0. Research topic

Normalizing Flows (vs **VAE** vs **GAN**)

1. Introduction

Statistical ML is all about probability.

训练一个概率模型，我们往往直接优化训练数据的最大似然概率$p(x;\theta)$. Normalizing Flows 能把简单的地摊货概率密度(比如高斯分布)形式转换成某种高大上的分布形式

Jacobian Matrix and Determinant

对于一个函数组映射: $f:\mathbb{R}^n \to \mathbb{R}^n$, 有$y_{1:n}=f_{1:n}(x_{1:n})$, **Jacobian Matrix**为该映射函数组的一阶偏导，其中$J_{ij}=\frac{\partial f_i}{\partial x_j}$,

$$\mathbf{J}=\begin{bmatrix}\frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n}\\ \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_n}\end{bmatrix}$$

其行列式写作 $det(\mathbf{J})$为一个具体的数（计算方法略）, 注意只有方阵**sqare matrix** 才有行列式。

Change of Variable Theorem

先考虑一维变量的情况，有随机变量 $z\in\mathbb{R}$，满足概率分布 $z\sim p(z)$。 变量 $y=f(z)$，当 f 可逆时，$z=f^{-1}(y)$。由于概率密度的积分总是1，有

$$\int p(y)dy=\int p(y)\frac{dy}{dz}dz=\int p(z)dz=1$$

$$p(y)=p(z)\left|\frac{dz}{dy}\right|=p(f^{-1}(y))\left|\frac{df^{-1}}{dy}\right|=p(f^{-1}(y))\left|(df^{-1})'(y)\right|$$

考虑多维的情况，当随机变量 $z\in\mathbb{R}^n$，满足概率分布 $z\sim p(z)$，同时$f:\mathbb{R}^n\to\mathbb{R}^n$ 是一个**可逆**的映射，变量 $y=f(z)$，则$z=f^{-1}(y)$。

$$p(x)=p(f^{-1}(y))\left|\det\frac{\partial f^{-1}}{\partial y}\right|$$

此处，$\det\frac{\partial f^{-1}}{\partial y}$ 表示函数 f 对 x 的雅可比行列式。

What is Normalizing Flows?

In Machine Learning problems, Gaussian distribution is widely used as assumption due to its good mathematical properties. However, exact density tends to be more complicated than Gaussian distribution and hence intractable.

Normalizing Flows (**NF**) is then coming to achieve complex distribution density estimation. A normalizing Flow are able to transform a simple distribution to a complex one by employing a successive suqence of invertiable tranformation functions. According to the change of variable theorem and computation of Jacobian determinant, we can acquire the distribution of target variable.

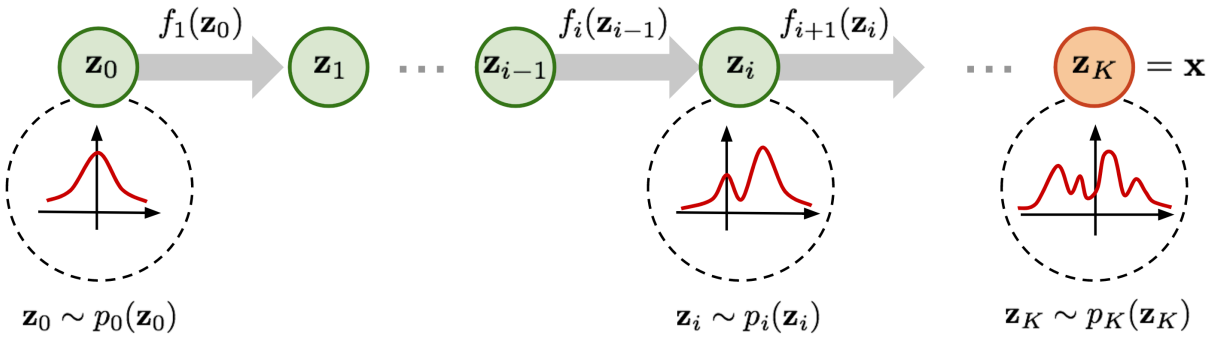

Figure 1. NF 模型的 forward propagation 示意图, 图来自[1]

（1） Figure 1中针对任意一次变量转换，

$$\mathbf{z}_i = f(\mathbf{z}_{i-1})$$
$$\mathbf{z}_{i-1} = f^{-1}(\mathbf{z}_i)$$
$$p_i(\mathbf{z}_i) = p_{i-1}(\mathbf{z}_{i-1}) \left| \det \frac{\partial f_i^{-1}}{\partial \mathbf{z}_i} \right| = p_{i-1}(\mathbf{z}_{i-1}) \left| \det \frac{\partial f_i^{-1}}{\partial \mathbf{z}_i} \right|$$

这里雅可比行列式中 ∂f_i^{-1} 没有显式的形式，计算起来并不方便，因此我们对其做一些变换，

$$\begin{aligned} \left| \det \frac{\partial f_i^{-1}}{\partial \mathbf{z}_i} \right| &= \left| \det \frac{\partial \mathbf{z}_{i-1}}{\partial \mathbf{z}_i} \right| = \left| \det \left(\frac{\partial \mathbf{z}_i}{\partial \mathbf{z}_{i-1}} \right)^{-1} \right| = \left| \det \left(\frac{\partial f_i}{\partial \mathbf{z}_{i-1}} \right)^{-1} \right| \\ &= \left| \det \frac{\partial f_i}{\partial \mathbf{z}_{i-1}} \right|^{-1} \qquad ; because \ \det \mathbf{M}^{-1} = (\det \mathbf{M})^{-1} \end{aligned}$$

因此，上述公式可写作，

$$p_i(\mathbf{z}_i) = p_{i-1}(f^{-1}(\mathbf{z}_i)) \left| \det \frac{\partial f_i^{-1}}{\partial \mathbf{z}_i} \right| = p_{i-1}(\mathbf{z}_{i-1}) \left| \det \frac{\partial f_i}{\partial \mathbf{z}_{i-1}} \right|^{-1}$$

两边取对数，

$$\log p_i(\mathbf{z}_i) = \log p_{i-1}(\mathbf{z}_{i-1}) - \log \left| \det \frac{\partial f_i}{\partial \mathbf{z}_{i-1}} \right|$$

（2） Figure 1中连续地运用一系列的可逆映射函数，最终的变量的密度函数可以由下式计算，

$$
\begin{aligned}
\mathbf{z}_K &= f_K \circ f_{K-1} \circ \cdots \circ f_1(\mathbf{z}_0) \\
\log p_K(\mathbf{z}_K) &= \log p_{K-1}(\mathbf{z}_{K-1}) - \log \left| \det \frac{\partial f_K}{\partial \mathbf{z}_{K-1}} \right| \\
&= \log p_{K-2}(\mathbf{z}_{K-2}) - \log \left| \det \frac{\partial f_{K-1}}{\partial \mathbf{z}_{K-2}} \right| - \log \left| \det \frac{\partial f_K}{\partial \mathbf{z}_{K-1}} \right| \\
&= \cdots \\
&= \log p_0(\mathbf{z}_0) - \sum_{i=1}^{K} \log \left| \det \frac{\partial f_i}{\partial \mathbf{z}_{i-1}} \right|
\end{aligned}
$$

由此可见，计算最终的概率密度函数的关键在于 映射函数 $f_{1:D}$ 的选择，而$f_{1:D}$需要满足以下条件:

1. 可逆
2. 方便计算其雅可比行列式

2. Related research papers

Simple Flows

Variational Inference with Normalizing Flows 首先介绍了 Normalizing Flows 的两种映射函数，分别为

Planar Flow

$$
f(\mathbf{z}) = \mathbf{z} + \mathbf{u}h(\mathbf{w}^T\mathbf{z} + b)
$$

其中，$\mathbf{u}, \mathbf{w} \in \mathbb{R}^d$ ，$b \in \mathbb{R}$ 是可学习的参数，h 为非线性函数。其雅可比行列式可写作，

$$
\begin{aligned}
\psi(\mathbf{z}) &= h'(\mathbf{w}^T\mathbf{z} + b)\mathbf{w} \\
\left| \det \frac{\partial f}{\partial \mathbf{z}} \right| &= \left| \det(\mathbf{I} + \mathbf{u}\psi(\mathbf{z})^T) \right| = \left| 1 + \mathbf{u}^T\psi(\mathbf{z}) \right|
\end{aligned}
$$

通过应用一系列的该函数，最终变量的概率密度函数可写作，

$$
\log p_K(\mathbf{z}_K) = \log p_0(\mathbf{z}_0) - \sum_{i=1}^{K} \log \left| 1 + \mathbf{u}_i^T \psi_i(\mathbf{z}_{i-1}) \right|
$$

此映射函数可以看作，用直线（或超平面）切割 z 空间，其中每条线收缩或扩展其周围的空间，见 **Figure 2**.

Radial Flow

$$
f(\mathbf{z}) = \mathbf{z} + \beta h(\alpha, r)(\mathbf{z} - \mathbf{z}_0)
$$

相似地，此映射函数可以看作，在 z 空间中引入球体，它们可以收缩或扩展球体内部的空间，见**Figure 2**.

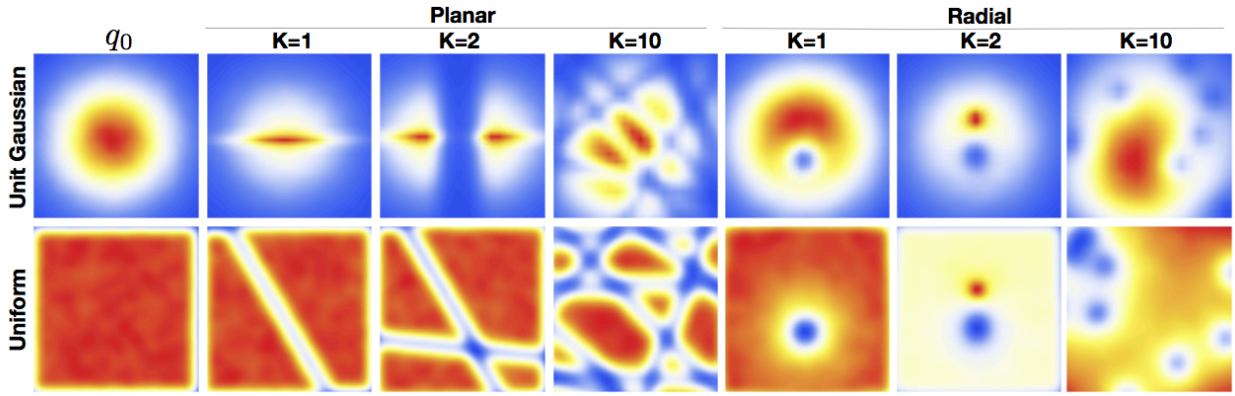

Figure 2. 平面和径向流对高斯分布和均匀分布的影响。图来自原文

Autoregressive Flows

TODO

##

8. References

[1] Flow-based Deep Generative Models

[2] Normalizing Flows