex function defined over some convex set $\Omega \subset \mathbb{R}^d$. Bregman divergence is important since their gradient with respect to the second argument is affine invariant:

$$\nabla_v D(au_1 + bu_1, v) = a\nabla_v D(u_1, v) + b\nabla_v D(u_2, v), \quad \text{for any } a + b = 1.$$

Affine invariance allows us to swap expected values with gradients:

$$\nabla_v D(\mathbf{E}[y], v) = \mathbf{E}[\nabla_v D(Y, v)], \quad \text{for any RV } Y \in \mathbb{R}^d.$$

The flow matchinbg loss employs a Bregman divergence to regress our learnable velocity $u_t^\theta(x)$ onto the target velocity $u_t(x)$ along the probability path p_t :

$$\mathcal{L}_{FM}(\theta) = \mathbf{E}_{t, X_t \sim p_t} D(u_t(X_t), u_t^\theta(X_t)),$$

where time $t \sim U[0,1]$. We mentioned before that the target velocity u_t is not tractable. It cannot be computed as is. Instead, we consider the simpler and tractable Conditional Flow Matching (CFM) loss:

$$\mathcal{L}_{CFM}(\theta) = \mathbf{E}_{t, Z, X_t \sim p_{t|Z}(\cdot|Z)} D(u_t(X_t \mid Z), u_t^\theta(X_t)). \tag{CFM-Loss}$$

The two losses are equivalent since their gradients coincide, and the proof of this follows by using the chain rule, Bayes' rule, and the affine invariance of the Bregman divergence.

The following theorem is generalization of the equivalence of the $\mathcal{F}_F M, \mathcal{L}_{CFM}$ losses, also used later in the manuscript.

Proposition 2.4 (Bregman divergence for learning conditional expectations)**.** *Let $X \in \mathcal{S}_X$, $Y \in \mathcal{S}_Y$ be RVs over state spaces $\mathcal{S}_X, \mathcal{S}_Y$, and $g : \mathbb{R}^p \times \mathcal{S}_X \to \mathbb{R}^n$, $(\theta, x) \mapsto g^\theta(x)$, where $\theta \in \mathbb{R}^p$ denotes learnable parameters. Let $D_x(u,v)$, $x \in \mathcal{S}_X$ be a Bregman divergence over a convex set $\Omega \subset \mathbb{R}^n$ that contains the image of f. Then*

$$\nabla_\theta \mathbf{E}_{X,Y} D_X(Y, g^\theta(X)) = \nabla_\theta \mathbf{E}_X(\mathbf{E}[Y \mid X], g^\theta(X)).$$

In particular, for all x with $p_X(x) > 0$, the global minimum of $g^\theta(x)$ w.r.t. θ satisfies

$$g^\theta(x) = \mathbf{E}[Y \mid X = x].$$

A useful variation of the FM loss is to sample times t from a distribution other than Uniform. Specifically, consider $t \sim \omega(t)$, where ω is a PDF over [0,1]. This leads to the weighted objective

$$\mathcal{L}_{CFM}(\theta) = \mathbf{E}_{t \sim \omega, Z, X_t} D(u_t(X_t \mid Z), u_t^\theta(X_t)) = \mathbf{E}_{t \sim U, Z, X_t} \omega(t) D(u_t(X_t \mid Z), u_t^\theta(X_t)).$$

ALthough mathematically equivalent, sampling $t \sim \omega$ leads to better performance than using weights $\omega(t)$ in larger scale image generation tasks.

2.6 Conditional Flows

We've reduced the problem of training a flow model u_t^θ to:

1. Find conditional probability paths $p_{t|Z}(x \mid z)$ yielding marginal probability paths $p_t(x)$ satisfying certain boundary conditions Eq. (Boundary-Cond).
2. Find conditional velocity fields $u_t(x \mid z)$ generating the conditional probability path.
3. Train using the conditional flow matching loss.

We will now discuss concrete options on how to do steps 1, 2. Specifically, we will discuss a flexible method to design conditional probability paths and velocity fields using a specific construction via conditional flows. The idea is as follows:

1. Define a flow model $X_{t|1}$ satisfying the boundary conditions Eq. (Boundary-Cond).
2. Extract the velocity field from $X_{t|1}$ by differentiation.

This process in turn defines both $p_{t|1}(x \mid x_1)$ and $u_t(x \mid x_1)$. In more detail, define the conditional flow model:

$$X_{t|1} = \psi_t(X_0 \mid x_1), \quad \text{where } X_0 \sim \pi_{0|1}(\cdot \mid x_1),$$

where $\psi : [0, 1) \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ is a conditional flow defined by

$$\psi_t(x \mid x_1) = \begin{cases} x & t = 0 \\ x_1 & t = 1, \end{cases} \tag{Conditional-Flow}$$

smooth in (t, x), and a diffeomorphism in x. (Smooth here means that all derivatives of $\psi_t(x \mid x_1)$ with respect to t and x exist and are continuous: $C^\infty([0, 1) \times \mathbb{R}^d, \mathbb{R}^d)$. These conditions can be further relaxed to $C^2([0, 1) \times \mathbb{R}^d, \mathbb{R}^d)$ at the expense of simplicity.) The push forward formula Eq. (Push-Forward) defines the probability density of $X_{t|1}$ as

$$p_{t|1}(x \mid x_1) := [\psi_t(\cdot \mid x_1)_\sharp \pi_{0|1}(\cdot \mid x_1)](x). \tag{Cond-Push-Forward}$$

This expression is unneeded in practical optimization of the CFM loss, instead it is used theoretically to show that $p_{t|1}$ satisfies the two boundary conditions Eq. (Boundary-Cond). According to its definition, $\psi_0(\cdot \mid x_1)$ is the identity map, keeping $\pi_{0|1}$ intact at time $t = 0$. Second, $\psi_1(\cdot \mid x_1) = x_1$ is the constant map, concentrating all probability mass at x_1 as $t \to 1$. Furthermore note that $\psi_t(\cdot \mid x_1)$ is a smooth diffeomorphism for $t \in [0, 1)$. Therefore, by the equivalence of flows and velocity fields, there exists a unique smooth conditional velocity field taking form:

$$u_t(x \mid x_1) = \dot{\psi}_t(\psi_t^{-1}(x \mid x_1) \mid x_1). \tag{Cond-Smooth-Vel}$$

To summarize, we have further reduced the task of finding the conditional path and a corresponding generating velocity to simply building a conditional flow $\psi_t(\cdot \mid x_1)$ satisfying Eq. (Conditional-Flow).

2.7 CFM Loss

We revisit the CFM loss Eq. (CFM-Loss) by setting $Z = X_1$, and using the conditional flows way of defining the conditional probability path and velocity:

$$\begin{aligned}\mathcal{L}_{CFM}(\theta) &= \mathbf{E}_{t,X_1,X_t\sim p_t(\cdot|X_1)} D(u_t(X_t \mid X_1), u_t^\theta(X_t)) \\ &= \mathbf{E}_{t,(X_0,X_1)\sim\pi_{0,1}} D(\dot{\psi}(X_0 \mid X_1), u_t^\theta(X_t)),\end{aligned} \tag{*}$$

where we have used LOTUS with $X_t = \psi_t(X_0 \mid X_1)$, and

$$u_t(X_t \mid X_1) = \dot{\psi}_t(\psi_t^{-1}(\psi_t(X_0 \mid X_1) \mid X_1) \mid X_1) = \dot{\psi}_t(X_0 \mid X_1).$$

The minimizer of $(*)$ according to Proposition 2.4 takes the form:

$$u_t(x) = \mathbf{E}\left[\dot{\psi}_t(X_0 \mid X_1) \mid X_t = x\right]. \tag{CFM-Minimizer}$$

2.8 Marginalization Trick for Conditional Flows

Next, we introduce a version of the marginalization trick for probability paths that are built from conditional flows. To this end, note that if $\pi_{0|1}(\cdot \mid x_1)$ is C^1, then $p_t(x \mid x_1)$ is also C^1 by construction. Moreover, $u_t(x \mid x_1)$ is conditionally integrable if

$$\mathbf{E}_{t,(X_0,X_1)\sim\pi_{0,1}} \left\|\dot{\psi}(X_0 \mid X_1)\right\| < \infty. \tag{Cond-Int}$$

Therefore, by setting $Z = X_1$, the following corollary of the marginalization trick is obtained.

Corollary 2.4.1. *Assume that q has bounded support, $\pi_{0|1}$ is $C^1(\mathbb{R}^d)$, and strictly positive for some x_1 with $g(x_1) > 0$, and $\psi_t(x \mid x_t)$ is a conditional flow satisfying Eq. (Conditional-Flow) and Eq. (Cond-Int). Then $p_{t|1}(x \mid x_1)$ and $u_t(x \mid x_1)$ defined in Eq. (Cond-Push-Forward) and Eq. (Cond-Smooth-Vel), respectively, define a marginal velocity field $u_t(x)$ generating the marginal probability path $p_t(x)$ interpolating p, q.*

This theorem will be used as a tool to show that particular choices of conditional flows lead to marginal velocity $u_t(x)$ generating the marginal probability path $p_t(x)$.

2.9 Conditional Flows with Other Conditions

Different conditioning chocies Z exist but are all essentially equivalent.

2.10 Optimal Transport and Linear Conditional Flow

We arrive to asking the question: how do we find a useful conditional flow $\psi_t(x \mid x_1)$? One approach is to choose it as a minimizer of a natural cost functional, ideally with some desirable properties. One popular example of such a cost functional is the dynamic optimal transport problem with quadratic cost, formalized as

$$\begin{cases} (p_t^*, u_t^*) & = \operatorname*{argmin}_{p_t,u_t} \int_0^1 \int \|u_t(x)\|^2 p_t(x)\,dx\,dt \\ & \quad \text{s.t. } p_0 = p, p_1 = q, \\ \frac{d}{dt}p_t + \operatorname{div}(p_t u_t) & = 0. \end{cases} \tag{Optimal-Transport}$$

The (p_t^*, u_t^*) above defines a flow via Eq. (Flow-ODE), with the form

$$\psi_t^*(x) = t\phi(x) + (1-t)x,$$

called the OT displacement interpolant, where $\phi : \mathbb{R}^d \to \mathbb{R}^d$ is the optimal transport map. The OT displacement interpolant solves the Flow Matching problem by defining the random variable

$$X_t = \psi_t^*(X_0) \sim p_t^*, \text{ when } X_0 \sim p.$$

The optimal transport formulation promotes straight sample trajectories

$$X_t = \psi_t^*(X_0) = X_0 + t(\phi(X_0) - X_0),$$

with a constant velocity $\phi(X_0) - X_0$, which are in general easier to sample with ODE solvers, in particular, the target sample X_1 is here perfectly solvable with a single step of the Euler method. We plug our maginal velocity formula Eq. (CFM-Minimizer) into the optimal transport problem Eq. (Optimal-Transport) and search for an optimal $\psi_t(x \mid x_1)$. While this seems like a challenge, we can instead bound the Kinetic Energy, for which a minimizer is readily found:

$$\begin{aligned} \int_0^1 \mathbf{E}_{X_t \sim p_t} \|u_t(X_t)\|^2 \, dt &= \int_0^1 \mathbf{E}_{X_t \sim p_t} \left\| \mathbf{E}\left[\dot{\psi}_t(X_0 \mid X_1) \mid X_t \right] \right\|^2 dt \\ &\leq \int_0^1 \mathbf{E}_{X_t \sim p_t} \mathbf{E}\left[\|\dot{\psi}_t(X_0 \mid X_1)^2\| \mid X_t \right] dt \\ &= \mathbf{E}_{(X_0, X_1) \sim \pi_{0,1}} \int_0^1 \|\dot{\psi}_t(X_0 \mid X_1)\|^2 \, dt, \end{aligned} \qquad \text{(OT-Bound)}$$

where in the first inequality, we applied Jensen's inequality, and in the second, we used the tower property of conditional expectation, and switch integration of t and expectation. The integrand can be minimized individually for each (X_0, X_1), leading to the variational problem $\gamma_t = \psi_t(x \mid x_1)$:

$$\min_{\gamma:[0,1]\to\mathbb{R}^d} \int_0^1 \|\dot{\gamma}_t\|^2 \, dt$$
$$\text{s.t.}\gamma_0 = x, \gamma_1 = x_1.$$

We use Euler-Lagrange equations, which in this case take the form $\frac{d^2}{dt^2}\gamma_t = 0$. Incorporating the boundary conditions, we obtain the minimizer

$$\psi_t(x \mid x_1) = tx_1 + (1-t)x. \qquad \text{(Linear-Conditional-Flow)}$$

Note that although not constrained to be, this choice of $\psi_t(x \mid x_1)$ is a diffeomorphism in x for $t \in [0, 1)$, and smooth in t, x, as required from conditional flows.

We can draw several concludsions from this.

1. The linear conditional flow minimizes a bound of the Kinetic Energy among all conditional flows.

2. In case the target q consists of a single data point $q(x) = \delta_{x_1}(\cdot)$, we have that the linear conditional flow Eq. (Linear-Conditional-Flow) is the Optimal Transport. Indeed, in this case, $X_t = \psi_t(X_0 \mid x_1) \sim p_t$ and $X_0 = \psi^{-1}(X_t \mid x_1)$ is a function of X_t, which makes $\mathbf{E}\left[\dot{\psi}_t(X_0 \mid x_1) \mid X_t\right] = \dot{\psi}_t(X_0 \mid x_1)$, and therefore, our bound Eq. (OT-Bound) becomes an equality.

3. Plugging in the linear conditional flow Eq. (Linear-Conditional-Flow) into Eq. (OT-Bound), we get
$$\int_0^1 \mathbf{E}_{X_t\sim p_t}\|u_t(X_t)\|^2\,dt \le \mathbf{E}_{(X_0,X_1)\sim\pi_{0,1}}\int_0^1 \|X_1 - X_0\|^2\,dt,$$
showing that the kinetic energy of the marginal velocity $u_t(x)$ is not bigger than that of the original coupling $\pi_{0,1}$.

The conditional flow in Eq. (Linear-Conditional-Flow) is in particular affine, and consequently motivates investigating the family of affine conditional flows.

2.11 Affine Conditional Flows

In the previous section, we discovered the linear Eq. (Linear-Conditional-Flow) was a minimizer to a bound of the kinetic energy among all conditional flows. The linear conditional flow is a partiuclar instance of the wider family of affine conditional flows, explored in this section. Consider

$$\psi_t(x \mid x_1) = \alpha_t x_1 + \sigma_t x, \qquad \text{(Conditional-Affine)}$$

where $\alpha_t, \sigma_t : [0,1] \to [0,1]$ are smooth functions satisfying

$$\alpha_0 = 0 = \sigma_1, \alpha_1 = 1 = \sigma_0, \text{ and } \dot\alpha_t, -\dot\sigma_t > 0 \text{ for } t \in (0,1).$$

We call the pair (α_t, σ_t) a scheduler. The derivative condition above ensures that α_t is strictly monotonically increasing, while σ_t is strictly monotonically decreasing. The conditional flow Eq. (Conditional-Affine) is a simple affine map in x for each $t \in [0,1)$, which satisfies Eq. (Conditional-Flow). The associated marginal velocity field Eq. (CFM-Minimizer) is

$$u_t(x) = \mathbf{E}\left[\dot\alpha_t X_1 + \dot\sigma_t X_0 \mid X_t = x\right]. \qquad \text{(Marginal-Velocity-Affine)}$$

By virtue of Corollary 2.4.1, we can prove that, if using independent coupling, and a smooth and strictly positive source density p with finite second moments, for example, a Gaussian $p = N(\cdot \mid 0, I)$, then u_t generates a probability path p_t interpolating p and q.

Theorem 2.5. *Assume that q has bounded support, and p is $C^1(\mathbb{R}^d)$, with strictly positive density with finite second moments, and these two relate by independent coupling $\pi_{0,1}(x_0, x_1) = p(x_0)p(x_1)$. Let $p_t(x) = p_{t|1}(x \mid x_1)q(x_1)\,dx_1$ be defined by Eq. (Cond-Push-Forward), with ψ_t defined by Eq. (Conditional-Affine). Then the marginal velocity Eq. (Marginal-Velocity-Affine) generates p_t interpolating p, q.*

In this affine case, the CFM loss takes the form

$$\mathcal{L}_{CFM}(\theta) = \mathbf{E}_{t,(X_0,X_1)\sim\pi_{0,1}} D(\dot\alpha_t X_1 + \dot\sigma_t X_0, u_t^\theta(X_t)).$$