Lecture 8: Denoising diffusion probabilistic models and normalizing flows

2025–06–18

1.1 Denoising diffusion probabilistic models

The content of this section is largely based on the work [2].

1.1.1 Basic

We denote by $\mathcal{N}(x;\mu,\Sigma)$ the density of the Gaussian random variable with mean $\mu\in\mathbb{R}^d$ and covariance matrix $\Sigma\in\mathbb{R}^{d\times d}$. Precisely,

$$\mathcal{N}(x;\mu,\Sigma)=(2\pi)^{-\frac{d}{2}}(\det\Sigma)^{-\frac{1}{2}}\mathrm{e}^{-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)}\,,\quad x\in\mathbb{R}^d\,. \tag{1.1}$$

Let π be a probability density on $\mathbb{R}^d$ and $f:\mathbb{R}\to\mathbb{R}$ be a convex function. Recall the Jensen's inequality

$$f\big(\mathbb{E}_\pi(g(x))\big)=f\Big(\int_{\mathbb{R}^d}g(x)\pi(x)dx\Big)\le\mathbb{E}_\pi\big(f(g(x))\big)\,, \tag{1.2}$$

which holds for any function $g:\mathbb{R}^d\to\mathbb{R}$.

Let us also recall the Kullback-Leibler (KL) divergence (or relative entropy).

Definition 1. Let Q_1 and Q_2 be two probability densities on the same measure space, where Q_1 is absolutely continuous with respect to Q_2. The KL divergence from Q_2 to Q_1 is defined as

$$D_{KL}(Q_1\,|\,Q_2):=\mathbb{E}_{Q_1}\Big(\log\frac{Q_1}{Q_2}\Big)\,. \tag{1.3}$$

Note that, in general, KL divergence is not symmetric, that is, $D_{KL}(Q_1\,|\,Q_2)$ and $D_{KL}(Q_2\,|\,Q_1)$ are not equal. Moreover, Jensen's inequality implies that $D_{KL}(Q_1\,|\,Q_2)\ge 0$ and it equals zero if and only if the two densities Q_1 and Q_2 are identical.

1.1.2 Theoretical background

Assume that the data distribution is $q_0(x)dx$ on $\mathbb{R}^d$. Denoising diffusion probabilistic models (DDPMs) are a class of generative models built on Markov chains. Specifically, given data $x^{(0)}\in\mathbb{R}^d$, states $x^{(1)},\dots,x^{(N)}\in\mathbb{R}^d$ are generated by evolving a Markov chain, which is called the forward process. The joint probability density of $x^{(1)},\dots,x^{(N)}$ given $x^{(0)}$ is

$$q(x^{(1:N)}\,|\,x^{(0)})=\prod_{k=1}^{N}q(x^{(k)}\,|\,x^{(k-1)})\,, \tag{1.4}$$

where $q(x^{(k)}\,|\,x^{(k-1)})$ is the transition density of the forward process, chosen as (Gaussian density)

$$q(x^{(k)}\,|\,x^{(k-1)})=\mathcal{N}(x^{(k)};\sqrt{1-\beta_k}x^{(k-1)},\beta_k\mathrm{I}_d)\,, \tag{1.5}$$

and $\beta_k > 0$ is a constant.

The generative process, also called the reverse process, is a Markov chain on $\mathbb{R}^d$ that is learnt to reproduce the data by reversing the forward process. Its joint probability density is

$$p_\theta(x^{(0:N)}) = p(x^{(N)}) \prod_{k=1}^{N} p_\theta(x^{(k-1)} \,|\, x^{(k)}) \,, \tag{1.6}$$

where $p(x^{(N)})$ is a (fixed) prior density, θ is the parameter to be learnt, and $p_\theta(x^{(k-1)} \,|\, x^{(k)})$ is the transition density of the reverse process, chosen as

$$p_\theta(x^{(k-1)} \,|\, x^{(k)}) = \mathcal{N}\Big(x^{(k-1)}; \mu_\theta(x^{(k)}, k), \Sigma_\theta(x^{(k)}, k)\Big) \,. \tag{1.7}$$

The probability density of $x^{(0)}$ generated by the reverse process is therefore $p_\theta(x^{(0)}) = \int p_\theta(x^{(0:N)})\, dx^{(1:N)}$. The learning objective is based on the standard variational bound on the negative log-likelihood. Specifically, using (1.4) and (1.6), and applying Jensen's inequality (for the convex function $f(x) = -\log(x)$), we can derive

$$\begin{aligned}
\mathbb{E}_{q_0}\big(-\log p_\theta(x^{(0)})\big) =& \mathbb{E}_{q_0}\Big[-\log\Big(\int p_\theta(x^{(0:N)})\, dx^{(1:N)}\Big)\Big] \\
=& \mathbb{E}_{q_0}\Big[-\log\Big(\int \frac{p_\theta(x^{(0:N)})}{q(x^{(1:N)} \,|\, x^{(0)})} q(x^{(1:N)} \,|\, x^{(0)})\, dx^{(1:N)}\Big)\Big] \\
\le& \mathbb{E}_{q_0}\Big[\int -\log\Big(\frac{p_\theta(x^{(0:N)})}{q(x^{(1:N)} \,|\, x^{(0)})}\Big) q(x^{(1:N)} \,|\, x^{(0)})\, dx^{(1:N)}\Big] \\
=& \mathbb{E}_{\mathbb{Q}}\Big(-\log \frac{p_\theta(x^{(0:N)})}{q(x^{(1:N)} \,|\, x^{(0)})}\Big) \\
=& \mathbb{E}_{\mathbb{Q}}\Big(-\log p(x^{(N)}) - \sum_{k=1}^{N} \log \frac{p_\theta(x^{(k-1)} \,|\, x^{(k)})}{q(x^{(k)} \,|\, x^{(k-1)})}\Big) =: L \,,
\end{aligned} \tag{1.8}$$

where $\mathbb{E}_{q_0}, \mathbb{E}_{\mathbb{Q}}$ denote the expectation with respect to the data distribution and the expectation with respect to the joint density $q(x^{(0:N)})$ under the forward process, respectively.

We can learn the parameter θ by minimizing the upper bound L in the last line of (1.8). However, as we see below, we can obtain a simplified objective function based on analyzing the upper bound L.

First, since each step of the forward process follows the Gaussian distribution in (1.5), the state $x^{(k)}$ of the forward process at any step k given $x^{(0)}$ also follows a Gaussian distribution, whose density is provided by the following result.

Lemma 1. Let $x^{(k)}$ be the forward process with transition density in (1.5). Then, we have

$$q(x^{(k)}|x^{(0)}) = \mathcal{N}(x^{(k)}; \sqrt{\bar{\alpha}_k} x^{(0)}, (1 - \bar{\alpha}_k)\mathrm{I}_d) \,, \tag{1.9}$$

where

$$\alpha_k = 1 - \beta_k\,, \quad \text{and} \quad \bar{\alpha}_k = \prod_{i=1}^{k} \alpha_i\,. \tag{1.10}$$

Proof. We prove (1.9) by induction. When $k = 1$, from (1.10) we have $\bar{\alpha}_k = \alpha_1 = 1 - \beta_1$. Therefore, (1.9) holds since it coincides with (1.5). Now, consider $k > 1$ and assume that (1.9) holds for $k-1$. Then, we have

$$x^{(k-1)} = \sqrt{\bar{\alpha}_{k-1}}x^{(0)} + \sqrt{1-\bar{\alpha}_{k-1}}r^{(k-1)}\,, \quad r^{(k-1)} \sim \mathcal{N}(0, \mathrm{I}_d)\,.$$

Similarly, from (1.5), we have

$$x^{(k)} = \sqrt{1-\beta_k}x^{(k-1)} + \sqrt{\beta_k}\tilde{r}^{(k-1)}\,, \quad \tilde{r}^{(k-1)} \sim \mathcal{N}(0, \mathrm{I}_d)\,.$$

Combining the two identities above, we obtain

$$\begin{aligned} x^{(k)} =& \sqrt{1-\beta_k}\big(\sqrt{\bar{\alpha}_{k-1}}x^{(0)} + \sqrt{1-\bar{\alpha}_{k-1}}r^{(k-1)}\big) + \sqrt{\beta_k}\tilde{r}^{(k-1)} \\ =& \sqrt{(1-\beta_k)\bar{\alpha}_{k-1}}x^{(0)} + \sqrt{(1-\beta_k)(1-\bar{\alpha}_{k-1})}r^{(k-1)} + \sqrt{\beta_k}\tilde{r}^{(k-1)}\,, \end{aligned}$$

which clearly shows that, given $x^{(0)}$, $x^{(k)}$ is Gaussian. Calculating its mean and covariance and also using the fact that $r^{(k-1)}$ and $\tilde{r}^{(k-1)}$ are independent standard Gaussian random variables, we have

$$\begin{aligned} & \mathbb{E}(x^{(k)}|x^{(0)}) = \sqrt{(1-\beta_k)\bar{\alpha}_{k-1}}x^{(0)} = \sqrt{\bar{\alpha}_k}x^{(0)}\,, \\ \text{and} \quad & \mathbb{E}\Big(\big(x^{(k)} - \sqrt{\bar{\alpha}_k}x^{(0)}\big)\big(x^{(k)} - \sqrt{\bar{\alpha}_k}x^{(0)}\big)^\top \Big| x^{(0)}\Big) \\ =& \Big((1-\beta_k)(1-\bar{\alpha}_{k-1}) + \beta_k\Big)\mathrm{I}_d \\ =& (1-\bar{\alpha}_k)\mathrm{I}_d\,, \end{aligned}$$

where we have used (1.10). This shows that (1.9) holds for k. □

Second, we can rewrite the variational upper bound L as follows.

Proposition 1. We have

$$L = L_N + \sum_{k=2}^{N} L_{k-1} + L_0\,, \tag{1.11}$$

where

$$\begin{aligned} L_0 =& -\mathbb{E}_{\mathbb{Q}}\Big(\log p_\theta(x^{(0)}|x^{(1)})\Big)\,, \\ L_{k-1} =& \mathbb{E}_{\mathbb{Q}}\Big[D_{KL}\Big(q\big(x^{(k-1)} \,|\, x^{(k)}, x^{(0)}\big)\,\Big|\, p_\theta\big(x^{(k-1)} \,|\, x^{(k)}\big)\Big)\Big]\,, \quad k = 2, \dots, N\,, \\ L_N =& \mathbb{E}_{\mathbb{Q}}\Big[D_{KL}\Big(q\big(x^{(N)}|x^{(0)}\big)\,\Big|\, p(x^{(N)})\Big)\Big]\,. \end{aligned} \tag{1.12}$$

Proof. For $k > 1$, we have the following identity

$$q(x^{(k)} \,|\, x^{(k-1)})\, q(x^{(k-1)}|x^{(0)}) = q(x^{(k-1)} \,|\, x^{(k)}, x^{(0)})\, q(x^{(k)}|x^{(0)})\,, \quad (1.13)$$

from which we obtain

$$q(x^{(k)} \,|\, x^{(k-1)}) = q(x^{(k-1)} \,|\, x^{(k)}, x^{(0)}) \frac{q(x^{(k)}|x^{(0)})}{q(x^{(k-1)}|x^{(0)})}\,. \quad (1.14)$$

Substituting (1.14) into the expression of L in (1.8) and using the definition of KL divergence in Definition 1, we can derive

$$\begin{aligned}
L =& \mathbb{E}_{\mathbb{Q}}\Big(-\log p(x^{(N)}) - \sum_{k=1}^{N} \log \frac{p_\theta(x^{(k-1)} \,|\, x^{(k)})}{q(x^{(k)} \,|\, x^{(k-1)})} \Big) \\
=& \mathbb{E}_{\mathbb{Q}}\Big(\log \frac{q(x^{(N)}|x^{(0)})}{p(x^{(N)})} + \sum_{k=2}^{N} \log \frac{q(x^{(k-1)} \,|\, x^{(k)}, x^{(0)})}{p_\theta(x^{(k-1)} \,|\, x^{(k)})} - \log p_\theta(x^{(0)}|x^{(1)}) \Big) \\
=& \mathbb{E}_{\mathbb{Q}}\Big[D_{KL}\Big(q(x^{(N)}|x^{(0)}) \Big| p(x^{(N)}) \Big) \\
& + \sum_{k=2}^{N} D_{KL}\Big(q(x^{(k-1)} \,|\, x^{(k)}, x^{(0)}) \,\Big|\, p_\theta(x^{(k-1)} \,|\, x^{(k)}) \Big) \\
& - \log p_\theta(x^{(0)}|x^{(1)}) \Big]\,,
\end{aligned}$$

which implies (1.11) and (1.12). □

In (1.11), the term L_N is independent of θ, while the conditional density $q(x^{(k-1)}|x^{(k)}, x^{(0)})$ in L_{k-1} is Gaussian with explicit expression given by the following result.

Lemma 2. Let $x^{(k)}$ be the forward process with transition density in (1.5). Then, we have

$$q(x^{(k-1)}|x^{(k)}, x^{(0)}) = \mathcal{N}(x^{(k-1)}; \widetilde{\mu}_k(x^{(k)}, x^{(0)}), \widetilde{\beta}_k \mathrm{I}_d)\,, \quad (1.15)$$

where

$$\begin{aligned}
\widetilde{\mu}_k(x^{(k)}, x^{(0)}) &= \frac{\sqrt{\bar{\alpha}_{k-1}}\beta_k}{1-\bar{\alpha}_k} x^{(0)} + \frac{\sqrt{\alpha_k}(1-\bar{\alpha}_{k-1})}{1-\bar{\alpha}_k} x^{(k)}\,, \\
\widetilde{\beta}_k &= \frac{1-\bar{\alpha}_{k-1}}{1-\bar{\alpha}_k}\beta_k\,.
\end{aligned} \quad (1.16)$$

Proof. Using (1.13), (1.5), and Lemma 1, we can derive

$$\begin{aligned}
&q(x^{(k-1)} \,|\, x^{(k)}, x^{(0)}) \\
=&q(x^{(k)} \,|\, x^{(k-1)}) \frac{q(x^{(k-1)}|x^{(0)})}{q(x^{(k)}|x^{(0)})} \\
=&\mathcal{N}\big(x^{(k)}; \sqrt{1-\beta_k}x^{(k-1)}, \beta_k \mathrm{I}_d\big) \frac{\mathcal{N}\big(x^{(k-1)}; \sqrt{\bar{\alpha}_{k-1}}x^{(0)}, (1-\bar{\alpha}_{k-1})\mathrm{I}_d\big)}{\mathcal{N}\big(x^{(k)}; \sqrt{\bar{\alpha}_k}x^{(0)}, (1-\bar{\alpha}_k)\mathrm{I}_d\big)} \\
=&C \exp\Big(-\frac{|x^{(k)} - \sqrt{1-\beta_k}x^{(k-1)}|^2}{2\beta_k}\Big) \cdot \exp\Big(-\frac{|x^{(k-1)} - \sqrt{\bar{\alpha}_{k-1}}x^{(0)}|^2}{2(1-\bar{\alpha}_{k-1})}\Big) \\
&\cdot \exp\Big(\frac{|x^{(k)} - \sqrt{\bar{\alpha}_k}x^{(0)}|^2}{2(1-\bar{\alpha}_k)}\Big),
\end{aligned} \tag{1.17}$$

where C is a normalizing constant independent of $x^{(0)}$, $x^{(k-1)}$, and $x^{(k)}$.

We obtain (1.15) by completing the square in the last line of (1.17) and using the relations in (1.10). □

1.1.3 Practical choices

We discuss the choices for the density $p_\theta(x^{(k-1)} \,|\, x^{(k)})$ in (1.7) of the reverse process. For the covariance matrix, we choose

$$\Sigma_\theta(x^{(k)}, k) = \sigma_k^2 \mathrm{I}_d\,, \text{ where } \sigma_k^2 = \beta_k \text{ or } \sigma_k^2 = \widetilde{\beta}_k = \frac{1-\bar{\alpha}_{k-1}}{1-\bar{\alpha}_k}\beta_k. \tag{1.18}$$

For the parametrization of the mean $\mu_\theta(x^{(k)}, k)$, we are motivated by the following result on the term L_{k-1} in (1.12).

Lemma 3. Assume that

$$p_\theta(x^{(k-1)} \,|\, x^{(k)}) = \mathcal{N}\Big(x^{(k-1)}; \mu_\theta(x^{(k)}, k), \sigma_k^2 \mathrm{I}_d\Big)\,. \tag{1.19}$$

Then, the term L_{k-1} in (1.12), where $2 \le k \le N$, can be written as

$$L_{k-1} = \frac{1}{2\sigma_k^2}\mathbb{E}_{\mathbb{Q}}\Big(\big|\mu_\theta(x^{(k)}, k) - \widetilde{\mu}_k(x^{(k)}, x^{(0)})\big|^2\Big) + C\,, \tag{1.20}$$

where C is a constant independent of θ and $\widetilde{\mu}_k$ is defined in (1.16).

Proof. Let us denote by C a generic constant that is independent of θ. From (1.12) and Definition 1, we have

$$\begin{aligned}
L_{k-1} =&\mathbb{E}_{\mathbb{Q}}\Big[D_{KL}\Big(q(x^{(k-1)} \,|\, x^{(k)}, x^{(0)}) \,\Big|\, p_\theta(x^{(k-1)} \,|\, x^{(k)})\Big)\Big] \\
=&\mathbb{E}_{\mathbb{Q}}\Big[\mathbb{E}_{x^{(k-1)} \sim q(x^{(k-1)} \,|\, x^{(k)}, x^{(0)})}\Big(\log \frac{q(x^{(k-1)} \,|\, x^{(k)}, x^{(0)})}{p_\theta(x^{(k-1)} \,|\, x^{(k)})}\Big)\Big] \\
=&-\mathbb{E}_{\mathbb{Q}}\Big[\mathbb{E}_{x^{(k-1)} \sim q(x^{(k-1)} \,|\, x^{(k)}, x^{(0)})}\Big(\log p_\theta(x^{(k-1)} \,|\, x^{(k)})\Big)\Big] + C\,.
\end{aligned}$$

Using (1.19) and the fact that the mean of $x^{(k-1)} \sim q(x^{(k-1)} \,|\, x^{(k)}, x^{(0)})$ is

$\widetilde{\mu}_k(x^{(k)}, x^{(0)})$ (see (1.16)), we can derive

$$\begin{aligned}
L_{k-1} =& \frac{1}{2\sigma_k^2}\mathbb{E}_{\mathbb{Q}}\Big[\mathbb{E}_{x^{(k-1)}\sim q(x^{(k-1)}\,|\,x^{(k)},x^{(0)})}\Big(\big|x^{(k-1)} - \mu_\theta(x^{(k)},k)\big|^2\Big)\Big] + C \\
=& \frac{1}{2\sigma_k^2}\mathbb{E}_{\mathbb{Q}}\Big[\mathbb{E}_{x^{(k-1)}\sim q(x^{(k-1)}\,|\,x^{(k)},x^{(0)})}\Big(\big|\mu_\theta(x^{(k)},k)\big|^2 - 2x^{(k-1)}\cdot\mu_\theta(x^{(k)},k)\Big)\Big] + C \\
=& \frac{1}{2\sigma_k^2}\mathbb{E}_{\mathbb{Q}}\Big(\big|\mu_\theta(x^{(k)},k)\big|^2 - 2\widetilde{\mu}_k(x^{(k)},x^{(0)})\cdot\mu_\theta(x^{(k)},k)\Big) + C \\
=& \frac{1}{2\sigma_k^2}\mathbb{E}_{\mathbb{Q}}\Big(\big|\mu_\theta(x^{(k)},k) - \widetilde{\mu}_k(x^{(k)},x^{(0)})\big|^2\Big) + C\,,
\end{aligned}$$

which gives (1.20). □

The expression (1.20) suggests that $\mu_\theta(x^{(k)}, k)$ should predict $\widetilde{\mu}_k$. Note that (1.9) implies the parametrization

$$x^{(k)}(x^{(0)}, \varepsilon) = \sqrt{\bar{\alpha}_k}x^{(0)} + \sqrt{1-\bar{\alpha}_k}\varepsilon\,,\ \text{where } \varepsilon \sim \mathcal{N}(0, \mathrm{I}_d)\,. \tag{1.21}$$

With this, as well as the expression of $\widetilde{\mu}_k$ in (1.16), we can further write (1.20) as

$$\begin{aligned}
&L_{k-1} - C \\
=&\mathbb{E}_{x^{(0)},\varepsilon}\Big[\frac{1}{2\sigma_k^2}\Big|\mu_\theta\big(x^{(k)}(x^{(0)},\varepsilon),k\big) - \widetilde{\mu}_k\Big(x^{(k)}(x^{(0)},\varepsilon), \frac{1}{\sqrt{\bar{\alpha}_k}}\big(x^{(k)}(x^{(0)},\varepsilon) - \sqrt{1-\bar{\alpha}_k}\varepsilon\big)\Big)\Big|^2\Big] \\
=&\mathbb{E}_{x^{(0)},\varepsilon}\Big[\frac{1}{2\sigma_k^2}\Big|\mu_\theta\big(x^{(k)}(x^{(0)},\varepsilon),k\big) - \frac{1}{\sqrt{\alpha_k}}\Big(x^{(k)}(x^{(0)},\varepsilon) - \frac{\beta_k}{\sqrt{1-\bar{\alpha}_k}}\varepsilon\Big)\Big|^2\Big]\,.
\end{aligned} \tag{1.22}$$

Equation (1.22) suggests that $\mu_\theta(x^{(k)}, k)$ must predict $\frac{1}{\sqrt{\alpha_k}}\big(x^{(k)} - \frac{\beta_k}{\sqrt{1-\bar{\alpha}_k}}\varepsilon\big)$, given $x^{(k)}$. Therefore, we choose the parametrization

$$\mu_\theta(x^{(k)}, k) = \frac{1}{\sqrt{\alpha_k}}\Big(x^{(k)} - \frac{\beta_k}{\sqrt{1-\bar{\alpha}_k}}\varepsilon_\theta(x^{(k)}, k)\Big)\,, \tag{1.23}$$

where ε_θ is a function that predicts ε from $x^{(k)}$. Sampling $x^{(k-1)} \sim p_\theta(x^{(k-1)}|x^{(k)})$ can be achieved by

$$x^{(k-1)} = \frac{1}{\sqrt{\alpha_k}}\Big(x^{(k)} - \frac{\beta_k}{\sqrt{1-\bar{\alpha}_k}}\varepsilon_\theta(x^{(k)}, k)\Big) + \sigma_k z\,,\ \text{where } z \sim \mathcal{N}(0, \mathrm{I}_d)\,. \tag{1.24}$$

Furthermore, with (1.21) and (1.23), equation (1.22) simplifies to

$$\begin{aligned}
&L_{k-1} - C \\
=&\mathbb{E}_{x^{(0)},\varepsilon}\Big[\frac{\beta_k^2}{2\sigma_k^2\alpha_k(1-\bar{\alpha}_k)}\big|\varepsilon_\theta(\sqrt{\bar{\alpha}_k}x^{(0)} + \sqrt{1-\bar{\alpha}_k}\varepsilon, k) - \varepsilon\big|^2\Big]\,.
\end{aligned} \tag{1.25}$$

Empirically, we drop the factor in the expression (1.25) of L_{k-1} and use (instead of L) the simplified loss function

$$L_{\text{simple}}(\theta) = \mathbb{E}_{k,x^{(0)},\varepsilon}\Big[\big|\varepsilon_\theta(\sqrt{\bar{\alpha}_k}x^{(0)} + \sqrt{1-\bar{\alpha}_k}\varepsilon, k) - \varepsilon\big|^2\Big]\,. \tag{1.26}$$

The training and sampling algorithms are summarized in Algorithms 1-2.

Algorithm 1 Training

repeat
  $x^{(0)} \sim q_0(x^{(0)})$
  $k \sim \text{Uniform}(\{1, 2, \dots, N\})$
  $\varepsilon \sim \mathcal{N}(0, \mathrm{I}_d)$.
  take gradient descent step on
    $\nabla_\theta \left| \varepsilon_\theta(\sqrt{\bar{\alpha}_k} x^{(0)} + \sqrt{1-\bar{\alpha}_k}\varepsilon, k) - \varepsilon \right|^2$
until converged

Algorithm 2 Sampling

$x^{(N)} \sim \mathcal{N}(0, \mathrm{I}_d)$
for $k = N, \dots, 1$ **do**
  $z \sim \mathcal{N}(0, \mathrm{I}_d)$ if $k > 1$, else $z = 0$
  $x^{(k-1)} = \frac{1}{\sqrt{\alpha_k}} \left(x^{(k)} - \frac{\beta_k}{\sqrt{1-\bar{\alpha}_k}} \varepsilon_\theta(x^{(k)}, k) \right) + \sigma_k z$
end for
return $x^{(0)}$

1.2 Normalizing flows

The content of this section is based on the references [3, 1].

1.2.1 Basic

The idea of normalizing flows is to express data $x \in \mathbb{R}^d$ as a transformation of z of the same dimension that follows a simpler distribution $p_z(z)$.

Let us first recall the change of variables formula. Assume that $z \in \mathbb{R}^d$ is a random variable with probability density $p_z(z)$ and $f : \mathbb{R}^d \to \mathbb{R}^d$ is a transformation that is both invertible and differentiable. Denote by p_x the probability density of $x = f(z)$, i.e. the image of z under f. Then, we have the change of variables formula

$$p_x(x) = p_z(z) |\det J_f(z)|^{-1}, \tag{1.27}$$

where $z = f^{-1}(x)$ and the Jacobian $J_f(z)$ is the $d \times d$ matrix defined as

$$J_f(z) = \begin{pmatrix} \frac{\partial f_1}{\partial z_1} & \cdots & \frac{\partial f_1}{\partial z_d} \\ \vdots & & \vdots \\ \frac{\partial f_d}{\partial z_1} & \cdots & \frac{\partial f_d}{\partial z_d} \end{pmatrix}. \tag{1.28}$$

Equivalently, (1.27) can be expressed in terms of Jacobian of f^{-1} as

$$p_x(x) = p_z(f^{-1}(x)) |\det J_{f^{-1}}(x)| . \tag{1.29}$$

Assume that f_1, f_2 are two transformations that are both invertible and differentiable. Then, their composition $f_2 \circ f_1$ is also invertible and differentiable. Moreover, we have

$$\begin{aligned} (f_2 \circ f_1)^{-1} &= f_1^{-1} \circ f_2^{-1} \\ \det J_{f_2 \circ f_1}(z) &= \det J_{f_2}(f_1(z)) \cdot \det J_{f_1}(z) . \end{aligned} \tag{1.30}$$

In practice, p_z is often chosen as a normal distribution and the transformation f is constructed as the composition of elementary transformations $f_1, f_2, \cdots, f_K$. The term "normalizing" refers to the fact that the data distribution of x becomes normal distribution under the transformation f^{-1}, while "flow" refers to the passing of data through the transformations $f_1, f_2, \cdots, f_K$.

1.2.2 Generative modeling with normalizing flows

Given the target density $p_x^*(x)$, we want to learn a transformation $f(\cdot;\theta)$ with parameters θ, such that the density of $x = f(z;\theta)$, denoted by $p_x(x;\theta)$, is close to $p_x^*(x)$. We use the (forward) KL divergence in Definition 1 to quantify the closeness between $p_x(x;\theta)$ and $p_x^*(x)$:

$$\begin{aligned}
&D_{KL}\Big(p_x^*(x) \mid p_x(x;\theta)\Big) \\
=&\mathbb{E}_{x\sim p_x^*(x)}\Big(\log\frac{p_x^*(x)}{p_x(x;\theta)}\Big) \\
=&-\mathbb{E}_{x\sim p_x^*(x)}(\log p_x(x;\theta)) + C \\
=&-\mathbb{E}_{x\sim p_x^*(x)}\Big[\log p_z(f^{-1}(x;\theta)) + \log|\det J_{f^{-1}}(x;\theta)|\Big] + C\,.
\end{aligned}$$

where C is a constant independent of θ and the last equality follows from (1.29). In practice, the transformation f is modeled by a neural network that is trained with the loss function

$$\text{Loss}(\theta) = -\frac{1}{N}\sum_{n=1}^{N}\Big[\log p_z(f^{-1}(x_n;\theta)) + \log|\det J_{f^{-1}}(x_n;\theta)|\Big]\,. \tag{1.31}$$

In order for (1.31) to be computable, the neural network architecture has to be designed such that

1. the input and output dimensions are the same;
2. f is invertible;
3. computing f^{-1} and its Jacobian determinant is efficient.

1.2.3 Normalizing flow models

As mentioned in the previous section, we can design the model f as the composition of building blocks $f_1, f_2, \cdots, f_K$. Assume that each f_k satisfies the requirement at the end of the previous section, then their composition $f = f_K \circ \cdots \circ f_1$ is invertible and the logarithmic of its Jacobian determinant can be computed as

$$\log|\det J_{f^{-1}}(x)| = \log\Big|\prod_{k=1}^{K}\det J_{f_k^{-1}}(z_k)\Big| = \sum_{k=1}^{K}\log|\det J_{f_k^{-1}}(z_k)|\,, \tag{1.32}$$

where $z_K = x$ and

$$z_{k-1} = f_k^{-1}(z_k), \quad k = K, \ldots, 1\,. \tag{1.33}$$

In the following, we briefly mention two types of building blocks, which are called Nonlinear Independent Components Estimation (NICE) model and Real Non-Volume Preserving (RealNVP) model.

1. NICE. The transformation $x = (x_1, x_2) = f(z)$ is defined as

$$\begin{aligned} x_1 =& z_1 \,, \\ x_2 =& z_2 + \mu_\theta(z_1) \,, \end{aligned} \tag{1.34}$$

where $z_1 \in \mathbb{R}^{d'}$ and $z_2 \in \mathbb{R}^{d-d'}$ are two disjoint subsets that form a partitions of z, d' is an integer between 1 and $d-1$, and $\mu_\theta : \mathbb{R}^{d'} \to \mathbb{R}^{d-d'}$ is a neural network. Its inverse is

$$\begin{aligned} z_1 =& x_1 \,, \\ z_2 =& x_2 - \mu_\theta(x_1) \,. \end{aligned} \tag{1.35}$$

It is straightforward to verify that the Jacobian determinant of f is one. Therefore, f defines a volume preserving transformation.

2. RealNVP. The transformation $x = (x_1, x_2) = f(z)$ is defined as

$$\begin{aligned} x_1 =& z_1 \,, \\ x_2 =& \exp(\sigma_\theta(z_1)) \odot z_2 + \mu_\theta(z_1) \,, \end{aligned} \tag{1.36}$$

where $\odot$ denotes elementwise product, and $\sigma_\theta, \mu_\theta : \mathbb{R}^{d'} \to \mathbb{R}^{d-d'}$ are two neural networks. The inverse of f is

$$\begin{aligned} z_1 =& x_1 \,, \\ z_2 =& \exp(-\sigma_\theta(x_1)) \odot (x_2 - \mu_\theta(x_1)) \,, \end{aligned} \tag{1.37}$$

The Jacobian determinant is $\det J_{f^{-1}}(x) = \exp(-\sum_{j=1}^{d-d'} (\sigma_\theta(x_1))_j)$, where $(\sigma_\theta(x_1))_j$ denotes the jth component of $\sigma_\theta(x_1)$.

Bibliography

[1] L. Dinh, J. Sohl-Dickstein, and S. Bengio. Density estimation using real NVP. In *International Conference on Learning Representations*, 2017.

[2] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020.

[3] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *J. Mach. Learn. Res.*, 22(1), 2021.