Flow Matching

samrudhdhi.rangrej

July 2024

1 Normalizing Flows

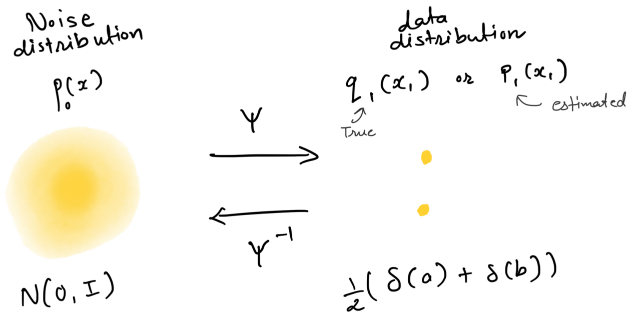


Figure 1: **Normalizing flows** map noise distribution to data distribution using a bijection.

Normalizing flows map a gaussian distribution $p_0(x) = \mathcal{N}(0, I)$ to a data distribution $p_1(x)$ using a bijection ψ (with inverse ψ^{-1}).

$$\begin{aligned} x_0 &\sim p_0(x) \\ x_1 &= \psi(x_0) \end{aligned} \tag{1}$$

According to change of variable formula:

$$\begin{aligned} p_1(x_1) &= \frac{p_0(x_0)}{|\det(J_\psi(x_0))|} \\ &= p_0(\psi^{-1}(x_1)) \left|\det\left(J_{\psi^{-1}}(x_1)\right)\right| \end{aligned} \tag{2}$$

Note on Jacobian J:

Jacobian quantifies change in the function (f) at a given point.

$$|\det(J_f)| = \begin{cases} > 1 & \text{expansion; density goes down} \\ < 1 & \text{contraction; density goes up} \end{cases}$$

To sample x_1 according to eq 1, we learn ψ by maximizing log-likelihood:

$$\begin{aligned} \log p_1(x_1) &= \log p_0(\psi^{-1}(x_1)) + \log(|\det(J_{\psi^{-1}}(x_1))|) \\ &\propto -\frac{x_0^2}{2} + \log(|\det(J_{\psi^{-1}}(x_1))|) \qquad (\because p_0 = \mathcal{N}(0, I)) \end{aligned} \tag{3}$$

Above, the first term causes ψ^{-1} to contract towards the origin and the second term causes it to expand away from the origin. Tension between these two terms avoids a degenerate solution, i.e. well-behaved and stable training.

Generally p_1 is a complex distibution, requiring an expressive ψ. We can get complex thus expressive ψ by composing simple and less expressive $\{\phi_k\}$.

$$\psi = \phi_1 \circ \dots \phi_k \dots \circ \phi_0 \tag{4}$$

Replacement of eq 4 in eq 3 results in following log-likelihood.

$$\log p_1(x_1) = -\frac{x_0^2}{2} + \sum_{k=1}^{K} \log(|\det(J_{\phi_k^{-1}}(x_1))|) \tag{5}$$

2 Continuous Normalizing Flow

Although there exist many instantiations of ϕ, let us consider discrete residual flows.

$$\phi_k(x) = x + \delta u_k(x) \tag{6}$$

Here, ϕ_k is invertible if u_k is a contraction with the Lipschitz constant $< \frac{1}{\delta}$. Then, ϕ_k^{-1} can be found using fixed point theorem.

Let's verify Lipschitz constant $< \frac{1}{\delta}$ ensures invertibility.

for ϕ_k to be invertible, following should hold true.

$$\begin{aligned}
\phi_k(a) &\neq \phi_k(b) \qquad \forall a \neq b \\
a + \delta u_k(a) &\neq b + \delta u_k(b) \\
\delta(u_k(a) - u_k(b)) &\neq b - a \\
-\frac{(u_k(a) - u_k(b))}{(a-b)} &\neq \frac{1}{\delta} \\
-\frac{\partial u_k}{\partial x} &\neq \frac{1}{\delta} \qquad \text{(when } |a-b| \to 0\text{)}
\end{aligned}$$

which holds true since $\left|\frac{\partial u_k}{\partial x}\right| < \frac{1}{\delta}$.

Re-writing eq 6,

$$\frac{\phi_k(x) - x}{\delta} = u_k(x) \tag{7}$$

If $\delta = \frac{1}{k}$ and $k \to \infty$, flows $\psi = \phi_1 \circ \dots \phi_t \dots \circ \phi_0$ can be given by ODE of the following form. Notice the switch from discrete variable k to continuous variable t:

$$\begin{aligned}
\frac{dx_t}{dt} &= \lim_{\delta \to 0} \frac{x_{t+\delta} - x_t}{\delta} \\
&= \lim_{\delta \to 0} \frac{\phi_t(x_t) - x_t}{\delta} \\
&= u_t(x_t) \\
\text{with } \ x_t = x_0 &\text{ at } \ t = 0 \text{ (i.e. initial condition).}
\end{aligned} \tag{8}$$

Thus, we get the *continuous* alternative of the change of variable formula from eq 2.

$$\begin{aligned}
x_t = \phi_t \circ \dots \circ \phi_0(x_0) &= x_0 + \int_0^s \frac{dx_s}{ds} ds \\
&= x_0 + \int_0^s u_s(x_s) ds
\end{aligned} \tag{9}$$

Let us denote $\phi_t \circ \cdots \circ \phi_0 = \psi_k$ (i.e. $\psi_1 = \psi$).

$$\begin{aligned}\psi_t(x_0) &= x_0 + \int_0^s u_s(\psi_s(x_0))ds \\ \frac{d\psi_t}{dt} &= u_t(\psi_t(x_0)) \qquad\qquad \text{(Note: Same as eq 8)}\end{aligned} \tag{10}$$

Change in the likelihood of x_t due to ψ_t (or u_t):

$$\frac{\partial}{\partial t}p_t(x_t) = -(\nabla(u_t p_t)(x_t)) \tag{11}$$

Above equation is also known as *'Transport Equation' for conserved quantities* or *'Law of conservation'*.

Law of conservation

An instantaneous change in the amount of quantity in a unit volume is equal to the amount of quantity that enters or exits that volume. In other words, the conserved quantity cannot be created or destroyed, only transferred.

Let's imagine a particle of a given quantity 'flowing' from position x_0 to x_1 from time $t = 0$ to 1. Then,

$$\begin{aligned}\psi_t(x_0) &= \text{A vector field denoting the position of a particle at time } t \text{ given the initial position of } x_0. \\ u_t(x_t) &= \text{A vector field denoting the velocity (direction and amount) with which the particle positioned at } x_t \text{ is flowing} \\ &\quad \text{at time } t. \\ p_t(x_t) &= \text{The density of the particles at position } x_t \text{ at time } t. \\ u_t p_t(x_t) &= \text{Flux describing the expected velocity of the fellow particles flowing 'away' from position } x_t \text{ at time } t.\end{aligned}$$

Let's derive log-likelihood in three steps by calculating $\frac{d}{dt}p_t(x_t)$, $\frac{d}{dt}\log p_t(x_t)$, and finally $\log p_1(x_1)$.

First, the total derivative (as x_t also depends on t) of p_t,

$$\begin{aligned}\frac{d}{dt}p_t(x_t) &= \frac{\partial}{\partial t}p_t(x_t) + \langle \nabla_{x_t} p_t(x_t), \frac{d}{dt}x_t \rangle \\ &= \underline{-\nabla(u_t p_t)(x_t)} + \langle \nabla_{x_t} p_t(x_t), u_t(x_t) \rangle \qquad\qquad (\because \text{eq 8 and eq 11}) \\ &= -p_t(x_t)(\nabla u_t)(x_t) - \langle \nabla_{x_t} p_t(x_t), u_t(x_t) \rangle + \langle \nabla_{x_t} p_t(x_t), u_t(x_t) \rangle \\ &= -p_t(x_t)(\nabla u_t)(x_t)\end{aligned} \tag{12}$$

Second, total derivative of $\log p_t(x_t)$,

$$\begin{aligned}\frac{d}{dt}\log p_t(x_t) &= \frac{1}{p_t(x_t)}\frac{d}{dt}p_t(x_t) \\ &= -\frac{p_t(x_t)(\nabla u_t)(x_t)}{p_t(x_t)} \qquad\qquad (\because \text{eq 12}) \\ &= -\nabla u_t(x_t)\end{aligned} \tag{13}$$

Finally, let's calculate the log-likelihood. Recall,

$$\begin{aligned}x_1 &= x_0 + \int_0^1 \frac{dx_t}{dt}dt \\ \log p_1(x_1) &= \log p_0(x_0) + \int_0^1 \frac{d}{dt}\log p_t(x_t)dt \\ &= \log p_0(x_0) - \int_0^1 \nabla u_t(x_t)dt\end{aligned} \tag{14}$$

Training with this objective requires approximation of i) integral which is intractable, and ii) divergence which is very expensive.
Can we design alternate objective?

3 Flow Matching

Let's say we knew $u_t(x_t)$ that allows us to flow from $p_0(x_0)$ to $p_1(x_1)$. Then the simple objective would be:

$$L_{FM}(\theta) = \mathbb{E}_{t,p_t(x)}||u_\theta(t,x) - u_t(x)||^2 \tag{15}$$

But, we don't know what $p_t(x_t)$ and $u_t(x_t)$ are.
One way is to construct p_t and u_t using probability paths and vector fields defined *per sample* (which is easy to do), followed by marginalization.

$$p_t(x) = \int p_t(x|x_1)q(x_1)dx_1 \tag{16}$$

(Note, q_1 is the true unknown data distribution and p_1 is our approximation of q_1.)
where conditional probability paths must satisfy:

$$p_0(x|x_1) = p_0(x) \qquad \text{(Generally chosen to be Normal distribution)} \tag{17}$$

$$p_1(x|x_1) = \mathcal{N}(x|x_1, \sigma_{min}I) \xrightarrow[\sigma_{min}\to 0]{} \delta_{x_1}(x) \qquad (p_1 \text{ is a mixture-of-Gaussian estimation of } q_1) \tag{18}$$

Similarly, we can construct vector field,

$$\begin{aligned} u_t(x) &= \int u_t(x|x_1)p_1(x_1|x)dx_1 \\ &= \int u_t(x|x_1)\frac{p_t(x|x_1)}{p_t(x)}q_1(x_1)dx_1 \end{aligned} \tag{19}$$

Verify validity of eq 19

We can verify above definition of marginal field using transport equation.

$$\begin{aligned} \frac{\partial p_t(x)}{\partial t} &= -\nabla(u_t(x)p_t(x)) \\ \frac{\partial}{\partial t}\int p_t(x|x_1)q_1(x_1)dx_1 &= -\nabla(u_t(x)p_t(x)) && (\because \text{ eq 16}) \\ \int \left[\frac{\partial}{\partial t}p_t(x|x_1)\right] q_1(x_1)dx_1 &= -\nabla(u_t(x)p_t(x)) \\ \int \left[-\nabla u_t(x|x_1)p_t(x|x_1)\right] q_1(x_1)dx_1 &= -\nabla(u_t(x)p_t(x)) && (\because \text{ conditional transport equation}) \\ -\nabla\left[\int u_t(x|x_1)\frac{p_t(x|x_1)q_1(x_1)}{p_t(x)}dx_1\right] p_t(x) &= -\nabla(u_t(x)p_t(x)) \\ \int u_t(x|x_1)\frac{p_t(x|x_1)q_1(x_1)}{p_t(x)}dx_1 &= u_t(x) \end{aligned}$$

Let's rewrite the FM objective using conditional vector field,

$$L_{CFM}(\theta) = \mathbb{E}_{t,p_t(x|x_1),q_1(x_1)}||u_\theta(t,x) - u_t(x|x_1)||^2 \tag{20}$$

Below we prove that gradients of L_{CFM} and L_{FM} are same in expectation. Hence, we are essentially learning $u_\theta(t,x) = u_t(x)$ without direct access, via marginal $u_t(x|x_1)$.

Gradients of L_{CFM} and L_{FM} are same

$$
\begin{aligned}
||u_\theta(t,x) - u_t(x|x_1)||^2 &= ||u_\theta(t,x)||^2 + ||u_t(x|x_1)||^2 - 2\langle u_\theta(t,x), u_t(x|x_1)\rangle && (21)\\
||u_\theta(t,x) - u_t(x)||^2 &= ||u_\theta(t,x)||^2 + ||u_t(x)||^2 - 2\langle u_\theta(t,x), u_t(x)\rangle && (22)
\end{aligned}
$$

Note, the middle term is independent of θ. Expectation of the first and the last term is same.

$$
\begin{aligned}
\mathbb{E}_{p_t}||u_\theta(t,x)||^2 &= \int ||u_\theta(t,x)||^2 \underline{p_t(x)} dx \\
&= \int\int ||u_\theta(t,x)||^2 p_t(x|x_1)q(x_1)dx_1 dx && (\because \text{eq } 16)\\
&= \mathbb{E}_{q(x_1),p_t(x|x_1)}||u_\theta(t,x)||^2 && (23)
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}_{p_t}\langle u_\theta(t,x), u_t(x)\rangle &= \int \Big\langle u_\theta(t,x), \int u_t(x|x_1)\frac{p_t(x|x_1)q(x_1)}{p_t(x)}dx_1\Big\rangle p_t(x)dx && (\because \text{eq } 19)\\
&= \int \Big\langle u_\theta(t,x), \int u_t(x|x_1)p_t(x|x_1)q(x_1)dx_1\Big\rangle dx \\
&= \int\int \Big\langle u_\theta(t,x), u_t(x|x_1)\Big\rangle p_t(x|x_1)q(x_1)dxdx_1 \\
&= \mathbb{E}_{q(x_1),p_t(x|x_1)}\langle u_\theta(t,x), u_t(x|x_1)\rangle && (24)
\end{aligned}
$$

Let's define $p_t(x_t|x_1)$ (and derive $u_t(x_t|x_1)$) to compute L_{CFM},

$$
p_t(x_t|x_1) = \mathcal{N}(x_t|\mu_t(x_1), \sigma_t(x_1)^2 I) \tag{25}
$$

This can be achieved with a simple flow.

$$
\psi_t(x_0|x_1) = \sigma_t(x_1)x_0 + \mu_t(x_1) \qquad \text{(i.e. affine map)} \tag{26}
$$

Further, based on eq 10:

$$
\begin{aligned}
u_t(\psi_t(x_0)) &= \frac{d}{dt}\psi_t(x_0) && (27)\\
&= \frac{d}{dt}(\sigma_t(x_1)x_0 + \mu_t(x)) && (28)\\
&= \sigma_t'(x_1)x_0 + \mu_t'(x) && (29)\\
&= \sigma_t'(x_1)\Big(\frac{\psi_t(x_0) - \mu_t(x_1)}{\sigma_t(x_1)}\Big) + \mu_t'(x) \qquad (\because \text{eq } 26) && (30)\\
u_t(x_t) &= \frac{\sigma_t'(x_1)}{\sigma_t(x_1)}(x_t - \mu_t(x_1)) + \mu_t'(x) && (31)
\end{aligned}
$$

Diffusion

$$
\begin{aligned}
\mu_t(x_1) &= x_1; \qquad \sigma_t(x_1) = \sigma_{1-t}; && \text{(Variance Exploding)} && (32)\\
\mu_t(x_1) &= \alpha_{1-t}x_1; \qquad \sigma_t(x_1) = \sqrt{1-\alpha_{1-t}^2}; && \text{(Variance Preserving)} && (33)
\end{aligned}
$$

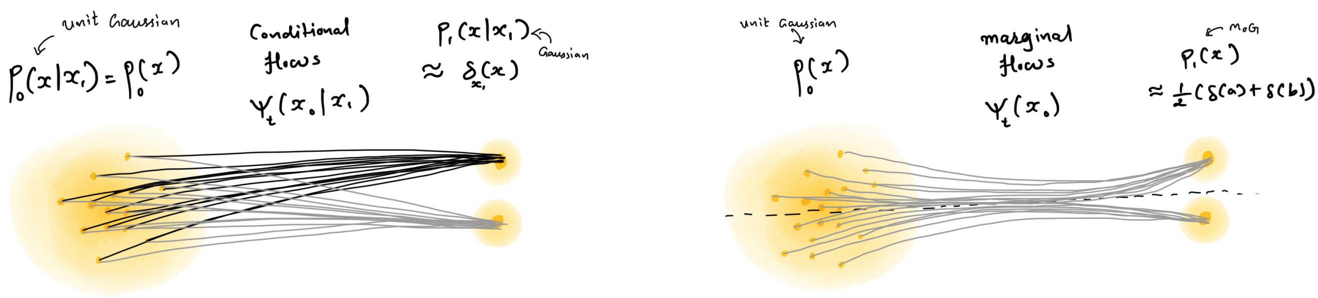
Optimal Transport
Here, $\sigma_t(x)$ and $\mu_t(x)$ can be any function that meet boundary condition.
A simple choice for μ_t and σ_t:

$$
\begin{aligned}
\mu_t(x_1) = tx_1; \qquad \sigma_t(x_1) &= (1-t) + t\sigma_{min} && (34)\\
t = 0 \longrightarrow \mu_0 = 0 \ ; \quad \sigma_0 &= 1 \\
t = 1 \longrightarrow \mu_1 = x_1; \quad \sigma_1 &= \sigma_{min}
\end{aligned}
$$

$$
\begin{aligned}
u_t(x_t|x_1) &= \frac{d}{dt}\psi_t(x_0|x_1) && (\because \text{eq } 10) \\
&= \frac{d}{dt}(tx_1 + (1-t)x_0 + t\sigma_{min}x_0) \\
&= x_1 - x_0 + \sigma_{min}x_0 && (35)
\end{aligned}
$$

Note, for a given pair of x_0 and x_1, u_t is constant for all t. x_0 ‘flows’ to x_1 in a straight line with a constant velocity. Also, note that eq 26 is a formula of a line.
Recall,

$$
\begin{aligned}
L_{CFM}(\theta) &= \mathbb{E}_{t,p_t(x_t|x_1),q_1(x_1)}||u_\theta(t,\psi_t(x_0)) - u_t(x_t|x_1)||^2 \\
&= \mathbb{E}_{t,p_0(x_0),q_1(x_1)}\Big|\Big|u_\theta(t,((1-t)+t\sigma_{min})x_0 + tx_1) - (x_1 - (1-\sigma_{min})x_0)\Big|\Big|^2 && (\because \text{eq } 26, 34) && (36) \\
&= \mathbb{E}_{t,p_0(x_0),q_1(x_1)}\Big|\Big|u_\theta(t,(1-t)x_0 + tx_1) - (x_1 - x_0)\Big|\Big|^2 && (\text{if } \ \sigma_{min} = 0) && (37)
\end{aligned}
$$



(a) (b)

Figure 2: (a) **Conditional flows** map a noise distribution to a narrow Gaussian ($\approx$ delta distribution) centered at the given data sample. Per-sample condition flows are straight paths that intersect each other. (b) **Marginal flows** map a noise distribution to a mixture of narrow Gaussian, with each Gaussian centered at one of the data sample. Marginal flows are curved and do not intersect.

Training and Sampling Algorithms

Algorithm 1 Training

repeat
    $x_1 \sim q_1(x_1)$; $t \sim U(0,1)$; $x_0 \sim \mathcal{N}(0,I)$;
    Take gradient descent step on
        $\nabla_\theta||u_\theta(t,(1-t)x_0 + tx_1) - (x_1 - x_0)||^2$ ▷ eq. 37
until converged

Algorithm 2 Sampling

$x_0 \sim \mathcal{N}(0,I)$;
for $t = 0, \ldots, 1-\delta$ **do**
    $x_{t+\delta} = x_t + \delta u_\theta(t, x_t)$ ▷ eq. 6
end for
return x_1

Prompt-guided CNF: We condition probablility paths and vector fields on prompt p (e.g. class id, caption, etc).

$$L_{CFM}(\theta) = \mathbb{E}_{t,p_t(x_t|x_1,p),q_1(x_1|p)}||u_\theta(t,x_t,p) - u_t(x_t|x_1,p)||^2 \tag{38}$$

(39)

Classifier-free guidance for FM: Replace $u_\theta(t,x_t)$ during sampling with,

$$\tilde{u}_\theta(t,x_t,p) = (1-w)\ u_\theta(t,x_t,\Phi) + w\ u_\theta(t,x_t,p); \quad \text{where guidance scale } w > 1 \tag{40}$$

Further Reading

- Lipman, Yaron, et al. “Flow Matching for Generative Modeling.” ICLR, 2023.
- `https://mlg.eng.cam.ac.uk/blog/2024/01/20/flow-matching.html`