# 标准化流模型 (Normalizing Flow)

## 1. Motivation of Normalizing Flow

标准化流模型 (Normalizing Flow, NF) [1] 是一种用于建模复杂分布的方法。在变分推断框架内，一般使用变分分布$q(\mathbf{z};\boldsymbol{\phi})$来近似真实后验分布$p(\mathbf{z}|\mathbf{x})$，优化目标是如下证据下界(Evidence Lower Bound, ELBO)：

$$\begin{aligned}\text{ELBO}(\boldsymbol{\phi},\boldsymbol{\theta}) &= E_{q(\mathbf{z};\boldsymbol{\phi})}[\log p(\mathbf{x}|\mathbf{z};\boldsymbol{\theta})] - \text{KL}\big(q(\mathbf{z};\boldsymbol{\phi})||p(\mathbf{z})\big) \\ &= \log p(\boldsymbol{x}) - \text{KL}(q(\mathbf{z};\boldsymbol{\phi})||p(\mathbf{z}|\mathbf{x}))\end{aligned} \tag{1}$$

其中，$\boldsymbol{\phi}$是神经网络参数，其输入是数据$\mathbf{x}$，输出是变分分布$q(\mathbf{z};\boldsymbol{\phi})$的参数；$\boldsymbol{\theta}$是神经网络参数，其输入是隐变量$\mathbf{z}$的采样结果，输出是数据$\mathbf{x}$的预测分布$p(\mathbf{x}|\mathbf{z};\boldsymbol{\theta})$；$p(\mathbf{z})$是隐变量$\mathbf{z}$的先验分布；$\log p(\boldsymbol{x})$是对数边缘似然。

通常，为了方便，变分分布$q(\mathbf{z};\boldsymbol{\phi})$会假定为均值为$\boldsymbol{\mu}_q$，方差$\boldsymbol{\Sigma}_q$为的高斯分布 (这个参数是通过神经网络)，先验分布$p(\mathbf{x}|\mathbf{z};\boldsymbol{\theta})$会假定为标准高斯分布：

$$q(\mathbf{z};\boldsymbol{\phi}) = \mathcal{N}\big(\mathbf{z};\boldsymbol{\mu}_q,\boldsymbol{\Sigma}_q\big),\ ,p(\mathbf{z}) = \mathcal{N}(\mathbf{z};\mathbf{0},\mathbf{I}) \tag{2}$$

事实上，使用变分推断作为极大似然估计手段的最初目的是，利用变分分布$q(\mathbf{z};\boldsymbol{\phi})$强大的表示能力，建模“任意复杂”的后验分布$p(\mathbf{x}|\mathbf{z};\boldsymbol{\theta})$。然而，公式 (2) 直接将$q(\mathbf{z};\boldsymbol{\phi})$简单地假定为高斯分布$\mathcal{N}\big(\mathbf{z};\boldsymbol{\mu}_q,\boldsymbol{\Sigma}_q\big)$，可能会限制其表示能力。

为此，NF 尝试将上述高斯分布$q(\mathbf{z};\boldsymbol{\phi}) = \mathcal{N}\big(\mathbf{z};\boldsymbol{\mu}_q,\boldsymbol{\Sigma}_q\big)$，转化为 (概率密度函数形式) 更加复杂的分布$q(\mathbf{z};\boldsymbol{\phi})$，从而提升$q(\mathbf{z};\boldsymbol{\phi})$对于后验分布的建模能力。

## 2. Normalizing Flow

### 2.1 对于随机变量$\mathbf{z}_0$的一次非线性变换

首先以一个简单的例子说明标准化流模型的工作过程。

给定一个随机变量$\mathbf{z}_0$的高斯分布 (可以是非标准高斯分布)：

$$q_0(\mathbf{z}_0) = \mathcal{N}(\mathbf{z};\boldsymbol{\mu},\boldsymbol{\Sigma}) \tag{3}$$

可以使用如下非线性函数$f$对随机变量$\mathbf{z}_0$做变换 (变换后变量的维度不变)：

$$\mathbf{z}_1 = f(\mathbf{z}_0) \tag{4}$$

这时，相较于$\mathbf{z}_0$，随机变量$\mathbf{z}_1$的“定义域”与“概率密度函数$q_1(\mathbf{z}_1)$”均发生了改变，这一点在材料 [2,3] 中做了详细的说明。需要说明的是，我们要求变换函数$f$是可逆的。

我们关心$\mathbf{z}_1$概率密度函数$q_1(\mathbf{z}_1)$的形式，其可表示如下：

$$q_1(\mathbf{z}_1) = q_0(\mathbf{z}_0)\left|\det\frac{\partial f^{-1}(\mathbf{z}_1)}{\partial \mathbf{z}_1}\right| = q_0(\mathbf{z}_0)\left|\det\frac{\partial f(\mathbf{z}_0)}{\partial \mathbf{z}_0}\right|^{-1} \tag{5}$$

其中，$f^{-1}(\mathbf{z}_1)$是$f(\mathbf{z}_0)$的反函数，$\det(\cdot)$表示矩阵的行列式。

相较公式 (3)，公式 (5) 的形式是更加复杂的，且$q_0(\mathbf{z}_0)$的缩放因子$\left|\det\frac{\partial f(\mathbf{z}_0)}{\partial \mathbf{z}_0}\right|^{-1}$会使不同位置采样点概率密度值发生变化。例如，对于在高斯分布$q_0(\mathbf{z}_0) = \mathcal{N}(\mathbf{z};\boldsymbol{\mu},\boldsymbol{\Sigma})$均值$\boldsymbol{\mu}$周围的点$\mathbf{z}_0$来说，$q_0(\mathbf{z}_0)$是比较大的，而若缩放因子$\left|\det\frac{\partial f(\mathbf{z}_0)}{\partial \mathbf{z}_0}\right|^{-1} \to 0$，那么，整体概率密度值$q_0(\mathbf{z}_0)\left|\det\frac{\partial f(\mathbf{z}_0)}{\partial \mathbf{z}_0}\right|^{-1}$将趋于 0。

尽管已知$\mathbf{z}_1$的概率密度函数$q_1(\mathbf{z}_1)$的形式，但相比于高斯分布$q_0(\mathbf{z}_0) = \mathcal{N}(\mathbf{z};\boldsymbol{\mu},\boldsymbol{\Sigma})$ (可以通过“重参数化技巧”完成采样操作)，对于$q_1(\mathbf{z}_1)$的采样操作仍是困难的。

根据 LOTUS 定理 (Law of the Unconscious Statistician)，对于$q_1(\mathbf{z}_1)$的采样操作，可以使用关于$q_0(\mathbf{z}_0)$的采样操作来做模拟，具体如下：

$$\mathbf{z}_1^{(k)} \sim q_1(\mathbf{z}_1) \Leftrightarrow \mathbf{z}_1^{(k)} = f\left(\mathbf{z}_0^{(k)}\right),\ \mathbf{z}_0^{(k)} \sim q_0(\mathbf{z}_0) \tag{6}$$

这表示，对于$q_1(\mathbf{z}_1)$的采样操作可以分为两个过程：首先，从原高斯分布$q_0(\mathbf{z}_0)$中采样的得到样本$\mathbf{z}_0^{(k)}$。然后，通过非线性变换$\mathbf{z}_1^{(k)} = f\left(\mathbf{z}_0^{(k)}\right)$得到$\mathbf{z}_1^{(k)}$，作为$q_1(\mathbf{z}_1)$的采样样本。

现在，假定随机变量$\mathbf{z}_0 = [z_{0,1}, z_{0,2}]^\top \in \mathbb{R}^{2\times 1}$服从如下高斯分布：

$$q_0(\mathbf{z}_0) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}\left(\mathbf{z}; \begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right) \tag{7}$$

其概率密度函数图像如图 1(a) 所示。

公式 (4) 所定义的非线性变换函数如下：

$$\begin{gathered}\mathbf{z}_1 = \begin{bmatrix}z_{1,1}\\z_{1,2}\end{bmatrix} = f(\mathbf{z}_0) = \mathbf{z}_0 + \mathbf{u}\mathrm{Tanh}(\boldsymbol{w}^\top \mathbf{z}_0 + b)\\ \text{其中，}\ \mathbf{u} = \begin{bmatrix}u_1\\u_2\end{bmatrix} = \begin{bmatrix}-3\\1\end{bmatrix} \in \mathbb{R}^{2\times 1},\ \boldsymbol{w} = \begin{bmatrix}w_1\\w_2\end{bmatrix} = \begin{bmatrix}-1\\5\end{bmatrix} \in \mathbb{R}^{2\times 1},\ , b = 1 \in \mathbb{R}\end{gathered} \tag{8}$$

后面我们会看到，以这种形式函数对随机变量做非线性变换的标准化流也被称为“平面流” (Planar flows)。

此时，变换函数关于$\mathbf{z}_0$的导数为：

$$\begin{gathered}\frac{\partial f(\mathbf{z}_0)}{\partial \mathbf{z}_0} = \mathbf{I} + \mathbf{u}\psi(\mathbf{z}_0)^\top\\ \text{其中，}\ \psi(\mathbf{z}_0) = \mathrm{Tanh}'(\boldsymbol{w}^\top \mathbf{z}_0 + b)\boldsymbol{w} = \begin{bmatrix}1 - \mathrm{Tanh}^2(w_1 z_{1,1} + b) & 0\\ 0 & 1 - \mathrm{Tanh}^2(w_2 z_{1,2} + b)\end{bmatrix}\boldsymbol{w}\end{gathered} \tag{9}$$

其行列式为：

$$\det \frac{\partial f(\mathbf{z}_0)}{\partial \mathbf{z}_0} = \det(\mathbf{I} + \mathbf{u}\psi(\mathbf{z}_0)^\top) = 1 + \mathbf{u}^\top \psi(\mathbf{z}_0) \tag{10}$$

利用公式 (5)，$\mathbf{z}_1$的概率密度函数$q_1(\mathbf{z}_1)$表示为：

$$q_1(\mathbf{z}_1) = q_0(\mathbf{z}_0)\left|\det \frac{\partial f(\mathbf{z}_0)}{\partial \mathbf{z}_0}\right|^{-1} = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \cdot |1 + \mathbf{u}^\top \psi(\mathbf{z}_0)|^{-1} \tag{11}$$

其概率密度函数图像如图 1(b) 所示。

注意到，图中有部分区域是“空白的”，表示$\mathbf{z}_1$在对应区间内无有效取值。这是由于非线性变换$\mathbf{z}_1 = f(\mathbf{z}_0)$会使得函数的定义域发生改变。同时，相较于$\mathbf{z}_0$，随机变量$\mathbf{z}_1$的概率密度分布是更加复杂的。

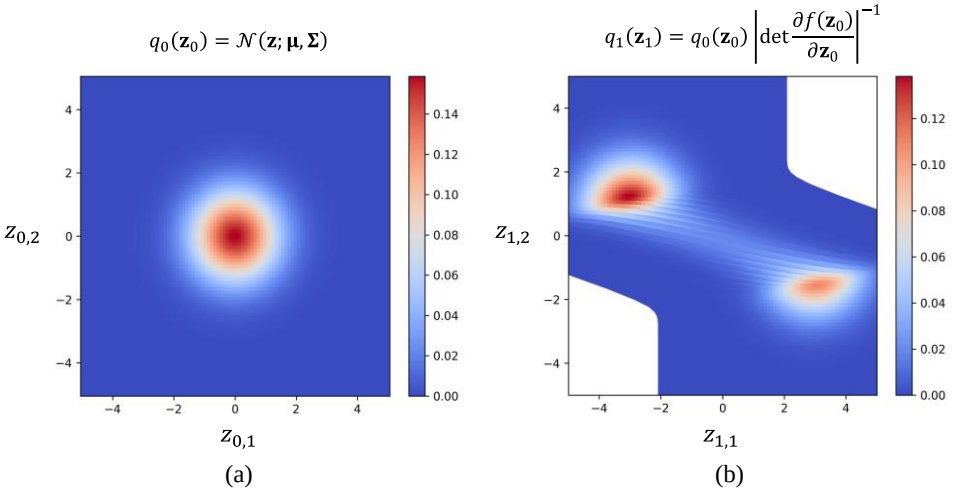
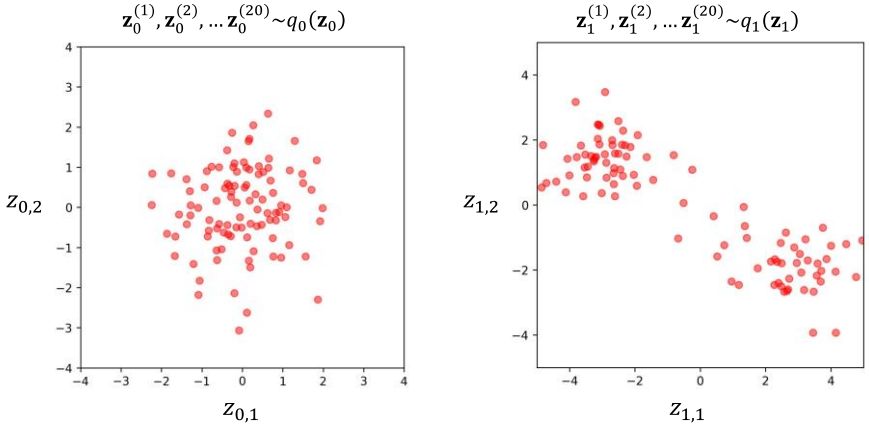



图 1：(a) 高斯分布$q_0(\mathbf{z}_0) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$的概率密度函数图像。(b) 复杂分布$q_1(\mathbf{z}_1) = q_0(\mathbf{z}_0)\left|\det \frac{\partial f(\mathbf{z}_0)}{\partial \mathbf{z}_0}\right|^{-1}$的概率密度函数图像。

现在，利用公式 (6)，我们对随机变量$\mathbf{z}_1$做采样操作。

首先对分布$q_0(\mathbf{z}_0)$做 20 次采样操作 (可通过重参数化技巧做模拟)，得到随机变量$\mathbf{z}_0$的样本$\mathbf{z}_0^{(1)}, \mathbf{z}_0^{(2)}, \ldots, \mathbf{z}_0^{(20)}$，如图 2(a)所示。然后，通过公式 (4)，对上述样本做非线性变换，得到随机变量$\mathbf{z}_1$的样本

$\mathbf{z}_1^{(1)} = f\left(\mathbf{z}_0^{(1)}\right), \mathbf{z}_1^{(2)} = f\left(\mathbf{z}_0^{(2)}\right), \ldots, \mathbf{z}_1^{(20)} = f\left(\mathbf{z}_0^{(20)}\right)$，如图 2(b)所示。




(a) (b)

图 2：(a) 高斯分布$q_0(\mathbf{z}_0) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$的采样结果。(b) 复杂分布$q_1(\mathbf{z}_1)$的采样结果。

## 参考文献


[1] Rezende D, Mohamed S. Variational inference with normalizing flows[C]//International conference on machine learning. PMLR, 2015: 1530-1538.

[2] https://deepgenerativemodels.github.io/assets/slides/cs236_lecture7.pdf

[3] https://deepgenerativemodels.github.io/assets/slides/cs236_lecture8.pdf