# VAE（AEVB）原文笔记

在存在具有难解的后验分布的连续潜在变量和大型数据集的情况下，我们如何在有向图模型中进行有效的推理和学习？我们介绍了一种随机变分推断和学习算法，该算法可扩展到大型数据集，并且在某些温和的微分条件下，甚至可以在难解的情况下工作。

### 我们的贡献是双重的：

（1）首先，我们证明了变分下界的再参数化产生了一个下界估计量，该估计量可以使用标准随机梯度方法直接优化。

（2）其次，我们表明，对于每个数据点具有连续潜在变量的（iid）数据集，通过使用提出的下界估计器将近似推理模型（也称为识别模型）拟合到难解的真实后验，可以实现特别有效的实现后验推断。理论上的优势可以体现在实验结果上。

我们如何使用连续的潜在变量和/或参数具有难以处理的后验分布的定向概率模型进行有效的近似推理和学习？变分贝叶斯（VB）方法涉及对难解性后验的近似优化。不幸的是，普通的均值场方法需要对期望值进行近似后验的解析解，这在一般情况下也是很难解决的。我们展示了变分下界的重新参数化如何产生下界的一个简单的可微分的无偏估计量。该SGVB（随机梯度变分贝叶斯）估计器可用于几乎任何具有连续潜在变量和/或参数的模型中的有效近似后验推断，并且可以使用标准随机梯度上升技术直接进行优化。

对于iid数据集和每个数据点连续的潜在变量的情况，我们提出了自动编码变分贝叶斯（AEVB）算法。在AEVB算法中，我们通过使用SGVB估计器优化识别模型来使推理和学习特别有效，该模型使我们能够使用简单的祖先采样执行非常有效的近似后验推理，从而使我们能够高效地学习模型参数，而无需每个数据点需要昂贵的迭代推理方案（例如MCMC）。所学习的近似后验推理模型也可以用于许多任务，例如识别，去噪，表示和可视化目的。当将神经网络用于识别模型时，我们得出了变分自动编码器。

### 假设的情形：

（1）先生成z再生成x。

（2）难解性：EM算法核心在于p（z|x）是可解的，这样p（x）必须是可解的，p（x|z;selta）一般可解。甚至平均场变分贝叶斯都是很难算的。（平均场变分贝叶斯是lda的变分推断的核心）

（3）大量数据：我们有大量数据，以至于批处理优化的成本太高；我们希望使用小型批次甚至单个数据点进行参数更新。另一方面，基于采样的解决方案（例如蒙特卡洛EM）太慢，因为它涉及每个数据点通常昂贵的采样循环。

### 三个问题：

1.对参数θ的有效近似ML或MAP估计。参数本身可能很有趣，例如，如果我们正在分析某个自然过程。它们还使我们能够模仿隐藏的随机过程，并生成类似于真实数据的人工数据（VAE）。

2.给定观测值x的潜在变量z的有效近似后验推断用于选择参数θ。这对于编码或数据表示任务很有用。

3.变量x的有效近似边际推断。这使我们能够执行所有需要先于x的推理任务。计算机视觉的常见应用包括图像去噪，修复和超分辨率。

### 解决的方案

为此我们提出一个识别模型qφ（z | x）：近似于难处理的真实后验pθ（z | x）。请注意，与近似在平均场变分推断后验不同，它不一定是阶乘的，并且它的参数φ并不是从某种封闭形式的期望中计算出来的。相反，我们将介绍一种与生成模型参数θ一起学习识别模型参数的方法。

从编码理论的角度来看，未观察到的变量z具有潜在表示或代码的解释。因此，在本文中，我们还将参考识别模型qφ（z | x）作为一个概率编码器，由于给定了一个数据点x，因此它产生了一个分布（例如高斯分布）编码z可能从中生成数据点x的可能值。同样，我们将pθ（x | z）称为概率解码器，因为给定z会产生一个x的分布。

批注：对EM而言，证据下界容易最大化的原因在于是完整数据。

而vae则是在于隐变量的假设可以相对随意与p（x|z）好算。

此处需要

SGVB：期望转采样求和转求导。但是要注意蒙特卡洛法本身近似的期望标志（测度论下的表示是什么）是有导数的。。

AEVB：

先采M个样本，推导M个z的分布参数。

再采噪音，再用这个噪音集体去生成M个z。（确切而言是L个噪音只不过这里L=1）

对应就有M个x的分布（概率）就可以计算对应似然，以让他最大化。

（直接的想法是：

一个样本

采L个噪音（L在有些论文设置为50）

对应就有M个x的分布（概率）就可以计算对应似然，以让他最大化。

）

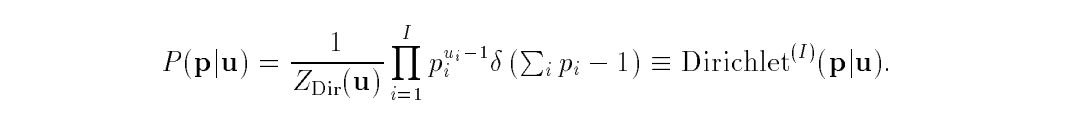
经验：如果成绩很高，建议py拿国奖。。钱可以分，甚至可以白给，但奖得拿。。（算错一步。。导师打分。。。）

至少应该有差异，否则会有大量低不成高不就的人。。

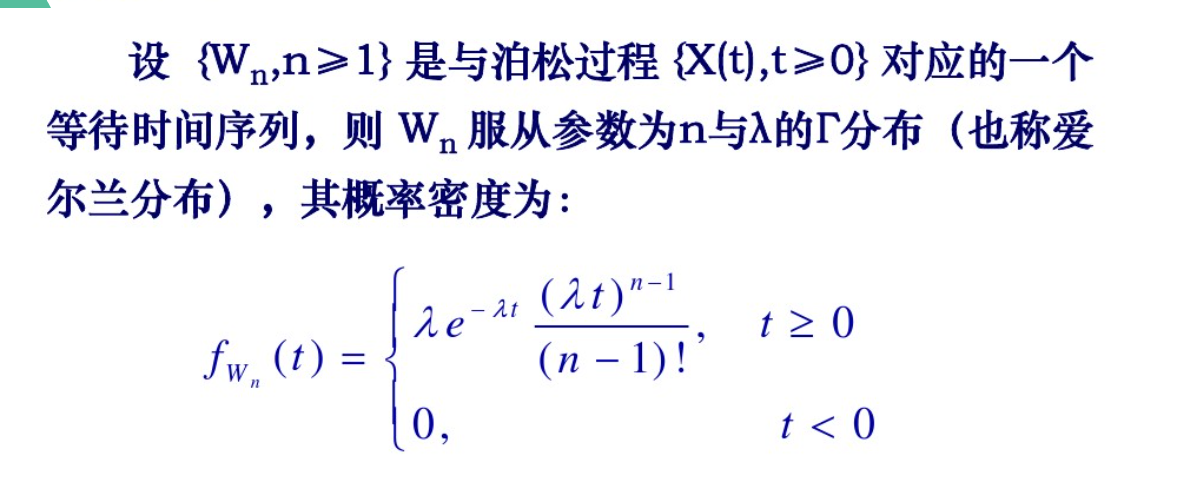
topic coherence怎么算的。。

实际多项式分布的形式

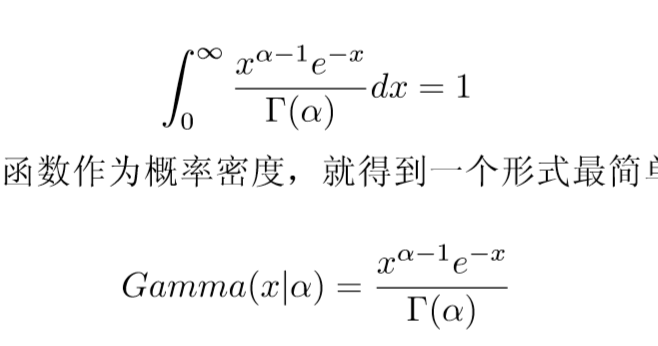
带限制的建议用delta表示。



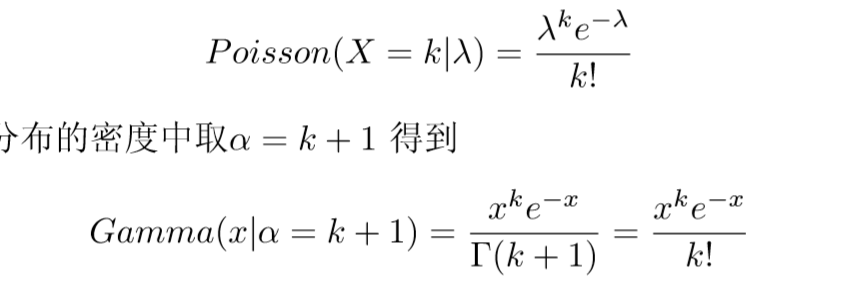




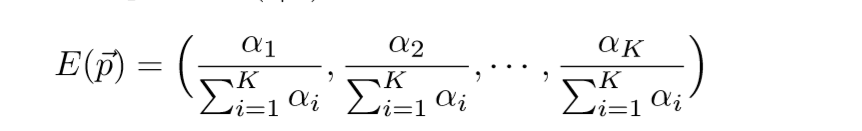
n代表来了几个，一个时是指数分布。



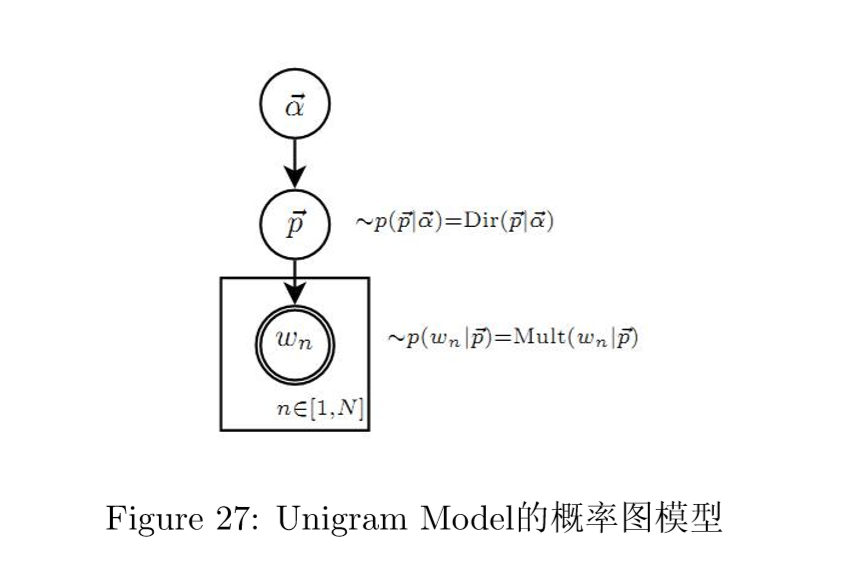
注意伽马分布直观上是x的函数。



固定x时对任意k求和



注意是对不同p算期望。

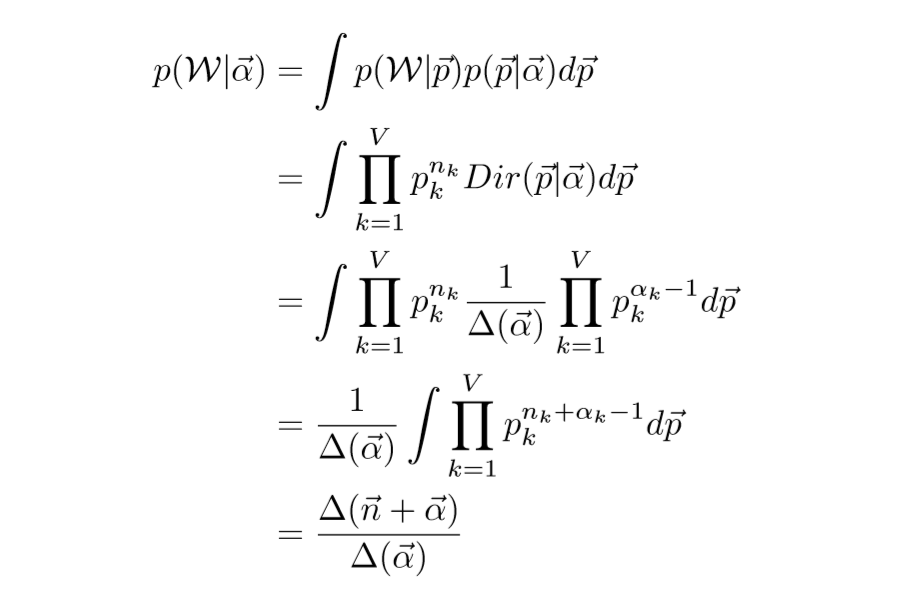


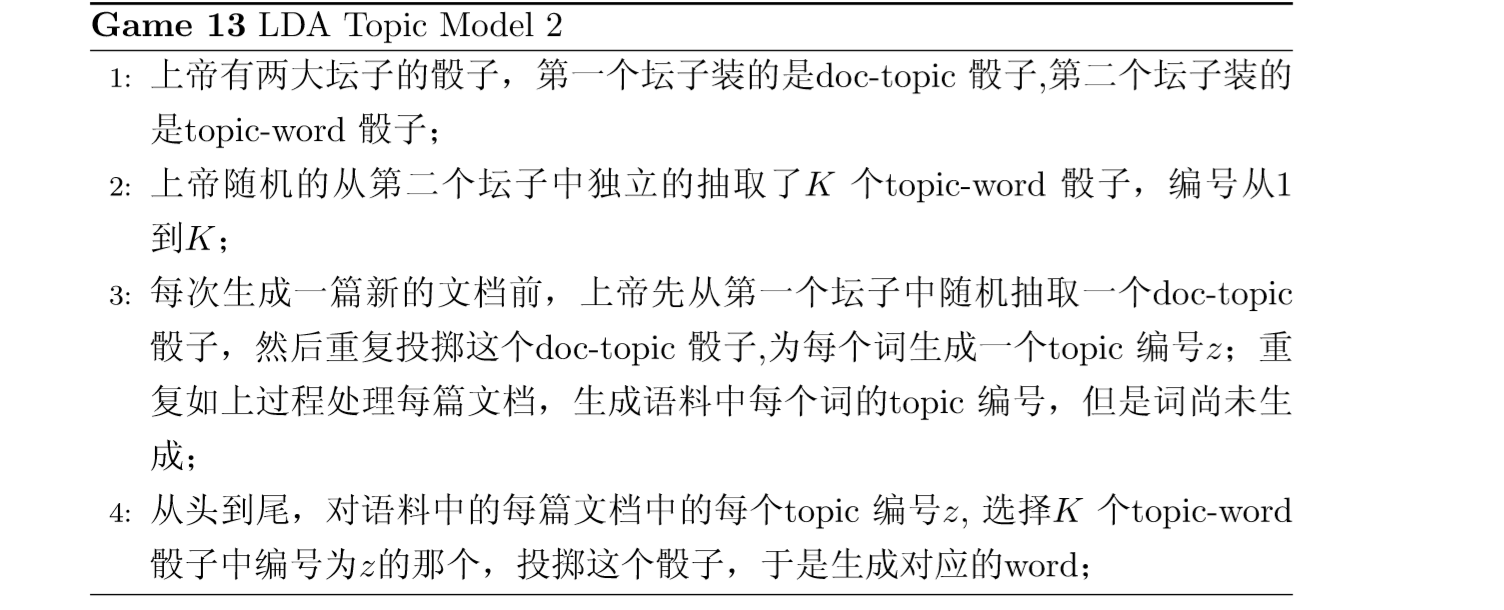
可以理解为简化版本，一个文档先对每个词语生成一个主题，而主题对词的映射是固定。

LDA是文档(α)生成主题混合，主题混合生成词语的主题，

Beta生成主题----词 混合 上面两个生成词语。

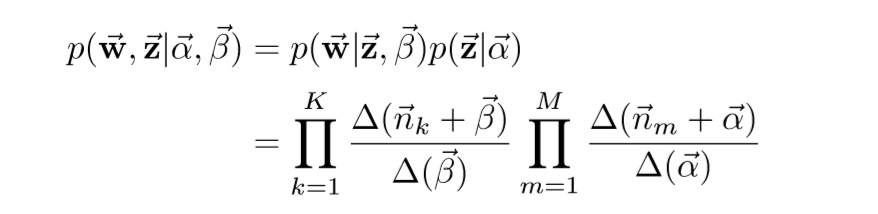
可以假设先生成一个“主题”（p即为对应向量），再根据主题不断采样。



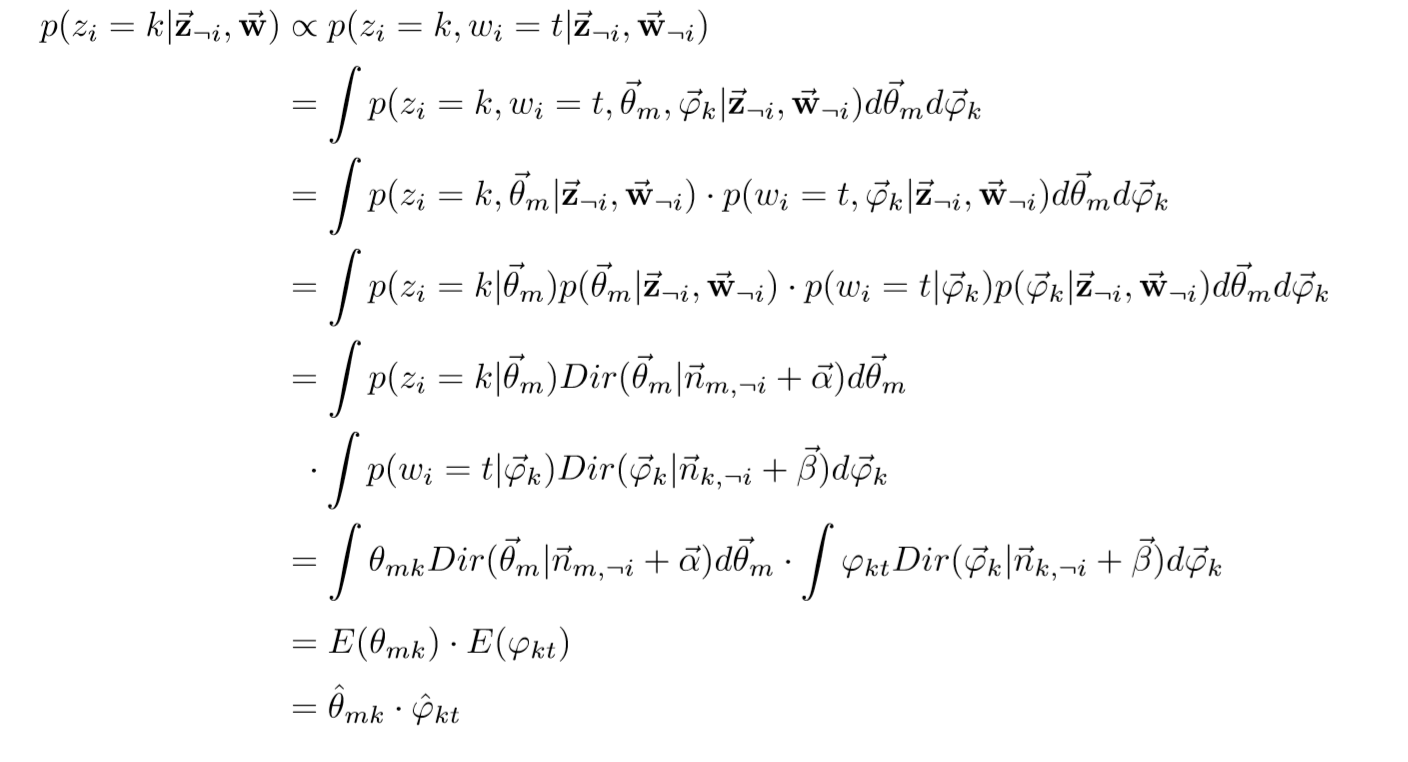


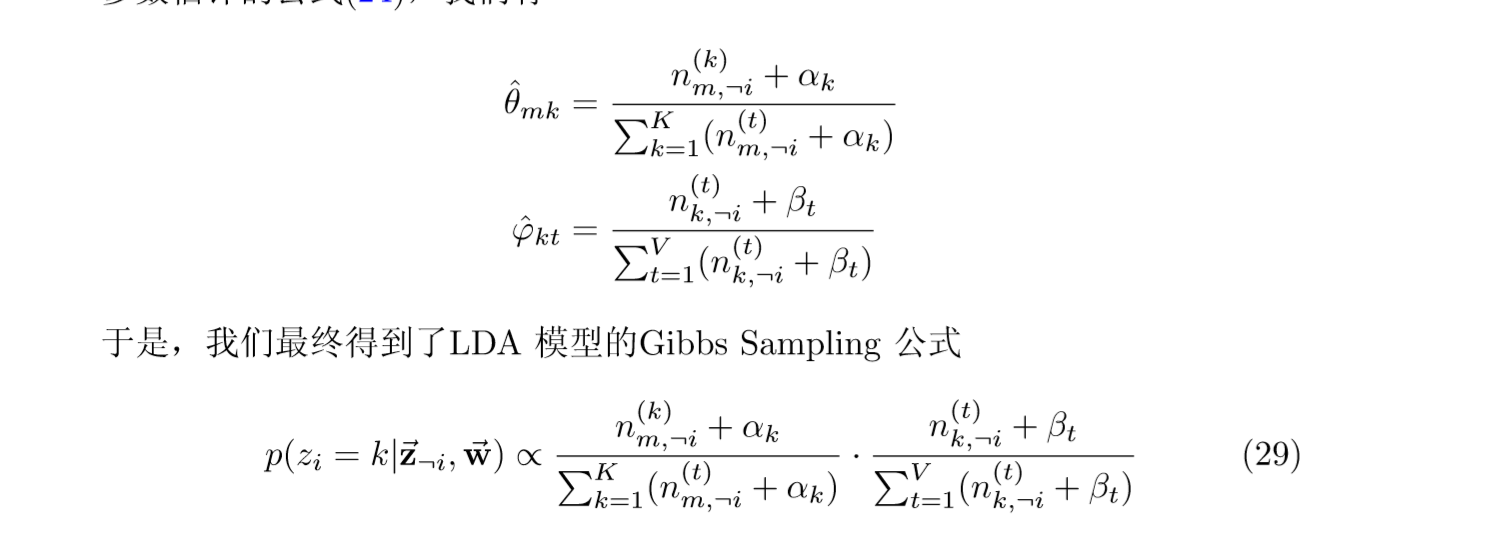
1：Dir，Dir

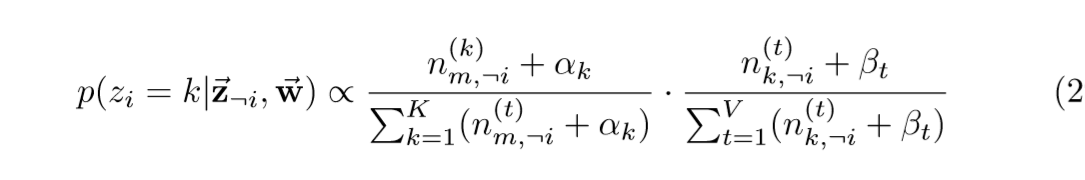
1. 得到beta
2. 得到α，得到z
3. 得到w



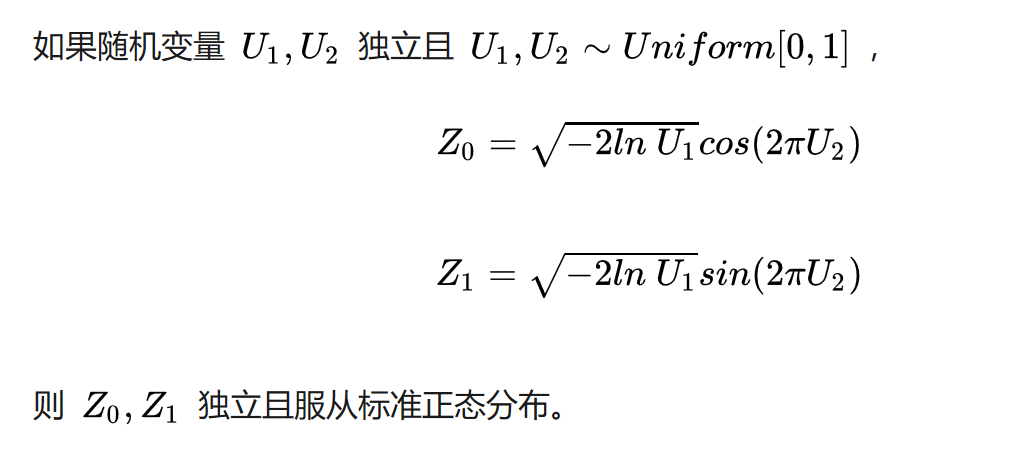
先拿到topic-word色子，再获得doc-topic色子。





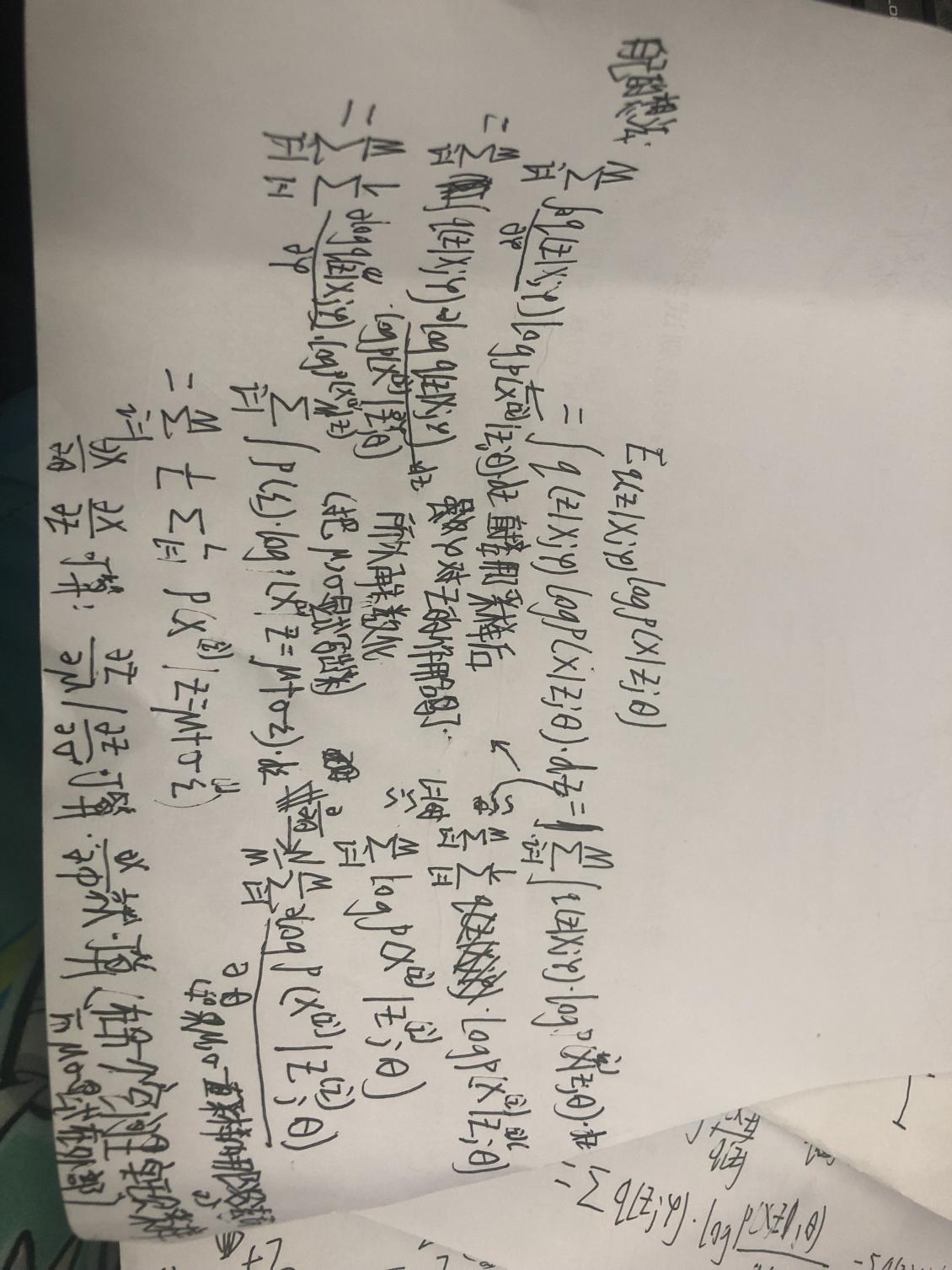


两类多项式分布对着混合即成矩阵。



Gibbs采样其实并不用轮换抽样。

再参数化：

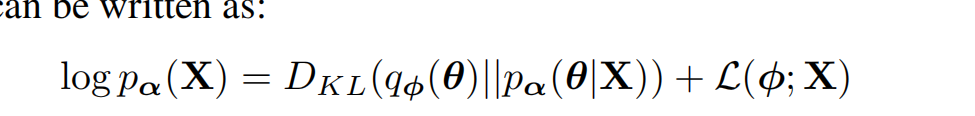


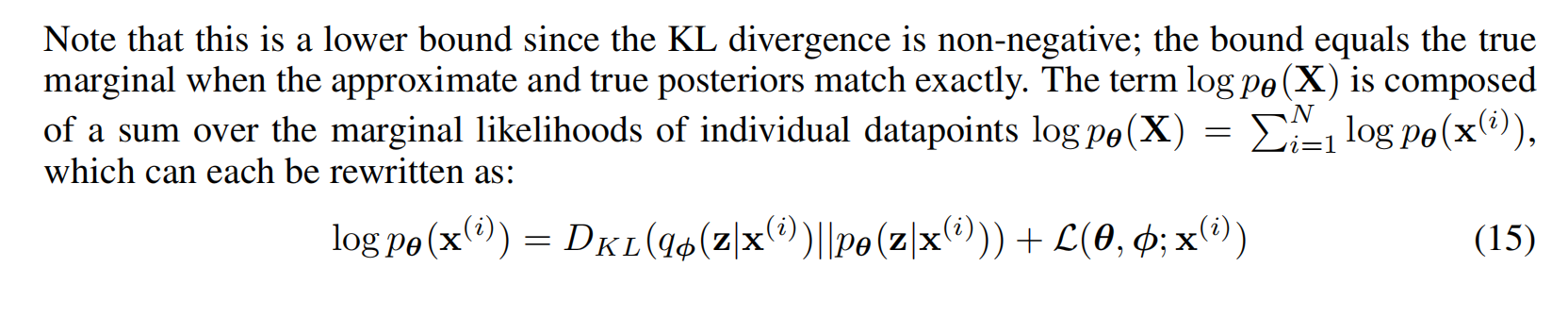


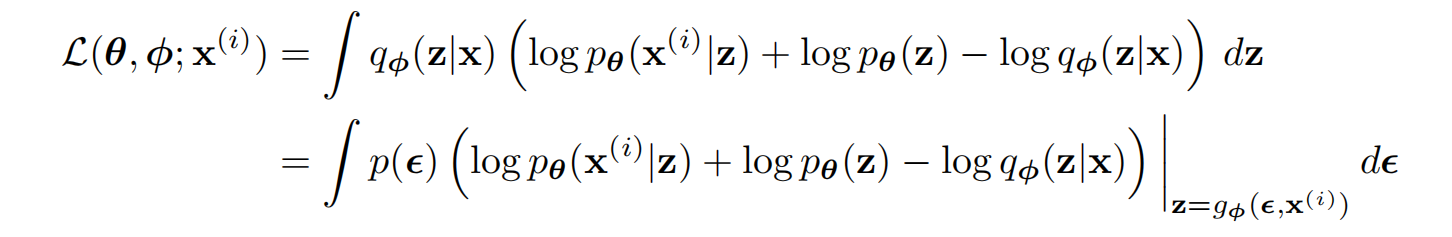
很多时候我们只会选m个再乘以M/N做近似。。极端情况M=1

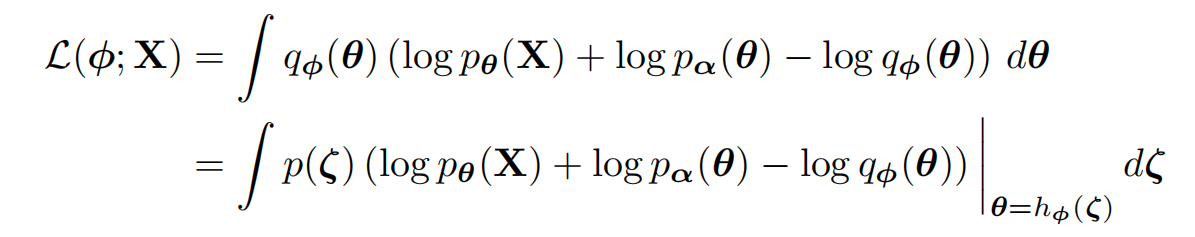
即随意取一个点然后乘以N

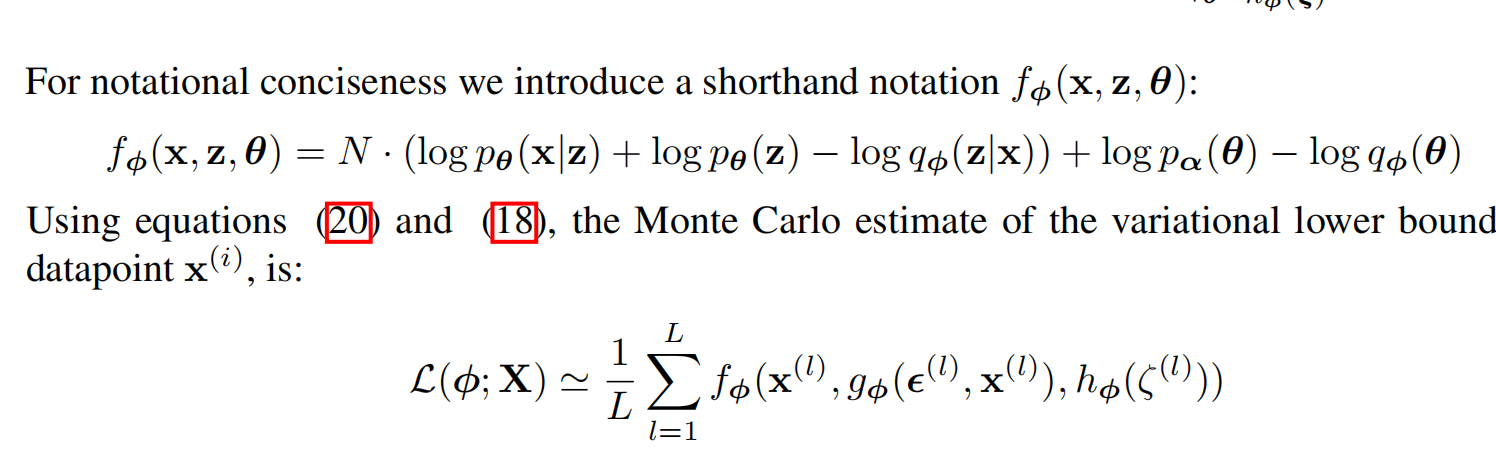
我们也可以用来推导selta。



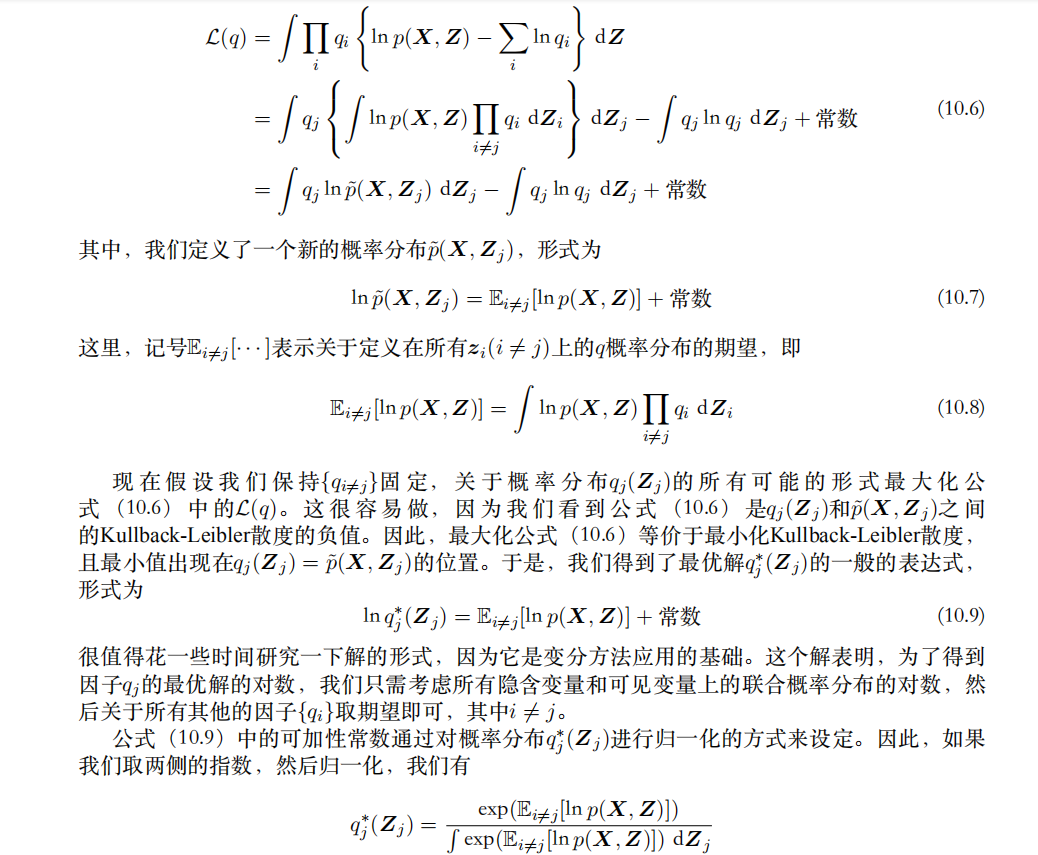








1. 使用用于编码器的深度神经网络（例如卷积网络）学习分层生成体系结构和解码器，由AEVB联合培训；（2）时间序列模型（即动态贝叶斯网络）；（3）将SGVB应用于全局参数；（4）具有潜在变量的监督模型，对于学习复杂的噪声分布很有用。



318.