# VAE简单推导

详细推导建议看原文

# 1. 准备知识

### 1.1 Monte Carlo

蒙特卡洛算法是随机算法中的一类算法的总称，另一类算法是拉斯维加斯算法，两者都是以著名的赌城命名的。通俗地理解，蒙特卡洛算法在采样数量有限制的情况下，尽可能地接近最优解的一种抽样方法。随着采样次数的增多，可以保证结果是越来越接近最优解的，但是除非对全局样本都进行采样，否则无法判断当前有没有找到最优解。 从数学上理解，最早的MC方法，是为了解决积分不好求的问题，转而用随机化的方法来计算积分。

$$\int_a^b h(x)dx$$

假设无法通过数学推导来求积分，而且对取值区间内的所有x进行枚举也是不现实的，可以将h(x)分解为某个函数f(x)和一个定义在(a,b)上的概率密度函数p(x)的乘积。这样整个积分就可以写成

$$\int_a^b f(x)p(x)dx = E_{p(x)}[f(x)]$$

原来的积分问题等同于函数f(x)在概率密度函数p(x)这个分布上的均值。如果我们在p(x)对应的分布式上抽样大量的数据点，那么就可以通过这些样本来逼近f(x)在这个分布上的均值：

$$\int_a^b h(x)dx = E_{p(x)}[f(x)] \approx \frac{1}{n}\sum_{i=1}^{n} f(x_i)$$

MC通过这种方法，近似最终的积分，并且随着采样数量的增加，越来越接近真实的积分值。除非对(a,b)内的所有样本采样，否则我们无法知道真实的积分值。
Ref: [https://www.zhihu.com/question/20254139](https://www.zhihu.com/question/20254139)

### 1.2 Intractability

在贝叶斯分析中，经常会说后验分布是intractable的，因此必须要近似推断。那么intractability到底是什么来历呢？ 因此贝叶斯分析中，经常要用到积分(integral)，在实际问题中，经常要对多维变量进行积分，这种积分理论上都是intractable的。 而对于非贝叶斯分析，很多都是基于最大似然的，这些方法主要是用到了导数。 而导数一般来说比积分更容易求：differentiation is more tractable than integration。 Ref: https://stats.stackexchange.com/questions/4417/intractable-posterior-distributions

### 1.3 Reparameterize

在深度学习中，我们经常想要把样本\(x\sim p_{\theta}(x)\)的梯度做反向传播，例如VAE中，我们需要将样本的梯度通过z反向传播给参数\(\phi\)。 举个具体的例子，我们想要利用梯度下降的方法最小化期望损失：

$$L(\theta, \phi) = E_{x\sim p_\phi(x)}[f_\theta(x)]$$

需要分别计算对参数$\theta, \phi$的导数。

$$\nabla_\theta E_{x\sim p_\phi(x)}[f_\theta(x)] = E_{x\sim p_\phi(x)}[\nabla_\theta f_\theta(x)]$$

Ref: https://gabrielhuang.gitbooks.io/machine-learning/reparametrization-trick.html

### 1.4 高斯分布与多元高斯分布

高斯分布是常见的概率分布，经常用来拟合自然界的一些连续随机变量产生的样本。通常可以记为：

$$x \sim N(\mu, \sigma^2)$$

它的概率密度函数PDF为：

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

# 2. VAE经典文章解读

### 2.1. 问题背景

对于观测到的服从独立同分布的样本集合$X = \{x^{(i)}\}_{i=1}^N$，假设每一个样本$x_i$都是由某个随机过程生成的，这个随机过程由连续的随机变量z控制。同时随机变量z也服从先验分布$p_\theta(z)$,参数为$\theta$。并且假设先验分布$p_\theta(z)$和似然函数$p_\theta(x|z)$都是指数族分布，并且概率密度函数PDF处处可导。以手写数字的图片识别问题为例，某一幅图像的隐含变量z需要包含如下信息：图像所代表的数字（0~9），笔画的粗细，写字的角度等等，在确定上述隐含变量之后，经过图像生成的算法，最终得到图像。

z一般是多维向量，并且可以从多维空间中按照概率密度函数$p(z)$抽取，这个概率密度函数的参数是变量$\theta$。此时的目标就是优化参数$\theta$得到概率密度函数$p(z)$，并从中抽取z来生成样本x。在VAE中，假设z的各个维度没有直接简单的关系，并且假设z是可以通过简单的多元高斯分布来抽取。这个假设背后的原理是**在d维空间中的任意分布，都可以表示成服从正态d个变量的组合**。基于这个原理，z的真实分布可以通过若干个相互独立的高斯分布来拟合。

VAE的主要目标还是学出隐含变量z，从而能够在无监督的情况下生成样本X。也就是最大化生成样本x的边缘概率

$$P(x) = \int P(X|z;\theta)P(z)dz$$

一种直观的方法是在z所在的空间中进行抽样，从而逼近生成X的概率

$$p(x) \approx \frac{1}{n}\sum_i p(x|z)$$

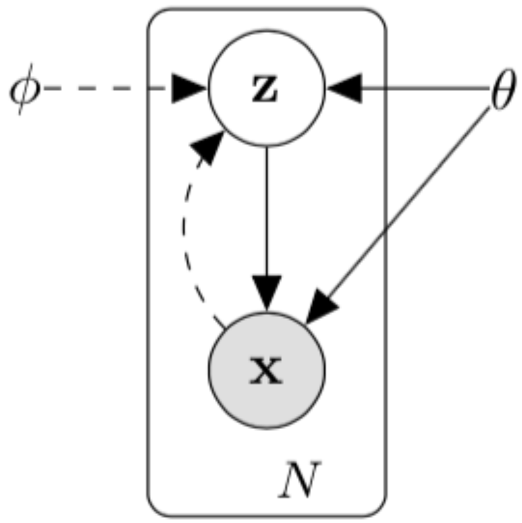
但是在高维空间中，如果对z进行随机抽样，那么得到的$p(x|z)$大多数都是0，这种抽出来的z并不能帮助我们提高学习的准确率，而且计算复杂度非常高。所以VAE的想法是通过样本X来学习出z的分布，然后在这个分布的基础上对z进行抽样，而不是从z的整个潜在空间中抽取。

传统方法的问题：

1. 边缘似然$p_\theta(x)$需要对z进行积分，$p_\theta(x) = \int p_\theta(z)p_\theta(x|z)dz$，因为涉及到了高维变量的积分，所以是intractable，我们不能对边缘似然进行求导。

2. 后验概率$p_\theta(z|x) = p_\theta(z|x)p_\theta(z)/p_\theta(x)$同样也是intractable，因此不能用EM的方法来拟合。
3. 所有需要积分的平均场变分贝叶斯方法都是intractable的。
4. 在大规模数据集上，MCMC的方法做抽样太耗时，算法的时间复杂度非常高。
   AEVB要解决的问题：
5. 拟合参数$\theta$的最大似然ML或者最大后验概率MAP
6. 拟合参数$z$的后验概率$p(z|x, \theta)$
7. 拟合x的边缘分布$p(x|\theta)$

为了解决上述问题，因为后验概率$p_\theta(z|x)$的形式未知，引入recognigion model $q_\phi(z|x)$来近似真实的后验概率。这里引入的参数$\phi$会和$\theta$一起学习，而不是想变分推断里面一样用期望的近似值来计算。
用概率图模型表示：




和Auto-Encoder的关系
潜在的变量z可以看作是一种code，因为每个样本x都可以看作是在z上的一个分布，因此对于每个样本x，将其表示成z上的分布，即前面引入的recognition model $q_\phi(z|x)$, 就是一种encode的过程。而给定隐含变量z，生成样本x，即$p_\theta(x|z)$则是一个decode的过程。

### 2.2 AEVB模型推导

我们引入recognition model $q_\phi(z|x)$的目的是逼近真实的后验概率$p_\theta(z|x)$，衡量两个分布的相似程度，我们用KL散度来描述：

$$KL(q_\phi(z|x^{(i)})||p_\theta(z|x^{(i)}))amp;= E_{q_\phi(z|x^{(i)})} log \frac{q_\phi(z|x^{(i)})}{p_\theta(z|x^{(i)})}$$
$$amp;= E_{q_\phi(z|x^{(i)})} log \frac{q_\phi(z|x^{(i)})p_\theta(x^{(i)})}{p_\theta(z|x^{(i)})p_\theta(x^{(i)})}$$
$$amp;= E_{q_\phi(z|x^{(i)})} log \frac{q_\phi(z|x^{(i)})}{p_\theta(z, x^{(i)})} + E_{q_\phi(z|x^{(i)})} logp_\theta(x^{(i)})$$
$$amp;= E_{q_\phi(z|x^{(i)})} log \frac{q_\phi(z|x^{(i)})}{p_\theta(z, x^{(i)})} + logp_\theta(x^{(i)})$$

将等式两边调换一下，每个样本的边缘似然函数，可以写为：

$$
\begin{aligned}
logp_\theta(x^{(i)}) amp;= D_{KL}(q_\phi(z|x^{(i)}) \| p_\theta(z|x^{(i)})) - E_{q_\phi(z|x^{(i)})} log \frac{q_\phi(z|x^{(i)})}{p_\theta(z, x^{(i)})} \\
amp;= D_{KL}(q_\phi(z|x^{(i)}) \| p_\theta(z|x^{(i)})) + L(\theta, \phi; x^{(i)})
\end{aligned}
$$

每个样本的边缘似然函数，可以写为：

$$logp_\theta(x^{(i)}) = KL(q_\phi(z|x^{(i)}) \| p_\theta(z|x^{(i)})) + L(\theta, \phi; x^{(i)})$$

第一项描述的是近似的后验分布和真实的后验分布之间的KL散度，因为KL散度大于等于0，所以L被称为variational lower bound，即以下不等式恒成立：

$$logp_\theta(x^{(i)}) \geq L(\theta, \phi; x^{(i)})$$

这个lower bound又可以进一步分解为：

$$L(\theta, \phi; x^{(i)}) = E_{q_\phi(z|x)}[logp_\theta(x^{(i)}, z) - logq_\phi(z|x^{(i)})]$$

或者

$$
\begin{aligned}
L(\theta, \phi; x^{(i)}) amp;= E_{q_\phi(z|x)}[logp_\theta(x^{(i)}, z) - logq_\phi(z|x^{(i)})] \\
amp;= E_{q_\phi(z|x)}[logp_\theta(x^{(i)}|z) + logp_\theta(z) - logq_\phi(z|x^{(i)})] \\
amp;= -KL(q_\phi(z|x^{(i)}) \| p_\theta(z)) + E_{q_\phi(z|x^{(i)})}[logp_\theta(x^{(i)}|z)]
\end{aligned}
$$

我们通过优化这个lower bound来实现优化边缘似然$p_\theta(x^{(i)})$,这个lower bound中同时包含了两个参数$\theta, \phi$,也就是前面说的将两个参数一起学习（优化）。因为对于后验概率$q_\phi(z|x)$,我们不知道具体的形式，因此不能求导，不能生成蒙特卡洛梯度。

将上面的式子重新组织一下，得到如下的形式：

$$logp_\theta(x^{(i)}) - KL(q_\phi(z|x^{(i)}) \| p_\theta(z|x^{(i)})) = -KL(q_\phi(z|x^{(i)}) \| p_\theta(z)) + E_{q_\phi(z|x^{(i)})}[logp_\theta(x^{(i)}|z)]$$

这个式子就是整个VAE模型的核心了，分别观察等式的左右两边，等式左边是我们想要最大化的目标，样本X的边缘概率$p(x)$加上一个损失项，损失项使得由后验概率Q生成的z能够最大可能生成样本X。在最优情况下，KL散度等于0，Q和P的概率是相同的。真实的z的后验概率$p_\theta(z|x)$描述了z有多大的概率生成X，因为我们不知道具体的形式，所以不能积分，但是好处在于我们可以设定q为可积分的形式，这样就可以用q的计算来代替p。最理想的情况下，q和p相等，则等式的左边就变成了直接优化$p(x)$。等式的右边则可以看作是auto-encoder的形式，$p_\theta(x^{(i)}|z)$是decode的过程，而$q_\phi(z|x^{(i)})$是encode的过程。最终，我们的核心目标：优化样本X的边缘概率$p(x)$变成了一个优化auto-encoder框架。

通过上述式子的转化，优化目标转化为了最大化等式右边的两项，可以通过随机梯度下降来进行优化。z的先验分布$p_\theta(z)$和z的后验拟合分布$q_\phi(z|x^{(i)})$的形式也是给定的,因此第一项是可以直接求的。第二项是一个似然函数的形式，可以通过抽样和最大似然来进行拟合，但是因为设计到了抽样的过程，所以无法将导数进行反向传播。
这里用到了一个很好玩的reparameterize trick来让表达式对$\phi$可导，引入一个可导的函数$g_\phi(\epsilon, x)$，$\epsilon$是辅助噪声参数，它有独立于数据的边缘概率$p(\epsilon)$，$g_\phi()$是受参数$\phi$控制的向量函数。reparameterize之后的隐含变量z可以表示为

$$\tilde{z} = g_\phi(\epsilon, x), \epsilon \sim p(\epsilon)$$

利用蒙特卡洛方法，我们生成关于$q_\phi(z|x)$的期望函数$f(z)$:

$$E_{q_\phi(z|x^{(i)})}[f(z)] = E_{p(\epsilon)}[f(g_\phi(\epsilon, x^{(i)}))] \approx \frac{1}{L}\sum_{l=1}^{L} f(g_\phi(\epsilon^{(l)}, x^{(i)}))\ where\ \epsilon^{(l)} \sim p(\epsilon)$$

此时的积分变量从z变成了输入的噪声参数$\epsilon$,z也从随机变量变成了确定变量($\epsilon$决定)，对模型变量的求导也就可以实现。

利用reparameterize的方法，前面提到的两种lower bound的形式转化为

$$\begin{aligned}
L(\theta, \phi; x^{(i)}) amp;&= E_{q_\phi(z|x)}[logp_\theta(x^{(i)}, z) - logq_\phi(z|x^{(i)})] \\
amp;&= \frac{1}{L}\sum_{l=1}^{L}[logp_\theta(x^{(i)}, z^{(i,l)}) - logq_\phi(z^{(i,l)}|x^{(i)})] \\
L(\theta, \phi; x^{(i)}) amp;&= -KL(q_\phi(z|x^{(i)}) \| p_\theta(z)) + E_{q_\phi(z|x^{(i)})}[logp_\theta(x^{(i)}|z)] \\
amp;&= -KL(q_\phi(z|x^{(i)}) \| p_\theta(z)) + \frac{1}{L}\sum_{l=1}^{L} logp_\theta(x^{(i)}, z^{(i,l)}) \\
where\ z^{(i,l)} amp;&= g_\phi(\epsilon^{(i,l)}, x^{(i)})\ and\ \epsilon^{(l)} \sim p(\epsilon)
\end{aligned}$$

这里出现的两种形式的lower bound，区别在于第一项一个是联合分布，一个是KL散度，而KL散度经常可以求积分，这样的话拟合出来的结果方差会更小，文章中的附录B举例了高斯分布的KL散度的积分的结果，可以直接表示为均值和方差的函数。

原始数据集中的样本数量为N，我们可以用minibatch的方法随机抽样M个样本来更新参数，这也是我们常用的Minibatch的方法。这个时候相当于有两层抽样，外层的抽样是从整个样本集中抽取batch，内层的抽样是从当前的batch中抽取若干个样本。

$$L(\theta, \phi; X) \simeq \frac{N}{M}\tilde{L}(\theta, \phi; x^{(i)})$$

如何挑选可导变换$g_\phi(\epsilon, x)$以及辅助噪声参数$\epsilon \sim p(\epsilon)$, 文章中提供了三种选择方案：

1. 可导的概率密度函数的积分CDF (cumulative distribution function)的倒数。噪声参数为高斯分布$\epsilon \sim U(0, I)$,$g_\phi(\epsilon, x)$是后验概率$q_\phi(z|x)$的CDF的倒数，例如Exponential, Cauchy, Logistic, Rayleigh, Pareto, Weibull, Reciprocal, Gompertz, Gumbel and Erlang distributions。
2. 噪声参数为高斯分布$\epsilon \sim U(0, I)$，g是均值+方差*$\epsilon$,例如Laplace, Elliptical, Student's t, Logistic, Uniform, Triangular and Gaussian distributions.
3. 有时候可以直接将随机变量表示为辅助噪声变量的变形，例如Log-Normal (exponentiation of normally distributed variable), Gamma (a sum over exponentially distributed variables), Dirichlet (weighted sum of Gamma variates), Beta, Chi-Squared, and F distributions.

## 2.3 VAE

AEVB的框架确定之后，需要确定的参数以及分布有：噪声参数$\epsilon$及其先验分布$p(\epsilon)$,引入的近似后验概率(encoder) $q_\phi(z|x)$,潜在变量z的先验分布$p_\theta(z)$,可导变换$g_\phi(\epsilon, x)$以及生成样本的似然函数(decoder)$p_\theta(x|z)$。
VAE就是将上述参数和分布实例化的一个例子，具体的

$$
\begin{aligned}
p(\epsilon) &= N(\epsilon; 0, I) \\
p_\theta(z) &= N(z; 0, I) \\
q_\phi(z|x^{(i)}) &= N(z; \mu^{(i)}, \sigma^{2(i)} I) \\
g_\phi(\epsilon^{(l)}, x^{(i)}) &= \mu^{(i)} + \sigma^{(i)} \odot \epsilon^{(l)} \\
p_\theta(x_i|z) &= y_i^{x_i}(1-y_i)^{1-x_i}\ or\ N(x^{(i)}; \mu'^{(i)}, \sigma'^{2(i)} I)
\end{aligned}
$$

使用高斯分布，一方面是因为高斯分布很容易拟合现实的数据，另一方面是便于计算。结合前面的推导，给定样本集X，我们可以训练出VAE的模型，模型得到了一个逼近的后验分布$q_\phi(z|x^{(i)})$，根据这个分布，我们可以随机抽样z来生成新的样本。

# 3. VAE的应用

### 3.1 手写字体生成

### 3.2 网络生成