6.7900 Fall 2024: Lecture Notes 23

1 Normalizing Flow Model

Recall that our task is to translate a simple latent distribution z to a complex distribution x through invertible transformations. In some cases, the dimension of x could be much larger than z - but for today's purpose, we will assume their dimensions are equal.

We can approach this problem by introducing a series of invertible layers $f_1, \cdots, f_L$ so that $x = f_L \circ \cdots \circ f_1(z)$. Because all layers are invertible, we can recover z from x with

$$z = g_1 \circ \cdots \circ g_L(x)$$

where $g_i = f_i^{-1}$. This is advantageous since recovering z allows us to explicitly evaluate the log-likelihood of the observed x:

$$P(x \mid \theta) = N(z(x) \mid 0, I)\frac{dz}{dx}$$

$$= N(z(x) \mid 0, I) \prod_{j=1}^{L} \left| \frac{\partial h_{j-1}}{\partial h_j} \right|$$

Note that the Jacobian transformation is to take into account that some of the transformations might be "stretching" or "pressing" the space, so it is not sufficient to only look at probability density

The downside of normalizing flows, of course, is that the transformations have to invertible; a simple linear transformation might not be invertible if unconstrained.

2 Continuous Flows

We can view each transformation of the normalizing flow model as one time step. A natural extension is to turn this into continuous time. Assume $t = 0$ is the simple distribution $p_0(x)$, and $t = 1$ is the complex distribution $p_1(x)$. We are going to learn a time dependent vector field $v_t(x)$ (with same dimension as dx) to specify how the distribution should move in time. With a continuous model, we can actually give an

"intermediate" distribution between the latent and complex distributions with the continuity equation:

$$\frac{d}{dt}p_t(x) = -\nabla_x \cdot (p_t(x)v_t(x)).$$

There are several approaches one might attempt. First, we could have specified time-dependent probability flow $p_t(x)$. The challenge is the vector field will be very challenging to compute. Another attempt is to start by specifying a vector field; but finding a vector field that gives $p_1(x)$ is very challenging. The approach we take is to specify a simple trajectory interpolating between $p_0(x)$ and $p_1(x)$ and use it as a guidance to train the vector field.

As one simple but important special case, assume for now that p_1 is a point mass (only one sample), and we move in a straight line from p_0 to the target sample. Then if we are at p_t at time t, we have

$$\frac{d}{dt}x_t = \frac{x_1 - x_t}{1 - t}$$

as the time-dependent vector field because we need to move to x_t in $1-t$ time. Note that in this case, we can explicitly write the distribution at time t:

$$p_t(x) \sim N(x \mid tx_1, (1-t)^2 I).$$

Exercise: Show the continuity equation holds in this case.

Going back to the general case, given t and x_t, there are multiple pairs of (x_0, x_1) whose linear interpolation at time t would result in x_t: each of them suggests going in a different direction - the vector field we want is simply the conditional expectation of these directions, i.e.,

$$\frac{d}{dt}x_t = \mathbb{E}_{x_1}\left(\frac{x_1 - x_t}{1 - t} \mid x_t, t\right).$$

We train this as follows.

1. Sample $x_0 \sim N(0, I)$.
2. Sample $x_1 \sim q(x_1)$.
3. Sample $t \sim U(0, 1)$.
4. Compute $x_t = (1-t)x_0 + tx_1$.
5. Take a gradient step to minimize $\left|\left|\frac{x_1 - x_t}{1-t} - v_\theta(x_t, t)\right|\right|^2$.

To show why taking the conditional expectation gives us a vector field that transports $p_0(x)$ to $p_1(x)$, we propose the following evolution of probability distribution:

$$p_t(x) = \int p_t(x \mid x_1) q(x_1) dx_1.$$

We can check that plugging $t = 0$ and $t = 1$ gives the correct distribution (p_0 and p_1). Thus, it remains to show that this satisfies the continuity equation with respect to the conditional expectation vector field.

Integrating both sides over x_1 with respect to $q(x)$ on the continuity equation gives

$$\int q(x_1) \frac{d}{dt} p_t(x \mid x_1) dx_1 = -\int q(x_1) \nabla_x \cdot (p_t(x \mid x_1) v(x \mid t, x_1)) dx_1.$$

Rearranging gives

$$\frac{d}{dt} p_t(x) = -\nabla_x \cdot \left(p_t(x) \int \frac{q(x_1) p_t(x \mid x_1)}{p_t(x)} v(x \mid t, x_1) dx_1 \right).$$

Now, $\frac{q(x_1)p_t(x|x_1)}{p_t(x)}$ is the conditional probability of x_1 given we are at x at time t, so $\int \frac{q(x_1)p_t(x|x_1)}{p_t(x)} v(x \mid t, x_1)$ is simply the conditional expectation of the simple vector fields for each data point.

To recap, we constructed a simple vector field if there is only one target sample; we then proposed to simply averaging the vector fields constructed in this manner for all target samples; lastly, we proved this conditional expectation vector field is the vector field we are looking for.