CONTINUOUS NORMALIZING FLOW

Table of Contents

Discrete Normalizing Flow (DNF)

Goal: Imagine want turn a simple distribution into a more complex one

Apply a sequence of invertible tranformations

$\{f_1, \ldots, f_k\}$ $f_i : \mathbb{R}^d \to \mathbb{R}^d$ $z_{i-1} \to z_i$

Train: $z_k = f_k \circ f_{k-1} \circ \cdots \circ f_1(x)$

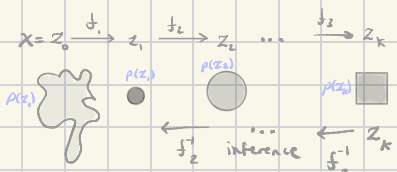
Infer: $x = f_k^{-1} \circ f_{k-1}^{-1} \circ \cdots \circ f_1^{-1}(z_k)$

an image $x = z_0, z_k \in \mathbb{R}^d$, $z_k \sim N(0, I)$

Let $p_X(x)$, $p_Z(z)$ be PDFs

$\Rightarrow$ Since f_i invertible,

$f_2 : z_1 \to z_2$



Discrete Change-of-variables (CoV) (and log CoV) formula

$$\lg p(x) = \lg p(z_K) + \sum_{i=1}^{K} \lg \left| \det \frac{\partial f_i(z_{i-1})}{\partial z_{i-1}} \right| \quad (A)$$

Gaussian; J_3: Jacobian

What's the problem with discrete?

1) Invertibility: You need specially designed invertible blocks
2) Expressivity: Finite stacks struggle with complex dists unless made deep/structured which gets expensive
3) Optimization: More layers you stack → more computationally expensive training gets

From DNF → Continuous Normalizing Flow (CNF)

Each discrete step becomes an infinitesimal transformation and the composition converges to an ODE flow

$$\frac{dx}{dt} = f_\theta(x_t, t)$$

Learn a vector field that continuously drags a simple dist into the data dist

No more finite invertible layers

What is an ODE?

It tells you how something changes over time

$$\frac{dx}{dt} = f(x, t)$$

The instantaneous ROC wrt time

tells you the slope at that point

Continuous Normalizing Flow (CNF)

$x_0 \sim p_0$ is the transformed sample we want match data

$x_T \sim N(0, I)$ is base distribution (say Gaussian)

Instantaneous Change-of-variables in continuous time

$$\frac{d}{dt} \lg p_t(x_t) = -\mathrm{Tr}\left(\frac{\partial f_\theta(x_t, t)}{\partial x_t}\right)$$

J_{f_θ}

$$\lg p(x_0) = \lg p(x_T) - \int_0^T \mathrm{Tr}\left(\frac{\partial f_\theta}{\partial x_t}\right) dt$$

p_t = PDF indexed by time, so

$p_t(x_t) = p(t)(x_t)$

$= p(x_t, t)$

whole flow is now a continuous trajectory

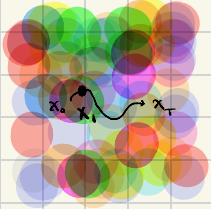
$x_t : t \longrightarrow x$

$x_t = x(t)$

$= (1-t)x + a$

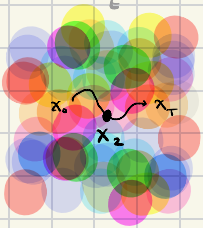
Problem: Trace is the computational bottleneck!

$p_t(x) \neq p_t(x_t)$

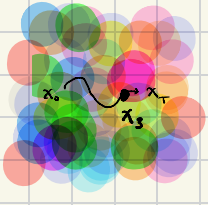


$t = 1 \quad p_1(x_1) = a \in \mathbb{R}$

p_t



$t = 2$



$t = 3$

Think of p_t like temperature

Discrete Flow	Continuous Flows
Product / log-sum of Jacobian determinants	Limit of infinitely many infinitesimal Jacobians / integral of Jacobian trace
Transformation: A warp of space (invertible layer)	Continuous tranformation: a time-dependent velocity field that drags probability mass around

Total Derivative

x a coordinate $\mathbb{R}^d$

x_t a trajectory

and $g(x_t, t)$

When we differentiate we have to account for 2 things

i) How the field g changes with time at a fixed point $\frac{\partial g}{\partial t}$

ii) How the field changes because you're moving along the trajectory through space

$$\nabla_x g \cdot \frac{dx_t}{dt}$$

That's why, for $x \in \mathbb{R}^d$ (Lagrangian, dot product, Eulerian)

$$\frac{d}{dt} g(x_t, t) = \frac{\partial g}{\partial x}(x_t, t) \cdot \frac{dx_t}{dt} + \frac{\partial g}{\partial t}(x_t, t)$$

$g(x(t), t)$

where $\frac{\partial g}{\partial x}(x_t, t) = \left(\frac{\partial g_1}{\partial x_1}, \frac{\partial g_2}{\partial x_2}, \dots, \frac{\partial g_d}{\partial x_d} \right) = \nabla_x g(x_t, t)$

and $\frac{dx_t}{dt} = \left(\dot{x}_{t,1}, \dot{x}_{t,2}, \dots, \dot{x}_{t,d} \right)$ a vector

$\Rightarrow \frac{\partial g}{\partial x}(x_t, t) \cdot \frac{dx_t}{dt}$ a dot product $\rightarrow$ scalar

Total Derivative $\rho_t(x_t) = \rho(x_t, t)$

$$\frac{d}{dt} g(x_t, t) = \underbrace{\nabla_x g(x_t, t)}_{\text{Lagrangian}} \cdot \frac{dx_t}{dt} + \underbrace{\frac{\partial g}{\partial t}(x, t)}_{\text{Eulerian}} \Big|_{x = x_t}$$

$$\frac{d}{dt} g_t(x_t) = \nabla_x g_t(x_t) \frac{dx}{dt} + \frac{\partial}{\partial t} g_t(x) \Big|_{x = x_t}$$

So we take the derivative of $\frac{d}{dt} \log p_t(x_t)$

$$\frac{d}{dt} \log p_t(x_t) = \frac{\partial}{\partial t} \log p_t(x) \Big|_{x=x_t} + \nabla_x \log p_t(x_t, t) \frac{dx_t}{dt}$$

$$= \frac{\partial}{\partial t} \log p_t(x) \Big|_{x=x_t} + \nabla_x \log p_t(x_t, t) \, f_\theta(x_t, t)$$

How does log density of x evolve as we push it through a flow?

Recall $\frac{dx}{dt} = f_\theta(x_t, t)$

Why the Total Derivative?

Because 1) The density field p_t changes wrt time $\left(\frac{\partial}{\partial t}\right)$

2) The sample x_t also moves through space as time evolves $\left(\nabla_x \cdot \frac{dx}{dt}\right)$

Now we have the formula

$$\frac{d}{dt} \log p_t(x_t) = \frac{\partial}{\partial t} \log p_t(x) \Big|_{x=x_t} + \nabla_x \log p_t(x_t, t) \, f_\theta(x_t, t)$$

Why the $|_{x=x_t}$ Notation?

When applying the chain rule, must be explicit that

- $\frac{\partial}{\partial t} \log p_t(x)$ means derivative wrt t while treating x constant
- Then evaluate the expression at the actual moving location x_t

$$\frac{d}{dt} \log p_t(x_t) = \frac{\partial}{\partial t} \log p_t(x) \Big|_{x=x_t} + \nabla_x \log p_t(x_t, t) \, f_\theta(x_t, t)$$

Continuity equation

$$\frac{\partial}{\partial t} \lg p_t(x) = -\nabla_x f_\theta(x_t, t) - f_\theta(x_t, t) \nabla_x \lg p_t(x)$$

$$= \left[-\nabla_x f_\theta(x_t, t) - f_\theta(x_t, t) \nabla_x \log p_t(x) \right]_{x=x_t} + \nabla_x \log p_t(x_t) f_\theta(x_t, t)$$

$$\therefore \quad \frac{d}{dt} \log p_t(x_t) = -\nabla_x \cdot f_\theta(x_t, t)$$

How Do We Get $\frac{\partial}{\partial t} p_t(x)$?

Conservation of Mass (Continuous)

Continuity Equation (Fluid Dynamics)

$$\frac{\partial \rho}{\partial t} = -\nabla \cdot (\rho v)$$

- $\frac{\partial \rho}{\partial t}$: local rate of ρ
- $-\nabla \cdot (\rho v)$: divergence of flux; dot product

$\rho(x, t)$ density at point x and time t

$v(x, t)$ the velocity field moving that density at that point and time

"Stuff can't vanish or appear, only move"

State	mass
diverge	out
converge	in

$$\frac{\partial \rho}{\partial t} = -\nabla(\rho v)$$

Continuity Equation using PDFs

$$\frac{\partial}{\partial t} p_t(x) + \nabla_x \cdot \left(f_\theta(x_t, t) p_t(x) \right) = 0$$

$$\Rightarrow \frac{\partial}{\partial t} p_t(x) = -\nabla_x \left(f_\theta(x_t, t) p_t(x) \right)$$

log density $\rightarrow$ $\frac{\partial}{\partial t} \log p_t(x)$ (by chain rule)

$$= \frac{\partial}{\partial t} \log p_t(x) \frac{\partial p_t(x)}{\partial t}$$

$$= \frac{1}{p_t(x)} \frac{\partial}{\partial t} p_t(x)$$

$$= \frac{1}{p_t(x)} \left[-\nabla_x \cdot \left(f_\theta(x_t, t) p_t(x) \right) \right]$$

$$= -\frac{1}{p_t(x)} \left(\nabla_x f_\theta(x_t, t) p_t(x) + f_\theta(x_t, t) \nabla_x p_t(x) \right)$$ (product rule)

$$= -\frac{\nabla_x \cdot f_\theta(x_t, t)\,\cancel{p_t(x)}}{\cancel{p_t(x)}} - \frac{f_\theta(x_t, t)\,\nabla_x p_t(x)}{p_t(x)}$$

$$= -\nabla_x \cdot f_\theta(x_t, t) - \frac{f_\theta(x_t, t)\,\nabla_x p_t(x)}{p_t(x)}$$

∇_x is gradient operator wrt $x = (x_1, \dots, x_d)$

$$\nabla_x g(x) = \sum \frac{\partial g_i(x)}{\partial x_i}$$

$$\frac{\nabla_x p_t(x)}{p_t(x)} = \frac{1}{p_t(x)} \nabla_x p_t(x) = \nabla_x \log p_t(x)$$

Multivariate Chain Rule

$$\nabla_x (g \circ h)(x) = g'(h(x)) \nabla_x h(x)$$

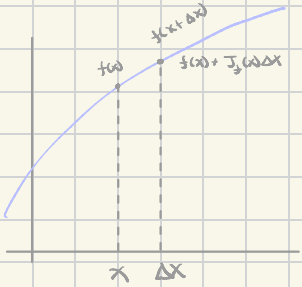
$$\frac{\partial}{\partial t} \log p_t(x) = -\nabla_x \cdot f_\theta(x_t, t) - \frac{f_\theta(x_t, t)\,\nabla_x p_t(x)}{p_t(x)}$$

$$\boxed{\frac{\partial}{\partial t} \log p_t(x) = -\nabla_x \cdot f_\theta(x_t, t) - f_\theta(x_t, t)\,\nabla_x \log p_t(x)}$$

Jacobian & Trace

$x \in \mathbb{R}^d \quad f: \mathbb{R}^d \to \mathbb{R}^d$

Jacobian: linear approximation of f around a small neighborhood of x



$$\boxed{f(x + \Delta x) \approx f(x) + J_f(x)\,\Delta x}$$

Jacobian Matrix

$$J_f(x) = \frac{\partial f(x)}{\partial x}$$

$$\begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_d} \\ \vdots & \frac{\partial f_i}{\partial x_j} & \vdots \\ \frac{\partial f_d}{\partial x_1} & \cdots & \frac{\partial f_d}{\partial x_d} \end{pmatrix}$$

Each diagonal term measures local stretching / compression along its own axis

Entry (i, j): How much the i-th velocity component changes when you nudge the j-th coordinate

$\frac{\partial f_i}{\partial x_i}$ → If you nudge x_i, how does the flow in the f_i direction speed up or slow down

A Little More to Get Trace

Divergence

$$\nabla \cdot f(x) = \sum_{i=1}^{d} \frac{\partial f_i}{\partial x_i}$$

dot product

Total Derivative

$$\frac{d}{dt} \log p_t(x_t) = -\nabla_x \cdot f_\theta(x_t, t)$$

$\mathbb{R}^d \cdot \mathbb{R}^d$

$$= -\sum \frac{\partial f_i(x_t, t)}{\partial x_i} = -\mathrm{Tr}\left(\frac{\partial f_\theta}{\partial x_t}\right)$$

Off diagonals of Jacobian tell us rotation / sheer

The diagonals of the Jacobian are the only entries of the Jacobian that directly tells us how volumes expand/contract along each axis

Fundamental Theorem of Calculus (FTC)

$$F'(x) = f(x) \Rightarrow F(b) - F(a) = \int_a^b f(x)\,dx$$

equivalently

$$\frac{dy}{dt} = g(t) \Rightarrow y(T) - y(0) = \int_0^T g(t)\,dt$$

So now we have

$$\frac{d}{dt} \log p_t(x_t) = -\nabla_x \cdot f_\theta(x_t, t) = -\mathrm{Tr}\left(\frac{\partial f_\theta}{\partial x_t}\right)$$

And

$$\log p(x_T) - \log p(x_0) = -\int_0^T \mathrm{Tr}\left(\frac{\partial f_\theta}{\partial x_t}\right)$$

By FTC

Instantaneous Change-of-variables in continuous time

$$\log p(x_0) = \log p(x_T) - \int_0^T \mathrm{Tr}\left(\frac{\partial f_\theta}{\partial x_t}\right)$$

Problem: Trace is the computational bottleneck!

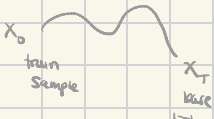
Solution: Hutchison Estimator → still expensive

Objective/Loss Function

Let x_T be base density (Gaussian)

x_0 be a training image

$$\ell(x;\theta) = \log p(x_T) - \int_0^T \mathrm{Tr}\left(\frac{\partial f_\theta}{\partial x_t}\right)$$



$$\text{MLE} \quad \max_\theta \mathbb{E}_{x \sim \text{data}}[\log p(x)]$$

Training

- $X_0 \in \mathbb{R}^d$ sample
- Define trajectory via ODE (NN)

$$\frac{dx_t}{dt} = f_\theta(x_t, t) \qquad x(0) = x_0$$

- Numerically integrate with ODE solver (PyTorch) to get x_T
- Choose simple base PDF $p(x_T) = N(0, I)$
- Compute

$$\log p(x_T)$$

- Compute instantaneous log-density change

$$\frac{d}{dt} \log p(x_t) = -\mathrm{Tr}\left(\frac{\partial f_\theta}{\partial x_t}\right)$$

- Integrate $\int_0^T \mathrm{Tr}\left(\frac{\partial f_\theta}{\partial x_t}\right) dt$
- Compute

$$\log p(x_0) = \log p(x_T) - \int_0^T \mathrm{Tr}\left(\frac{\partial f_\theta}{\partial x_t}\right) dt$$