About Flow Matching

Emails from Ying Nian Wu, transcribed by Andrew Lizarraga

March 2024

1

Let $x_0 \sim q(x_0)$, and let $x_1 \sim \mathrm{N}(0, I)$, independently. Let $(x_t, t \in [0, 1])$ be a trajectory that connects x_0 and x_1. You can imagine a particle moves from x_1 to x_0 over time t in $[0, 1]$. If we repeat 1 billion times, you can imagine we have 1 billion particles that trace out 1 billion trajectories. We call each such original trajectory as a path.

At each time t, we have a snapshot of 1 billion particles that form a marginal distribution q_t.

Define

$$Q(x_0, x_{dt}, x_{2dt}, ..., x_{t-dt}, x_t, ..., x_1)$$

to be the distribution of the 1 billion trajectories. Now we learn $P = \prod_t p(x_{t-dt}|x_t)$, which is a Markov process going from x_1 to x_0.

We treat the 1 billion trajectories Q as the data, and we learn P by MLE, which amounts to minimizing $D_{KL}(Q|P)$, which is still the same as the original 2015 paper, except that we can make Q, i.e., the trajectories, arbitrary.

In the 2015 paper, they re-write $Q = q(x_0)q(x_T|x_0)\prod_t q(x_{t-1}|x_t, x_0)$, where $q(x_{t-1}|x_t, x_0)$ is Gaussian. The $D_{KL}(q(x_{t-1}|x_t, x_0)|p(x_{t-1}|x_t))$ expression only means that we can write the least squares loss from the above MLE on the trajectory data (augmented data) Q as minimizing

$$\mathrm{E}_q(x_{t-1}|x_t, x_0)[|x_{t-1} - f_\theta(x_t, t)|^2]$$

which is equivalent to minimizing

$$|\mathrm{E}(x_{t-1}|x_t, x_0) - f_\theta(x_t, t)|^2$$

where f is Unet or transformer.

You can imagine 100 trajectories going from the same x_0 to the same x_t. They go through 100 different x_{t-1}. We use $f_\theta(x_t, t)$ to predict these 100 different x_{t-1} by least squares, then it is the same as we use $f_\theta(x_t, t)$ to predict the average of these 100 x_{t-1}. The average is in closed form. So we use the average of x_{t-1} as the output data ,which leads to more accurate training by variance reduction.

2

The above $q(x_{t-1}|x_t, x_0)$ are derived from the Markovian $Q = \prod_t q(x_t|x_{t-1})$, but we can design arbitrary $q(x_{t-1}|x_t, x_0)$. The trajectory Q does not need to be Markovian. **The trajectories can be any arbitrary continuous trajectories from sampled x_0 to sampled x_1 to serve as data**.

For non-Markovian trajectories, the Markovian model P cannot model the data Q exactly, i.e., $D_{KL}(Q|P)$ cannot be zero, so that we do not have ELBO. But this is not a very big issue as there can be other ways to approach the log-likelihood.

The point is that as long as $D_{KL}(q(x_{t-dt}|x_t)|p(x_{t-dt}|x_t)) = 0$, then we are good, i.e., our Markovian model only needs to model the local transition ($x_t \to x_{t-dt}$) in the trajectories, then marginally we have $p_t = q_t$. We can prove this by induction. If $p_t = q_t$, then $p_{t-dt} = q_{t-dt}$ because $p_{t-dt}(x_{t-dt}) = \int p_t(x_t)p(x_{t-dt}|x_t)dx_t = \int q_t(x_t)q(x_{t-dt}|x_t)dt = q_{t-dt}(x_{t-dt})$.

Imagine you follow the paths in the following way: at time t, suppose you are at $x_t = x$. Suppose there are 100 paths going through x at time t. You can randomly pick a path to go to x_{t-dt}. That means for the 1 billion particles, at any moment t, if multiple particles meet at x, then they switch identities by a random permutation. This will not change the 1 billion paths, and will not change the distribution of the particles of the snapshot at time t. Of course the whole trajectory of a particle will not be one of the original trajectories, because the particle keeps changing paths tracked by the original trajectories. In other words, the particles become memoryless, they do not remember where they come from, or they do not remember their identities.

3

In the original diffusion paper, the trajectory is zig-zag and non-differentiable, because $x_t = x_{t-dt} + \sqrt{dt}e_t$, where $e_t \sim N(0, I)$, i.e., the velocity $= (x_t - x_{t-dt})/dt \to \infty$.

For flow, the trajectory is smooth and differentiable, i.e., $x_t = a_t x_0 + b_t x_1$, so that velocity $v(x_t, t) = a'_t x_0 + b'_t x1$, where a'_t and b'_t are time derivatives, i.e.,

$$x_{t-dt} = x_t + v(x_t, t)dt.$$

Again at each time t and $x_t = x$, there are 100 particles meeting at x, so we have 100 velocities $v(x, t)$. Then we can assume our model $p(x_{t-dt}|x_t)$ to be $x_{t-dt} \sim \mathrm{N}(x + u(x, t)dt, \sigma^2)$, where $u(x, t)$ is the drift, and σ^2 is the variance (the normal assumption can be changed to a more accurate distribution, but it does not matter as we will see next).

Then clearly the MLE of $u(x, t)dt$ is the average of the 100 $v(x, t)dt$, and σ^2 is the variance of the 100 $v(x, t)dt$, i.e.,

$$u(x, t) = \mathrm{E}(v(x, t)|x_0, x_1, \text{ so that trajectory from } x_1 \text{ to } x_0 \text{ passes } x \text{ at time } t).$$

But

$$\sigma^2 = \mathrm{var}(v(x, t)|\dots)dt^2 = \mathrm{var}_t dt^2,$$

where var_t is the variance of the 100 $v(x, t)$. Thus $p(x_{t-dt}|x_t)$ is

$$x_{t-dt} = x_t + u(x, t)dt + e(x, t)dt,$$

where $e(x, t) = v(x, t) - u(x, t)$, and $v(x, t)$ is random (one of 100), and $\mathrm{E}(e(x, t)) = 0$.

The above path is not diffusion anymore, because we have $e(x, t)dt$, instead of $e(x, t)\sqrt{dt}$. Even though $\mathrm{E}(e(x, t)) = 0$, we have $\mathrm{var}(e(x, t)) = \mathrm{var}_t dt^2$. That is, $e(x, t)dt$ is a random drift. Each path is still differentiable.

Okay, now consider we move according to the above model for N steps, so that $Dt = Ndt$, e.g., dt is 1 milli-second, and Dt is 1 second, so that $N = 1000$. Then we have

$$x_{t-Dt} = x_t + \sum u(x, t)dt + \sum e(x, t)dt,$$

and

$$\sum e(x,t)dt = \sum e(x,t)Dt/N \to 0,$$

because $\sum e(x,t)/N \to 0$ according to the law of large number, where $e(x,t)$ are independent due to the Markovian property, i.e., at each time t, we randomly pick a path that passes through x. Equivalently

$$\text{var}(\sum e(x,t)dt) = \sum \text{var}_t dt^2 = (\sum \text{var}_t dt)dt \to 0$$

.

So if we look at the trajectories at the scale Dt, they are deterministic $x_{t-Dt} = x_t + u(x,t)Dt$. In our infinitesimal analysis, we can let $Dt \to 0$ and for each Dt, we let $N \to \infty$.

The above argument generally applies to ODE/SDE, where **the random drift can be replaced by the deterministic average drift**. For instance, in Langevin dynamics, the conditional score can be replaced by the marginal score.

Now back to the path $x_t = a_t x_0 + b_t x_1$, recall x_0 and x_1 are high-dimensional vectors. So x_t is generally a curve that connects x_0 and x_1. We can make it straight line as Qiang Liu did. For learning, we can minimize

$$\text{E}[|v(x,t) - u_\theta(x,t)|^2 | x_0, x_1, \text{ trajectory passes } x \text{ at } t].$$

This is actually the v-prediction of Jonathan Ho, except they use spherical interpolation, so that x_t is a circle. The v-prediction can be translated to the epsilon-prediction, which is about score. So flow matching still follows $D_{KL}(Q|P)$ of the original paper, i.e., MLE on the augmented data or trajectory data Q. We only need to make the trajectory in data Q straight.