自动编码变分推理用于主题模型

阿卡什·斯里瓦斯塔瓦（Akash Srivastava）
爱丁堡大学 10 信息中心 Crichton St
英国爱丁堡（EH89AB）
Akas.SristaVaA.E.A.U.K.

查尔斯·萨顿*
爱丁堡大学 10 信息中心 Crichton St
英国爱丁堡（EH89AB）
联合国

摘要

主题模型是学习文本表示的最受欢迎的方法之一，但是一个主要的挑战是，对主题模型的任何更改都需要在数学上推导新的推理算法。解决此问题的一种有前途的方法是自动编码变分贝叶斯（AEVB），但事实证明很难将其应用于主题模型。我们提供了我们所知的第一个有效的基于 AEVB 的潜在狄利克雷分配（LDA）的推理方法，我们将其称为主题模型的自动编码变分推理（AVITM）。该模型解决了 Dirichlet 之前和组件崩溃对 AEVB 造成的问题。我们发现 AVITM 在准确性上与传统方法相匹配，推理时间要短得多。确实，由于有推理网络，我们发现没有必要支付对测试数据进行变分优化的计算成本。由于 AVITM 是黑匣子，因此很容易将其应用于新主题模型。为了说明这一点，我们展示了一个称为 ProdLDA 的新主题模型，该模型用专家产品代替了 LDA 中的混合模型。通过仅从 LDA 更改一行代码，我们发现 ProdLDA 产生了更多可解释的主题，即使 LDA 是通过折叠的吉布斯采样进行训练的。

1 介绍

主题模型（布莱， 2012) 是用于学习无监督文本表示的最广泛使用的模型之一，在文献中有数百种不同的模型变体，并且已经发现了从探索科学文献到应用的各种应用（布莱和拉菲蒂， 2007) 到计算机视觉（菲菲和佩罗纳， 2005), 生物信息学（Rogers 等， 2005), 和考古（明诺， 2009). 应用主题模型和开发新模型的主要挑战是计算后验分布的计算成本。因此，大量工作考虑了近似推理方法，最流行的方法是变分方法，尤其是均值场方法，以及马尔可夫链蒙特卡罗方法，尤其是基于折叠吉布斯采样的方法。

均值场和塌陷的吉布斯都具有以下缺点：即使建模假设只有很小的变化，将它们应用于新的主题模型也需要重新推导推理方法，这在数学上可能是费时费力的，并且存在局限性从业者自由探索不同建模假设空间的能力。这激励了黑匣子推理方法的发展（Ranganath 等，2014; MNIH 和格雷戈， 2014; Ku 波段 cukelbir 等， 2016; 金马威林 2014) 它仅需非常有限且易于从模型计算信息，因此，只要生成过程的声明性规范简单，就可以自动应用于新模型。

自动编码可变贝叶斯（AEVB）（金马威灵， 2014; Rezende 等， 2014) 对于主题模型来说是一个特别自然的选择，因为它训练了推理网络（达扬等人， 1995), 直接将文档映射到近似后验分布的神经网络，

*隶属关系：艾伦·图灵学院，大英图书馆，伦敦尤斯顿路 96 号，NW1 2DB

无需运行其他变体更新。这在直观上很吸引人，因为在主题模型中，我们期望从文档到后验分布的映射行为良好，也就是说，文档中的少量更改只会产生主题的较小更改。这正是像神经网络这样的通用函数逼近器应该擅长表示的映射类型。从本质上讲，推理网络学会了模仿概率推理的效果，因此在测试数据上，我们可以享受概率建模的好处，而无需付出更多的推理费用。

但是，尽管潜在高斯模型取得了一些显著成功，但黑盒推理方法要应用于主题模型则要困难得多。例如，在最初的实验中，我们尝试应用 ADVI (Kucukelbir 等， 2016)，最近的黑盒变分方法，但是很难获得任何有意义的主题。两个主要挑战是：首先，Dirichlet 先验不是位置标度系列，这阻碍了重新参数化；其次，众所周知的组件折叠问题 (Dinh 和 Dumoulin，2016)，其中，推理网络陷入所有主题均相同的不良局部最优中。

在本文中，据我们所知，什么是第一种有效的主题模型 AEVB 推理方法，我们将其称为主题模型的自动编码变分推理或 AVITM[1]。在几个数据集上，我们发现 AVITM 产生的质量与标准平均场推论的质量相当，而训练时间却大大减少。我们还发现，推理网络学会了高度精确地模拟近似推理的过程，因此根本不需要对测试数据进行变分优化。

但是也许更重要的是，AVITM 是一种黑盒方法，易于应用于新模型。为了说明这一点，我们提出了一个新的主题模型 ProdLDA，其中各个单词的分布是专家的产物，而不是 LDA 中使用的混合模型。我们发现，无论通过自动确定的主题一致性或定性检查来衡量，ProdLDA 始终比标准 LDA 产生更好的主题。此外，由于我们使用神经网络执行概率推理，因此我们可以在 80 分钟内在单个 GPU 上将主题模型拟合到大约一百万个文档上，并且由于我们使用的是黑匣子推理方法，因此实现 ProdLDA 需要进行以下更改：我们执行标准 LDA 的代码只有一行。

总而言之，我们方法的主要优点是：

1. *主*题连贯性：即使使用 Gibbs 采样训练了 LDA，ProdLDA 始终比 LDA 返回更好的主题。
2. *计*算效率：训练 AVITM 像标准均值场一样快速高效。在新数据上，AVITM 比标准均值场快得多，因为它仅需要通过神经网络的一次前向传递。
3. *黑*匣子：AVITM 不需要严格的数学推导即可处理模型中的更改，并且可以轻松地应用于各种主题模型。

总体而言，我们的结果表明，AVITM 已准备好与均值字段并列，而 Gibbs 崩溃了，成为主题模型的主要推论方法之一。

2 背景

为了修正符号，我们首先描述主题建模和 AVITM。

2.1 潜在狄利克雷分配

我们描述了最流行的主题模型，潜在狄利克雷分配 (LDA)。在 LDA 中，集合的每个文档都表示为主题的混合，其中每个主题 β_k 是整个词汇表的概率分布。我们还使用 β 表示矩阵 $\beta = (\beta_1 \ldots \beta_K)$。生成过程如算法中所述 1. 在这种生成模型下，边际可能性

[1] 代码位于 https://github.com/akashgit/autoencoding_vi_for_topic_models

为*每个文件做*
  绘制主题分布 θ∽Dirichlet（α）；
  对*于位置 n 处的每个单词*
    样本主题 zn∽多项式（1，θ）；样本词 wn-多项式（1，βzn）；
  结束

算法 1：LDA 作为生成模型。

文件是

$$p(\mathbf{w}|\alpha,\beta) = \int_{\theta} \prod_{n=1}^{N} \sum_{z_n=1}^{k} P(\mathrm{Wn}, \mathrm{Zn}, \beta)\, P(\mathrm{Zn}\times\mathrm{Th})\, P(\theta/\alpha)\, D\theta。 \tag{1}$$

由于变量之间的耦合，隐藏变量 θ 和 z 的后验推论是棘手的。
多项式假设下的 θ 和 β（迪基，1983).

2.2 平均场和 aevb

在主题模型中有效推理的一种流行近似方法是平均场变分推理，它通过在 θ 上引入自由的变分参数 γ 和在 z 上引入 φ 并掉落它们之间的边缘来打破 θ 和 z 之间的耦合。这导致近似的变化后验 q（θ，z |γ，φ）=qγ（θ）神经网络 qφ（zn），经过优化可最佳地逼近真实后验 p(θ，z|w，α，β).优化问题是要最小化

$$L(\gamma,\varphi/\alpha,\beta) = DKL[q(\theta,z/\gamma,\varphi)\,||\,p(\theta,z/w,\alpha,\beta)] - \log p(w/\alpha,\beta)。 \tag{2}$$

实际上，以上等式是边际对数似然的下界，有时也称为证据下界（ELBO），这一事实可以通过乘以除法轻松地得到验证。(1) 通过变分后验，然后在其对数上应用詹森不等式。请注意，平均场方法会针对每个文档在一组独立的变分参数上进行优化。为了强调这一点，我们将以非标准名称“去耦均值场变分推论（DMFVI）”来引用此标准方法。

对于 LDA，由于 Dirichlet 和多项式分布之间的共轭性，因此此优化具有封闭形式的坐标下降方程。尽管这是 DMFVI 的计算方便方面，但它也限制了它的灵活性。将 DMFVI 应用于新模型依赖于从业者获得封闭式更新的能力，这可能不切实际，有时甚至是不可能的。

装甲兵部队（金马威灵，2014; Rezende 等，2014) 是旨在“黑匣子”推理方法规避此问题的几种最新方法之一。首先，将 ELBO 改写为

$$L(\gamma,\varphi/\alpha,\beta) = -DKL[q(\theta,z/\gamma,\varphi)\,||\,p(\theta,z/\alpha)] + E_{q(\theta,z/\gamma,\varphi)}[\log p(w/z,\theta,\alpha,\beta)] \tag{3}$$

这种形式很直观。第一项试图将潜在变量上的后验变量与潜在变量上的先验匹配，而第二项确保变量后验有利于潜在变量的值，该值擅长解释数据。类似于自动编码器，该第二项称为重构项。

使这种方法“自动编码”（实际上与 DMFVI 的主要区别）的是变量分布的参数化。在 AEVB 中，通过使用称为推理网络的神经网络来计算变化参数，该神经网络将观察到的数据作为输入。例如，如果模型先验 p（θ）是高斯模型，我们可以将推理网络定义为前馈神经网络（μ（w），v（w））= f（w，γ），其中 μ（w）和 v（w）都是长度为 k 的向量，而 γ 是网络的参数。然后我们可以选择一个高斯变分分布 qγ（θ）= N（θ；μ（w），diag（v（w））），其中 diag（●●●）从列 vec-产生对角矩阵 TOR。然后可以通过优化 ELBO 选择变化参数 γ (3). 请注意，我们有

现在，与 DMFVI 不同，它们耦合了不同文档的变异参数，因为它们都是从同一神经网络计算得出的。计算关于 q in 的期望(3), 金马与威灵 (2014); Rezende 等。 (2014) 使用他们称为"重新参数化技巧" (RT; 也出现在威廉姆斯 (1992)). 在 RT 中，我们定义了具有简单分布的变量 U，该变量独立于所有变量参数（如均匀或标准正态），并且具有重新参数化函数 F，使得 F (U, γ) 具有分布 qγ。这总是可能的，因为我们可以选择 F 作为 qγ 的逆累积分布函数，尽管我们还将希望 F 易于计算和可微。如果我们可以确定一个合适的 F，那么我们可以近似(3) 通过获取 U 的蒙特卡洛样本，并使用随机梯度下降法优化 γ。

3 在潜在狄利克雷中自动编码变分贝叶斯分配

尽管从概念上讲很简单，但将 AEVB 应用于主题模型提出了一些实际挑战。首先是需要确定 q (θ) 和 q (zn) 的重新参数化函数以使用 RT。zn 很容易处理，但是 θ 更加困难；如果选择 q (θ) 为 Dirichlet，则很难应用 RT，而如果选择 q 为高斯或逻辑对数，则 KL 散度为(3) 变得更加成问题。第二个问题是组件崩溃的众所周知的问题 (Dinh 和 Dumoulin, 2016), 这是不良的局部最优类型，尤其是 AEVB 和类似方法所特有。在接下来的几个小节中，我们将描述针对每个问题的解决方案。

3.1 收拾 z'S

使用重新参数化处理像 z 这样的离散变量可能会遇到问题，但是幸运的是，在 LDA 中，可以方便地将变量 z 求和。通过折叠 z，我们只需要从 θ 采样，就可以减少(1)至

$$p(\mathbf{w}|\alpha,\beta) = \int_\theta \prod_{n=1}^{N} P(w_n|\beta,\theta)\, P(\theta|\alpha)\, d\theta. \quad (4)$$

其中 wn |β, θ 的分布为 Multinomial (1, βθ), 请记住，β 表示所有主题词概率向量的矩阵。

3.2 处理狄里克雷信念：拉普拉斯近似

LDA 从主题比例 θ 的狄里克雷特先验得名，而狄里克雷特先验的选择对于获得可解释的主题很重要 (Wallach 等， 2009). 但是，由于难以为 RT 开发有效的重新参数化功能，因此难以在 AEVB 中处理 Dirichlet。幸运的是，对于高斯分布确实存在 RT，并且已经证明在变分自动编码器 (VAE) 的情况下，RT 具有很好的性能 (金马威灵， 2014).

我们通过构造 Dirichlet 先验的 Laplace 近似来解决此问题。以下 麦凯 (1998), 我们以 softmax 为基础，而不是单纯形。这种选择有两个好处。首先，狄利克雷分布在 softmax 基础上是单峰的，其模式与变换后的密度均值一致。其次，softmax 基础还允许在没有单纯形约束的情况下对成本函数进行无约束的优化。在此基础上，关于 softmax 变量 h 的 Dirichlet 概率密度函数由下式给出：

$$P(\theta(\mathbf{h})|\alpha) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \theta_k^{\alpha_k}\, g(\mathbf{1}^T\mathbf{h}). \quad (5)$$

θ=σ(h)，其中 σ(。) 表示 softmax 函数。回想一下，σ 的雅可比行列式与 θk 成正比，而 g (●) 是一个任意密度，可以通过约束多余的自由度。我们使用的拉普拉斯近似 Hennig 等。 (2012)， 哪一个 自然杀伤细胞

具有以下特性：对于大 k（主题数），协方差矩阵变为对角线。Dirichlet 先验 p（θ|α）的近似值导致 softmax 变量的分布
h 是具有均值 μ1 和协方差矩阵 Σ 1 的多元法线

$$\mu 1k = \text{对数}\,\alpha k - \frac{1}{k}\sum_i \text{对数}\,\alpha i$$

$$1kk = \frac{1}{\alpha_k}\left(1 - \frac{2}{K}\right) + \frac{1}{K^2}\sum_i \frac{1}{\alpha_k}. \tag{6}$$

最终，我们以 pˆ（θ | μ1，Σ1）= LN（θ | μ1，Σ1）在单纯形基础上近似 p（θ|α），其中 LN 是参数 μ1，Σ1 的对数正态分布。尽管我们用逻辑对数来近似 LDA 中的 Dirichlet 优先级，但这与相关主题模型的想法不同（布莱
拉弗蒂， 2006)，因为我们使用对角协方差矩阵相反，它是标准 LDA 的近似值。

3.3 变分目标

现在我们可以编写修改后的变分目标函数。我们使用对角协方差在 θ 上的对数正态变化分布。更准确地说，我们将两个推理网络定义为参数 δ 的前馈神经网络 fμ 和 fΣ。每个网络的输出是 R^k 中的向量。然后对于文档 w，我们将 q（θ）定义为对数正态，均值 μ0 = fμ（w，δ），对角协方差 Σ0= diag（fΣ（w，δ））， 其中 diag 将列向量转换为对角线
矩阵。注意，我们可以通过采样 f∽N（0，I）并从 q（θ）生成样本
$\theta = \sigma(\mu_0 + \Sigma_0^{1/2} f)$.

我们现在可以将 ELBO 编写为

$$\underset{=}{I}(\theta) \underset{d=1}{\overset{D}{-}} \frac{1}{2}\left(Tr\left(^{-I}0\right) + \quad - \mu\right)^T \Sigma^{-1}(\mu \; / \; \mu) - K + \text{对数}\frac{\Sigma|}{|\Sigma_0|} \tag{7}$$

$$+ e_{E\sim N(0,I)}\, W_d \text{日志}\left(\sigma(\beta)\sigma(\mu_0 + \Sigma_0^{1/2} f), 1^1\right),$$

其中 Θ 表示所有模型和变分参数的集合，w1。..wD 是语料库中的文档。该方程式的第一行来自两个逻辑正态分布 q 和 pˆ之间的 KL 散度，而第二行是重构误差。

为了在优化过程中对 β 矩阵施加单纯形约束，我们应用 softmax 变换。也就是说，每个主题 $\beta k \in R$ 是不受约束的，并且符号 σ（β）表示
将 softmax 函数分别应用于矩阵 β 的每一列。请注意，然后可以将每个单词 wn 的多项式混合表示为 p（wn |β，θ）= [σ（β）θ]wn，
点积(7). 优化(7)，我们根据无意识统计学家的定律，使用来自 f 的蒙特卡洛样本的随机梯度下降法。

3.4 培训和实践考虑：处理组件崩溃

AEVB 容易崩溃（Dinh 和 Dumoulin， 2016)，这是训练初期的一种非常接近先验信念的局部最优类型。随着模型潜在维度的增加，变分目标中的 KL 正则化占主导地位，因此对于潜在变量中接近先验变量的分量，输出解码器权重将崩溃，并且不会表现出任何后验差异。在我们的情况下，由于包含了 softmax 变换以生成 θ，因此特别发生了崩溃。结果是，k 个推断的主题与表中所示的相同 7.

通过调整优化，我们能够解决此问题。具体来说，我们使用 ADAM 优化器来训练网络（金马和爸， 2015) 使用高力矩权重（β1）和学习率（η）。通过以更高的速率进行训练，可以轻松避免功能空间中的早期高峰。该

问题在于，基于动量的训练加上较高的学习率会导致优化器出现差异。虽然显式梯度裁剪在一定程度上有所帮助，但我们发现批量归一化 (艾菲和塞格迪， 2015) 通过平滑功能空间并因此抑制突然的差异，可以做得更好。

最后，我们还发现，当应用到 θ 强制网络使用更多容量时，丢弃单元的性能会提高。

尽管在 AEVB 框架中更为突出，但如果学习偏移（称为 τ 参数），则也可能在 DMFVI 中发生崩溃 (霍夫曼， 1999)) 设置不正确。有趣的是，在训练的早期迭代中也可以使用类似的基于学习偏移或基于退火的方法来降低 KL 项的权重，以避免局部最优。

4 prodlda：具有以下乘积的潜在狄利克雷分配专家

在 LDA 中，分布 p（w |θ， β）是多项式的混合。这种假设的问题在于，它永远无法做出比被混合的成分更准确的预测 (欣顿和萨拉赫特迪诺夫， 2009)。这可能会导致某些主题的质量较差并且与人为判断不符。解决此问题的一种方法是用专家加权的产品代替这个单词级混合词，根据定义，比任何组成专家都可以做出更准确的预测 (韩丁， 2002)。在本节中，我们介绍了一种新颖的主题模型 PRODLDA，该模型用专家的加权产品替换了 LDA 中单词级别的混合假设，从而极大地改善了主题一致性。这很好地说明了像 AVITM 这样的黑匣子推断方法的优点，可以探索新模型。

4.1 模型

PRODLDA 模型可以简单地描述为潜在 Dirichlet 分配，其中主题之间的单词级混合是在自然参数空间中进行的，即在混合之前，主题矩阵不限于存在于多项式单纯形中。换句话说，LDA 的唯一变化
是 β 未归一化，并且 wn 的条件分布定义为 wn |β， θ ∽
多项式（1， σ（β θ））。

与专家产品的连接非常简单，因为对于多项式，自然参数的混合对应于平均参数的加权几何平均值。即，考虑由均值向量 p 和 q 参数化的两个 N 维多项式。定义相应的
自然参数为 p =σ（r）和 q =σ（s），令 δ ∈[0，1]。然后很容易证明

$$p(x \mid \delta r + (1-\delta) s) \propto \prod_{i=1}^{N} \sigma(\delta r_i + (1-\delta) s_i)^{x_i} \propto \prod_{i=1}^{N} [r_i^{\delta} \cdot s_i^{(1-\delta)}]^{x_i}。$$

因此，PRODLDA 模型可以简单地描述为专家的产品，即 $p(w_n \mid \theta, \beta) \propto \prod_k p(w_n \mid z_n = k, \beta)^{\theta_k}$ 的。PRODLDA 是指数族 PCA 的一个实例 (Collins 等，2001) 类，并与指数家庭和声有关 (Welling 等， 2004) 但非高斯先验。

5 相关工作

有关主题建模的概述，请参见布莱 (2012)。有几个基于神经网络和神经变分推理的主题模型示例 (欣顿和萨拉赫特迪诺夫， 2009; 拉罗切尔和劳利， 2012; MNIH 和格雷戈， 2014; 苗等 2016) 但是我们不知道将 AEVB 一般应用于分析人员指定的主题模型的方法，甚至还没有成功将 AEVB 成功应用于最广泛使用的主题模型潜在的 Dirichlet 分配。

最近，苗等。 (2016) 介绍了一个密切相关的模型，称为神经变异文档模型 (NVDM)。该方法对主题使用潜在的高斯分布，例如概率潜在语义索引，并在 logit 空间中对主题词分布进行平均。然而，

他们没有使用我们工作的两个关键方面之一：在使用高斯或高动量训练之前显式逼近 Dirichlet。在实验中，我们证明了这些方面可以带来更好的培训和更好的话题。

6 实验与结果

对主题模型进行定性评估是一项艰巨的任务，因此，大量工作已开发出自动评估指标，试图与人类对主题质量的判断相匹配。传统上，困惑度被用来衡量模型的拟合优度，但一再表明，困惑度并不是主题定性评估的良好指标（新人 等 2010). 因此，提出了几个新的主题一致性评估指标。看到刘 等。 (2014) 进行比较审查。Lau 等。 (2014) 结果表明，在所有竞争指标中，一组主题中所有成对单词之间的归一化点向互信息（NPMI）与人类判断最接近，因此我们在工作中采用了它。我们还报告了困惑，主要是作为评估不同优化器功能的一种方式。遵循标准惯例 (Blei 等， 2003), 对于变分方法，我们使用 ELBO 来计算困惑度。对于 AEVB 方法，我们使用与训练相同的蒙特卡洛近似法来计算 ELBO。

我们在 20 个新闻组（12,000 个训练实例，具有 2000 个单词的词汇量）和 RCV1 第 2 卷（800K 个训练实例，具有 10000 个单词的词汇量）数据集上进行实验。我们的预处理包括标记化，删除 20 个新闻组的一些非 UTF-8 字符以及英语停用词。我们首先将 AVITM 推理方法与标准的在线平均场变分推理进行比较（霍夫曼等人， 2010) 并崩溃了吉布斯采样（格利菲斯和斯泰弗斯，2004) 在 LDA 模型上。我们使用两种方法的标准实现，DMFVI 和槌的 scikit-learn（麦卡勒姆， 2002) 倒塌的吉布斯。然后，我们在三种不同的主题模型上比较两种自动编码推理方法：使用我们的推理方法的标准 LDA，PRODLDA 和神经变异文档模型（NVDM） (Miao 等，2016), 使用本文所述的推论。[2]

# 主题	普达尔达 VAE	LDA VAE	LDA 二甲基亚	LDA 倒塌的吉布斯	NVDM
50	**0.24**	0.11	0.11	0.17	0.08
200	**0.19**	0.11	0.06	0.14	0.06

表 1：20 个新闻组数据集的平均主题连贯性。越高越好。

表 1 和 2 显示两个不同设置 k（主题数）下所有模型的平均主题一致性值。比较 LDA 的不同推理方法，我们发现，与以前的工作一致，折叠的 Gibbs 采样比均值场方法产生的主题更好。在变分方法中，我们发现 VAE-LDA 模型 (AVITM) [3] 产生与标准 DMFVI 相似的主题连贯性和困惑性（尽管在某些情况下，VAE-LDA 产生更好的主题）。但是，AVITM 的训练速度比 DMFVI 快得多。20 个新闻组需要 46 秒，而 DMFVI 需要 18 分钟。而对于 RCV1 的一百万个文档语料库而言，它仅在 1.5 小时之内即可完成，而 scikit-learn 的 DMFVI 的实现即使在运行 24 小时后仍无法返回任何结果。[4]

将新主题模型比 LDA 进行比较，很明显，即使通过折叠的 Gibbs 采样进行训练，PRODLDA 也会比 LDA 找到更好的主题。为了对此进行定性验证，我们在表格中显示所有模型的主题示例 6. ProdLDA 的主题在视觉上比 NVDM 或 LDA 更加连贯。不幸的是，NVDM 的性能不及 LDA

[2] 我们都使用了 https://github.com/carpedm20/variational-text-tensorflow 和 NVDM 作者的 (Miao 等， 2016) 实现。

[3] 我们最近发现，“增白”主题矩阵可以显着提高 VAE-LDA 的主题一致性。正在准备手稿。

[4] 因此，我们无法报告 RCV1 上 DMFVI 的主题一致性

# 主题	普达尔达 VAE	LDA VAE	LDA 二甲基亚	LDA 倒塌的吉布斯	NVDM
50	**0.14**	0.07	-	0.04	0.07
200	**0.12**	0.05	-	0.06	0.05

表 2：RCV1 数据集的平均主题一致性。越高越好。由于推理未能在 24 小时内收敛，因此未报告 LDA DMFVI 的结果。

# 主题	普达尔达 VAE	LDA VAE	LDA 二甲基亚	LDA 倒塌的吉布斯	NVDM
50	1172	1059	1046	**728**	837
200	1168	1128	1195	**688**	884

表 3：20 个新闻组的困惑分数。越低越好。

对于任何 k 值。为了避免任何训练上的差异，我们训练所有竞争模型，直到达到先前工作中报告的困惑。这些报告在表中 3[5].

AVITM 推理的主要好处在于，它不需要运行变异优化，对于新数据而言，变异优化可能会非常昂贵。相反，推理网络可用于获取新文档中新数据点的主题比例，而无需运行任何优化。我们评估这种近似是否正确，即神经网络是否有效地学习了模拟概率推断。我们通过在训练集上训练模型，然后在测试集上训练模型，固定主题（β 矩阵），并且如果我们直接通过运行推理神经网络或通过标准方法来获得主题比例来比较测试的困惑，就可以对此进行验证。测试集上推理网络的变量优化。如表所示 4，困惑几乎保持不变。这样的计算优势非常明显。在这两个数据集上，使用标准变分算法都需要在不到一分钟的时间内使用神经网络计算复杂度
即使使用较小的 20 个新闻组数据，大约也需要 3 分钟。最后，我们调查
PRODLDA 中主题一致性得到改善的原因。一，表 5 探索的影响
我们两个主要思想中的每一个分别。在此表中，“ Dirichlet”表示先验是 Dirichlet 的 Laplace 近似值（α = 0.02），而“ Gaussian”则表示我们使用标准 Gaussian 作为先验。“高学习率”训练意味着我们使用 β1> 0.8 和 0.1>η> 0.001[6] 使用批量归一化，而“低学习率”是指未经批量归一化的 β1> 0.8 和 0.0009>η> 0.00009。（对于这两个参数，精确值是由贝叶斯优化选择的。我们发现“ with BN”情况下的这些值接近于 Adam 优化器中的默认设置。）我们发现，我们在此实现的高主题一致性只有同时使用这两种技巧，才有可能开展工作。实际上，需要高的学习速度和动量来避免在训练开始时出现局部最小值，并且需要批量标准化以能够以这些值训练网络而不会发散。如果以较低的动量值或较低的学习率训练，则 PRODLDA 会显示组件崩溃。有趣的是，如果我们选择高斯先验，而不是 ProdLDA 或 NVLDA 中使用的逻辑正态近似，那么即使在学习率较低的情况下，也无需动量或批量归一化，该模型也更易于训练。

与 NVDM 相比，AVITM 主题模型的主要优点是 Laplace 近似使我们能够匹配感兴趣的特定 Dirichlet。正如所指出的 Wallach 等。(2009), Dirichlet 超参数的选择对于 LDA 的主题质量很重要。根据这种推理，我们假设 AVITM 主题比 NVDM 主题质量更高，因为它们更加集中，即适用于感兴趣的文档的更特定子集。我们为图中的这一假设提供了支持 1，通过评估主题后验比例的稀疏性，即通常使用多少个模型主题来解释每个文档。为了估计主题比例的稀疏性，我们将 PRODLDA 和 NVDM 的高斯潜在空间中的样本投影为单纯形，并将其平均化为文档。我们比较话题

[5] 我们注意到接下来有很多新工作辛顿和萨拉胡迪诺夫 (2009) 仅在测试数据的一小部分上报告 LDA Gibbs 采样器的困惑。我们的结果有所不同，因为我们使用了整个测试数据集。

[6] β1 是前一时间步的平均梯度权重，η 是学习率。

# 主题	仅推理网络	推理网络+优化
50	1172	1162
200	1168	1151

表 4：在 20 个新闻组测试集上对 VAE-LDA 推理网络的评估。**“仅推理网络”**显示了在训练集上训练推理网络时的测试复杂性，但未对测试集执行变型优化。**“推理网络+优化”**显示了在测试集上优化 ELBO 的标准方法。神经网络有效地学习有效地近似概率推断。

NVDM 使用的标准高斯先验的稀疏性与具有不同超参数的 Dirichlet 先验的 Laplace 近似。显然，狄拉克雷特先验的拉普拉斯近似值显着提高了稀疏性，为我们的假设提供了支持，即保持狄拉克雷特先验说明了我们方法中主题一致性的提高。

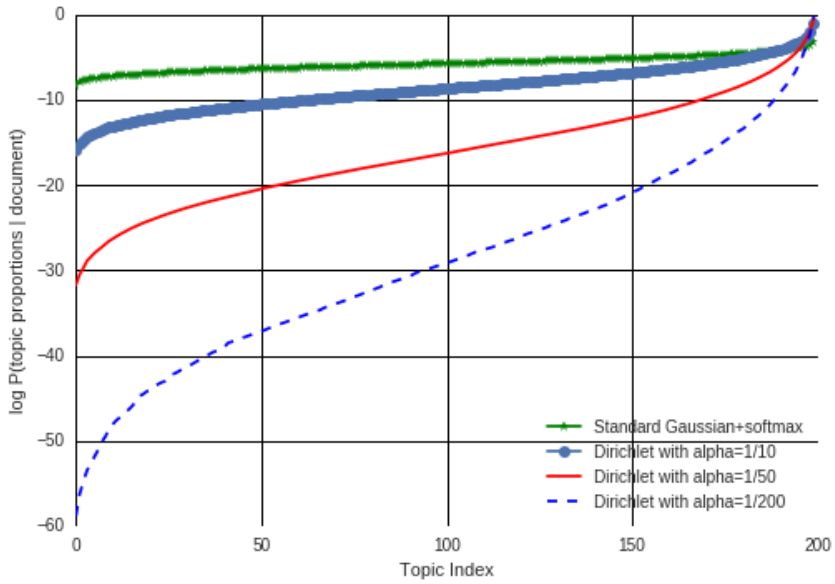

图 1：先前对 θ 的假设对神经主题模型中 θ 稀疏度的影响。

	狄氏 +高学习率	狄氏 +低学习率	高斯先验 +高学习率	高斯先验 +低学习率
主题连贯	**0.24**	0.016	0.08	0.08

表 5：20 个新闻组上 k = 50 时 PRODLDA 的先验和优化策略的不同选择的平均主题连贯性

推论网络架构如图 2 在附录中。

7 讨论与未来工作

我们提供了我们已知的第一个有效的潜在 Dirichlet 分配 AEVB 推理算法。尽管此组合原则上看似简单，但实际上由于 Dirichlet 先验问题以及部件坍塌问题，这种方法很难训练。通过解决这两个问题，我们提出了一种用于主题模型的黑匣子推理方法，其显着优势是神经网络无需计算任何变型优化就可以计算新文档的主题比例。作为优势的例证

模型	主题
普达尔达	主板 MEG 打印机 Quadra HD Windows 处理器 VGA MHZ 连接器 亚美尼亚种族灭绝土耳其人土耳其穆斯林穆斯林大屠杀土耳其亚美尼亚亚 美尼亚希腊电压电压输出插座电路电缆接线电线面板电机安装 赛季 NHL 球队曲棍球季后赛冰球联赛传单防守球员 以色列以色列黎巴嫩阿拉伯阿拉伯阿拉伯平民领土巴勒斯坦民兵
LDA NVLDA	db 文件输出程序行输入写入位 int 返回 驱动器磁盘获取卡 scsi 使用硬控制器比赛玩赢一 年球员获得认为良好使得使用法律状态健康文件枪 公共发行控制火器人们说一个认为生活使知道神人 看到
LDA DMFV I	撰写文章 dod 骑权去得到夜间经销商喜欢 枪法使用毒品犯罪政府法院刑事火器控制月球传单击剑飞船 力量我们存在上帝去刻薄 stephanopoulos 加密航天器成熟 rsa 密码土星违反月球加密文件程序 可用服务器版本包括软件输入 ftp 使用
LDA 倒塌的吉布斯	获得右背光侧，如看时间一 列表邮件发送发布匿名互联网文件信息用户消息谢谢请知道任 何人帮助看欣赏得到需要电子邮件耶稣教会神法律说基督教基 督一天来 自行车道奇骑狗摩托车写文章宝马头盔得到
NVDM	光死烧伤母亲体内的生命 保险药物不同体育朋友银行业主温哥华购买祷告输入包接口输出 磁带报价组件通道级模型价格四方曲棍球插槽圣季后赛约瑟夫交 易市场经销商 基督教教会网关天主教基督教同性恋复活调制解调器鼠标星期日

表 6：从所有模型中随机选择的五个主题。

1. 写文章得到像任何人一样的感谢，请知道看一个 **2. 文章写一个请喜欢任何人都知道要得到** **3. 写文章谢谢任何人都喜欢看一下** **4. 文章写一个像知道谢谢任何人都需要** **5. 文章写谢谢，请成为任何人一次**

表 7：当组件崩溃时，VAE-LDA 无法学习任何有意义的主题。该表显示了从无 BN 和高动量训练的情况下训练 VAE-LDA 模型起的五个随机采样的主题（本质上是彼此的细微变化）。

黑盒推理技术，我们提出了一个新的主题模型 ProdLDA，该主题模型比 LDA 的主题要好得多，而 AVITM for LDA 只需更改一行代码。我们的结果表明，AVITM 推理已准备好与均值场并列，而 Gibbs 崩溃了，成为主题模型的主要推理方法之一。未来的工作可能包括扩展我们的推理方法以处理动态和相关主题模型。

确认

感谢 Andriy Mnih，Chris Dyer，Chris Russell，David Blei，Hannah Wallach，Max Welling，Mirella Lapata 和 Yishu Miao 的宝贵意见，讨论和反馈。

引用

大卫●布雷概率主题模型。ACM 通讯，55（4）：77-84，2012 年。

David M. Blei 和 John D. Lafferty。相关主题模型。神经信息处理系统进展，2006 年。

David M. Blei 和 John D. Lafferty。相关的科学主题模型。应用统计年鉴，1（1）：17-35，2007 年。

David M Blei，Andrew Y Ng 和 Michael I Jordan。潜在狄利克雷分配。机器学习研究杂志，2003 年 3 月 1 日：993–1022。

Michael Collins, Sanjoy Dasgupta 和 Robert E Schapire。主成分分析到指数族的概括。《神经信息处理系统的进展》，第 13 卷，第 23 页，2001 年。

Peter Dayan, Geoffrey E Hinton, Radford M Neal 和 Richard S Zemel。亥姆霍兹机器。*神*经计算，7（5）：889-904，1995 年。

詹姆斯·M·迪基。多种超几何功能：概率解释和统计用途。美国统计协会杂志，78（383）：628-637，1983 年。

Laurent Dinh 和 Vincent Dumoulin。 训练神经贝叶斯网络。 http://www.IOM.UnTural.Ca//~BEGIOY/CIVAR/NCAP2014 暑期学校/幻灯片/ 劳伦斯迪尼希法法罗，2016 年 8 月。

李飞飞和彼得罗·佩罗纳。用于学习自然场景类别的贝叶斯分层模型。在 IEEE 计算机协会计算机视觉和模式识别会议（CVPR'05）中，第 2 卷，第 524–531 页。IEEE，2005 年。

Thomas L Griffiths 和 Mark Steyvers。寻找科学课题。美国国家科学院院刊，101（增刊 1）：5228-5235，2004 年。

Philipp Hennig, David H Stern, Ralf Herbrich 和 Thore Graepel。内核主题模型。在 AISTATS，第 511–519 页，2012 年。

杰弗里·欣顿（Geoffrey E Hinton）。通过最大程度地减少对比差异来培训专家的产品。神经计算，14（8）：1771-1800，2002 年。

Geoffrey E Hinton 和 Ruslan R Salakhutdinov。复制的 softmax：无向主题模型。在 *《神经信息处理系统的进展》，第 1607-1614 页，2009 年。*

Matthew Hoffman, Francis R Bach 和 David M Blei。在线学习潜在的狄利克雷分配。《神经信息处理系统的进展》，第 856–864 页，2010 年。

托*马斯·霍夫曼。概率潜在语义索引。在第 22 届国际 ACM SIGIR 信息检索研究与开发会议论文集中，第 50-57 页。ACM，1999 年。*

谢尔盖·艾菲（Sergey Ioffe）和克里斯蒂安●塞格迪（Christian Szegedy）。批量归一化：通过减少内部协变量偏移来加速深度网络训练。第 448–456 页，2015 年。

Diederik Kingma 和 Jimmy Ba。亚当：一种随机优化方法。第三届学习代表国际会议（ICLR），2015 年。

Diederik P Kingma 和 Max Welling。自动编码可变贝叶斯。国际学习代表大会（ICLR），班夫，2014 年。

Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman 和 David M Blei。自动微分变分推理。arXiv 预印本 arXiv：1603.00788，2016。

雨果·拉罗谢尔（Hugo Larochelle）和斯坦尼斯拉斯●劳利（Stanislas Lauly）。神经自回归主题模型。《神经信息处理系统的进展》，第 2708–2716 页，2012 年。

刘 Han 汉，大卫●纽曼和蒂莫西●鲍德温。机器阅读茶叶：自动评估主题一致性和主题模型质量。在 EACL 中，第 530–539 页，2014 年。

David JC MacKay。选择拉普拉斯近似的基础。机器学习，33（1）：77-86，1998。

安德鲁·麦卡勒姆（Andrew McCallum）。Mallet：用于语言工具包的机器学习。http://mallet.cs. 乌梅,2002.

苗逸树，雷宇和菲尔·布卢森（Phil Blunsom）。用于文本处理的神经变分推理。第 1727 至 1736 页，2016 年。

大卫●米诺（David Mimno）。重建庞贝家庭。在主题模型应用研讨会上，NIPS，2009 年。

Andriy Mnih 和 Karol Gregor。信念网络中的神经变异推理和学习。第 1791-1799 页，2014 年。

David Newman，Jey Han Lau，Karl Grieser 和 Timothy Baldwin。自动评估主题的连贯性。在人类语言技术中：计算语言学协会北美分会 2010 年年会，第 100–108 页。计算语言学协会，2010 年。

Rajesh Ranganath，Sean Gerrish 和 David M Blei。黑匣子变异推理。在 AISTATS，第 814–822 页，2014 年。

Danilo Jimenez Rezende，Shakir Mohamed 和 Daan Wierstra。深度生成模型中的随机反向传播和近似推断。第 1278-1286 页，2014 年。

Simon Rogers，Mark Girolami，Colin Campbell 和 Rainer Breitling。cdna 芯片数据集的潜在过程分解。IEEE / ACM 关于计算生物学和生物信息学的交易（TCBB），2（2）：143-156，2005 年。

Hanna Wallach，David Mimno 和 Andrew McCallum。重新思考 LDA：为什么先验很重要。在*尼*普斯，2009 年。

Max Welling，Michal Rosen-Zvi 和 Geoffrey E Hinton。指数家庭和声及其在信息检索中的应用。《神经信息处理系统的发展》，第 4 卷，第 1481–1488 页，2004 年。

罗纳德·威廉姆斯。用于连接主义强化学习的简单统计梯度跟踪算法。机器学习，8（3-4）：229-256，1992 年。

A 网络架构

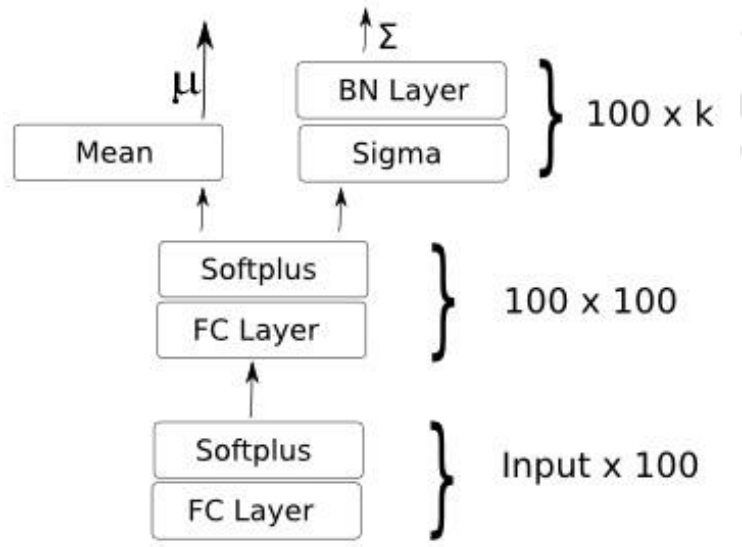

图 2：实验中使用的推理网络架构。