arXiv:2305.02930v2 [stat.ML] 1 Feb 2024

分段归一化流

Harry Thomas Jones Bevins [1 2] **Will Handley** [1 2] **Thomas Gessey-Jones** [1 2]

摘要

归一化流通过从基分布进行可逆变换来建模复杂概率密度，是一种成熟的方法。然而，归一化流捕捉目标分布的精度，会受到基分布拓扑结构的显著影响。目标分布与基分布的拓扑结构不匹配会导致性能不佳，多模态问题通常就是如此。许多研究尝试修改基分布的拓扑结构，使其更好地匹配目标分布，例如采用高斯混合模型（Izmailov 等人，2020年；Ardizzone 等人，2020年；Hagemann 和 Neumayer，2021年）或学习接受／拒绝采样（Stimper 等人，2022年）。我们提出分段归一化流，将目标分布划分为多个簇，使其拓扑结构更好地匹配标准正态基分布，并训练一系列流来建模复杂的多模态目标分布。我们使用一些标准基准测试展示分段流的性能，并将其精度与 Stimper 等人（2022年）所采用的多模态分布建模方法进行比较。我们发现，在标准基准测试中，我们的方法始终优于 Stimper 等人（2022年）的方法，模拟精度更高。

1. 引言

归一化流（NF）是一种有效的工具，可根据一组目标样本模拟概率分布。它们是基分布（通常为标准正态分布）与目标分布之间的双射可微变换（Papamakarios 等人，2019）。它们已成功应用于诸多领域，例如玻尔兹曼分布的近似（Wirnsberger 等人，2020）、宇宙学（Alsing & Handley，2021；Bevins & Hassan，2022；Friedman & Hassan，2022）、图像生成（Ho 等人，2019；Grci´c 等人，2021）、音频合成（van den Oord 等人，2018）、密度估计（Papamakarios 等人，2017；Huang 等人，2018）和变分推断（Rezende & Mohamed，2015）等。

当目标分布为多模态时，正规化流往往需要很长时间才能得到良好训练，而且通常会在不同模态之间形成训练数据中不存在的“桥接”（见图1）。这些桥接源于基分布的拓扑与通常较为复杂的目标分布之间的差异（Cornish 等，2019），也是正规化流的同胚性质所导致的结果（Runde，2007）。

处理多模态分布的一种方法是修改基分布的拓扑。这一方法见于（Stimper 等人，2022年），作者提出了一种通过学习型拒绝采样构造的基分布。解决基分布与目标分布拓扑不匹配问题的其他方法还包括：连续索引流（Cornish 等人，2019）、扩展模型拟合空间（Huang 等人，2018）以及添加随机层（Wu 等人，2020年）等。

一些研究尝试将高斯混合模型用于基分布（例如，Papamakarios 等，2017；Izmailov 等人，2020年），并扩展了各种 NF 结构，使其能够在训练过程中执行分类。例如，Izmailov 等人 (2020年) 表明，只需少量带标签样本，就能依据潜在空间中的模态对数据进行分类，并取得很高的准确率。 Ardiz- zone 等人(2020年) 利用信息瓶颈，通过高斯混合模型进行分类。 Hagemann 和 Neumayer (2021 年) 探讨了类似的思路。最近有研究表明，可以在训练过程中利用沃罗诺伊聚类有效地学习不同模态之间的边界（Chen 等人，2022年）。

我们的主要贡献 是对来自我们的样本进行分类。

[1]卡夫利宇宙学研究所，剑桥大学，马丁利路，剑桥，CB3 0HA [2]卡文迪许天体物理学，剑桥大学，卡文迪许实验室，J. J. 汤姆孙大道，剑桥，CB3 0HE。通信作者：Harry Thomas Jones Bevins <htjb2@cam.ac.uk>、Will Handley <wh260@cam.ac.uk>、Thomas Gessey-Jones< tg400@cam.ac.uk>。

《 41^{st} 国际机器学习会议》，维也纳，奥地利。PMLR 235，2024 年。版权所有 2024，归作者所有。

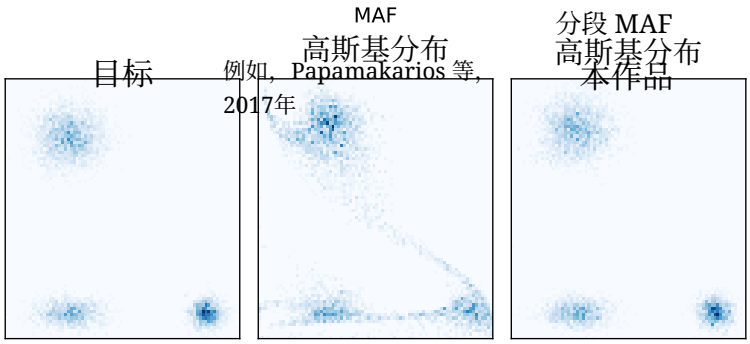


图*1.* 本文所述分段归一化流（NF）与在相同多模态目标分布上训练的单个 NF 的简单示意图。我们使用提前停止来训练不同的归一化流。对于单个 NF，我们可以看到不同簇之间出现了桥接现象，而分段方法中则没有。MAF 和分段 MAF 组件的架构经过选择，使两者的超参数数量大致相同。

使用聚类算法对目标分布进行处理训练，并采用分段方法，在所得簇上训练归一化流，其中每一段都由一个具有高斯基分布的归一化流构成。这样做能够提高流的准确性，并减少实现精准训练所需的计算资源，原因有二。首先，该方法可并行化，每一段都可以同时训练；其次，如果使用足够多的簇，目标分布中每一段的拓扑都会更接近高斯基分布。我们使用一系列基准多模态分布，将所得结果的准确性与（Stimper 等人，2022年）提出的方法进行比较。

2. 归一化流

如果目标分布为 X，基分布为$Z \sim \mathcal{N}(\mu = 0,\ \sigma = 1)$，那么通过变量变换，我们可以写成

$$p_X(x) = p_Z(z)\left|\det\left(\frac{\partial z}{\partial x}\right)\right|, \tag{1}$$

其中，基分布选为高斯分布，以便于采样和求值。如果两者之间存在双射变换 f，那么 z 由 $f_\theta^{-1}(x)$ 给出，变量变换公式变为

$$p_X(x) = p_Z(f_\theta^{-1}(x))\left|\det\left(\frac{\partial f_\theta^{-1}(x)}{\partial x}\right)\right|, \tag{2}$$

其中，θ 是描述从基分布到目标分布的变换的参数。在这里，雅可比行列式的绝对值表示从 p_Z 到 p_X的体积收缩。变换 f 对应于 p_X 和p_Z 之间的可逆归一化流，而 θ 则对应于流的超参数。在我们的工作中，我们使用掩码自回归流（MAFs，Papamakarios 等人，2017），其基于用于分布估计的掩码自编码器（MADE，Germain 等人，2015）。

对于 MAF，多维目标分布被分解为一系列一维条件分布

$$p_X(x) = \prod_i p(x_i|x_1, x_2, ..., x_{i-1}), \tag{3}$$

其中，索引 i表示维度。每个条件概率都建模为一个高斯分布，其标准差为 σ_i，均值为 μ_i，可写作

$$P(x_i|x_1,x_2, ..., x_{i-1}) = \mathcal{N}(\mu_i(x_1, x_2, ..., x_{i-1}), \sigma_i(x_1, x_2, ..., x_{i-1})). \tag{4}$$

σ_i 和 μ_i 是 θ 的函数，并由网络输出。通过训练网络，使目标分布与重构分布之间的 Kullback-Leibler（KL 散度）最小化，从而得到网络权重和偏置的最优值

$$\mathcal{D}(p_X(x)||p_\theta(x)) = -\mathbb{E}_{p_X(x)}[\log p_\theta(x)]+\mathbb{E}_{p_X(x)}[\log p_X(x)]. \tag{5}$$

KL 散度中的第二项与网络无关，因此在最小化问题中可以忽略这一项，并且由于

$$-\mathbb{E}_{p_X(x)}[\log p_\theta(x)] = -\frac{1}{N}\sum_{j=0}^{N}\log p_\theta(x_j), \tag{6}$$

其中，j 对应于 N 个样本集合中的一个样本，我们的最小化问题变为

$$\underset{\theta}{\operatorname{argmin}} -\sum_{j=0}^{N}\log p_\theta(x_j). \tag{7}$$

实际上，训练流是为了将来自目标分布的样本转换为来自基分布的样本，并通过变量变换来评估损失。

$$\underset{\theta}{\operatorname{argmin}} -\sum_{j=0}^{N}[\log p_Z(z'_j) + \log\left|\frac{\partial z'_j}{\partial x_j}\right|], \tag{8}$$

其中 z' 是 θ(Alsing 等，2019；Stimper 等人，2022年）。关于 MADE 网络的示意图，请参见Bevins 等(2023) 的图 1。通常，会将多个 MADE 网络串联起来，以提高 MAF 的表达能力（Papamakarios 等，2017）。

3. 聚类与归一化流

我们提出一种分段方法来学习复杂的多模态分布：在训练前将目标分布分解为多个簇，然后在每个簇上分别训练独立 MAF。对于 k 个样本簇 x_k，因此我们有一组

$$x = \{x_1, x_2, ..., x_k\} \text{ and } p_\theta = \{p_{\theta_1}, p_{\theta_2}, ..., p_{\theta_k}\}. \tag{9}$$

通常，如果用于目标分布的样本是依据某些样本权重抽取的，w_j，（即使用 MCMC（例如 Foreman-Mackey 等，2013）或嵌套采样（Skilling，2006；Handley 等，2015）算法），那么每个簇中的每组样本都有相应的权重

$$w = \{w_1, w_2, ..., w_k\}. \tag{10}$$

我们的总损失函数在 N_{clusters} 上的表达式为

$$\underset{\theta}{\operatorname{argmin}} -\Bigg[\sum_{j=0}^{N_0} w_{0,j}\log p_{\theta_1}(x_{0,j}) + \sum_{j=0}^{N_1} w_{1,j}\log p_{\theta_2}(x_{1,j}) + ... + \sum_{j=0}^{N_k} w_{k,j}\log p_{\theta_k}(x_{k,j})\Bigg], \tag{11}$$

可写为

$$\underset{\theta}{\operatorname{argmin}} -\sum_{k=0}^{N_{\text{cluster}}}\sum_{j=0}^{N_k} w_{k,j}\log p_{\theta_k}(x_{k,j}). \tag{12}$$

因此，整个网络的损失函数就是各个簇 MAF 的损失之和。为此，我们使用 ADAM 优化器（Kingma 与 Ba，2014）分别训练每个簇的 MAF。由于每个簇构成的目标分布比整体分布更简单，我们发现，与在同一问题上训练单个 NF 相比，训练分段 NF 所需的计算资源更少（见第4 节和表 2）。可以并行训练 PNF 的各个部分，从而进一步减少计算时间（墙钟时间）。

为了确定最能描述目标分布所需的簇数，我们使用平均轮廓系数。其中，单个样本的轮廓系数用于衡量该样本与其所在簇中的样本之间的相似程度，并将其与该样本和簇外样本之间的相似程度进行比较（Rousseeuw，1987）。该系数的取值范围为 -1 到 1；当各簇彼此清晰分离且定义明确时，其值达到峰值，如图 2所示。

我们的方法不受聚类算法选择的影响（参见第 4.2节），不过本文使用 k-均值算法（Steinhaus，1957；MacQueen，1967）。

我们可以从经过训练的分段归一化流的每个簇中抽取样本，其权重由目标簇的总权重决定，$\mathcal{W}_k = \sum_{j=0}^{N_k} w_{j,k}$。这一过程如图3所示，同时展示了分类和训练的一般流程。

对于样本 x，我们的分段 MAF 的概率为

$$\log p_\theta(x) = \log\left(\sum_k \mathcal{W}_k \exp\left(\log p_{\theta_k}(x)\right)\right) - \log\left(\sum_k \mathcal{W}_k\right), \tag{13}$$

或等价地

$$\log p_\theta(x) = \log\left(\sum_k\left[\exp\left(\log(\mathcal{W}_k) + \log(p_{\theta,k})\right)\right]\right) - \log\left(\sum_k \mathcal{W}_k\right), \tag{14}$$

并且，如果我们对 $\mathcal{W}_k$ 进行归一化，使其总和等于一，就得到

$$\log p_\theta(x) = \log\left(\sum_k\left[\exp\left(\log(\mathcal{W}_k) + \log(p_{\theta,k})\right)\right]\right). \tag{15}$$

由此，我们可以计算 KL 散度及其相关的蒙特卡罗误差

$$D_{KL} = \left\langle \log\frac{p_\theta(x)}{p_X(x)}\right\rangle_{p_\theta(x)} \pm \frac{1}{\sqrt{N}}\sigma\left(\log\frac{p_\theta(x)}{p_X(x)}\right)_{p_\theta(x)}, \tag{16}$$

其中，$p_X(x)$ 是解析目标分布，N 样本$x \sim p_\theta(x)$，而 $\sigma(f)_g$ 是 f 相对于 g 的标准差。

4. 基准测试

4.1. 玩具模型

我们在一系列多模态问题上测试分段归一化流，这些问题在（Stimper 等人，2022年）中进行了探讨。

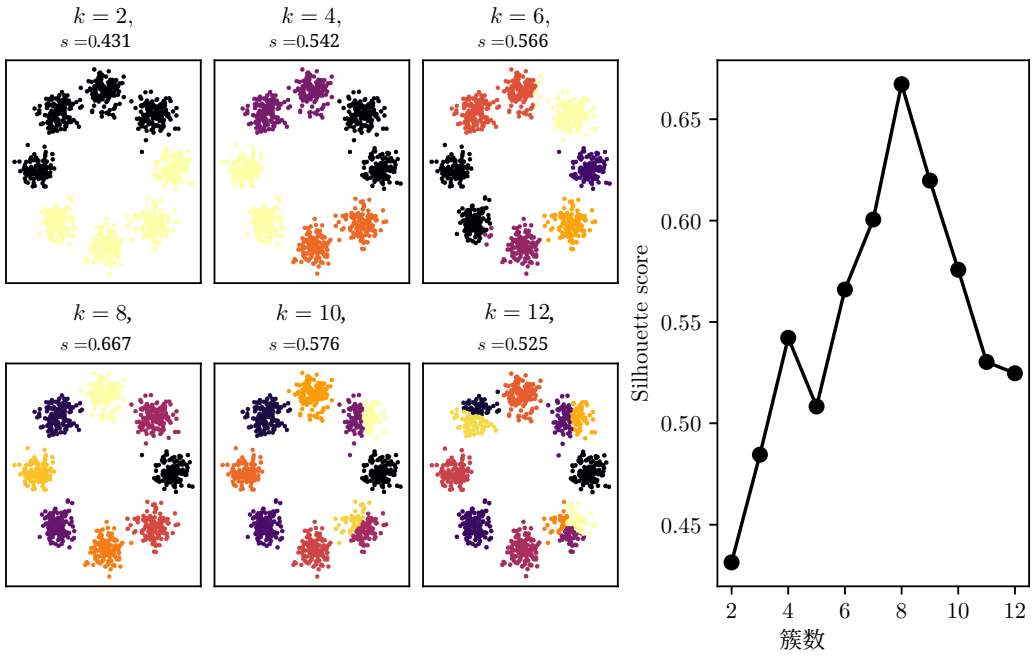


图*2.* 该图展示了如何使用轮廓系数 s 来确定描述数据所需的簇数。在此示例中， s 在八个簇时达到峰值，每个簇对应圆中的一个高斯分布。

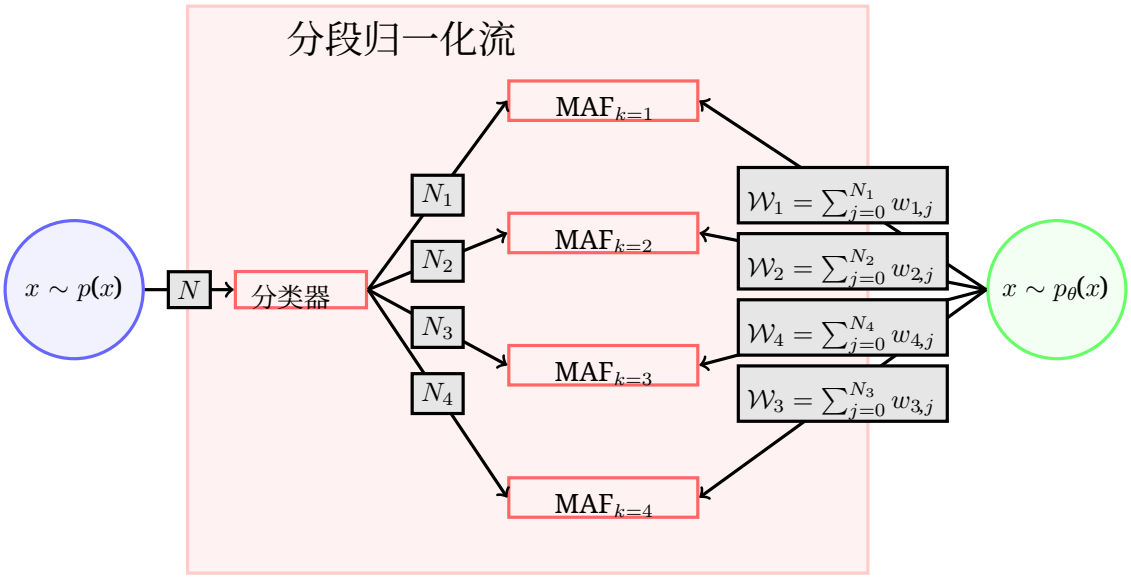


图*3.* 该图展示了分段归一化流的训练和采样过程。从左到右，首先从 $p_X(x)$中抽取一组样本 x ，并将样本划分为若干簇（例如使用 k-means），再根据轮廓系数确定簇数。然后，我们在每个簇 k 上训练一个掩码自回归流，并将标准正态分布作为基分布。从 $p_\theta(x)$中抽取样本时，我们根据各簇的总权重定义权重，并据此选择要从哪个MAF中抽取样本。

分布	KL 散度		
	MAF 高斯基分布 例如（Papamakarios 等，2017）	RealNVP 重采样基分布 （Stimper 等人，2022年）	分段 MAF 高斯基分布 本文
双月	-0.054 ± 0.009	0.028 ± 0.055	$\mathbf{0.004 \pm 0.022}$
高斯圆环	-0.046 ± 0.050	0.231 ± 0.138	$\mathbf{0.001 \pm 0.005}$
双环	0.015 ± 0.078	0.080 ± 0.106	$\mathbf{0.031 \pm 0.032}$

表 *1*。 该表展示了一系列多模态分布的目标样本与其等效表示之间的 KL 散度，这些等效表示分别采用单个 MAF、具有重采样基分布的 RealNVP（例如 Stimper 等人， 2022年）以及分段 MAF。我们在目标分布上对每种类型的流分别训练十次。报告的 KL 散度为十次训练结果的均值。每次训练的误差按照公式（16）计算，并在各次训练之间以平方和开根的方式合并误差，从而得到 KL 均值的误差。图 4展示了这些分布，以及从每种流中抽取的一组样本示例。我们在训练过程中采用了提前停止，并且每种流的超参数数量大致相同。我们发现，对于多模态概率分布的建模，我们的分段方法与（Stimper 等人，2022年）中详细介绍的方法具有竞争力，而且由于 KL 散度的误差更小，在重复训练过程中表现得更加一致。负 KL 散度是蒙特卡罗计算方法所导致的结果。

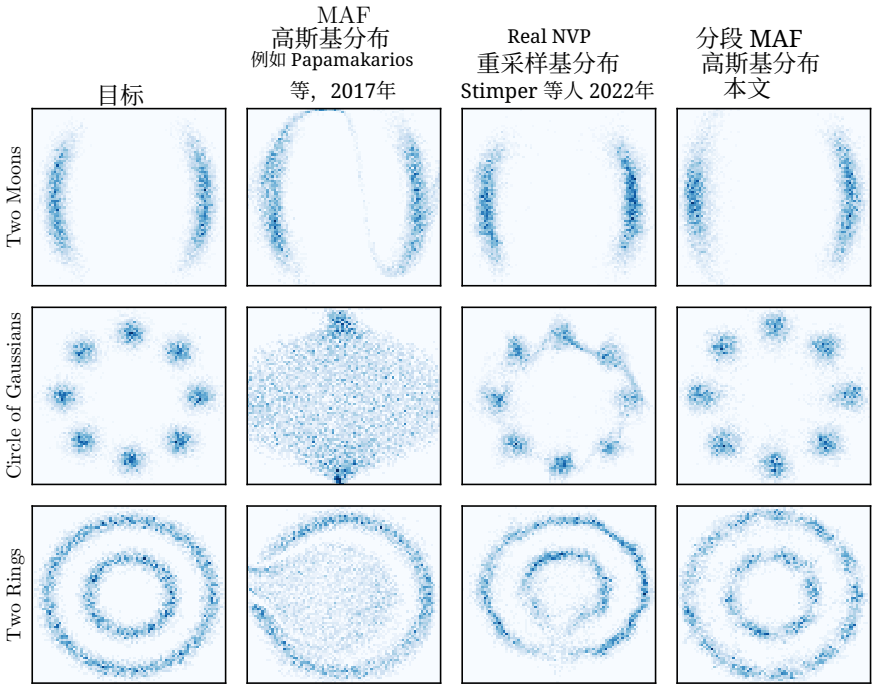


图 *4*。 图中展示了一系列多模态目标分布：第一列为样本的二维直方图，其余列为使用三种不同类型的归一化流生成的该分布的三种表示。我们按照正文所述，使用提前停止来训练所有归一化流。第二列展示了一个简单的掩码自回归流，其高斯基分布基于（Papamakarios 等， 2017）中的工作；后两列则展示了两种提高模型准确性的方法。第一种方法来自（Stimper 等人， 2022年），使用 RealNVP 流对分布建模，同时对基分布重新采样，使其更好地匹配目标分布的拓扑。最后一列展示了我们的 PNF 的结果；通过聚类，我们有效地修改了目标分布的拓扑，使其更好地匹配高斯基分布。

对于每种情况，我们分别使用单个 MAF 和高斯基分布、Stimper 等人 Stimper 等人 (2022年) 中的 RealNVP 和重采样基分布，以及我们提出的分段 MAF 方法来建模分布。虽然有许多不同的方法可以准确地学习多模态分布，但绝大多数方法都不是双射的，这与 Stimper 等人 (2022年) 中采用的方法不同，因此我们将结果与该方法进行比较。我们计算原始目标分布与训练所得流之间的 Kullback-Leibler 散度，以评估不同方法的性能。

图 4 展示了目标分布，以及从每种流中抽取的样本：MAF、分段 MAF，以及采用重采样基分布的 RealNVP；相应的 KL 散度列于表 1。我们对每种流分别训练十次，并计算每种流在每个分布上的平均 KL 散度。将相应的蒙特卡罗误差平方后相加，再开平方，得到平均 KL 散度的误差。我们发现，对于所有多模态基准，我们的分段方法都优于Stimper 等人(2022年) 提出的重采样基分布方法。我们使用

h
t
t
p
s://
github.com/VincentStimper/resampled-base-flows 上的代码拟合 RealNVP 流，并使用 MARGARINE (Bevins 等，2023年; 2022年) 提供的 https://github.com/htjb/margarine 代码拟合 MAF 和 PNF。我们对 MAF、PNF 和 RealNVP 的架构进行了选择，使它们的超参数数量大致相同。

我们使用提前停止来训练论文中的所有流，包括来自 Stimper 等人 (2022年) 的 RealNVP 重采样基分布模型。对于单个 MAF 和 RealNVP，我们将目标分布中 20% 的样本用作测试样本。当经过最大请求迭代次数的 2%（本文所有示例均为 10,000 次）后，最小测试损失仍未改善时，我们便提前停止，并回滚至最优模型。对于 PNF，我们将每个簇中的样本划分为训练样本和测试样本，并根据相同的标准对每个簇执行提前停止，这意味着每个簇的训练轮数各不相同。本文中的所有实验均在配备 M2 芯片和 8GB 内存的 Apple MacBook Pro 上运行。

在表格 2中，我们报告了在基准分布上分别对 MAF、RealNVP 和 PNF 进行五次训练的训练成本、所需训练轮数以及平均耗时。我们将训练 MAF 和 RealNVP 的成本定义为训练轮数、 E训练数据和测试数据合计的长度 N 以及网络中的超参数数量的乘积 h

$$\mathcal{C} = ENh, \tag{17}$$

其思路是，在每个轮次中，我们执行 h 次计算，共 N 次（因为我们会同时评估测试损失和训练损失）。

对于 PNF，分布的每个部分都使用一部分数据进行训练，训练轮数各不相同，并且 PNF 中的每个 MAF 都有 $h_k < h$，使得 $\sum_k h_k \approx h$。这种“分而治之”方法的成本为

$$\mathcal{C} = \sum_k E_k N_k h_k. \tag{18}$$

为便于比较，我们报告了 RealNVP 和 PNF 的成本及相关误差，并将其归一化为 MAF 的对应数值。我们发现，RealNVP 所需的训练轮数最少，墙钟时间也最短。不过，报告的 PNF 墙钟时间对应于按顺序训练每个部分；通过并行化，可以使其与 RealNVP 更具竞争力。这一点也体现在 PNF 训练成本显著降低上。

4.2. 选择聚类算法

接下来，我们将说明聚类算法的选择如何影响 PNF 的准确性。我们针对双环分布，使用五种不同的聚类算法对目标样本进行初始划分，并重复上一节的分析。

迄今为止，我们一直采用 k-均值聚类算法来识别目标分布中的簇。这里，我们测试小批量 k-均值（Sculley，2010）、均值漂移算法（Cheng，1995）、谱聚类（Shi & Malik，2000）、凝聚聚类（Nielsen，2016）以及 BIRCH 算法（Zhang 等人，1996）。所有这些算法都可以通过 SKLEARN PYTHON 软件包轻松使用，我们采用了默认设置。更多详情请参阅 Saxena 等人(2017) 的综述。

对于每种聚类算法，我们运行训练十次，并根据公式（16）计算目标分布与 PNF 之间的 KL 散度。训练运行中的 KL 散度均值及相应误差见表 3。我们可以清楚地看到，k-均值算法优于其他算法，但均值漂移和 BIRCH 聚类彼此不相上下，k-均值和 RealNVP 重采样基分布方法。

本研究并非旨在穷尽所有可用的聚类算法，但它说明了可以使用不同的聚类算法构建 PNF。根据此处展示的结果，我们推荐使用 k-means、均值漂移和 BIRCH 聚类，但强调聚类算法的选择可能取决于数据。

4.3. 物理示例

通常会在 UCI 机器学习数据集上比较提出的归一化流架构。

Cost			
分布	MAF	RealNVP	PNF
双月	1.000 ± 0.191	0.241 ± 0.003	$\mathbf{0.225 \pm 0.037}$
高斯圆环	1.000 ± 0.206	0.794 ± 0.092	$\mathbf{0.226 \pm 0.005}$
双环	1.000 ± 0.265	0.800 ± 0.098	$\mathbf{0.236 \pm 0.007}$

训练轮数			
分布	MAF	RealNVP	PNF
双月	3131 ± 598	$\mathbf{655 \pm 9}$	2118 ± 351
高斯圆环	1765 ± 364	$\mathbf{1217 \pm 141}$	19130 ± 412
双环	1896 ± 503	$\mathbf{1317 \pm 162}$	42444 ± 1720

运行时间 [秒]			
分布	MAF	RealNVP	PNF
双月	72.80 ± 12.72	$\mathbf{4.05 \pm 0.05}$	22.72 ± 3.31
高斯圆环	42.13 ± 8.49	$\mathbf{7.62 \pm 0.86}$	26.47 ± 1.15
双环	46.02 ± 12.14	$\mathbf{8.33 \pm 1.04}$	44.55 ± 1.82

表 *2*。 该表总结了本文所研究的三个玩具示例分布上，MAF、RealNVP 和 PNF 的训练成本、总训练轮数以及墙钟时间。这些数值是五次训练运行的平均值，并附有相应的蒙特卡罗误差。此处，正文中定义的成本及其相应误差均以 MAF 的数值进行归一化，以便比较。在测试的三种算法中，RealNVP 所需的训练轮数最少，墙钟时间也最短。然而，PNF 的训练成本显著低于 RealNVP 和 MAF，这是因为数据被分配到多个较小的 MAF 中，而这些 MAF 所需的超参数更少。需要注意的是，PNF 的总墙钟时间对应于按串行方式训练各个部分；通过并行化，可以使 PNF 的墙钟时间达到与 RealNVP 相当的水平。这一点已反映在成本中。

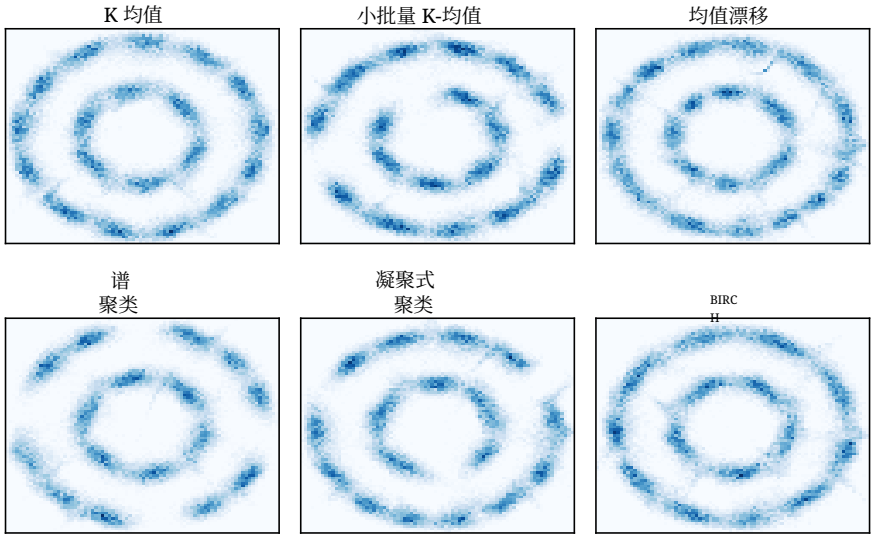


图 *5*。 该图展示了使用不同聚类算法构建的多个不同 PNF 所生成的样本，以及来自真实分布的样本和来自单个掩码自回归流的样本。该图表明，PNF 可以使用不同的聚类算法有效构建；结合表 3 可知，某些聚类算法的表现优于其他算法。我们为 PNF 和 MAF 选择了网络架构，使它们都具有大致相同数量的超参数。

Flow Type	D_{KL}
MAF	0.015 ± 0.078
RealNVP 重采样基分布	0.080 ± 0.106
PNF k-均值	$\mathbf{0.031 \pm 0.032}$
PNF 小批量 k-均值	0.847 ± 0.038
PNF 均值漂移	0.040 ± 0.007
PNF 谱聚类	0.814 ± 0.035
PNF 凝聚聚类	0.419 ± 0.028
PNF BIRCH	0.036 ± 0.007

表 3。该表展示了目标分布、双环玩具示例与一系列采用不同聚类算法构建的 PNF、单个 MAF，以及采用重采样基分布的 RealNVP 之间的 KL 散度。MAF、RealNVP 和 k-均值 PNF 的数值摘自表 1。我们发现，k-均值是测试过的聚类算法中表现最好的，但均值漂移和 BIRCH 聚类的表现也能与 RealNVP 相媲美。

流类型	POWER (6)	GAS (8)
MAF	$\mathbf{-0.05 \pm 0.08}$	$\mathbf{5.10 \pm 0.09}$
PNF	$\mathbf{-0.09 \pm 0.07}$	$\mathbf{4.32 \pm 0.68}$
流类型	HEPMASS (21)	MINIBOONE (43)
MAF	$\mathbf{-20.11 \pm 0.02}$	$\mathbf{-11.63 \pm 0.44}$
PNF	-20.71 ± 0.02	-12.40 ± 0.45

表 4. 我们比较了本文提出的 PNF 与单个 MAF 在 UCI 机器学习资料库中的四个数据集上的准确率（Kelly 等人，2023）。表中显示了在 POWER、GAS、HEPMASS 和 MINIBOONE 这四个 UCI 数据集的测试数据集上评估得出的平均对数似然。我们在括号中标出了每个数据集的维度数。总体而言，MAF 的表现优于 PNF；然而，Papamakarios 等(2017) 此前发现，通常使用标准正态基分布的 MAF 比使用高斯混合基分布的 MAF 更能拟合这些数据集，这表明这些数据集并非显著的多模态分布。同样，我们在训练期间使用提前停止，并指出 PNF 和 MAF 的超参数数量大致相同。误差为两个标准差，与 Papamakarios 等(2017) 中相同。

代码库（Kelly 等人，2023）。我们通过评估流在一组未见过的测试样本上的平均对数概率来衡量模拟的准确性。平均对数概率越高，表示网络越准确；第 2 节对损失函数的讨论对此有所说明。表 4报告了在四个 UCI 数据集（POWER、GAS、HEPMASS 和 MINIBOONE）上进行训练的结果，并附有相应的蒙特卡洛误差条。

POWER 和 GAS 数据集远大于 HEP-MASS 和 MINIBOONE，因此训练时计算成本也更高。因此，我们从这两个数据集中各选取了具有代表性的子集用于训练和测试。

总体而言，与 PNF 相比，MAF 在这些物理数据集上的表现更好，不过 PNF 在 POWER 和 GAS 数据集上也具有竞争力。我们注意到，文献中没有迹象表明这些数据集是多模态的。例如，Papamakarios 等(2017) 发现，与采用高斯混合模型基分布的 MAF 相比，采用标准正态基分布的 MAF 通常在这些目标上的表现更好。因此，我们并不一定期望 PNF 的表现优于 MAF，正如玩具多模态案例中所示。事实上，对于所有数据集，轮廓系数都倾向于选择测试过的最小簇数，即 2。不过令人欣慰的是，PNF 在这些示例中的表现并没有比 MAF 差很多。

5. 局限性

归一化流之所以有用，是因为它们具有双射性且可微；然而，在尝试处理多模态分布时，这些性质有时会丧失。我们的分段方法是可微的，前提是用于将目标分布中的样本划分到各簇的优化算法也是可微的，并且 k 簇数在训练过程中保持固定。

如果我们知道样本来自哪个簇，那么我们的分段方法就是双射的。换句话说，从 N 维基分布中抽取的一组实数，与目标分布上的一组 N 维实数加上一个簇编号之间存在一一映射，$n \in \mathbb{N}$

$$\mathbb{R}^N_{\mathcal{N}(\mu=0,\sigma=1)} \Longleftrightarrow \mathbb{R}^N_{p_\theta(x)} \times \mathbb{N}. \tag{19}$$

6. 结论

在这项工作中，我们展示了一种新方法：借助对目标分布进行初始分类并划分为多个簇，从而利用正则化流准确地学习多模态概率分布。

归一化流通常难以处理多模态分布，因为基分布与待学习的目标分布在拓扑结构上存在显著差异。通过将目标分布划分为多个聚类，我们实际上是在每个分段中使目标分布的拓扑结构更接近基分布的拓扑结构，从而提升表达能力。Stimper 等人(2022年)则通过对基分布重新采样，使其拓扑结构与目标分布更加相似。我们发现，这两种方法在基准测试中表现都很好，但在采用提前停止进行训练时，我们的分段方法更善于捕捉这些分布。我们方法的分段特性使 MAF 能够并行训练。

通过结合聚类与归一化流，我们能够准确表示复杂的多模态概率分布。

7. 致谢

HTJB 感谢剑桥卡弗里宇宙学研究所、卡弗里基金会以及剑桥圣埃德蒙学院提供的支持。WJH 感谢英国皇家学会通过大学研究奖学金给予支持。TGJ 感谢科学技术设施理事会（英国）通过资助编号 ST/V506606/1 持续提供支持。

本研究使用了剑桥大学的 DiRAC 数据密集型服务（CSD3）。该服务由剑桥大学信息服务部代表 STFC DiRAC 高性能计算设施管理（www.dirac.ac.uk）。剑桥 CSD3 的 DiRAC 部分由商业、能源与产业战略部、英国研究与创新署和科学技术设施理事会的资本资金以及科学技术设施理事会的运营拨款资助。DiRAC 是英国研究与创新署数字研究基础设施的一部分。

8. 影响声明

机器学习中的偏差目前是一个备受关注和研究的活跃领域（例如 Jiang和Nachum，2020年；Mehrabi等人，2021年）。虽然本研究聚焦于用于概率密度估计的归一化流，但我们认为这一通用方法具有更广泛的应用前景。可以设想将训练数据集划分为不同类别，并训练一系列生成模型来生成新数据。之后，可以根据用户设定的权重而非数据定义的权重抽取样本，正如本文所采用的方法。因此，在训练机器学习模型时，可以开始克服一些选择偏差。当然，权重由“用户决定”这一特性也可能被用来在建模中引入更多偏差，因此，权重的推导方法必须明确且符合伦理规范。

参考资料

Alsing, J. 和 Handley, W. 使用任意先验的嵌套采样。《皇家天文学会月报》，505(1):L95–L99，2021年7月。doi: 10.1093/mnrasl/ slab057.

Alsing、J.、Charnock、T.、Feeney、S. 和 Wandelt、B. 使用神经密度估计器和主动学习实现快速无似然宇宙学。《月刊皇家天文学会会刊》，488(3)：4440–4458，2019年9月。doi：10.1093/mnras/stz1960。

Ardizzone, L.、Mackowiak, R.、Rother, C. 和 K¨othe, U。使用信息瓶颈训练归一化流，以实现具有竞争力的生成式分类。神经信息处理系统进展，33: 7828–7840，2020年。

Bevins、H.、Handley、W.、Lemos、P.、Sims、P.、de Lera Acedo、E. 和 Fialkov、A. 使用掩码自回归流和核密度估计器进行边际贝叶斯统计，并附有宇宙学领域的示例。*arXiv* 电子预印本，论文编号：arXiv:2207.11457，2022年7月。

Bevins, H. T. J., Handley, W. J., Lemos, P., Sims, P. H., de Lera Acedo, E., Fialkov, A., and Alsing, J. 使用归一化流和核密度估计器对贝叶斯推断结果进行边缘化后处理。《皇家天文学会月刊》，526(3):4613– 4626，2023年12月。doi: 10.1093/mnras/stad2997。Chen, R. T., Amos, B., and Nickel, M. 通过可微分镶嵌实现半离散归一化流。《神经信息处理系统进展》，35: 14878–14889，2022年。Cheng, Y. 均值漂移、模式搜索与聚类。*IEEE*模式分析与机器智能汇刊，17(8):790–799，1995年。doi: 10.1109/34.400568。Cornish, R., Caterini, A. L., Deligiannidis, G., and Doucet, A. 通过连续索引归一化流放宽双射约束。载于国际机器学习会议，2019年。Foreman-Mackey, D., Hogg, D. W., Lang, D., and Goodman, J. emcee：MCMC锤子。《太平洋天文学会刊》，125(925):306，2013年3月。doi: 10.1086/670067。Friedman, R. and Hassan, S. HIGlow：用于高保真HI图建模的条件归一化流。*arXiv*电子预印本，文章编号 arXiv:2211.12724，2022年11月。doi: 10.48550/arXiv.2211.12724。Germain, M., Gregor, K., Murray, I., and Larochelle, H. MADE：用于分布估计的掩码自编码器。*arXiv*电子预印本，文章编号 arXiv:1502.03509，2015年2月。doi: 10.48550/arXiv.1502.03509。ˇ Grci´c, M., Grubi˘si´c, I.,and Segvi´c, S. 密集连接归一化流。载于 Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W.（编），《神经信息处理系统进展》，第34卷，第23968–23982页。Curran Associates, Inc., 2021年。网址：https://proceedings.neurips.cc/paper_files/paper/2021/file/c950cde9b3f83f41721788e3315a14a3-Paper.pdf.

Hagemann, P. 和 Neumayer, S. 使用混合模型稳定可逆神经网络。《反问题》，37(8): 085002，2021年。

Handley, W. J., Hobson, M. P. 和 Lasenby, A. N. POLY-CHORD：新一代嵌套采样。《皇家天文学会月刊》，453(4):4384–4398，2015年11月。doi: 10.1093/mnras/stv1911。

Ho, J., Chen, X., Srinivas, A., Duan, Y., 和 Abbeel, P. 流++：通过变分去量化和架构设计改进基于流的生成模型。载于 Chaudhuri, K. 和 Salakhutdinov, R.（编），第*36*届国际机器学习会议论文集，第97卷，收录于机器学习研究论文集，第2722–2730页。PMLR，2019年6月9日至15日。网址：
h
t
tps://proceedings.mlr.press/v97/ho19a.html。

Huang, C.-W., Krueger, D., Lacoste, A 和 Courville, A. 神经自回归流。载于 Dy, J. 和 Krause, A.（编），第*35*届国际机器学习会议论文集，第80卷，《机器学习研究论文集》，第2078–2087页。PMLR，2018年7月10–15日。网址：
https://proceedings.mlr.press/v80/huang18d.html。

Izmailov、P.、Kirichenko、P.、Finzi、M. 和 Wilson、A. G. 使用归一化流进行半监督学习。载于国际机器学习会议，第 4615–4630 页。PMLR，2020年。

Jiang, H. 和 Nachum, O. 识别并纠正机器学习中的标签偏差。载于国际人工智能与统计学大会，第 702–712 页。PMLR，2020年。

Kelly, M., Longjohn, R. 和 Nottingham, K. UCI 机器学习资料库，2023。网址https://archive.ics.uci.edu。

Kingma, D. P. 和 Ba, J. Adam：一种随机优化方法。*CoRR*，abs/1412.6980，2014 年。

MacQueen, J. B. 多变量观测的分类与分析的一些方法。载于 Cam, L. M. L. 和 Neyman, J.（编），第五届伯克利数学统计与概率研讨会论文集，第 1 卷，第 281–297 页。加利福尼亚大学出版社，1967 年。

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. 和 Galstyan, A. 机器学习中的偏差与公平性综述。*ACM* 计算机调查，54(6)，2021年7月。ISSN 0360-0300。doi: 10.1145/3457607。
URLhttps://doi.org/10.1145/3457607。

Nielsen, F. 层次聚类，第195–211页。2016年2月。ISBN 978-3-319-21902-8。doi: 10.1007/ 978-3-319-21903-5 8。

Papamakarios, G.、Pavlakou, T. 和 Murray, I. 用于密度估计的掩码自回归流。收录于 Guyon, I.、Luxburg, U. V.、Bengio, S.、Wallach, H.、Fergus, R.、Vishwanathan, S. 和 Garnett, R.（编），神经信息处理系统进展，第30卷。Curran Associates, Inc.，2017年。网址
https://proceedings.neurips.cc/paper_
f
iles/paper/2017/file/6c1da886822c67822bcf3679d04369fa-Paper
.
pdf.

Papamakarios, G., Nalisnick, E. T., Rezende, D. J., Mohamed, S. 和 Lakshminarayanan, B. 用于概率建模与推断的归一化流。《机器学习研究杂志》，22:57:1–57:64，2019年。

Rezende, D. 和 Mohamed, S. 使用归一化流进行变分推断。载于 Bach, F. 和 Blei, D.（编），第*32*届国际机器学习会议论文集，第37卷，收录于《机器学习研究论文集》，第1530–1538页，法国里尔，2015年7月7日至9日。PMLR。网址：
https://proceedings.mlr.press/v37/rezende15.html。

Rousseeuw, P. J. 轮廓系数：用于解释和验证聚类分析的图形辅助工具。计算与应用数学杂志，20:53–65，1987年。ISSN 0377-0427。doi: https://doi.org/10.1016/0377-0427(87)90125-7。网址：
h
t
tps://www.sciencedirect.com/science/article/pii/0377042787901257。

Runde, V.《拓扑学入门》. Springer, 2007.

Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., Er, M. J., Ding, W., and Lin, C.-T. 聚类技术及其进展综述。神经计算，267:664–681，2017。ISSN0925- 2312。doi: https://doi.org/10.1016/j.neucom.2017.06. 053。
URLhttps://www.sciencedirect.com/science/article/pii/S0925231217311815。

Sculley, D. Web 规模的 k均值聚类。WWW '10，第1177–1178 页，纽约州纽约市，2010 年。计算机协会。ISBN 9781605587998。doi: 10. 1145/1772690.1772862。
URLhttps://doi.org/10.1145/1772690.1772862。

Shi, J. 和 Malik, J. 归一化割与图像分割。*IEEE* 模式分析与机器智能汇刊，22(8):888–905，2000。doi: 10.1109/34.868688。

Skilling, J. 用于通用贝叶斯计算的嵌套采样。贝叶斯分析，1(4):833 – 859，2006。doi: 10.1214/06-BA127。
URLhttps://doi.org/10.1214/06-BA127。

Steinhaus, H. 关于将物质体划分为若干部分。波兰科学院院刊，第三辑，4:801–804，1957。ISSN 0001-4095。

Stimper, V., Sch¨olkopf, B., 和 Miguel Hernandez-Lobato, J. 归一化流的基分布重采样。第*25*届国际人工智能与统计学大会论文集，151:4915–4936，2022年3月28日至30日。网址：
h
t
tps://proceedings.mlr.press/v151/stimper22a.html。

van den Oord, A., Li, Y., Babuschkin, I., Simonyan, K., Vinyals, O., Kavukcuoglu, K., van den Driessche, G., Lockhart, E., Cobo, L., Stimberg, F., Casagrande, N., Grewe, D., Noury, S., Dieleman, S., Elsen, E., Kalchbrenner, N., Zen, H., Graves, A., King, H., Walters, T., Belov, D., 和 Hassabis, D. Parallel WaveNet：快速高保真语音合成。载于 Dy, J. 和 Krause, A.（编），第 *35* 届国际机器学习会议论文集，第 80 卷，《机器学习研究论文集》，第 3918–3926 页。PMLR，2018 年 7 月 10–15 日。网址：
h
ttps://proceedings.mlr.press/v80/oord18a.html。

Wirnsberger, P., Ballard, A.J., Papamakarios, G., Aber-crombie, S., Racani`ere, S., Pritzel, A., Jimenez Rezende, D. 和 Blundell, C. 通过学习映射进行定向自由能估计。《化学物理学杂志》，153(14):144112，2020年。doi: 10.1063/5.0018903。网址：
https://doi.org/10.1063/5.0018903。

Wu, H., K¨ohler, J. 和 No´e, F. 随机归一化流。《神经信息处理系统进展》，33: 5933–5944，2020年。

Zhang, T., Ramakrishnan, R. 和 Livny, M. BIRCH：一种适用于超大型数据库的高效数据聚类方法。《*SIGMOD* 记录》，25(2):103–114，1996年6月。ISSN 0163- 5808。doi: 10.1145/235968.233324。网址：
https://doi.org/10.1145/235968.233324。