Lecture 7: Generative modeling by flow-matching

2025–06–11

1.1 Dirac delta function

Definition 1. We define the Dirac delta function $\delta(x)$ such that

$$\int_{\mathbb{R}^d} \delta(x) f(x) dx = f(0) \tag{1.1}$$

for any bounded continuous function $f : \mathbb{R}^d \to \mathbb{R}$.

The following lemma, which follows directly from the definition above, summarizes basic properties of δ function.

Lemma 1. The following two identities hold.

1. For any $x_0 \in \mathbb{R}^d$, we have

$$\int_{\mathbb{R}^d} \delta(x - x_0) f(x) dx = f(x_0) \,. \tag{1.2}$$

2. $\int_{\mathbb{R}^d} \delta(x) dx = 1$.

Proof. Applying a simple change of variables $x \to x + x_0$ and using (1.1) we obtain

$$\int_{\mathbb{R}^d} \delta(x - x_0) f(x) dx = \int_{\mathbb{R}^d} \delta(x) f(x + x_0) dx = f(x_0) \,.$$

Choosing $f \equiv 1$ in (1.1), we get the second identity. □

Remark.

1. Lemma 1 suggests that $\delta(x - x_0)dx$ can be viewed as the probability distribution where the probability is one at the single point $x = x_0$ and the probability is zero elsewhere.

2. Intuitively, delta function can be thought as the limit of Gaussian density

$$\psi_\sigma(x) = (2\pi\sigma^2)^{-\frac{d}{2}} \mathrm{e}^{-\frac{|x|^2}{2\sigma^2}} \,, \quad x \in \mathbb{R}^d \tag{1.3}$$

as $\sigma \to 0+$. In fact, we can prove that, for a bounded smooth function $f : \mathbb{R}^d \to \mathbb{R}$,

$$\lim_{\sigma \to 0+} \int_{\mathbb{R}^d} f(x) \psi_\sigma(x) dx = f(0) = \int_{\mathbb{R}^d} \delta(x) f(x) dx \,. \tag{1.4}$$

Consider a C^1-smooth map $\xi : \mathbb{R}^d \to \mathbb{R}^k$, where $1 \le k < d$. We will often encounter the integral

$$\int_{\mathbb{R}^d} f(x) \delta(z - \xi(x)) dx \,, \tag{1.5}$$

where $z \in \mathbb{R}^k$. The integral (1.5) can be rigorously defined as the limit of integrations with respect to Gaussian densities (see (1.4)). A useful identity related to (1.5) is given in the following lemma.

Lemma 2. For two test functions $f : \mathbb{R}^d \to \mathbb{R}$ and $g : \mathbb{R}^k \to \mathbb{R}$, we have

$$\int_{\mathbb{R}^d} f(x)g(\xi(x))dx = \int_{\mathbb{R}^k} \Big(\int_{\mathbb{R}^d} f(x)\delta(z-\xi(x))dx\Big)g(z)dz\,. \tag{1.6}$$

Proof. Exchanging the order of two integrals and using the first identity (1.2) in Lemma 1, we can derive

$$\begin{aligned}
&\int_{\mathbb{R}^k} \Big(\int_{\mathbb{R}^d} f(x)\delta(z-\xi(x))dx\Big)g(z)dz \\
=&\int_{\mathbb{R}^d} f(x)\Big(\int_{\mathbb{R}^k} g(z)\delta(z-\xi(x))dz\Big)dx \\
=&\int_{\mathbb{R}^d} f(x)g(\xi(x))dx\,.
\end{aligned}$$

□

From Lemma 2, it is not difficult to verify that, for any fixed $z \in \mathbb{R}^k$,

$$\int_{\mathbb{R}^d} g(\xi(x))f(x)\delta(z-\xi(x))dx = g(z)\int_{\mathbb{R}^d} f(x)\delta(z-\xi(x))dx\,. \tag{1.7}$$

Example (Linear map ξ)**.** Let $x = (y, z) \in \mathbb{R}^d$, where $y \in \mathbb{R}^{d-k}$ and $z \in \mathbb{R}^k$ denote the first $d-k$ components and the last k components of x, respectively. Consider the linear map

$$\xi(x) = \xi(y, z) = z\,, \quad \forall x = (x, y) \in \mathbb{R}^d\,. \tag{1.8}$$

Then, the integral (1.5) can be expressed as

$$\begin{aligned}
&\int_{\mathbb{R}^d} f(x')\delta(z-\xi(x'))dx' \\
=&\int_{\mathbb{R}^k}\int_{\mathbb{R}^{d-k}} f(y',z')\delta(z-z')dy'dz' \\
=&\int_{\mathbb{R}^{d-k}}\Big(\int_{\mathbb{R}^k} f(y',z')\delta(z-z')dz'\Big)dy' \\
=&\int_{\mathbb{R}^d} f(y',z)dy'\,,
\end{aligned} \tag{1.9}$$

where we have exchanged the order of integrals and used the identity (1.2) of Lemma 1.

To summarize, in the linear case, the integral (1.5) is simply the integration of $f(y, z)$ with respect to its first variable y.

1.2 Conditional expectation

Let X be a random variable in $\mathbb{R}^d$ whose probability density is $p(x)$. The following result provides the probability density of $Z = \xi(X)$, which is a random variable taking values in $\mathbb{R}^k$.

Lemma 3. The probability density of $Z = \xi(X)$ is given by

$$Q(z) = \int_{\mathbb{R}^d} p(x)\delta(z - \xi(x))dx\,, \quad z \in \mathbb{R}^k\,. \tag{1.10}$$

Proof. Let $g : \mathbb{R}^k \to \mathbb{R}$ be a test function. Using the fact that the probability density of X is p and $Z = \xi(X)$, let us derive

$$\begin{aligned} &\int_{\mathbb{R}^k} g(z)Q(z)dz \\ =&\mathbb{E}_{Z\sim Q}(g(Z)) \\ =&\mathbb{E}_{X\sim p}\big(g(\xi(X))\big) \\ =&\int_{\mathbb{R}^d} g(\xi(x))p(x)dx \\ =&\int_{\mathbb{R}^k} g(z)\Big(\int_{\mathbb{R}^d} p(x)\delta(z-\xi(x))dx\Big)dz\,, \end{aligned} \tag{1.11}$$

where we have used Lemma 2 to derive the last equality. Since the derivation (1.11) above is true for a general test function g, we conclude that the density of $Z = \xi(X)$ is given by (1.10). □

Choosing $g \equiv 1$ in (1.11), we have that

$$\int_{\mathbb{R}^k} Q(z)dz = \int_{\mathbb{R}^d} p(x)dx = 1\,. \tag{1.12}$$

Therefore, Q is indeed a probability density. Informally, the expression (1.10) states that the density of $Z = \xi(X)$ at z is the "sum" of the density $p(x)$ for all $x \in \mathbb{R}^d$ such that $\xi(x) = z$. In literature, Q is often termed as "marginal density".

Now, we introduce conditional expectation.

Definition 2. Let X be a random variable in $\mathbb{R}^d$ whose probability density is p. Let $z \in \mathbb{R}^k$ and $f : \mathbb{R}^d \to \mathbb{R}$, we define the expectation of $f(X)$ conditioned on the event that $\xi(X) = z$ as

$$\mathbb{E}\big(f(X)\big|\xi(X) = z\big) = \frac{\int_{\mathbb{R}^d} f(x)\delta(\xi(x) - z)p(x)dx}{Q(z)}\,, \tag{1.13}$$

where $Q(z)$ is defined in (1.10).

Example. As in the previous example, consider again the linear map in (1.8). Using a similar argument as in (1.9), we can derive

$$\begin{aligned}&\int_{\mathbb{R}^d} f(x)\delta(\xi(x)-z)p(x)dx = \int_{\mathbb{R}^{d-k}} f(y,z)p(y,z)dy\\&Q(z) = \int_{\mathbb{R}^{d-k}} p(y,z)dy\,.\end{aligned} \tag{1.14}$$

Therefore, in this case, the conditional expectation can be expressed as

$$\mathbb{E}\big(f(X)\big|\xi(X)=z\big) = \frac{\int_{\mathbb{R}^{d-k}} f(y,z)p(y,z)dy}{\int_{\mathbb{R}^{d-k}} p(y,z)dy}\,. \tag{1.15}$$

In the following result, we state a useful identity related to the conditional expectation.

Proposition 1 (Law of total expectation)**.** For a test function $f:\mathbb{R}^d\to\mathbb{R}$, we have

$$\mathbb{E}(f(X)) = \mathbb{E}_{z\sim Q}\Big(\mathbb{E}\big(f(X)\big|\xi(X)=z\big)\Big)\,. \tag{1.16}$$

Proof. Using (1.13), we can compute the right hand side of (1.16) as

$$\begin{aligned}&\mathbb{E}_{z\sim Q}\Big(\mathbb{E}\big(f(X)\big|\xi(X)=z\big)\Big)\\&=\int_{\mathbb{R}^k}\Big(\mathbb{E}\big(f(X)\big|\xi(X)=z\big)\Big)Q(z)dz\\&=\int_{\mathbb{R}^k}\Big(\frac{\int_{\mathbb{R}^d} f(x)\delta(\xi(x)-z)p(x)dx}{Q(z)}\Big)Q(z)dz\\&=\int_{\mathbb{R}^k}\int_{\mathbb{R}^d} f(x)\delta(\xi(x)-z)p(x)dxdz\\&=\int_{\mathbb{R}^d} f(x)p(x)\Big(\int_{\mathbb{R}^k}\delta(\xi(x)-z)dz\Big)dx\\&=\int_{\mathbb{R}^d} f(x)p(x)dx\\&=\mathbb{E}(f(X))\,.\end{aligned} \tag{1.17}$$

□

Proposition 1 states that the (full) expectation of $f(X)$ can be computed in two separate steps, namely, by first taking expectation conditioning on $\xi(X)=z$, and then taking expectation with respect to the density $Q(z)$.

We conclude this section with a characterization of conditional expectation.

Proposition 2. Given a function $f:\mathbb{R}^d\to\mathbb{R}$, define

$$\widetilde{f}(z) = \mathbb{E}\big(f(X)\big|\xi(X)=z\big)\,,\quad z\in\mathbb{R}^k. \tag{1.18}$$

Then, for any $g : \mathbb{R}^k \to \mathbb{R}$, we have

$$\mathbb{E}\Big(\big|f(X) - \widetilde{f}(\xi(X))\big|^2\Big) \leq \mathbb{E}\Big(|f(X) - g(\xi(X))|^2\Big) . \tag{1.19}$$

Proof. Let us compute the right hand side of (1.19). Using Proposition 1, we have

$$\begin{aligned}
&\mathbb{E}\Big(\big|f(X) - g(\xi(X))\big|^2\Big) \\
=&\mathbb{E}(f^2(X)) + \mathbb{E}\Big(- 2f(X)g(\xi(X)) + g^2(\xi(X))\Big) \\
=&\mathbb{E}(f^2(X)) + \mathbb{E}_{z\sim Q}\Big[\mathbb{E}\Big(- 2f(X)g(\xi(X)) + g^2(\xi(X))\Big|\xi(X) = z\Big)\Big] \\
=&\mathbb{E}(f^2(X)) + \mathbb{E}_{z\sim Q}\Big(- 2g(z)\mathbb{E}\big(f(X)\big|\xi(X) = z\big) + g^2(z)\Big) \\
=&\mathbb{E}(f^2(X)) + \mathbb{E}_{z\sim Q}\Big[\Big|g(z) - \mathbb{E}\big(f(X)\big|\xi(X) = z\big)\Big|^2 - \Big(\mathbb{E}\big(f(X)\big|\xi(X) = z\big)\Big)^2\Big] \\
=&\mathbb{E}(f^2(X)) - \mathbb{E}_{z\sim Q}\big(\widetilde{f}^2(z)\big) + \mathbb{E}_{z\sim Q}\Big(\big|g(z) - \widetilde{f}(z)\big|^2\Big) .
\end{aligned} \tag{1.20}$$

Since the above derivation holds for a general test function g, taking $g = \widetilde{f}$, we see that the left hand side of (1.19) can be written as

$$\mathbb{E}\Big(\big|f(X) - \widetilde{f}(\xi(X))\big|^2\Big) = \mathbb{E}(f^2(X)) - \mathbb{E}_{z\sim Q}\big(\widetilde{f}^2(z)\big) . \tag{1.21}$$

Substituting (1.21) into (1.20), we obtain, for a general test function g,

$$\mathbb{E}\Big(\big|f(X) - g(\xi(X))\big|^2\Big) = \mathbb{E}\Big(\big|f(X) - \widetilde{f}(\xi(X))\big|^2\Big) + \mathbb{E}_{z\sim Q}\Big(\big|g(z) - \widetilde{f}(z)\big|^2\Big) ,$$

which implies (1.19). □

Proposition 2 states that the conditional expectation (1.18) minimizes the mean square error from $f(x)$ among all functions of the form $g(\xi(x))$.

1.3 Flow-based generative models

Assume that a dataset is given, which is sampled from a target density $p_1 = p_{\text{target}}$ on $\mathbb{R}^d$. Also assume that a prior density p_0 on $\mathbb{R}^d$, typically a Gaussian density, is chosen, such that one can directly generate samples according to p_0.

For $t \in [0, 1]$, we define $p(\cdot, t)$ as the probability density of the random variable

$$X_t = (1 - t)X_0 + tX_1 , \quad \text{where } X_0 \sim p_0, \quad X_1 \sim p_1 . \tag{1.22}$$

Because $X_t = X_0$ at $t = 0$ and $X_t = X_1$ at $t = 1$, we clearly have

$$p(\cdot, 0) = p_0 , \quad p(\cdot, 1) = p_1 . \tag{1.23}$$

Our goal is to learn a vector field $u : \mathbb{R}^d \times [0, 1] \to \mathbb{R}^d$ such that, starting from a random initial state $Y_0 \sim p_0$, the transition density of the ODE

$$\frac{dY_t}{dt} = u(Y_t, t) , \quad t \in [0, 1] \tag{1.24}$$

at time t coincides with $p(\cdot,t)$ for any $t \in [0,1]$. In particular, by construction we have $Y_1 \sim p(\cdot,1) = p_1$. Therefore, if we are able to learn the vector field u, we can sample the initial state Y_0 according to the prior p_0 and then simulate (1.24) to get samples Y_1 that are distributed according to the target density.

To achieve this goal, let us first recall that the probability density of Y_t under the ODE (1.24), denoted by $q(\cdot,t)$, satisfies the following continuity equation

$$\frac{\partial q}{\partial t} + \operatorname{div}(uq) = 0\,. \tag{1.25}$$

The following result provides the equation that is satisfied by the density $p(\cdot,t)$.

Proposition 3. The probability density $p(x,t)$ of X_t in (1.22) solves the equation

$$\frac{\partial p(x,t)}{\partial t} + \operatorname{div}\Big(\mathbb{E}\big(X_1 - X_0 \big| X_t = x\big) p(x,t)\Big) = 0 \tag{1.26}$$

in a weak sense.

Proof. Let $f \in \mathbb{R}^d \times [0,1] \to \mathbb{R}$ be a C^1-smooth test function with compact support and

$$f(\cdot,0) = f(\cdot,1) = 0\,. \tag{1.27}$$

Integrating by parts and using (1.27), we have

$$\int_0^1 \int_{\mathbb{R}^d} f(x,t) \frac{\partial p(x,t)}{\partial t} dx dt = -\int_0^1 \int_{\mathbb{R}^d} \frac{\partial f(x,t)}{\partial t} p(x,t) dx dt\,. \tag{1.28}$$

For the right hand side of (1.28), we can derive

$$\begin{aligned} & -\int_0^1 \int_{\mathbb{R}^d} \frac{\partial f(x,t)}{\partial t} p(x,t) dx dt \\ =& -\int_0^1 \mathbb{E}\Big(\frac{\partial f}{\partial t}(X_t,t)\Big) dt \\ =& -\int_0^1 \mathbb{E}\Big(\frac{df}{dt}(X_t,t) - \nabla f(X_t,t) \cdot \frac{dX_t}{dt}\Big) dt\,, \end{aligned} \tag{1.29}$$

where the last equality follows from the chain rule. Using (1.22) and (1.27), we can continue to derive

$$\begin{aligned} & -\int_0^1 \int_{\mathbb{R}^d} \frac{\partial f(x,t)}{\partial t} p(x,t) dx dt \\ =& -\int_0^1 \mathbb{E}\Big(\frac{df}{dt}(X_t,t) - \nabla f(X_t,t) \cdot \frac{dX_t}{dt}\Big) dt \\ =& -\mathbb{E}\Big(f(X_1,1) - f(X_0,0)\Big) + \int_0^1 \mathbb{E}\Big(\nabla f(X_t,t) \cdot (X_1 - X_0)\Big) dt \\ =& \int_0^1 \mathbb{E}\Big(\nabla f(X_t,t) \cdot (X_1 - X_0)\Big) dt\,. \end{aligned} \tag{1.30}$$

To proceed, we use the law of total expectation (see Proposition 1) and derive

$$
\begin{aligned}
& -\int_0^1 \int_{\mathbb{R}^d} \frac{\partial f(x,t)}{\partial t} p(x,t) dx dt \\
=& \int_0^1 \mathbb{E}\Big(\nabla f(X_t,t)\cdot (X_1 - X_0)\Big) dt \\
=& \int_0^1 \mathbb{E}_{x\sim p(x,t)} \Big[\mathbb{E}\Big(\nabla f(X_t,t)\cdot (X_1 - X_0)\Big| X_t = x\Big)\Big] dt \\
=& \int_0^1 \mathbb{E}_{x\sim p(x,t)} \Big(\nabla f(x,t)\cdot \mathbb{E}\big(X_1 - X_0\big| X_t = x\big)\Big) dt \\
=& \int_0^1 \int_{\mathbb{R}^d} \nabla f(x,t)\cdot \mathbb{E}\big(X_1 - X_0\big| X_t = x\big) p(x,t)\, dx dt \\
=& -\int_0^1 \int_{\mathbb{R}^d} f(x,t) \mathrm{div}\Big(\mathbb{E}\big(X_1 - X_0\big| X_t = x\big) p(x,t)\Big) dx dt\,,
\end{aligned}
\tag{1.31}
$$

where the last equality follows from integration by parts. The equation (1.26) is obtained after combining (1.28) and (1.31). □

Comparing (1.25) and (1.26), we can see that, in order to match the density $q(\cdot,t)$ of the ODE flow with the density $p(\cdot,t)$, we can choose the vector field u as the conditional expectation in (1.26). Specifically, we consider the objective

$$
\begin{aligned}
& \int_0^1 \mathbb{E}_{x\sim p(x,t)} \Big(\Big|u(x,t) - \mathbb{E}\big(X_1 - X_0\big| X_t = x\big)\Big|^2\Big) dt \\
=& \int_0^1 \mathbb{E}_{x\sim p(x,t)} \Big(|u(x,t)|^2 - 2u(x,t)\cdot \mathbb{E}\big(X_1 - X_0\big| X_t = x\big)\Big) dt + C \\
=& \int_0^1 \mathbb{E}_{x\sim p(x,t)} \Big[\mathbb{E}\Big(|u(X_t,t)|^2 - 2u(X_t,t)\cdot (X_1 - X_0)\Big| X_t = x\Big)\Big] dt + C \\
=& \int_0^1 \mathbb{E}_{X_0\sim p_0, X_1\sim p_1} \Big(\Big|u\big((1-t)X_0 + tX_1, t\big) - (X_1 - X_0)\Big|^2\Big) dt + C\,,
\end{aligned}
\tag{1.32}
$$

where C is a constant independent of u and we have used Proposition 1 to derive the last equality. The derivation above implies that we can learn the conditional expectation in (1.26) using the loss function

$$
\mathrm{Loss}(u) = \mathbb{E}_{t\sim U[0,1]} \mathbb{E}_{X_0\sim p_0, X_1\sim p_1} \Big(\Big|u\big((1-t)X_0 + tX_1, t\big) - (X_1 - X_0)\Big|^2\Big)\,. \tag{1.33}
$$

In fact, instead of the linear interpolation in (1.22), we can consider a more general "interpolant" $I_t : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$, for $t \in [0,1]$, which satisfies

$$
I_0(x,y) = x, \text{ and } \quad I_1(x,y) = y, \quad x, y \in \mathbb{R}^d\,, \tag{1.34}
$$

and define

$$
X_t = I_t(X_0, X_1)\,, \quad \text{where } X_0 \sim p_0, \quad X_1 \sim p_1\,. \tag{1.35}
$$

In this general setting, Proposition 3 can be generalized, using the same argument, to the following result.

Proposition 4. The probability density $p(x,t)$ of X_t in (1.35) solves the equation

$$\frac{\partial p(x,t)}{\partial t} + \operatorname{div}\Big(\mathbb{E}\big(\partial_t I_t(X_0,X_1)\big|X_t = x\big)p(x,t)\Big) = 0 \tag{1.36}$$

in a weak sense.

Accordingly, the vector field of the ODE flow can be learned with the following objective function

$$\text{Loss}(u) = \mathbb{E}_{t\sim U[0,1]}\mathbb{E}_{X_0\sim p_0, X_1\sim p_1}\Big(\big|u\big(I_t(X_0,X_1),t\big) - \partial_t I_t(X_0,X_1)\big|^2\Big). \tag{1.37}$$