超越高斯数据的学习：基于具有表达能力且累积量可控的数据模型研究神经网络的学习动态

¨Onat Ure, Samet Demir, Zafer Do˘gan

工程学院与KUIS人工智能中心，科ç, 大学，伊斯坦布尔，土耳其{oure25、sdemir20、zdogan}@ku.edu.tr

arXiv:2602.02153v1 [stat.ML] 2 Feb 2026

摘要

我们采用一种矩可控非高斯数据模型，研究数据的高阶统计量对神经网络（NNs）学习动态的影响。考虑到双层神经网络的表达能力，我们首先将该数据模型构建为生成式双层神经网络，并使用埃尔米特多项式展开激活函数。这样，在保持数据模型真实性的同时，我们便能通过埃尔米特系数对偏度和峰度等高阶累积量进行可解释的控制。利用从该数据模型生成的样本，我们针对双层神经网络开展了受控在线学习实验。结果揭示了训练过程中逐步深入的矩级学习过程：网络首先捕捉均值和协方差等低阶统计特征，随后逐渐学习高阶累积量。最后，我们在Fashion-MNIST数据集上预训练生成模型，并利用生成的样本开展后续实验。这些补充实验的结果验证了我们的结论，并展示了该数据模型在真实场景中的实用性。总体而言，我们提出的方法弥合了简化的数据假设与实际数据复杂性之间的差距，为研究机器学习和信号处理中的分布效应提供了一个有原则的框架。

索引词— 非高斯数据、高阶统计量、埃尔米特多项式、神经网络、学习动态

1. 引言

现代数据集——从图像、语音到传感器测量数据——呈现出复杂的统计依赖关系。实际数据很少符合高斯分布：多模态性、偏度、重尾和高峰度十分常见，并且会显著影响学习行为。然而，大多数关于泛化的理论分析都依赖高斯假设 [1–4]。这些假设虽在数学上便于处理，却从根本上忽略了影响现实应用性能的高阶结构。

由累积量描述的高阶统计量（作为矩的替代指标，同时也是衡量非高斯性的指标） [5, 6] 与学习动态密切相关。具体而言，理解高阶统计量如何影响泛化已成为当务之急 [7–13]。近期证据表明，采用随机梯度下降训练的神经网络不会同时利用所有统计特征；相反，它们会先与低阶矩（均值和协方差）对齐，然后再利用高阶累积量 [7–10]。若干合成模型试图刻画这一阶段性行为。例如，尖峰累积量模型 [11] 和混合累积量模型 [12]向高斯数据中注入结构化的非高斯性，以便开展受控实验。然而，这些方法从根本上仍有局限，因为它们将特定的非高斯结构硬编码，而未能提供广泛且可调节的控制。

本文通过引入一种通用且便于解析处理的生成框架来对非高斯数据进行建模，从而弥补了这一空白。我们的方法能够精细控制高阶累积量（包括偏度和峰度），并揭示它们对网络泛化的影响。具体而言，我们构建了一个双层生成式神经网络，其中激活函数通过埃尔米特多项式展开 [14]。通过调整埃尔米特系数，我们可以系统地调节非高斯性的程度和类型。这一设计首次实现了对各个累积量及其对学习动态的贡献进行清晰分离，从而弥合了过于简单的基于高斯分布的分析与实践中遇到的丰富分布之间的差距。

我们的主要贡献如下：(i) 我们提出了一个框架，能够独立控制高阶累积量（例如偏度、峰度），同时又具有足够的表达能力来表示广泛的非高斯分布族。这为系统研究机器学习和信号处理中的非高斯效应提供了一种强大的新工具。(ii) 通过在合成数据集和真实世界数据集上进行受控实验，我们表明，在该模型上训练的神经网络呈现出逐矩学习动态，在训练过程中逐步纳入高阶统计量。

我们谨此说明，本研究部分由土耳其科学技术研究委员会（TÜBİTAK）ARDEB 1001计划下的项目124E063资助。O. ¨U. 获得土耳其科学技术研究委员会（TÜBİTAK）硕士奖学金（B¨IDEB 2210）和KUIS人工智能研究中心提供的人工智能研究员奖学金资助。S.D. 获得KUIS人工智能研究中心提供的人工智能研究员奖学金，以及土耳其科学技术研究委员会（TÜBİTAK）博士奖学金（B¨IDEB 2211）资助。通讯作者为Zafer Do˘gan（zdo-gan@ku.edu.tr）。

2. 拟议框架

我们提出了一个框架，用于系统研究高阶累积量对双层神经网络学习动态的影响。其核心是一个生成模型，通过可调的非线性映射引入非高斯性。我们并非随意注入非高斯性，而是将该映射视为一个小型神经模块，其厄米特系数可作为“旋钮”，用来调节偏度、峰度及其他累积量。由此，我们得到一种简单且便于解析的机制，可以开启或关闭高阶效应，并观察其对下游学习的影响。

在实践中，生成模型——例如生成对抗网络（GAN） [15], 归一化流 [16],和扩散模型 [17] ——通常通过将潜在高斯向量 $\mathbf{z}$ 与非线性映射 $\Theta(\mathbf{z})$结合构建，该映射将高斯向量转换为非高斯向量。受此启发，我们采用相同的方法。形式上，设 $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathbb{R}^p$ 为潜在高斯向量。在本文中，我们重点关注以下非线性映射：

$$\Theta(\mathbf{z}; \mathbf{F}) := \sigma(\mathbf{F}\mathbf{z}) \in \mathbb{R}^k, \tag{1}$$

其中$\mathbf{F} \in \mathbb{R}^{k\times p}$ 表示映射的参数，$\sigma:\ \mathbb{R} \to \mathbb{R}$ 为非线性激活函数。于是，生成数据模型可简洁地表示为

$$\hat{\mathbf{x}} := \mathbf{W}\,\Theta(\mathbf{z}; \mathbf{F}) \in \mathbb{R}^d, \tag{2}$$

其中$\mathbf{W} \in \mathbb{R}^{d\times k}$ 表示最后一个线性层。需要注意的是，所有非高斯性均仅来自 Θ。

为了参数化并控制$\Theta(\cdot)$的高阶累积量，我们将 Θ 的每个分量展开为厄米特多项式 [14], ，它们在高斯测度下构成完备的正交基，并作出以下假设。

假设1。 对于$\mathbf{f}_i^\top$ 的每一行$\mathbf{F}$，其分量$\sigma(\mathbf{f}_i^\top\mathbf{z})$ 满足 $\mathbb{E}_{\mathbf{z}}\big[\sigma(\mathbf{f}_i^\top\mathbf{z})^2\big] < \infty.$

在此假设下， Θ 的每个分量都可进行厄米特展开

$$\Theta(\mathbf{z}; \mathbf{F}) = \lim_{\ell\to\infty} \sum_{i=0}^{\ell} c_i\, \mathrm{He}_i(\mathbf{F}\mathbf{z}), \tag{3}$$

其中，$\mathrm{He}_i(\cdot)$ 表示第 i 阶概率论厄米特多项式，而 c_i 是其展开系数。在厄米特展开中，每个系数 c_i 都决定了多项式表征中相关累积量的强度，从而能够直接且直观地控制生成数据的统计结构。

命题 1 （表达能力）。假设（3）成立，且 $p > d$。那么，在标准弱拓扑（例如，*Wasserstein* 距离）下，由（2）定义的数据模型类，在具有有限矩的 d 维概率测度空间中稠密。

证明（概念性说明）。根据通用逼近定理，任何足够正则的函数都可以由双层神经网络 [18] 逼近。因此，由（2）定义的类在具有有限矩的 d 维概率测度空间中稠密（例如，在 Wasserstein 距离下）。由于埃尔米特多项式构成了高斯测度 [14], 下平方可积函数的完备正交基， $\ell\to\infty$ 任意激活函数 σ，因此，任何应用于高斯输入的特征化映射 Θ 都可以展开为（3）的形式，而不会损失表达能力。由此，通过调整埃尔米特系数和网络参数，所诱导的模型可以在选定的弱拓扑／Wasserstein 意义下，逼近任何具有有限矩的 d 维目标。 □

因此，式（2）给出了一个兼具累积量可控性与高度表达能力的数据模型，可用于探究非高斯效应对学习的影响。此外，无需无穷多个系数 $\{c_i\}_{i=0}^{\infty}$ 也能实现精确的累积量控制：

命题 2（有限情况下累积量的可控性 ℓ）。 对于有限的 ℓ，给定 $\ell+1$ 个目标累积量，总能找到 $\ell+1$个能够实现这些累积量的厄米特系数。

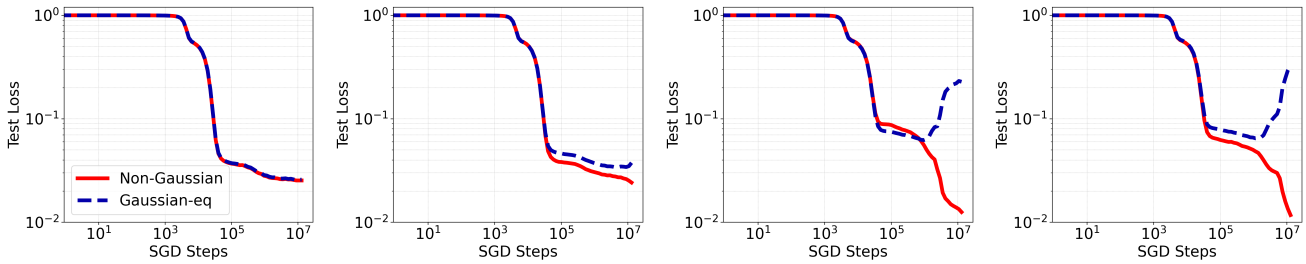
证明。 令 $\sigma_\ell(z) := \sum_{i=0}^{\ell} c_i\mathrm{He}_i(z)$ 表示具有 $\ell+1$ 个系数的截断展开。由于 (2) 中的 $\mathbf{W}$ 起仿射变换的作用，因此只需分析$\sigma_\ell(\mathbf{F}\mathbf{z})$的累积量。不失一般性，考虑标量情形$\sigma_\ell(\mathbf{f}^\top\mathbf{z})$，其中 $\mathbf{f} \in \mathbb{R}^p$。令 κ_i 表示其第 i 阶累积量。于是存在多项式映射 $g_i:\ \mathbb{R}^{\ell+1} \to \mathbb{R}$，使得 $\kappa_i = g_i(c_0, \ldots, c_\ell)$。给定 $\{\kappa_i\}_{i=0}^{\ell}$，求解所得的 $\ell+1$ 个多项式方程组，即可得到所需的系数$\{c_i\}_{i=0}^{\ell}$。因此，通过前 $\ell+1$ 个埃尔米特系数，可以精确设定最高至 ℓ 阶的累积量。□

由于控制 $(\ell+1)$ 阶累积量只需要第 ℓ 阶厄米特展开——即 $(\ell+1)$ 个系数$\{c_i\}_{i=0}^{\ell}$——因此本文余下部分将聚焦于以下有限阶生成模型：

$$\mathbf{x} = \mathbf{W}\Theta_\ell(\mathbf{z}; \mathbf{F}), \tag{4}$$

其中， Θ_ℓ 是由$\Theta_\ell(\mathbf{z},\mathbf{F}) := \sigma_\ell(\mathbf{F}\mathbf{z})$定义的 ℓ阶多项式映射，其中 $\sigma_\ell(z) := \sum_{i=0}^{\ell} c_i\mathrm{He}_i(z)$，其$c_i \in \mathbb{R}$系数用于控制累积量。

通过为 $i \geq 2$ 选择非零系数 c_i，我们明确地将高阶累积量引入数据分布。得益于埃尔米特多项式的正交性，可以独立估计每个累积量，而不受其他多项式项的干扰。尽管每个系数都会影响多个累积量，但这种映射仍然直接且易于解释，从而能够以有理论依据且精细的方式控制分布偏离高斯性的程度。



(a) 不含任何高阶累积量的高斯数据 $(c_2 = c_3 = 0)$

(b) 含有高阶累积量的非高斯数据（从第 2 个系数开始） $(c_2 = 0.2 > 0,\ \ c_3 = 0)$

(c) 含有高阶累积量的非高斯数据（从第 3 个系数开始） $(c_2 = 0,\ \ c_3 = 0.2 > 0)$

(d) 含有高阶累积量的非高斯数据（来自两个系数） $(c_2 = 0.2 > 0,\ \ c_3 = 0.2 > 0)$

图 1：高斯与非高斯二元分类（合成数据）：在一个经过训练、用于区分标准高斯分布数据与非高斯数据（4）的双层神经网络上，非高斯数据集及其具有匹配均值和协方差的高斯等效对应数据（第 3.1 节）的测试损失。对于非高斯数据模型（4），我们设定参数$\mathbf{W} = \mathbf{F} = \mathbf{I}$、$\boldsymbol{\mu} = \mathbf{0}$、$\boldsymbol{\Sigma} = \mathbf{I}_p$，其中厄米特展开的次数为 $\ell = 3$，潜在维度为 $p = 128$，埃尔米特系数取为 $c_0 = 0.4$ 和 $c_1 = 0.5$。所考虑的双层神经网络具有 512 个隐藏单元，并采用 ReLU 作为激活函数。网络使用在线随机梯度下降进行优化，损失函数为均方误差损失，学习率为 10^{-1}。结果取 5 次独立实验的平均值。

3. 实验结果

本节将介绍实验结果，评估我们的累积量可控的非高斯数据模型对泛化的影响。为直接评估高阶累积量的作用，我们在合成数据集和真实数据集上开展了一系列实验。

3.1. 实验设置

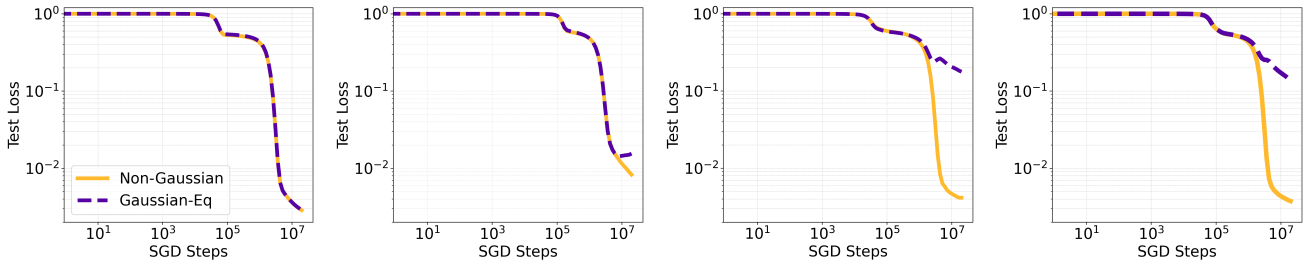
我们通过引入累积量可控数据模型，阐明高阶统计量在学习过程中的作用。该模型使我们能够分离非高斯结构对网络泛化行为的影响。具体而言，厄米特展开系数能够直接且直观地控制生成数据的分布结构，而不依赖于任何特定的下游任务。为展示这一能力，我们在实验中考虑一个二分类问题，其中两个类别的分布各不相同，具体如下：

非高斯数据集 –我们构建了一个包含 $2N$ 个样本的平衡数据集，每个类别有 N 个样本。类别分布如下：第一个类别（标签 -1）由从标准高斯分布中抽取的样本组成，第二个类别（标签 $+1$）则由通过我们的数据模型 (4) 生成的非高斯样本组成。两者的关键区别在于高阶统计量。第一个类别的分布仅具有一阶和二阶结构（由于高斯性，其所有高阶累积量均为零），而第二个类别的分布则具有由非线性数据模型 (4) 引入的丰富高阶依赖关系。**高斯等效评估数据集** – 与此同时，为了剥离高阶统计量在泛化中的作用，我们构建了一个辅助的高斯等效数据集，其中第一个类别保持不变，第二个类别则通过从高斯分布中抽样构造，该分布的一阶和二阶矩与我们的数据模型 (4) 相匹配。通过比较非高斯数据集与高斯等效数据集上的测试性能，我们可以分离高阶累积量对学习动态和泛化的特定贡献。**训练过程** – 在实验中，我们在非高斯数据集上训练神经网络模型，并在非高斯数据集和高斯等效数据集上评估其泛化性能，以展示高阶统计量对学习动态的影响。

3.2. 合成数据实验

为聚焦于非高斯性（完全由非线性函数引起）对学习动态的影响，我们首先进行合成数据实验，其中数据模型 (4) 的参数矩阵设为单位矩阵。具体而言，输入数据由式 (4) 中厄米特展开的三阶截断 ($\ell = 3$) 构造。固定 c_0 和c_1，以保持低阶结构一致，并使用 c_2 和c_3 调节高阶多项式；从而调节偏度和峰度等高阶累积量。

图 1 展示了测试损失随高阶厄米特系数变化的情况。这一结果证实，正如命题 2 所预期的那样，可以通过有限个系数 c_i 控制非高斯性。具体而言，当高阶系数为零时（$c_2 = c_3 = 0$如图 1a 所示），数据模型 (4) 就退化为纯高斯分布。随后，引入非零 c_2（如图 1b 所示）会通过增加偏度并改变峰度来改变分布，从而丰富



(a) $c_3 = c_5 = 0$ (b) $c_3 \neq 0,\ \ c_5 = 0$ (c) $c_3 = 0,\ \ c_5 \neq 0$ (d) $c_3 \neq 0,\ \ c_5 \neq 0$

图 2：高斯与非高斯二元分类（FASHION-MNIST数据）：对于一个经过训练、用于区分标准高斯分布样本与由我们的预训练数据模型（4）生成的非高斯 FASHION-MNIST 样本的双层神经网络，展示非高斯数据集及其均值和协方差匹配的高斯等效对应数据上的测试损失（第 3.1 节和第 3.3 节）。在此实验中，我们仅关注 T恤/上衣类别（标签 0）的样本。为了训练数据模型的参数，我们采用一个生成对抗网络模型，其生成器形式为$\mathbf{W}\tanh(\mathbf{F}\mathbf{z} + \mathbf{b})$，其中仅训练$\mathbf{F}$ 和$\mathbf{b}$ 以生成逼真的 T 恤样本，而$\mathbf{W}$ 设为单位矩阵；判别器则是一个具有最终 Sigmoid 输出的双层 ReLU 网络。生成器和判别器使用 Adam 进行对抗训练，共 100 个周期（学习率 10^{-4},$\beta_1 = 0.9$, $\beta_2 = 0.999$），批量大小为 48，损失函数为二元交叉熵损失。生成对抗网络训练完成后 [15], , 我们对 tanh 应用厄米特展开（截断至次数 $\ell = 5$），并将所得埃尔米特系数与参数$\mathbf{W}$、$\mathbf{F}$、$\mathbf{b}$结合，以得到我们的累积量可控数据模型（4）。对于二元分类，我们采用具有 512 个隐藏单元的双层神经网络，潜在维度为 $p = 100$。模型通过在线随机梯度下降（随机梯度下降）进行优化，使用均方误差损失，学习率固定为 3×10^{-4}。

数据，使其具备超越前两个矩的额外统计结构。相比之下，激活 c_3 （如图 1c-d 所示）会引入反对称分量，因为 He_3 是奇函数。这会注入仅凭均值或协方差无法捕捉的、对符号敏感的高阶依赖性。 c_2 与 c_3 之间由累积量驱动的差异凸显了特定高阶项如何引发不同的学习动态。结果验证了网络最初会捕捉低阶矩，但随着训练推进，会逐渐利用这些高阶统计量。偏度和峰度等非零累积量能够提供高斯等效数据中所不具备的信息，从而提升泛化能力。文献中将这种对高阶统计量的渐进式利用称为“分布简洁性偏差”[7–10], , 这一现象在我们的框架中自然涌现。

3.3. Fashion-MNIST 上的实验

为了证明我们的框架不仅适用于合成场景，也适用于真实世界场景，我们对数据模型 (4) 进行预训练，使其生成来自 FASHION-MNIST 数据集的样本。随后，为了明确分析高阶累积量的作用，我们使用厄米特展开，将非线性函数 $\sigma = \tanh$ 按照 (4) 展开，并截断至五阶 ($\ell = 5$)。由于 $\tanh(\cdot)$ 是奇函数，偶数阶系数为零 ($c_0 = c_2 = c_4 = 0$)。其余系数通过对 $\tanh(\cdot)$ 进行厄米特展开得到，其中 c_1 固定以保持一阶矩，而 c_3 和 c_5 则通过改变来控制高阶累积量。在使用预训练数据模型进行的实验中，我们观察到该预训练模型能够成功生成类似于 Fashion-MNIST 数据集的样本，证实了命题 1 的结论。

对于高斯与非高斯二分类实验（Fashion-MNIST 版本，对应图 1 中的实验），图 2 展示了高阶 Hermite 系数对学习动态的影响。当$c_3 = c_5 = 0$时，变换近似线性，分布仍接近高斯分布。引入非零c_3 会增加偏度，而高阶项则进一步丰富累积量结构。这些新增的矩提供了高斯等价数据中所不具备的统计信息，从而带来更强的泛化效应。与合成数据结果一致，结果证实了通过系数 c_i 控制非高斯性的可行性，以及网络对累积量的渐进式利用。

4. 结论

本研究提出了一个灵活且易于解析处理的框架，用于研究高阶累积量如何塑造学习动态。借助厄米特展开，我们的非线性数据模型能够精确控制累积量，同时保持高度的表达能力。合成数据和真实世界数据上的实验验证了这些特性，并揭示了逐阶矩的学习动态。我们进一步发现，神经网络会按顺序学习累积量：先捕捉低阶矩，再在训练过程中逐渐纳入高阶特征。这些发现凸显了高斯假设的局限性，并为在学习理论中研究非高斯数据分布提供了一种通用工具。

5. 参考文献

[1] Hugo Cui、Luca Pesce、Yatin Dandi、Florent Krzakala、Yue M. Lu、Lenka Zdeborov´a 和 Bruno Loureiro，“双层网络经过一次梯度更新后的特征学习渐近分析”，载于第 *41* 届国际机器学习会议论文集，2024年。

[2] Behrad Moniri、Donghwan Lee、Hamed Hassani 和 Edgar Dobriban，“双层神经网络中通过一步梯度步进行非线性特征学习的理论”，载于第*41*届国际机器学习会议论文集，2024年，《机器学习研究论文集》，第36106–36159页。

[3] Jimmy Ba、Murat A. Erdogdu、Taiji Suzuki、Zhichao Wang、Denny Wu 和 Greg Yang，“特征学习的高维渐近分析：一步梯度步如何改善表征”，载于第*36*届神经信息处理系统国际会议论文集，2022年。

[4] Samet Demir 和 Zafer Dogan，“结构化高斯混合数据下双层神经网络经过一步梯度步后的渐近分析”，载于国际学习表征会议，2025年。

[5] Jonathan Novak，“自由概率的三讲”，载于随机矩阵，第65卷，*MSRI* 出版物，第309–383页。2014年。

[6] Ronald A Fisher 和 John Wishart，《从较简单模式的公式推导双向划分的模式公式》，伦敦数学学会会刊，第 195–208 页，1931年。

[7] Alessandro Ingrosso 和 Sebastian Goldt，《数据驱动的神经网络卷积结构涌现》，美国国家科学院院刊，第 119 卷，第 40 期，第 e2201854119 页，2022 年。

[8]Maria Refinetti、Alessandro Ingrosso 和 Sebastian Goldt，《通过随机梯度下降训练的神经网络学习复杂度不断增加的分布》，载于第 *40* 届国际机器学习会议论文集，2023年，机器学习研究论文集第 202 卷，第 28843–28863 页。

[9] Sandra Nestler、Moritz Helias 和 Matthieu Gilson，《神经架构实现的统计时间模式提取》，物理评论研究，第 5 卷，第 033177 页，2023年。

[10] Nora Belrose、Quintin Pope、Lucia Quirke、Alex Mallen 和 Xiaoli Fern，《神经网络学习复杂性不断增加的统计特征》，arXiv 预印本 arXiv:2402.04362，2024年，arXiv:2402.04362。

[11] Eszter Sz´ekely、Lorenzo Bardone、Federica Gerace 和 Sebastian Goldt，《高效地从高阶相关性中学习：假设检验、随机特征和神经网络》，载于神经信息处理系统进展，2024年，第37卷，第78479–78522页。

[12] Lorenzo Bardone 和 Sebastian Goldt，“沿楼梯滑下：相关潜变量如何加速神经网络的学习”，载于第*41*届国际机器学习会议论文集，2024年，第235卷机器学习研究论文集，第3024–3045页。

[13] Fabiola Ricci、Lorenzo Bardone 和 Sebastian Goldt，“从非高斯输入中进行特征学习：高维独立成分分析的情形”，载于第*42*届国际机器学习会议论文集，2025年，第267卷机器学习研究论文集，第51614–51639页。

[14] Ryan O’Donnell，布尔函数分析，剑桥大学出版社，剑桥，2014年。

[15] Ian J. Goodfellow、Jean Pouget-Abadie、Mehdi Mirza、Bing Xu、David Warde-Farley、Sherjil Ozair、Aaron Courville 和 Yoshua Bengio，“生成对抗网络”，载于神经信息处理系统进展，2014年，第27卷。

[16]Ivan Kobyzev、Simon JD Prince 和 Marcus A Brubaker，《归一化流：当前方法简介与综述》，*IEEE*模式分析与机器智能汇刊，第 43 卷第 11 期，第 3964–3979 页，2020年。

[17] Ling Yang、Zhilong Zhang、Yang Song、Shenda Hong、Runsheng Xu、Yue Zhao、Wentao Zhang、Bin Cui 和 Ming-Hsuan Yang，《扩散模型：方法与应用综合综述》，*ACM*计算调查，第 56 卷第 4 期，第 1–39 页，2023年。

[18] Kurt Hornik、MaxwellStinchcombe 和 HalbertWhite，《多层前馈网络是通用逼近器》，神经网络，1989-1。