arXiv:2301.00537v1 [stat.ML] 2 Jan 2023

后验坍塌与潜变量不可辨识性

Yixin Wang
密歇根大学
yixinw@umich.edu

David M. Blei哥伦比亚大学
david.blei@columbia.edu

John P. Cunningham哥伦比亚大学 j
pc2181@columbia.edu

摘要

变分自编码器通过假设低维潜变量，并将其映射到由神经网络参数化的灵活分布中，来对高维数据建模。不幸的是，变分自编码器常常遭遇后验坍塌：潜变量的后验分布与其先验分布相同，导致变分自编码器无法生成有意义的表征。现有关于后验坍塌的研究往往将其归因于神经网络的使用，或变分近似带来的优化问题。本文将后验坍塌视为潜变量不可辨识性问题。我们证明，当且仅当生成模型中的潜变量不可辨识时，后验才会坍塌。这意味着，后验坍塌并非灵活分布或近似推断所特有的现象。相反，即使在采用精确推断的经典概率模型中，这种现象也可能发生，我们对此亦作了展示。基于这些结果，我们提出一类潜变量可辨识变分自编码器，即在不牺牲灵活性的前提下确保可辨识性的深度生成模型。该模型通过利用双射布雷尼耶映射，并使用输入凸神经网络对其参数化，解决潜变量不可辨识性问题，无需特殊的变分推断目标或优化技巧。在合成数据集和真实数据集上，潜变量可辨识变分自编码器在缓解后验坍塌以及提供有意义的数据表征方面均优于现有方法。

1 引言

变分自编码器（VAE）是用于高维数据的强大生成模型 [28, 46]。其核心思想是将概率建模的推断原理与神经网络的灵活性相结合。在变分自编码器中，每个数据点都由从先验分布中抽取的低维潜变量独立生成，然后映射到由神经网络参数化的灵活分布。

遗憾的是，变分自编码器经常遭受后验坍塌，这是一种重要且受到广泛研究的现象，即潜变量的后验分布等于先验分布 [6, 8, 38, 62]。这一现象也被称为潜变量坍塌、KL 消失和过度剪枝。后验坍塌使变分自编码器无法产生有意义的表征，因为其每个数据点对应的潜变量都具有完全相同的后验分布。

后验坍塌常见于生成模型高度灵活的变分自编码器，这引发了一种普遍猜测：后验坍塌的发生是因为变分自编码器的生成模型中包含灵活的神经网络 [11], , 或者是因为它使用了变分推断 [59]。基于这些假设，许多旨在缓解后验坍塌的策略因此着重于修改变分推断目标（例如 [44]），设计特殊的优化方案

第35届神经信息处理系统大会（NeurIPS 2021）。

用于变分自编码器中变分推断的优化方案（例如 [18, 25, 32]），或限制生成模型的容量（例如 [6, 16, 60]。）

本文将后验坍塌视为潜变量不可辨识性问题。我们证明，当且仅当生成模型中的潜变量不可辨识时，才会发生后验坍塌；通俗地说，这意味着似然函数不依赖于潜变量 [40, 42, 56]。下文将借助近期关于贝叶斯不可辨识性的研究结果，正式证明这种等价关系 [40, 42, 43, 49, 58]。

更广泛地说，后验坍塌与潜变量不可辨识性之间的关系意味着，后验坍塌并非神经网络或变分推断所特有的现象。事实上，在采用精确推断方法拟合的经典概率模型中也可能出现这种现象，例如高斯混合模型和概率主成分分析（PPCA）。这种关系也为理解现有的后验坍塌规避方法提供了新的视角，例如 Delta-VAE [44] 或 β-VAE [19]。这些方法通过启发式调整嵌入模型参数优化中的近似推断过程来发挥作用。尽管这些方法最初旨在修补变分目标，但本文的结果表明，这些调整之所以有用，是因为它们有助于避开潜变量不可辨识的参数，从而避免后验坍塌。

后验坍塌与不可识别性之间的关系指出了解决这一问题的直接方法：我们必须使潜变量可识别。为此，我们提出潜变量可识别变分自编码器，这是一类既与经典变分自编码器一样灵活、又具有可识别性的变分自编码器。潜变量可识别变分自编码器通过利用 Brenier映射 [36,39]并使用输入凸神经网络 [2, 35]对其进行参数化，解决了潜变量不可辨识性。可识别变分自编码器的推断采用标准的变分推断目标，无需特殊修改或优化技巧。在合成数据集和真实数据集上，我们表明，可识别变分自编码器能够缓解后验坍塌，同时不牺牲对数据的保真度。

相关工作。 现有避免后验坍塌的方法通常会修改变分推断目标、为变分自编码器设计新的初始化或优化方案，或者在每个数据点与其潜变量之间添加神经网络连接 [1, 3, 6, 8, 12, 15, 16, 17, 18, 21, 25, 27, 32,34, 38, 44, 50, 51, 52, 55, 61, 62, 63]。近期也有若干论文尝试解释后验坍塌。Chen 等人 [8]解释了不精确的变分近似如何导致变分自编码器中的编码效率低下，而这可能因某种信息偏好而导致后验坍塌。Dai 等人 [11] 认为，在使用深度神经网络训练变分自编码器时，后验坍塌部分可归因于局部最优。Lucas 等人 [33] 指出，后验坍塌并非变分推断训练目标所特有；即使没有变分近似，概率主成分分析的对数边际似然也存在可能导致后验坍塌的糟糕局部最优。Yacoby 等人 [59]讨论了当生成模型参数不可识别时，变分近似如何选择出不理想的生成模型。与这些工作不同，我们仅将后验坍塌视为潜变量不可辨识性问题，而非优化、变分近似或神经网络本身的问题。我们利用这一结果提出可识别变分自编码器，以直接避免后验坍塌。

在变分自编码器之外，潜变量在概率模型中的可识别性早已在统计学文献中得到研究 [40, 42, 42, 43, 49, 56, 58]。最近，Betancourt [5] 研究了潜变量可识别性对高斯混合模型贝叶斯计算的影响。Khemakhem 等人[23, 24]提出借助辅助数据来解决深度生成模型中的不可识别性。Kumar 和 Poole [29] 研究了变分族如何帮助解决变分自编码器的不可识别性。这些研究针对不同目标探讨可识别性问题：它们针对变分自编码器的不同子集提出可识别性条件，旨在恢复数据的真实因果因素，并改善解耦或分布外泛化。与这些论文相关，我们展示了后验坍塌这一现象，说明可识别性这一概念虽属经典，却能成为现代概率建模中的有力工具。考察可识别性有助于找到解决后验坍塌的新方法。

贡献。 我们证明，当且仅当生成模型中的潜变量不可识别时，才会发生后验坍塌。随后，我们提出潜变量可识别变分自编码器，这类变分自编码器与经典变分自编码器一样灵活，同时其潜变量在理论上可识别。在合成数据集和真实数据集上的实验表明，潜变量可识别变分自编码器能够缓解后验坍塌，且无需修改变分自编码器目标函数或采用特殊的优化技巧。

2 后验坍塌与潜变量不可识别性

考虑一个数据集 $\boldsymbol{x} = (x_1,...,x_n)$；每个数据点的维度为$m$。假设有$n$ 个潜变量$\boldsymbol{z} = (z_1,...,z_n)$，变分自编码器（VAE）假定每个数据点$x_i$ 由一个K维潜变量z_i生成：

$$z_i \sim p(z_i), \qquad x_i \,|\, z_i \sim p(x_i \,|\, z_i\,; \theta) = \mathrm{EF}(x_i \,|\, f_\theta(z_i)), \tag{1}$$

其中 x_i 服从参数为 $f_\theta(z_i)$的指数族分布；f_θ 对条件似然进行参数化。在深度生成模型中，f_θ 是一个参数化的神经网络。高斯混合模型 [45]和概率主成分分析 [10, 47, 48, 54] 也是式 1 的特例。

为了拟合模型，变分自编码器通过最大化对数边际似然的变分近似来优化参数 θ^。找到最优的 θ 后，我们可以使用近似后验 $q_{\hat{\phi}}(z \,|\, x)$ 以及变分参数 $\hat{\phi}$，或其期望 $\mathbb{E}_{q_{\hat{\phi}}(z \,|\, x)}[z \,|\, x]$，来构造数据的表示。

请注意，这里我们暂不考虑计算方面的问题，并讨论变分近似精确的理想情形。这样的选择是合理的：如果精确后验分布发生后验坍塌，那么近似后验也会如此（变分近似无法“解坍塌”已坍塌的后验分布）。不过，我们也指出，实践中确实存在仅由变分推断导致后验坍塌的情形。一个显著的例子是近似族过于受限：此时，精确后验可能不会坍塌，而近似后验却会坍塌。

2.1 后验坍塌 ⇔ 潜变量不可辨识性

我们首先定义后验坍塌和潜变量不可辨识性，然后证明二者之间的联系。

定义 1 (后验坍塌 [6, 8, 38, 62])。 给定一个概率模型 $p(\boldsymbol{x}, \boldsymbol{z}; \theta)$、一个参数值 $\theta = \hat{\theta}$，以及一个数据集 $\boldsymbol{x} = (x_1,...,x_n)$，如果潜变量的后验分布

$$p(\boldsymbol{z} \,|\, \boldsymbol{x}; \hat{\theta}) = p(\boldsymbol{z}). \tag{2}$$

后验坍塌现象可能出现在各种概率模型中，也可能涉及不同的潜变量。当概率模型是变分自编码器时，它只有局部潜变量$\hat{\boldsymbol{z}} = (z_1,...,z_n)$，且公式 2 等价于后验坍塌的常见定义：对所有$i$ [12, 17, 33, 44]，$p(z_i \,|\, x_i\,; \hat{\theta}) = p(z_i)$ 。高斯混合模型中也观察到了后验坍塌现象 [5]；当模型中的混合分量数量多于数据生成过程中的混合分量数量时，潜在混合权重的后验分布与其先验分布相似。无论模型为何种类型，后验坍塌都会导致潜变量无法为数据集提供有意义的概括。

定义 2 （潜变量不可辨识性 [42, 56]）。 给定似然函数 $p(\boldsymbol{x} \,|\, \boldsymbol{z}\,; \theta)$、参数值 $\theta = \hat{\theta}$以及数据集 $\boldsymbol{x} = (x_1,..., x_n)$，若满足以下条件，则潜变量 $\boldsymbol{z}$ 不可辨识：

$$p(\boldsymbol{x} \,|\, \boldsymbol{z} = \tilde{\boldsymbol{z}}'\,; \hat{\theta}) = p(\boldsymbol{x} \,|\, \boldsymbol{z} = \tilde{\boldsymbol{z}}\,; \hat{\theta}) \qquad \forall \tilde{\boldsymbol{z}}', \tilde{\boldsymbol{z}} \in \mathcal{Z}, \tag{3}$$

其中 $\mathcal{Z}$ 表示 z 的定义域，而 $\tilde{}z'$、$\tilde{}z$ 表示潜变量 z 可以取的任意两个值。因此，对于 z 上的任意先验分布 $p(z)$，条件似然等于边缘分布 $p(x \,|\, z = \tilde{}z\,; \hat{}\theta) = \int p(x \,|\, z\,; \hat{}\theta) p(z) dz = p(x; \hat{}\theta) \; \forall \tilde{}z \in \mathcal{Z}$。

定义 2 指出，当数据集 $\boldsymbol{x}$ 的似然不依赖于 $\boldsymbol{z}$时，潜变量 $\boldsymbol{z}$ 不可辨识。这也称为实际不可辨识性 [42, 56]，并且与给定 θ [40,42, 43, 49, 58] 时$\hat{\boldsymbol{z}}$ 条件不可辨识（或条件无信息性）的定义密切相关。要确保潜变量可识别性，只需确保似然$p(\boldsymbol{x} \,|\, \boldsymbol{z}, \theta)$ 对于所有 θ 都是 $\boldsymbol{z}$ 的单射（亦称一对一）函数。若此条件成立，则

$$\tilde{\boldsymbol{z}}' \neq \tilde{\boldsymbol{z}} \qquad \Rightarrow \qquad p(\boldsymbol{x} \,|\, \boldsymbol{z} = \tilde{\boldsymbol{z}}'\,; \hat{\theta}) \neq p(\boldsymbol{x} \,|\, \boldsymbol{z} = \tilde{\boldsymbol{z}}\,; \hat{\theta}). \tag{4}$$

请注意，潜变量不可辨识性只要求对于给定数据集$\boldsymbol{x}$ 和^参数值 θ，式 3 成立。因此，对于某个数据集，潜变量在模型中可能是可辨识的，但对于另一个数据集则不可辨识；在某个参数值下可能可辨识，但在另一个参数值下则不可辨识。示例见附录 A。

潜变量可识别性（定义 2） [42, 56]不同于模型可识别性 [41], , 后者是一种相关概念，也被认为是导致后验坍塌的因素之一 [59]。潜变量可识别性的要求较弱：它只要求潜变量$\boldsymbol{z}$ 在ˆ某个特定参数值 $\theta = \theta$下可辨识，而模型可识别性则要求$\boldsymbol{z}$ 和 θ 均可辨识。

We 现在证明后验坍塌与潜变量不可辨识性等价 .

定理1 (潜变量不可辨识性 $\Leftrightarrow$ 后验坍塌)。考虑一个概率模型 ˆ$p(\boldsymbol{x}, \boldsymbol{z}; \theta)$、一个数据集 $\boldsymbol{x}$以及参数值 $\theta = \theta$。当且仅当潜变量 $\boldsymbol{z}$ 的后验分布发生坍塌时，局部潜变量ˆ在 θ 处不可辨识，$\boldsymbol{p}(\boldsymbol{z}|\boldsymbol{x}) = p(\boldsymbol{z})$。

证明。 为证明不可辨识性意味着后验坍塌，注意根据贝叶斯法则，

$$p(\boldsymbol{z}|\boldsymbol{x}; \hat{\theta}) \propto p(\boldsymbol{z})p(\boldsymbol{x}|\boldsymbol{z}; \hat{\theta}) = p(\boldsymbol{z})p(\boldsymbol{x}; \hat{\theta}) \propto p(\boldsymbol{z}), \tag{5}$$

其中，中间的等式源于潜变量不可辨识性的定义。由于两者都是密度，这意味着 ˆp(z |x; θ) =p(z)。为证明后验坍塌意味着潜变量不可辨识性，我们再次运用贝叶斯法则。后验坍塌意味着 p(z) =p(z |x ; θ) $\propto$ p(z)· p(x|z ; ˆθ)，这进一步意味着 p(x|z ; ˆθ) 关于 z 为常数。如果 p(x|z ; ˆθ) 非平凡地依赖于 z，那么作为 z 的函数，p(z) 必须不同于 p(z)p(x|z ; θ)。 □

定理1的证明很直接，但定理1有一项重要推论。它表明，后验坍塌问题主要源于模型和数据，而非ˆ推断或优化。如果变分自编码器的最大似然参数 θ 使潜变量z 不可辨识，那么我们就会观察到后验坍塌。定理1还阐明了为何在拟合变分自编码器时，后验分布可能从未坍塌变为坍塌（或反过来）。拟合变分自编码器时，某些参数迭代值可能导致后验坍塌，其他参数迭代值则可能不会。

定理1指出了现有方法为何能够缓解后验坍塌。考虑 β- 变分自编码器 [19], 变分自编码器滞后编码器 [18], 以及半摊销变分自编码器 [25]。尽管这些方法的出发点各不相同，但它们会修改变分自编码器的优化目标或算法，以避免取到使潜变量不可辨识的参数值 θ 。由此得到的后验分布可能不会坍塌，不过这些算法的最优参数已不再是最大似然估计的近似值。

定理1也有助于我们理解实践中观察到的后验坍塌，其表现为后验分布近似（而非完全）等于先验分布，即 ˆp(z |x; θ) $\approx$p(z)。在若干关于变分自编码器的实证研究中（例如 [12, 18, 25]），我们观察到先验分布与后验分布之间的 Kullback-Leibler（KL）散度接近于零，但并不完全为零；这一特性源于似然 p(x|z) 对潜变量 z 几乎为常数。在这些情况下，定理1指出，潜变量几乎不可辨识，即 p(x |˜z′) $\approx$p(x|˜z),$\forall$˜z, ˜z′，因此式 2 近似成立。

2.2 潜变量不可辨识性与后验坍塌的示例

我们通过三个示例来说明定理1。这里讨论高斯混合变分自编码器（GMVAE）的示例。概率主成分分析（PPCA）和高斯混合模型（GMM）的内容见附录 A。

高斯混合变分自编码器 [13, 51] 是以下模型：

$$p(z_i) = \text{Categorical}(1/K), \quad p(w_i|z_i; \mu, \Sigma) = \mathcal{N}(\mu_{z_i}, \Sigma_{z_i}), \quad p(x_i|w_i; f, \sigma) = \mathcal{N}(f(w_i), \sigma^2 \cdot I_m),$$

其中， μ_k是 d 维的， Σ_k 是 d ×d 维的，参数为 $\theta = (\mu, \Sigma, \text{f}, \sigma^2)$。假设函数 f 完全灵活；因此，f ($\text{w}_i$) 可以捕捉数据的任意分布。我们感兴趣的潜变量是分类变量 z = (z_1,..., z_n)。如果其后验分布发生坍塌，那么对于所有 k = 1,..., K，p(z_i =k|x) = 1/K。

考虑将一个 K = 2 的高斯混合变分自编码器模型拟合到一个包含 5,000 个样本的数据集。该数据集取自一个同样具有 K = 2 个明显分离簇的高斯混合变分自编码器；不存在模型设定错误。通常通过优化最大对数边际似然 θ = $\text{argmax}_\theta \log p(\boldsymbol{x}|\theta)$ 来拟合高斯混合变分自编码器 ˆ。请注意，使该函数达到全局最优值的 θ 可能不止一个。

我们关注两个似然最大化参数。一个具有潜变量可识别性，且z_i 的后验分布不会坍塌。另一个不具备可识别性；其后验分布会坍塌。

ˆ1. 第一个似然最大化参数 θ_1 是真实参数；拟合得到的K 个簇的分布对应于K 个数据生成簇。给定该参数，潜变量z_i 是可辨识的，因为K 个数据生成簇各不相同；不同的簇成员归属z_i 必然会产生不同的似然$p(x_i \mid z_i; \hat{\theta}_1)$。$z_i$ 的后验分布不会坍塌。

ˆ 2. 然而，在第二个似然最大化参数 θ_2 下，所有 K 个拟合簇共享相同的分布，且每个分布都等于数据的边缘分布。具体而言，对于所有 k，$(\mu^* \mathrm{k}, \Sigma^* \mathrm{k}) = (0, \mathrm{I}_d)$，并且每个拟合簇都是 K 个原始数据生成簇的混合，即边缘分布。在此参数值下，模型仍然能够完全捕捉数据的混合分布。然而，所有 K 个混合分量都相同，因此潜变量 z_i 不可识别；不同的簇成员身份 z_i 不会导致不同的似然 $\mathrm{p}(\mathrm{x}_i \mid \mathrm{z}_i; \theta_2)$，因此 z_i 的后验分布会坍塌。图 1a 展示了该高斯混合变分自编码器（不可识别）对风车数据 [22] 的拟合。在第 3 节中，我们将构建一种潜变量可识别的变分自编码器（LIDVAE），以避免这种坍塌。

潜变量可识别性取决于模型和真实的数据生成分布。考虑使用$K = 2$ 拟合同一个高斯混合变分自编码器，但改用另一个包含 5,000 个样本的数据集；这些样本来自一个只有一个簇的高斯混合变分自编码器。（这里存在模型设定错误。）一个使参数达到最大值的取值ˆ θ_3 是：两个拟合得到的簇都对应于真实的数据生成ˆ簇。尽管这个参数取值与上文第一个极大值 θ_1 相似——二者都对应于真实的数据生成簇——但这个数据集导致了潜变量可识别性的另一种情况。两个拟合得到的簇相同，因此不同的簇成员归属不会导致ˆ不同的$p(x_i \mid z_i; \theta_3)$似然。潜变量$z_i$ 不可辨识，其后验分布会发生坍塌。

要点。 </style}本节中的高斯混合变分自编码器示例（以及附录 A 中的概率主成分分析和高斯混合模型示例）展示了潜变量在模型中不可辨识并发生后验坍塌的不同方式。这些示例表明，即使是真实后验分布——无需使用变分推断——在不可辨识的模型中也可能发生坍塌。它们还说明，潜变量是否可辨识可能同时取决于模型和数据。后验坍塌是模型和数据本身固有的问题，而非神经网络或变分推断所特有的问题。
定理1中后验坍塌与潜变量不可辨识性之间的等价关系也意味着，为了缓解后验坍塌，我们应尝试解决潜变量不可辨识性。在下一节中，我们将构建这样一类潜变量可识别变分自编码器。

3 通过布雷尼耶映射构建潜变量可识别变分自编码器

现在，我们构建潜变量可识别变分自编码器。这类变分自编码器能够保证其潜变量可辨识，因此后验分布不会坍塌。

3.1 潜变量可识别变分自编码器

构建潜变量可识别变分自编码器时，我们依据一项关键观察：为了保证潜变量可识别性，只需使似然函数 $P(x_i \mid z_i; \theta)$ 对 θ 的所有取值均为单射。若似然函数是单射，那么对于任意 θ，z_i 的每个取值都会对应不同的分布 $P(x_i \mid z_i; \theta)$。特别地，对于任意优化后的 ˆθ，这一事实都成立，因此无论数据如何，潜变量 z_i都必须是可辨识的。根据定理1，其后验分布不会坍塌。因此，构建潜变量可识别变分自编码器，就相当于为变分自编码器构建一个单射似然函数。该构建基于若干线性和非线性单射函数构件，再将它们组合成一个从 $\mathcal{Z}_d$ 到 $\mathcal{X}_m$的单射似然$p(x_i \mid z_i;\ \theta)$ 映射，其中$\mathcal{Z}$和 $\mathcal{X}$ 表示z_i 和x_i 可以取值的集合。例如，如果x_i 是一个 m 维二元向量，那么 $\mathcal{X} = \{0,1\}^m$；如果z_i 是一个K维实值向量，那么 $\mathcal{Z} = \mathbb{R}^d$。

LIDVAE 的构件：单射函数。对于从 $\mathbb{R}^{d_1}$ 到 $\mathbb{R}^{d_2}$ $(d_2 \geq d_1)$的线性映射，我们考虑通过一个$d_1 \times d_2$维矩阵 β进行矩阵乘法。对于一个d_1维变量z，当矩阵 $\beta^\top$ 具有满列秩 [53]时，左乘该矩阵是单射。例如，对角线上元素全为 1、其他元素全为 0 的矩阵具有满列秩。

对于非线性单射函数，我们关注Brenier映射 [4, 37]。一个d维Brenier映射是从 $\mathbb{R}^d$ 到 $\mathbb{R}$的凸函数梯度。也就是说，Brenier映射满足$g = \nabla T$ for

某个凸函数 T： $\mathbb{R}^d \to \mathbb{R}$。布雷尼耶映射也称为单调传输映射。它们保证是双射 [4, 37]，因为其导数是凸T的海森矩阵，必须是半正定的，且行列式非负 [4]。

要构建一个使用Brenier映射的变分自编码器，我们需要对Brenier映射进行神经网络参数化。由于Brenier映射是凸函数的梯度，我们从凸函数的神经网络参数化开始，即输入凸神经网络（ICNN） [2, 35]。这种对凸函数的参数化将使Brenier映射能够表示为ICNN的梯度。

一个L层的输入凸神经网络（ICNN）是一个从 $\mathbb{R}^d$ 映射到 $\mathbb{R}$的神经网络。给定输入$u \in \mathbb{R}^d$，其第l层为

$$\boldsymbol{z}_0 = \mathbf{u}, \qquad \boldsymbol{z}_{l+1} = h_l(\mathbf{W}_l \boldsymbol{z}_l + \mathbf{A}_l \mathbf{u} + \mathbf{b}_l), \qquad (l = 0, \ldots, L-1), \tag{6}$$

其中，最后一层z_L 必须是标量，$\{W_l\}$是满足$W_0 = \mathbf{0}$的非负权重矩阵。函数$\{h_l : \mathbb{R} \to \mathbb{R}\}$是第l层逐元素凸且逐项非递减的激活函数；它们逐元素应用于向量$(W_l z_l + A_l u + b_l)$。$h_0 : \mathbb{R} \to \mathbb{R}$的一种常见选择是带泄漏的ReLU的平方，即$h_0(x) = (\max(\alpha \cdot x, x))^2$，其中 $\alpha = 0.2$；其余的h_l均设为带泄漏的ReLU，$h_l(x) = \max(\alpha \cdot x, x)$。这种神经网络被称为“输入凸”，因为可以保证它是一个凸函数。

输入凸神经网络可以在紧致域上以 sup norm 逼近任意凸函数（Chen 等人的定理 1 [9]）。给定凸函数的神经网络参数化形式，我们可以将 Brenier 映射 $g_\theta(\cdot)$ 参数化为其关于输入的梯度 $g_\theta(u) = \partial z_L / \partial u$.这种 Brenier 映射的神经网络参数化形式是紧致域上所有 Brenier 映射的通用逼近器，因为输入凸神经网络是凸函数的通用逼近器 [9]。

潜变量可识别变分自编码器（LIDVAE）。我们通过将两个双射布雷尼耶映射与一次单射矩阵乘法复合，为 LIDVAE 构造单射似然。由于单射映射与双射映射的复合必定是单射，因此所得复合映射必定是单射。设 $g_{1,\theta} : \mathbb{R}^K \to \mathbb{R}^K$ 和 $g_{2,\theta} : \mathbb{R}^D \to \mathbb{R}^D$ 是两个 Brenier 映射，并且 β 是一个 $K \times D$ 维矩阵$(D \geq K)$，其主对角线元素均为 1，其他元素均为 0。矩阵$\beta^\top$ 满列秩，因此乘以 $\beta^\top$ 是单射。于是，复合映射 $g_{2,\theta}(\beta^\top g_{1,\theta}(\cdot))$必定是从低维空间 $\mathbb{R}^K$ 到高维空间 $\mathbb{R}^D$的单射函数。

定义 3（通过 Brenier 映射实现的潜变量可识别变分自编码器（LIDVAE））。通过 *Brenier* 映射实现的LIDVAE 按如下方式生成一个D维数据点x_i，$\in \{1, \ldots, n\}$：

$$z_i \sim p(z_i), \qquad x_i \mid z_i \sim \mathrm{EF}(x_i \mid g_{2,\theta}(\beta^\top g_{1,\theta}(z_i))), \tag{7}$$

其中，EF 表示指数族分布；z_i 是一个K维潜变量，可以是离散的，也可以是连续的。模型的参数为 $\theta = (g_{1,\theta}, g_{2,\theta})$，其中$g_{1,\theta} : \mathbb{R}^K \to \mathbb{R}^K$和$g_{2,\theta} : \mathbb{R}^D \to \mathbb{R}^D$ 是两个连续布雷尼耶映射。矩阵 β 是一个$K \times D$维矩阵$(D \geq K)$，其主对角线元素均为 *1*，其余元素均为 *0*。

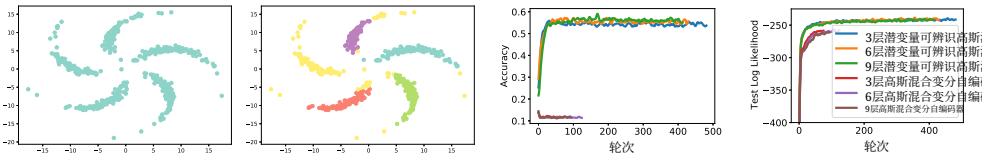
与经典变分自编码器（式 1）相比，LIDVAE（式 7）将函数 $f_\theta : \mathcal{Z}^K \to \mathcal{X}^D$ 替换为单射映射 $g_{2,\theta}(\beta^\top g_{1,\theta}(\cdot))$；该映射由双射布雷尼耶映射 $g_{1,\theta}$、$g_{2,\theta}$ 以及一个满列秩的零一矩阵 $\beta^\top$ 复合而成。由于指数族分布的似然函数是单射，LIDVAE 的似然函数 $p(x_i \mid z_i; \theta) = \mathrm{EF}(g_{2,\theta}(\beta^\top g_{1,\theta}(z_i)))$ 必须是单射。因此，将任意函数 $f_\theta : \mathcal{Z}^K \to \mathcal{X}^D$ 替换为单射映射 $g_{2,\theta}(\beta^\top g_{1,\theta}(\cdot))$，对于确保潜变量 z_i 的可识别性以及避免 LIDVAE 中的后验坍塌至关重要。由于 LIDVAE 中的潜变量 z_i 必须可辨识，其后验分布不会坍塌。

尽管似然是单射的，LIDVAE 与变分自编码器一样灵活；使用布雷尼耶映射和输入凸神经网络并不会限制生成模型的能力。粗略地说，LIDVAE 可以对 $\mathbb{R}^D$ 中的任意分布进行建模，因为布雷尼耶映射能够将 $\mathbb{R}^d$ 中任意给定的非原子分布映射到 $\mathbb{R}^d$ [37] 中的任意其他分布。此外，输入凸神经网络参数化是布雷尼耶映射的通用逼近器 [2]。我们在下述命题中总结 LIDVAE 的关键性质。

命题 2. 潜变量 z_i 在 LIDVAE 中是可辨识的，即对于所有$i \in \{1, \ldots, n\}$，我们有

$$p(x_i \mid z_i = \tilde{z}'; \theta) = p(x_i \mid z_i = \tilde{z}; \theta) \qquad \Rightarrow \qquad \tilde{z}' = \tilde{z}, \qquad \forall \tilde{z}', \tilde{z}, \theta. \tag{8}$$

此外，对于任何由VAE生成的数据分布，都存在一个能够生成相同分布的LIDVAE 。（证明见附录 B。）



(a) 非潜变量可辨识高斯混合变分自编码器 **(b)** 潜变量可辨识高斯混合变分自编码器 **(c)** 准确率 **(d)** 对数似然

图 1： （a）-（b）：经典高斯混合变分自编码器的后验分布 [13, 26, 51] 在风车数据集上拟合时发生了坍塌；所有数据点的潜变量都预测出相同的值。然而，潜变量可辨识高斯混合 VAE（LIDGMVAE）的后验分布并未坍塌，并提供了有意义的表征。（c）-（d）在 Fashion MNIST 上拟合时，潜变量可辨识高斯混合变分自编码器所产生的后验分布比高斯混合变分自编码器的信息量显著更高，同时也取得了更高的测试对数似然。

3.2 LIDVAE 中的推断

在 LIDVAE 中进行推断与经典 VAE 中的推断完全相同，因为这两种 VAE 的差异仅在于参数约束。为了拟合 LIDVAE，我们采用 VAE 的经典摊销推断算法；我们最大化对数边际似然 [28] 的证据下界（ELBO）。

一般而言，LIDVAE 可直接替代变分自编码器。二者的容量相同（命题 2），并采用相同的推断算法，但 LIDVAE 具有可辨识性，且不会遭受后验坍塌。使用 LIDVAE 的代价在于计算方面：生成模型（即解码器）通过神经网络的梯度进行参数化；因此，其优化需要计算神经网络梯度的梯度，这会增加变分自编码器推断的计算复杂度，有时也会给优化带来挑战。使用随机梯度下降拟合经典变分自编码器时，计算复杂度为 $O(k \cdot p)$，其中 k 为迭代次数，p 为参数数量；而拟合潜变量可辨识变分自编码器的计算复杂度可能为 $O(k \cdot p^2)$。

3.3 LIDVAE 的扩展

LIDVAE 的构造揭示了一种使生成模型中的潜变量可辨识的通用策略：用单射非线性映射替代非线性映射。我们可以运用这一策略，使许多其他 VAE 变体中的潜变量也可辨识。下面给出两个例子：混合 VAE 和序列变分自编码器。

混合 VAE（高斯混合变分自编码器是其特例）使用指数族混合对数据建模，再通过灵活的神经网络映射来生成数据。我们利用布雷尼耶映射构建了与之对应的潜变量可识别版本。

例 1 (潜变量可识别混合 VAE（LIDMVAE）)。一个LIDMVAE 生成一个D维数据点x_i，$i \in \{1,\dots,n\}$，其生成过程如下：

$$z_i \sim \text{Categorical}(1/K), \quad w_i \,|\, z_i \sim \text{EF}(w_i \,|\, \beta_1^\top z_i), \quad x_i \,|\, w_i \sim \text{EF}(x_i \,|\, g_{2,\theta}(\beta_2^\top g_{1,\theta}(w_i))), \tag{9}$$

其中，W_i 是一个 K 维 *one-hot* 向量，用于指示簇分配。模型的参数为 $\theta = (g1,\theta, g2,\theta)$，其中函数 $g1,\theta : \mathbb{R}^M \to \mathbb{R}^M$ 和 $g2,\theta : \mathbb{R}^D \to \mathbb{R}^D$ 是两个连续布雷尼耶映射。矩阵 β_1 和 β_2 分别是一个 $K \times M$ 维矩阵（$M \geq K$）和一个 $M \times D$ 维矩阵（$D \geq M$）；两者的主对角线元素均为 1，其余元素均为 0。

LIDMVAE 与经典混合 VAE 的区别在于 $p(x_i \,|\, z_i)$，其中我们将其神经网络映射替换为对应的单射映射，即两个 Brenier映射与矩阵乘法 $g_{2,\theta}(\beta_2^\top g_{1,\theta}(\cdot))$ 的复合。作为一个特例，将示例 1 中的两个指数族都设为高斯分布，即可得到 LIDGMVAE；我们将在第 4 节中使用它对图像进行建模。

接下来，我们推导序列变分自编码器的可辨识对应模型，该模型根据潜变量使用自回归模型对数据进行建模。

	Fashion-MNIST				Omniglot			
	AU	**KL**	**MI**	**LL**	**AU**	**KL**	**互信息**	**LL**
变分自编码器 [28]	0.1	0.2	0.9	-258.8	0.02	0.0	0.1	-862.1
SA-VAE [25]	0.2	0.3	1.3	-252.2	0.1	0.2	1.0	-853.4
滞后变分自编码器 [18]	0.4	0.6	1.6	-248.5	0.5	1.0	3.6	-849.4
β-变分自编码器 [19](β=0.2)	0.6	1.2	2.4	-245.3	0.7	1.4	5.9	-842.6
LIDGMVAE（本研究）	**1.0**	**1.6**	**2.6**	**-242.3**	**1.0**	**1.7**	**7.5**	**-820.3**

	合成				雅虎				Yelp			
	AU	**KL**	**MI**	**对数似然**	**AU**	**KL**	**MI**	**LL**	**AU**	**KL**	**MI**	**LL**
变分自编码器 [28]	0.0	0.0	0.0	-46.5	0.0	0.0	0.0	-519.7	0.0	0.0	0.0	-635.9
SA-VAE [25]	0.4	0.1	0.1	-40.2	0.2	1.0	0.2	-520.2	0.1	1.9	0.2	-631.5
滞后变分自编码器 [18]	0.5	0.1	0.1	-40.0	0.3	1.6	0.4	**-518.6**	0.2	3.6	0.1	**-631.0**
β-变分自编码器 [19](β=0.2)	**1.0**	0.1	0.1	**-39.9**	0.5	4.7	0.9	-524.4	0.3	**10.0**	0.1	-637.3
LIDSVAE	**1.0**	**0.5**	**0.6**	-40.3	**0.8**	**7.2**	**1.1**	-519.5	**0.7**	9.1	**0.9**	-634.2

Tab表 1： 在图像和文本数据集上，LIDVAE 在避免 g
post后验坍塌的同时，实现了与数据相近的拟合优度。

示例2 (潜变量可辨识的序贯 VAE (LIDSVAE))。 一个 LIDSVAE 通过以下方式生成一个 D 维数据点 x_i，$i \in \{1,...,n\}$

$$z_i \sim p(z_i), \qquad x_i \,|\, z_i, x_{<i} \sim \mathrm{EF}(g_{2,\theta}(\beta_2^\top g_{1,\theta}([z_i, f_\theta(x_{<i})]))),$$

其中，$x_{<i} = (x_1,.., x_{i-1})$ 表示 x 在第 i 个维度之前的历史。函数 $f_\theta : \mathscr{X}_{<i} \to \mathbb{R}^H$ 将历史 $X_{<i}$ 映射为一个 H 维向量。最后，$[z_i, f_\theta(x_{<i})]$ 是一个 $(K+H)\times 1$向量，表示向量 $(z_i)_{K\times 1}$ 和 $(f_\theta(x_{<i}))_{H\times 1}$按行堆叠后的结果。

与混合 VAE 类似，LIDSVAE 与序列变分自编码器的区别也仅在于它在$g_{2,\theta}(\beta_2^\top g_{1,\theta}(\cdot))$ 函数中的使用方式，即在 $p(x_i \,|\, z_i, x_{<i})$中使用该函数。我们将在第 4 节使用 LIDSVAE 对文本建模。

4 实证研究

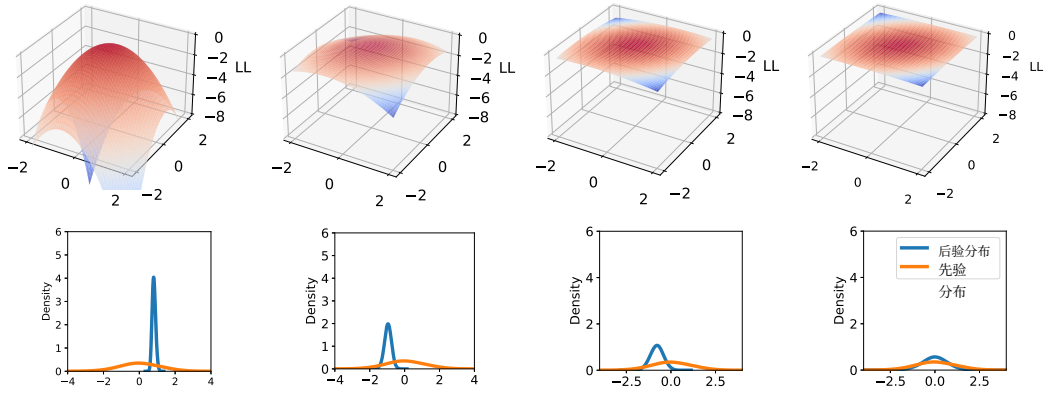
我们在图像和文本数据集上研究 LIDVAE，发现随着生成模型容量的增加，LIDVAE 不会遭受后验坍塌，同时对数据的拟合效果相近。我们进一步研究概率主成分分析（PPCA），说明当似然函数关于潜变量几乎恒定时，即使使用马尔可夫链蒙特卡罗（MCMC），后验分布也会发生坍塌。

4.1 LIDVAE 在图像和文本上的应用

我们采用三项指标来评估后验坍塌：（1）后验分布与先验分布之间的 KL 散度，KL(q(z|x)||p(z))；（2）活跃单元（AU）的百分比：AU $= \sum_{d=1}^{D} \mathbb{1}\{\mathrm{Cov}_{p(\boldsymbol{x})}(\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[z_d]) \geq \epsilon\}$，其中 $z_d = (z_{1d},..., z_{nd})$ 是所有 n 个数据点对应的潜变量 z 的第 d 个维度。计算 AU 时，我们按照 Burda 等人 [7] 的方法，计算所有数据点的后验均值 $(\mathbb{E}[z_{1d}|x_1],...,\mathbb{E}[z_{nd}|x_n])$，并据此计算 $\mathbb{E}[z_{id}|x_i]$ 在各个 i 上的样本方差。阈值 ϵ 设为 0.01 [7]；%AU 的理论最大值为 1；（3）x_i 与 z_i之间的近似互信息（MI），$I(x, z) = \mathbb{E}_{\boldsymbol{x}}[\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log(q(z|x))]] - \mathbb{E}_{\boldsymbol{x}}[\mathbb{E}_{q(\boldsymbol{z}|x)}[\log(q(z))]]$。我们还使用留出测试集 [7]上的对数似然重要性加权估计来评估模型拟合。对于混合 VAE，我们还会评估类别型潜变量相对于真实标签的预测准确率，以量化其信息量。

对比方法。我们将 LIDVAE 与经典变分自编码器 [28]、β-VAE（β=0.2）、半摊销变分自编码器 [25]，和滞后变分自编码器 [18]进行比较。在整个实证研究中，我们使用灵活的变分近似族（图像使用 RealNVP [14]，文本使用长短期记忆网络 [20]）。

结果：图像。我们首先在从风车数据 [22], MNIST、 [31], Fashion MNIST [57], 和 Omniglot [30]中抽取的四个子采样图像数据集上研究 LIDGMVAE。图 1a 和图 1b 展示了 GMVAE 和 LIDGMVAE 对风车数据 [22]的拟合结果。GMVAE 的后验分布



(a) $\sigma = 0.2$ **(b)** $\sigma = 0.5$ **(c)** $\sigma = 1.0$ **(d)** $\sigma = 1.5$

图2： 随着 PPCA 中噪声水平的增加，潜变量变得更接近不可识别，因为似然曲面变得更平坦，后验分布也更容易发生后验坍塌。其似然曲面变得更平坦，后验分布也更接近先验分布。上方面板：PPCA 的似然曲面，以两个潜变量 z_1,z_2为自变量。当 σ 增加时，似然曲面变得更平坦，潜变量 z_1,z_2 也更接近不可识别。下方面板：不同 σ 值下 z_1 的后验分布。当 σ 增加时，后验分布变得更接近先验分布。

潜变量发生坍塌，将所有数据点归因于同一个潜在簇。相比之下，LIDGMVAE 生成了忠实反映聚类结构的分类变量潜变量。图 1 研究了随着生成模型灵活性提高，LIDGMVAE 的表现。图 1c 显示，LIDGMVAE 的分类变量潜变量对真实标签的预测能力显著高于其经典对应方法。此外，随着生成模型变得更加灵活，其性能并未下降。图 1d 显示，LIDGMVAE 始终取得更高的测试对数似然。表1比较了9层生成模型中 VAE 的不同变体。在四个数据集上，LIDGMVAE 都缓解了后验坍塌。与其他 VAE 变体相比，它取得了更高的 AU、KL 和互信息。它还实现了更高的测试对数似然。

结果：文本。 我们将 LIDSVAE 应用于三个经过子采样的文本数据集，它们分别取自合成文本数据集、雅虎数据集和 Yelp 数据集 [60]。合成数据集由一个经典的两层序列 VAE 生成，其潜变量维度为五。表1比较了 LIDSVAE 与序列变分自编码器。在这三个文本数据集上，LIDSVAE 在缓解后验坍塌方面优于其他变分自编码器变体，通常能取得更高的 AU、KL 和互信息。

4.2 概率主成分分析中的潜变量不可辨识性与后验坍塌

这里我们表明，当潜变量趋于不可辨识时，概率主成分分析的后验分布会接近先验分布。我们使用哈密顿蒙特卡洛（HMC）进行推断，从而避免变分近似对后验坍塌的影响。

考虑一个具有两个潜在维度的概率主成分分析模型，$\mathrm{p}(\mathrm{z}_i) = \mathcal{N}(\mathrm{z}_i; 0, \mathrm{I}_2)$, $\mathrm{p}(\mathrm{x}_i|\mathrm{z}_i; \theta) = \mathcal{N}(\mathrm{x}_i; \mathrm{z}^\top \mathrm{i}\, \mathrm{w}, \sigma^2 \cdot \mathrm{I}_5)$，其中 σ^2 的值已知，z_i 是感兴趣的潜变量，而 w 是唯一的感兴趣参数。当噪声 σ^2 设为较大值时，潜变量 z_i 可能变得近乎不可辨识。原因是，似然函数 $\mathrm{p}(\mathrm{x}_i|\mathrm{z}_i)$ 随着 σ^2 增大而变化得越来越缓慢。例如，图2显示，似然曲面随着 σ^2 增大而变得更加平坦。因此，后验分布随着 σ^2 增大而越来越接近先验分布。当 $\sigma = 1.5$ 时，后验分布发生坍塌。这一不可辨识性论证解释了 [33] 的第6.2节中描述的密切相关现象。

5 讨论

在这项工作中，我们表明，后验坍塌现象是潜变量不可辨识性问题。它并非特定于在变分自编码器中使用神经网络或特定推断算法。

相反，这是模型和数据集固有的问题。为此，我们提出了一类基于Brenier映射的LIDVAE，以解决潜变量不可辨识性并缓解后验坍塌。通过实证研究，我们发现，LIDVAE在缓解后验坍塌方面优于现有方法。

LIDVAE 的潜变量保证是可辨识的。然而，这并不保证 LIDVAE 的潜变量与参数是联合可辨识的。换言之，即使 LIDVAE 的潜变量可辨识，LIDVAE 模型本身仍可能不可识别。潜变量可识别性与模型可识别性之间的差异看似微不足道，但解决潜变量可识别性问题的可处理性，在于使不可识别性成为理解后验坍塌的一种富有成效的视角方面，发挥着关键作用。要确保潜变量可识别性，只需确保 ˆ似然 $p(\boldsymbol{x}|\boldsymbol{z},\theta)$ 是关于 $\boldsymbol{z}$ 的单射函数。相比之下，解决一般变分自编码器类别的模型可识别性问题，仍是一个长期悬而未决的问题；近期的一些进展依赖于辅助变量 [23, 24]。解决潜变量可识别性问题的可处理性，是制定原则性方案以缓解后验坍塌的关键推动因素。

这项工作存在一些局限性。其中之一是，理论论证侧重于精确后验分布的坍塌。其理由是，如果精确后验分布发生坍塌，那么它的变分近似也必然发生坍塌，因为对后验分布进行变分近似无法使其“恢复”。尽管如此，变分近似也可能使后验分布“坍塌”，即精确后验分布并未坍塌，但变分近似后验却发生了坍塌。本文提出的理论论证和算法方法并不适用于这种情形，这仍是一个值得进一步研究的方向。

第二个局限性是，本文提出的潜变量可识别变分自编码器比经典变分自编码器具有更高的计算成本。虽然潜变量可识别变分自编码器能够确保其潜变量的可识别性并缓解后验坍塌，但也需要付出计算上的代价，因为其生成模型（即解码器）使用神经网络的梯度进行参数化。因此，拟合潜变量可识别变分自编码器需要计算神经网络的梯度的梯度，其计算复杂度远高于拟合经典变分自编码器。开发计算效率更高的潜变量可识别变分自编码器变体，是未来另一个值得探索的方向。

致谢。 感谢 Taiga Abe 和 Gemma Moran 富有帮助的讨论，也感谢匿名审稿人提出的建设性意见，这些意见改进了本文稿。David Blei 得到 ONR N00014-17-1-2131、ONR N00014-15-1-2209、NSF CCF-1740833、美国国防高级研究计划局 SD2 FA8750-18-C-0130、亚马逊和西蒙斯基金会的支持。John Cunningham 得到西蒙斯基金会、McKnightFoundation、扎克曼研究所、GrossmanCenter 和 GatsbyCharitableTrust 的支持。

参考文献

[1] Alemi, A. A., Poole, B., 等（2017）。修复损坏的证据下界。arXiv 预印本 arXiv:1711.00464。[2] Amos, B., Xu, L., & Kolter, J. Z.（2017）。输入凸神经网络。载于《第34届国际机器学习会议论文集》第70卷（第146–155页）。：JMLR。org。[3] Asperti, A.（2019）。变分自编码器与变量坍塌现象。《传感器与换能器》，234(6)，1–8。[4] Ball, K.（2004）。单调传输简介。载于《泛函分析的几何方面》（第41–52页）。Springer。[5] Betancourt, M.（2017）。识别贝叶斯混合模型。https://mc-stan.org/users/documentation/case-studies/identifying_mixture_models。访问日期：2021年5月4日。[6] Bowman, S., Vilnis, L., 等（2016）。从连续空间生成句子。载于《第20届SIGNLL计算自然语言学习会议论文集》（第10–21页）。[7] Burda, Y., Grosse, R., & Salakhutdinov, R.（2015）。重要性加权自编码器。arXiv 预印本 arXiv:1509.00519。[8] Chen, X., Kingma, D. P., 等（2016）。变分有损自编码器。arXiv 预印本 arXiv:1611.02731。[9] Chen, Y., Shi, Y., & Zhang, B.（2018）。通过神经网络实现最优控制：一种凸方法。arXiv 预印本 arXiv:1805.11835。[10] Collins, M., Dasgupta, S., & Schapire, R. E.（2001）。主成分分析向指数族的推广。载于《神经信息处理系统大会》，第13卷（第23页）。[11] Dai, B., Wang, Z., & Wipf, D.（2019）。罪魁祸首通常是谁？重新审视 VAE 后验坍塌的责任归因。arXiv 预印本 arXiv:1912.10702。[12] Dieng, A. B., Kim, Y., Rush, A. M., & Blei, D. M.（2018）。使用生成式跳跃模型避免潜变量坍塌。arXiv 预印本 arXiv:1807.04863。[13] Dilokthanakul, N., Mediano, P. A., 等（2016）。使用高斯混合变分自编码器进行深度无监督聚类。arXiv 预印本 arXiv:1611.02648。[14] Dinh, L., Sohl-Dickstein, J., & Bengio, S.（2016）。使用 Real NVP 进行密度估计。arXiv 预印本 arXiv:1605.08803。[15] Fu, H., Li, C., 等（2019）。循环退火计划：缓解 KL 消失的一种简单方法。arXiv 预印本 arXiv:1903.10145。[16] Gulrajani, I., Kumar, K., 等（2016）。PixelVAE：用于自然图像的潜变量模型。arXiv 预印本 arXiv:1611.05013。[17] Havrylov, S. & Titov, I.（2020）。使用莱文斯坦变分自编码器防止后验坍塌。arXiv 预印本 arXiv:2004.14758。[18] He, J., Spokoyny, D., Neubig, G., & Berg-Kirkpatrick, T.（2019）。变分自编码器中的滞后推断网络与后验坍塌。arXiv 预印本 arXiv:1901.05534。[19] Higgins, I., Matthey, L., 等（2016）。 β-VAE：使用受约束的变分框架学习基本视觉概念。[20] Hochreiter, S. & Schmidhuber, J.（1997）。长短期记忆。《神经计算》，9(8)，1735–1780。[21] Hoffman, M. D. & Johnson, M. J.（2016）。证据下界手术：划分变分证据下界的另一种方法。[22] Johnson, M. J., Duvenaud, D. K., Wiltschko, A., Adams, R. P., & Datta, S. R.（2016）。使用神经网络组合图模型，以实现结构化表示和快速推断。载于《神经信息处理系统进展》（第2946–2954页）。[23] Khemakhem, I., Kingma, D. P., & Hyvärinen, A.（2019）。变分自编码器与非线性独立成分分析：统一框架。arXiv 预印本 arXiv:1907.04809。[24] Khemakhem, I., Monti, R. P., Kingma, D. P., & Hyvarinen, A.（2020）。ICE-BeeM：基于非线性独立成分分析的可识别条件能量深度模型。[25] Kim, Y., Wiseman, S., Miller, A., Sontag, D., & Rush, A.（2018）。半摊销变分自编码器。载于《国际机器学习会议论文集》（第2678–2687页）。[26] Kingma, D. P., Mohamed, S., Rezende, D. J., & Welling, M.（2014）。使用深度生成模型进行半监督学习。载于《神经信息处理系统进展》（第3581–3589页）。[27] Kingma, D. P., Salimans, T., 等（2016）。使用逆自回归流改进变分推断。载于《神经信息处理系统进展》（第4743–4751页）。

[28] Kingma, D. P. & Welling, M. (2014). 自编码变分贝叶斯。载于《表征学习国际会议论文集》（ICLR），第1卷。[29] Kumar, A. & Poole, B. (2020). 关于 β-VAE中的隐式正则化。载于国际机器学习会议（第5480–5490页）。：PMLR。[30] Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). 通过概率程序归纳实现人类水平的概念学习。《科学》，350(6266)，1332–1338。[31] LeCun, Y., Cortes, C., & Burges, C. (2010). MNIST手写数字数据库。ATT实验室 *[*在线*]*。网址：http://yann.lecun.com/exdb/mnist, 2。[32] Li, B., He, J., Neubig, G., Berg-Kirkpatrick, T., & Yang, Y. (2019). 一种出人意料地有效的文本深层潜变量建模修正方法。arXiv预印本 arXiv:1909.00868。[33] Lucas, J., Tucker, G., Grosse, R. B., & Norouzi, M. (2019). 别怪证据下界！从线性变分自编码器的角度看后验坍塌。载于《神经信息处理系统进展》（第9403–9413页）。[34] Maaløe, L., Fraccaro, M., Liévin, V., & Winther, O. (2019). BIVA：用于生成式建模的超深层潜变量层级结构。载于 H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (编)，《神经信息处理系统进展》，第32卷：Curran Associates, Inc. [35] Makkuva, A. V., Taghvaei, A., Oh, S., & Lee, J. D. (2019). 通过输入凸神经网络实现最优传输映射。arXiv预印本 arXiv:1908.10962。[36] McCann, R. J. et al. (1995). 单调保测映射的存在性与唯一性。《杜克数学杂志》，80(2)，309–324。[37] McCann, R. J. & Guillen, N. (2011). 关于最优传输的五讲：几何、正则性与应用。《度量测度空间的分析与几何：高等数学研讨班（SMS）讲义》，蒙特利尔，（第145–180页）。[38] Oord, A. v. d., Vinyals, O., & Kavukcuoglu, K. (2017). 神经离散表征学习。arXiv预印本 arXiv:1711.00937。[39] Peyré, G., Cuturi, M., et al. (2019). 计算最优传输。《机器学习基础与趋势》，11(5-6)，355–607。[40] Poirier, D. J. (1998). 修正不可识别模型中的信念。《计量经济理论》，14(4)，483–509。[41] Rao, B. & Prakasa, R. (1992). 随机模型中的可识别性：概率分布的刻画。概率与数理统计。学术出版社。[42] Raue, A., Kreutz, C., et al. (2009). 利用剖面似然分析部分观测动态模型的结构可识别性与实际可识别性。《生物信息学》，25(15)，1923–1929。[43] Raue, A., Kreutz, C., Theis, F. J., & Timmer, J. (2013). 贝叶斯方法论与频率学派方法论的强强联合：对不可识别性条件下推断的研究。《皇家学会哲学汇刊A辑：数学、物理与工程科学》，371(1984)，20110544。[44] Razavi, A., Oord, A. v. d., Poole, B., & Vinyals, O. (2019). 使用 delta-VAE 防止后验坍塌。arXiv预印本 arXiv:1901.03416。[45] Reynolds, D. A. (2009). 高斯混合模型。《生物识别百科全书》，741，659–663。[46] Rezende, D. J., Mohamed, S., & Wierstra, D. (2014). 深度生成模型中的随机反向传播与近似推断。arXiv预印本 arXiv:1401.4082。[47] Roweis, S. (1998). 用于主成分分析和稀疏主成分分析的期望最大化算法。神经信息处理系统进展，（第626–632页）。[48] Roweis, S. & Ghahramani, Z. (1999). 线性高斯模型综述。《神经计算》，11(2)，305–345。[49] San Martın, E. & González, J. (2010). 贝叶斯可识别性：对一场无定论争论的贡献。《智利统计学杂志》，1(2)，69–91。[50] Seybold, B., Fertig, E., Alemi, A., & Fischer, I. (2019). 双解码器：对变分自编码器潜在空间进行正则化。arXiv预印本 arXiv:1905.07478。[51] Shu, R. (2016). 高斯混合变分自编码器：关于变分推断、生成模型与深度神经网络的经验。[52] Sønderby, C. K., Raiko, T., Maaløe, L., Sønderby, S. K., & Winther, O. (2016). 如何训练深度变分自编码器和概率阶梯网络。载于第33届国际机器学习会议（ICML 2016）。[53] Strang, G., Strang, G., Strang, G., & Strang, G. (1993). 线性代数导论，第3卷。Wellesley-Cambridge出版社，马萨诸塞州韦尔斯利。

[54] Tipping, M. E. & Bishop, C. M. (1999). 概率主成分分析。英国皇家统计学会会刊：*B*辑（统计方法学），61(3)，611–622。[55] Tomczak, J. M. & Welling, M. (2017). 带 VampPrior 的变分自编码器。*arXiv*预印本 *arXiv:1705.07120*。[56] Wieland, F.-G., Hauber, A. L., Rosenblatt, M., Tönsing, C., & Timmer, J. (2021). 关于结构可识别性与实际可识别性。《系统生物学最新观点》。[57] Xiao, H., Rasul, K., & Vollgraf, R. (2017). Fashion-MNIST：用于机器学习算法基准测试的新型图像数据集。*CoRR*，abs/1708.07747。[58] Xie, Y. & Carlin, B. P. (2006). 层次模型中的贝叶斯学习与可识别性度量。《统计规划与推断杂志》，136(10)，3458–3477。[59] Yacoby, Y., Pan, W., & Doshi-Velez, F. (2020). 刻画并避免变分自编码器中存在问题的全局最优解。[60] Yang, Z., Hu, Z., Salakhutdinov, R., & Berg-Kirkpatrick, T. (2017). 使用膨胀卷积改进用于文本建模的变分自编码器。载于第*34*届国际机器学习会议论文集（第*70*卷）（第3881–3890页）。[61] Yeung, S., Kannan, A., Dauphin, Y., & Fei-Fei, L. (2017). 应对变分自编码器中的过度剪枝。*arXiv*预印本 *arXiv:1706.03643*。[62] Zhao, T., Lee, K., & Eskenazi, M. (2018). 面向可解释神经对话生成的无监督离散句子表示学习。载于计算语言学协会第*56*届年会论文集（第*1*卷：长篇论文）（第1098–1107页）。[63] Zhao, Y., Yu, P., Mahapatra, S., Su, Q., & Chen, C. (2020). 变分自编码器中的离散化瓶颈：避免后验坍塌的序列到序列学习。*arXiv*预印本 *arXiv:2004.10603*。

补充材料
后验坍塌与潜变量不可辨识性

A 后验坍塌示例（续）

我们再介绍两个后验坍塌的示例：概率主成分分析和高斯混合模型。

A.1 概率主成分分析

我们考察经典的概率主成分分析（PPCA），并说明其局部潜变量在最大似然参数值处（即对数边际似然的全局极大值处）可能发生后验坍塌。这个示例进一步阐明了 Lucas 等人 [7],的观点；他们证明，在 PPCA 中，即使没有任何变分近似，后验坍塌也可能发生，但这是由对数边际似然的局部极大值所导致的。这里我们表明，即使在全局极大值处，也可能发生后验坍塌，而这并非由局部极大值导致的优化问题所致。

考虑一个具有两个潜在维度的概率主成分分析（PPCA），

$$p(z_i) = \mathcal{N}(z_i \mid 0, I_2),$$
$$p(x_i \mid z_i; \theta) = \mathcal{N}(x_i \mid z_i^\top w, \sigma^2 \cdot I_5),$$

其中，z_i 是感兴趣的潜变量，其他的 $\theta = (w, \sigma^2)$ 则是模型的参数 .

考虑将该模型拟合到两个数据集，每个数据集包含 500 个样本，并关注最大似然参数值。根据数据集的真实分布，PPCA 可能会发生后验坍塌，也可能不会。

1. 从一维概率主成分分析中采样数据，

$$x_i \sim \mathcal{N}(x_i \mid \mathcal{N}(0, I_1) \cdot \bar{w}_1, \bar{\sigma}_1 \cdot I_5). \tag{10}$$

（模型仍为二维。）在此$\hat{}$情况下，潜变量 z_i 不（完全）可辨识。原因在于，一组最大似然参数为 $\theta = (\hat{}w, \hat{}\sigma) = ([\mathbf{0}, \bar{}w_1], \bar{\sigma}_1)$，也就是说，将一个潜在维度设为零，另一个设为真实的数据生成方向。在此$\hat{}\theta$下，似然函数在潜变量的第一个维度上为常数，即z_{i1}；见图 3a。因此，z_{i1} 的后验分布会坍塌并与先验分布一致，而 z_{i2} 的后验分布则仍保持峰状（图 3b ）。

2. 从二维概率主成分分析中采样数据，

$$x_i \sim \mathcal{N}(x_i \mid \mathcal{N}(0, I_2) \cdot \bar{w}_2, \bar{\sigma}_2 \cdot I_5). \tag{11}$$

潜变量 z_i 是可辨识的。似然函数随z_{i1} 和z_{i2}变化；z_{i1} 和z_{i2} 的后验分布均呈峰状（图 3c 和 3d）。

A.2 高斯混合模型

虽然我们重点讨论了局部潜变量的后验坍塌，但模型也可能出现全局潜变量的后验坍塌。考虑一个包含两个簇的简单高斯混合模型（GMM），

$$p(\alpha) = \text{Beta}(\alpha \mid 5, 5),$$
$$p(x_i \mid \alpha; \theta) = \alpha \cdot \mathcal{N}(x_i \mid \mu_1, \sigma_1^2) + (1 - \alpha) \cdot \mathcal{N}(x_i \mid \mu_2, \sigma_2^2).$$

其中 α 是一个全局潜变量，而 $\theta = (\mu_1, \mu_2, \sigma_1, \sigma_2)$ 是该模型的参数。将此模型拟合到三个数据集，每个数据集包含 10^5 个样本。

1. 从两个不重叠的簇中采样数据，

$$x_i \sim 0.15 \cdot \mathcal{N}(-10, 1) + 0.85 \cdot \mathcal{N}(10, 1). \tag{12}$$

潜变量 α 是可辨识的。两个数据生成簇存在显著差异，因此在最大似然（ML）参数下，似然函数会随 $\alpha \in [0, 1]$而变化（图 4a）。 α 的后验分布也呈现尖峰（图 4b），并且与先验分布有很大差异。

2. 从两个重叠簇中采样数据，

$$x_i \sim 0.15 \cdot \mathcal{N}(-0.5,1) + 0.85 \cdot \mathcal{N}(0.5,1). \tag{13}$$

潜变量 α 是可辨识的。然而，它几乎不可辨识。虽然两个数据生成簇并不相同，但由于彼此重叠，它们非常相似。因此，在最大似然参数$\theta^* = (\mu_1^*, \mu_2^*, \sigma_1^*, \sigma_2^*) = (-0.5, 0.5, 1, 1)$下，似然函数 $p(x_i \mid \alpha;\ \theta^*)$ 变化缓慢；见图 4a。因此，α 的后验分布与先验分布非常接近；见图 4b。

3. 从单一高斯分布 $x_i \sim \mathcal{N}(-1,1)$中采样生成数据。潜变量 α 不可辨识。原因在于，一组最大似然参数为 $\theta^* = (\mu_1^*, \mu_2^*, \sigma_1^*, \sigma_2^*) = (-1, -1, 1, 1)$，即将两个混合分量都设为与真实数据生成高斯分布相同。

在这个 θ^* 下，潜变量 α 不可识别，其似然函数 $p(\{x_i\}^n{}_{i=1} \mid \alpha;\ \theta^*)$ 关于 α 是常数，因为两个混合分量相等；图 4a 展示了这一点。此外，α 的后验分布发生坍塌，$p(\alpha \mid \{x_i\}^n{}_{i=1};\ \theta^*) = p(\alpha)$. 图 4b 展示了这一点：针对 α 后验分布的哈密顿蒙特卡洛（HMC）样本与从先验分布中抽取的样本十分接近。（在这种情况下，精确推断难以实现，因此我们使用 HMC 作为精确推断的近似方法。）这个例子展示了不可识别性与后验坍塌之间的联系；同时也表明，后验坍塌并非变分推断所特有，而是模型和数据带来的问题。

与概率主成分分析（PPCA）一样，这些高斯混合模型（GMM）示例表明，潜变量在概率模型中是否可识别，不仅取决于模型，也取决于数据。尽管这三个示例都使用同一个 GMM 模型进行拟合，但由于样本的生成方式各不相同，它们的可识别性情况也有所不同。

B 命题 2 的证明

我们通过证明最一般形式的 LIDVAE 具有潜变量可识别性和灵活性，来证明命题 2 的一个一般化版本。LIDVAE、LIDMVAE 和 LIDSVAE（定义4以及示例1和2）都将是它的特例。因此，命题 2 也将是下文陈述的更一般结果（命题 3）的一个特例。

我们首先定义 LIDVAE 的最一般形式。

Defi定义4（通过 Brenier 映射构造的一般 LIDVAE）。 一般的 LIDVAE 通过 *Brenier* 映射生成 s *an* D维数据点 x_i，$\in \{1,\ldots,n\}$ </*style*，定义如下：

$$(z_i)_{K\times 1} \sim p(z_i), \tag{14}$$

$$(w_i)_{M\times 1} \mid z_i \sim \mathrm{EF}(w_i \mid \beta_1^\top z_i), \tag{15}$$

$$(x_i)_{D\times 1} \mid w_i, x_{<i} \sim \mathrm{EF}(x_i \mid h \circ g_{2,\theta}(\beta_2^\top g_{1,\theta}([w_i, f_\theta(x_{<i})]))), \tag{16}$$

其中，*EF* 表示指数族分布；z_i 是一个 K 维潜变量，可以是离散的，也可以是连续的。模型的参数为 $\theta = (g1,\theta, g2,\theta, f_\theta)$，其中 $f_\theta: \mathscr{X}_{<i} \to \mathbb{R}^H$ 是一个将所有先前数据点 $x_{<i}$ 映射为 H 维向量的函数，$g1,\theta : \mathbb{R}^{M+H} \to \mathbb{R}^{M+H}$和 $g2,\theta : \mathbb{R}^D \to \mathbb{R}^D$ 是两个连续单调传输映射。函数 $h(\cdot)$ 是指数族的双射链接函数，例如 *Sigmoid* 函数。矩阵 β_1 是一个 $K \times M$ 维矩阵（$M \geq K$），其所有主对角线元素均为 *1*，其余元素均为 *0*，因此它具有满行秩。类似地，β_2 是一个 $(M+H)\times D$ 维矩阵（$D \geq M+H$），其所有主对角线元素均为 *1*，其余元素均为 *0*，也具有满行秩。最后，$[w_i, f_\theta(x_{<i})]$是一个 $(M+H)\times 1$维向量，表示向量 $(w_i)_{M\times 1}$和 $(f_\theta(x_{<i}))_{H\times 1}$按行堆叠而成。

一般形式的 LIDVAE 与一般形式如下的经典变分自编码器不同

$$(z_i)_{K\times 1} \sim p(z_i), \tag{17}$$

$$(w_i)_{M\times 1} \mid z_i \sim \mathrm{EF}(w_i \mid \beta_1^\top z_i), \tag{18}$$

$$(x_i)_{D\times 1} \mid w_i, x_{<i} \sim \mathrm{EF}(x_i \mid h \circ g_\theta([w_i, f_\theta(x_{<i})])), \tag{19}$$

关键区别在于公式 19：经典变分自编码器在公式 19 中使用任意函数 $g: \mathbb{R}^{M+H} \to \mathbb{R}^{D}$。相比之下，LIDVAE 使用复合函数 $g_{2,\theta}(\beta_2^\top g_{1,\theta}(\cdot))$，并满足公式 16 中的附加约束。

一般的 LIDVAE 可以处理独立同分布数据和序列数据。对于独立同分布数据（例如图像），我们可以将 $f_\theta(\cdot)$ 设为零函数，这意味着 $P(x_i \mid w_i, x_{<i}) = P(x_i \mid w_i)$。对于序列数据（例如文本），我们可以将 $f_\theta(\cdot)$ 设为长短期记忆网络，将历史 $x_{<i}$ 嵌入到一个 H 维向量中。

一般的 LIDVAE 可以模拟许多现有的变分自编码器。令 z_i 为分类变量（独热）向量，则分布 $\mathbf{EF}(z^\top{}_i \ \beta_\theta)$ 是指数族混合。随后，可识别变分自编码器通过灵活函数 g_θ 映射该混合模型。当 z_i 为实数值时，它通过灵活函数映射指数族主成分分析，从而模拟经典变分自编码器。

当我们令 z_i 为分类变量（独热）向量，并在公式 15 和 16 中将指数族分布 EF 设为高斯分布时，LIDGMVAE 是一般 LIDVAE 的一个特例。在这种情况下，$w_i \sim \text{Gaussian}(z^\top{}_i \ \beta_\theta, \gamma_\theta)$ 是高斯混合。接着，我们将 $f_\theta(\cdot)$ 设为零函数，这意味着 $P(x_i \mid w_i, x_{<i}) = P(x_i \mid w_i)$，最后将 h 设为恒等函数。

这一通用的 LIDVAE 还涵盖伯努利混合模型，这是用于 MNIST 数据的 LIDGMVAE 的一种常见变体。具体而言，我们可以将 z_i 设为分类变量（独热）向量，然后在公式 15 中将指数族分布 EF 设为高斯分布，使得 $w_i \sim \text{Gaussian}(z^\top{}_i \ \beta_\theta, \gamma_\theta)$ 成为高斯混合。接下来，我们将 $f_\theta(\cdot)$ 设为零函数，这意味着 $P(x_i \mid w_i, x_{<i}) = P(x_i \mid w_i)$，然后将 h 设为 Sigmoid 函数，最后在公式 16 中将 EF 设为伯努利分布。

当我们在公式 15 中将 EF 设为点质量，并将 $\beta_{1,\theta}$ 设为单位矩阵时，LIDSVAE 是通用 LIDVAE 的另一种特殊情况，这意味着 $w_i = z_i$。然后在公式 16 中将 EF 设为分类变量分布，并将 h 设为恒等函数，即可得到与示例 2 相同的配置。

通过引入额外的满行秩矩阵 β_k（例如主对角线元素全为 1、其余元素全为 0 的矩阵）和额外的 Brenier 映射 $g_{k,\theta}$，可以为 LIDVAE 增加更多层，使其更深。例如，我们可以通过设置以下内容，为式 16 增加一层：

$$(x_i)_{D\times 1} \mid w_i, x_{<i} \sim \mathrm{EF}(g_{3,\theta}(\beta_3^\top g_{2,\theta}(\beta_2^\top g_{1,\theta}([w_i, f_\theta(x_{<i})])))).$$

接下来，我们将证明这一通用类 LIDVAE 的潜变量可识别性和灵活性，从而推出上述所有特殊情形均具有可识别性和灵活性。

命题 3. 在 LIDVAE 中，潜变量 z_i 是可辨识的，即对于所有 $i \in \{1,\ldots,n\}$，都有

$$p(x_i \mid z_i = \tilde{z}', x_{<i}; \theta) = p(x_i \mid z_i = \tilde{z}, x_{<i}; \theta) \quad \Rightarrow \quad \tilde{z}' = \tilde{z}, \quad \forall \tilde{z}', \tilde{z}, \theta. \tag{20}$$

此外，对于经典 变分自编码器 （式 *17* 至 *19*）生成的任意数据分布，都存在一个LIDVAE 能够生成相同的分布。

证明。我们首先证明潜变量可识别性。要证明潜变量 z_i 可辨识，只需证明对于所有 θ，从 z_i 到 $\mathrm{p}(\mathrm{x}_i \mid \mathrm{z}_i; \theta)$ 的映射是单射。该映射具有单射性，因为其中涉及的所有变换（β_1、β_2、$\mathrm{g1}_\theta$、$\mathrm{g2}_\theta$）都是单射，而它们的复合也必定是单射：线性变换（β_1、β_2）具有满行秩，因此是单射；非线性变换（$\mathrm{g1}_\theta$、$\mathrm{g2}_\theta$）是单调传输映射，并且保证为双射 [1, 9]；最后，指数族似然是单射。

接下来，我们通过证明任意由变分自编码器生成的 $p(\boldsymbol{x})$ 均可由 LIDVAE 生成，来确立 LIDVAE 的灵活性。证明分为两步：（1）我们证明，任意由变分自编码器生成的 $p(\boldsymbol{x})$，均可由具有单射似然 $p(x_i \mid z_i; \theta)$ 的变分自编码器生成；（2）我们证明，任意由具有单射似然的变分自编码器生成的 $p(\boldsymbol{x})$，均可由 LIDVAE 生成。

为证明 (1)，假设 β_1 不具有满行秩，且 g_θ 不是单射。于是存在某个 $\mathrm{Z}' \in \mathbb{R}^d$、$\mathrm{d} < \mathrm{K}$，以及单射 $\beta'_{1,\theta}$，g'_θ，使得新的变分自编码器能够表示相同的 $\mathrm{p}(\mathrm{x})$。这是因为，我们总可以通过考虑函数的商空间，将非单射函数转化为单射函数。具体而言，我们考虑这样的商空间：其中 z 与 z' 之间的等价关系定义为 $\mathrm{p}(\mathrm{x} \mid \mathrm{z}; \theta) = \mathrm{p}(\mathrm{x} \mid \mathrm{z}'; \theta)$，它会诱导出到 $\mathbb{R}^d$ 的一个双射。当 $\mathrm{p}(\mathrm{z}')$ 不再是标准高斯分布时，必然存在一个双射 Brenier映射 $\tilde{\mathrm{z}} = \mathrm{f}_t(\mathrm{z}')$，使得 $\mathrm{p}(\tilde{\mathrm{x}})$ 是标准高斯分布（McCann 等人的定理 6 [8]）。

	Pinwheel				MNIST			
	AU	**KL**	**MI**	**LL**	**AU**	**KL**	**MI**	**LL**
变分自编码器 [6]	0.2	1.4e-6	2.0e-3	**-6.2 (5e-2)**	0.1	0.1	0.2	-108.2 (5e-1)
SA-VAE [5]	0.2	1.6e-5	2.0e-2	-6.5 (5e-2)	0.4	0.4	0.6	-106.3 (7e-1)
滞后变分自编码器 [3]	0.6	0.7e-3	1.5e0	-6.5 (4e-2)	0.5	0.8	1.7	-105.2 (5e-1)
β-变分自编码器 [4](β=0.2)	**1.0**	**1.2e-3**	**2.3e0**	-6.6 (6e-2)	0.8	1.5	2.8	-100.4 (6e-1)
LIDGMVAE（本文工作）	**1.0**	**1.2e-3**	2.2e0	-6.5 (5e-2)	**1.0**	**1.8**	**3.9**	**-95.4 (7e-1)**

表 2： LIDGMVAE 不会遭受后验坍塌，并且在 9层生成模型中比其经典对应模型拟合得更好。报告的数值为十个不同随机种子下的均值（标准差）。（数值越高越好。）

为证明 (2)，我们说明任何具有单射映射的变分自编码器都可以重新参数化为 LIDVAE。要证明这一结论，只需说明任何单射函数 $\mathrm{l}_\theta : \mathbb{R}^{M+H} \to \mathbb{R}^D$ 都可以重新参数化为 $\mathrm{g2},_\theta(\beta_2^\top \mathrm{g1},_\theta(\cdot))$。下面我们通过求解 $\mathrm{l}_\theta(\mathrm{z}) = \mathrm{g2},_\theta(\beta_2^\top \mathrm{g1},_\theta(\mathrm{z}))$ 中的 $\mathrm{g}_1,\mathrm{g}_2$ 和 β，给出这样的重新参数化。我们将 $\mathrm{g1},_\theta$ 设为恒等映射，将 β_2 设为一个 $(\mathrm{M}+\mathrm{H}) \times \mathrm{D}$ 矩阵，其主对角线元素均为 1，其余元素均为 0，并将 $\mathrm{g2},_\theta$ 设为一个可逆映射，该映射在 z 的 (M +H) 维子空间上与 l_θ 一致。最后，我们指出，同样的论证也适用于变分自编码器的一种变体，其中 $w_i = z_i$。它与 Kingma 和 Welling [6] 中的经典变分自编码器一致。应用上述相同的论证即可证明命题 2。

□

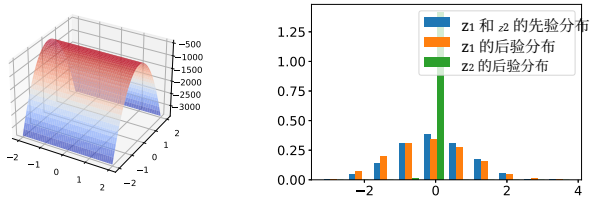
C 实验详情

在图像实验中，神经网络的所有隐藏层均有 512 个单元。我们将连续潜变量的数量设为 64，并将分类变量的维度设为真实标签的数量。随后，我们使用两层 RealNVP ([2]) 作为近似族，以探究变分推断的影响。

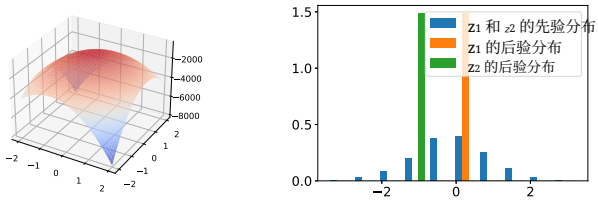
在文本实验中，神经网络的所有隐藏层均有 1024 个单元。我们将嵌入维度设为 1024。随后，按照拟合序列变分自编码器的常见做法，我们使用两层长短期记忆网络作为近似族。

D 其他实验结果

Tabl图 2 包含 LIDVAE 在图像数据集（Pinwheel 和 MNI
ST）

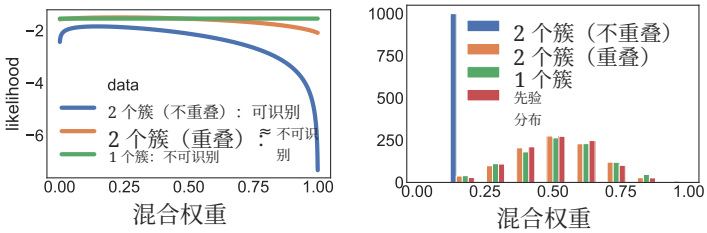


(a)似然（1D 概率主成分分析）　(b)后验分布（1D 概率主成分分析）



(c)似然（二维护主成分分析）　(d)后验分布（二维护主成分分析）

图 3：使用多于所需数量的潜在维度拟合概率主成分分析，会导致局部潜变量不可辨识，且后验分布坍塌。(a)-(b) 将二维概率主成分分析拟合到由一维概率主成分分析生成的数据。似然曲面沿潜变量的一个维度保持不变，即该潜变量不可辨识。因此，与之对应的后验分布会坍塌。(c)-(d) 将二维概率主成分分析拟合到由二维概率主成分分析生成的数据，不会出现后验坍塌；其似然曲面在所有维度上都会变化。



(a)似然函数　(b)后验直方图

图 4：当模型中的某个潜变量不可辨识（non-ID）时，其似然函数为常数函数，且其后验分布等于先验分布，即后验分布发生坍塌。考虑一个含有两个簇的高斯混合模型 $x \sim \alpha \cdot \mathcal{N}(\mu_1, \sigma_1^2) + (1-\alpha) \cdot \mathcal{N}(\mu_2, \sigma_2^2)$，将混合权重 α 视为潜变量，其他量视为参数。分别将模型拟合到由一个高斯簇（α 不可辨识）、两个重叠的高斯簇（α近乎不可辨识）和两个不重叠的高斯簇（α 可辨识）生成的数据集。在最优参数下，当潜变量 α （近乎）不可辨识时，似然函数 $p(x|\alpha)$（近乎）为常数；其后验分布也（近乎）等于先验分布。否则，似然函数不是常数，且后验分布会出现峰值。

参考文献

[1] Ball, K. (2004). 单调传输入门. 收录于泛函分析的几何方面（第 41–52 页）. Springer.
[2] Dinh, L., Sohl-Dickstein, J., & Bengio, S. (2016). 使用 Real NVP 进行密度估计. *arXiv* 预印本 *arXiv:1605.08803*.[3] He, J., Spokoyny, D., Neubig, G., & Berg-Kirkpatrick, T. (2019). 变分自编码器中的推断网络滞后与后验坍塌. *arXiv* 预印本 *arXiv:1901.05534*.[4] Higgins, I., Matthey, L., 等. (2016). β-VAE：使用受约束变分框架学习基本视觉概念.[5] Kim, Y., Wiseman, S., Miller, A., Sontag, D., & Rush, A. (2018). 半摊销变分自编码器. 收录于国际机器学习会议（第 2678–2687 页）.[6] Kingma, D. P. & Welling, M. (2014). 自编码变分贝叶斯. 收录于国际学习表征会议（*ICLR*）论文集，第 1 卷.[7] Lucas, J., Tucker, G., Grosse, R. B., & Norouzi, M. (2019). 别怪证据下界！从线性变分自编码器的角度看后验坍塌. 收录于神经信息处理系统进展（第 9403–9413 页）.[8] McCann, R. J. 等. (1995). 单调保测映射的存在性与唯一性.《杜克数学杂志》，80(2), 309–324.[9] McCann, R. J. & Guillen, N. (2011). 关于最优传输的五讲：几何、正则性与应用. 度量测度空间的分析与几何：蒙特利尔高等数学研讨班（*SMS*）讲义，（第 145–180 页）.