丢弃神经网络的信息平面分析

Linara Adilova波鸿鲁尔大学，计算机科学学院 linara.adilova@ruhr-uni-bochum.de

Bernhard C. Geiger Know-Center GmbHgeiger@ieee.org

Asja Fischer
Ruhr波鸿,ers计算机科学学院
asja.fischer@ruhr-uni-bochum.de

arXiv:2303.00596v1 [cs.IT] 1 Mar 2023

摘要

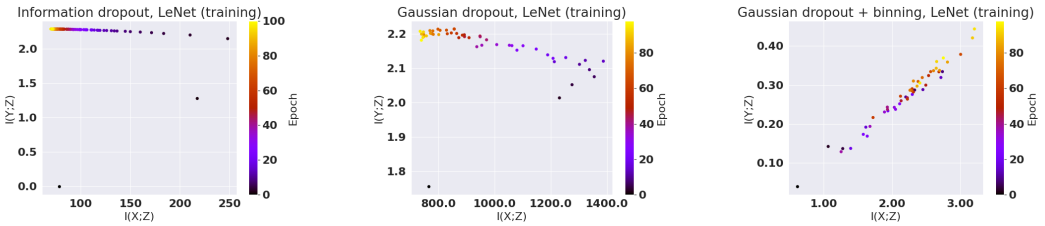
信息论框架有望解释神经网络的预测能力。具体而言，信息平面分析通过测量输入与表征之间，以及表征与输出之间的互信息（MI），应能深入揭示训练过程。然而，研究表明，这种方法高度依赖于互信息估计器的选择。对于确定性网络而言，如果输入与表征之间的互信息为无穷大，这一问题会更加严重。因此，估计值由不同的估计方法决定，却无法充分体现信息论视角下的训练过程。在本研究中，我们证明，加入连续分布噪声的丢弃机制能够确保互信息有限。我们通过一系列实验[1] 证明，这使得对一类在实践中广泛使用的丢弃神经网络进行有意义的信息平面分析成为可能。

1 引言

深度学习中的信息瓶颈假设推测，前馈神经网络的训练分为两个阶段（Shwartz-Ziv 和 Tishby，2017）：拟合阶段和压缩阶段。前者对应于从输入中提取信息并形成已学习的表征，其特征是输入与隐藏表征之间的互信息（MI）增加。后者对应于遗忘对预测目标无用的信息，这体现为已学习表征与输入之间的互信息下降，同时表征与目标之间的互信息保持不变或增加。通过信息平面（IP）分析可以观察这两个阶段，即分析训练过程中输入与表征之间，以及表征与目标之间的互信息如何变化（示例见图1）。有关信息平面分析的概述，请参阅（Geiger，2022）。

信息瓶颈假设优雅且看似合理，但要通过实证进行研究却颇具挑战。正如Amjad 和 Geiger (2020年，定理1）所示，如果输入分布是连续的，那么确定性神经网络所学习的输入与表征之间的互信息为无穷大。因此，标准做法是假设输入分布是离散的（例如，等同于当前数据集的经验分布 S），并通过分箱将实值隐藏表征离散化，以便进行有意义的测量，即避免互信息始终取最大值$\log(|S|)$ (Shwartz-Ziv 和 Tishby，2017年）。然而，在这种离散且确定性的设定下，互信息在理论上等同于表征的香农熵。考虑到分箱的影响，互信息的降低实质上等同于几何压缩（Basirat 等人，2021年）。此外，基于分箱的估计高度依赖于所选的分箱大小（Ross，2014年）。为改用连续输入分布，Goldfeld

[1]实验代码已公开发布于https://github.com/link-er/IP_丢弃。



（a）信息丢弃　　（b）高斯丢弃（我们的方法）　　（c）高斯丢弃（分箱）

图 1：LeNet 网络中采用信息丢弃或高斯丢弃的某一层激活值对应的信息平面。与基于分箱估计互信息的信息平面相比，我们的估计结果（无论是高斯丢弃还是信息丢弃）都清晰地显示出压缩。这表明，即使互信息是有限的，分箱估计器也无法收敛到真实的互信息（另见第 4 节）。

等人(2019年）建议通过向每个隐藏表征添加高斯噪声，将确定性神经网络替换为随机神经网络。这类随机网络在实践中很少使用，因此限制了分析所能带来的洞见。

相比之下，丢弃作为一种随机性来源，因其有效的正则化特性而在实践中被广泛使用。因此，本工作研究的核心问题是：i）我们能否在具有丢弃噪声的神经网络中获得准确且有意义的互信息估计？ii）如果可以，为丢弃网络构建的信息平面是否能够证实信息瓶颈假设？我们的主要贡献回答了这些问题，概括如下：我们提出了一项理论分析，表明由于噪声的离散性质，二元 dropout 并不能阻止互信息变为无穷大。相比之下，我们证明，具有任意连续分布的丢弃噪声不仅会产生有限的互信息，还提供了一种优雅的估计方法。这一点尤其适用于高斯丢弃，已知其对泛化的促进作用甚至强于二元 dropout（Srivastava 等人，2014年），也适用于信息丢弃（Achille 和 Soatto，2018年）。我们通过一系列在基准神经网络和数据集上进行的实验，实证分析了高斯丢弃和信息丢弃设置下互信息估计的质量。尽管我们的结果并未最终证实或否定信息瓶颈假设，但它们表明，使用我们的估计器获得的信息平面与使用分箱估计器获得的信息平面呈现出性质上不同的行为，并强烈表明压缩阶段确实正在发生。

2 神经网络的互信息估计

我们使用以下记号：小写字母表示随机变量（RVs）的取值，例如，b 表示随机变量 B 的一个取值；$H(A)$表示离散随机变量 A 的香农熵，其分布记为 p_a；$h(B)$表示连续随机变量 B 的微分熵，其分布由概率密度函数 p_b 描述；$I(A;B)$表示随机变量 A 和 B 之间的互信息；$X \in \mathcal{X} \subseteq \mathbb{R}^n$ 和 $Y \in \mathcal{Y}$ 分别表示神经网络的输入及其对应的目标；$f(X)$表示输入经过网络前向传播至所关注隐藏层后的结果；Z 是一个描述隐藏表征的 N 维随机变量。

关于度量神经网络的输入 X 与隐藏表示Z 之间互信息的不同方法存在诸多注意事项——例如，对于确定性神经网络和连续输入分布，互信息可能为无穷大；互信息估计值取决于估计器的参数化方式，等等。这些问题已在文献中得到广泛讨论（Saxe 等人，2019年；Geiger，2022），本节将简要回顾这些问题。对于表征 Z 与目标 Y之间的互信息，则不会出现这些问题，因为目标通常是离散随机变量（类别），其互信息始终有限。

估计 $I(X;Z)$ 的一种方法是假设输入来自离散分布。这一观点得到所用计算资源精度有限（Loren- zen 等人，2021年）这一事实的支持，并且便于使用有限数据集 S 来描述该分布。在这种设置下，假设 (X,Y) 在数据集 S上服从均匀分布，并以固定的分箱大小（例如对应于计算机精度的大小）对 Z进行离散化。两者之间的互信息

X 与离散化后的 $\hat{Z}$ 之间的互信息计算为 $I(X;\hat{Z}) = H(\hat{Z}) - H(\hat{Z}|X) = H(\hat{Z}) - 0 = H(\hat{Z})$，其中 $H(Z|X) = 0$，因为 $f(\cdot)$，且对 Z 的离散化是确定性的。因此，输入与表征之间估计得到的互信息对应于离散化表征的熵；当分箱大小较小时，该熵等于数据集上经验分布的熵 $H(X) = \log|S|$，除非 $f(\cdot)$ 将数据集中的不同点映射到潜在空间中的同一点。

一种更符合对真实世界数据常见描述的方案，是假设 X服从连续分布。如果网络变换 $f(\cdot)$ 使表征 Z服从离散分布，那么在样本量足够大的前提下，可以利用分解式 $I(X,Z) = H(Z) - H(Z|X) = H(Z)$，基于香农熵估计互信息（注意，$Z$ 的维度 N 可能很高，因此估计 $H(Z)$ 时可能会遭遇维数灾难）。然而，正如（Amjad 和 Geiger，2020）的定理 1 所示，对于使用常见激活函数的神经网络，潜在表示的分布并非离散分布。在这种情况下（即，$f(\cdot)$ 是确定性的，X 是连续的，而 Z 并非纯粹离散），X 与 Z 之间的互信息是无限的[2]。通过分箱，也就是将 Z 量化为离散随机变量 Z，互信息 $I(X;Z) = H(Z)$仍然有限，但这种熵的定性行为取决于激活函数的性质和所选的分箱大小（Saxe 等人，2019年）。

由上述讨论可知，在确定性神经网络中估计 $I(X;Z)$ 是一个定义不良的问题，而且估计结果揭示的并非信息论图景，往往反映的反而是由所选估计器的属性决定的几何图景。为解决上述挑战，一些作者建议转而研究随机神经网络的信息平面（Amjad 和 Geiger，2020年；Goldfeld 等人，2019年）。Goldfeld 等人(2019年) 提出在训练期间向表征中添加零均值高斯噪声 D。这会将确定性神经网络转变为随机神经网络，而且研究表明，这种做法能够获得与原模型相似的训练结果和预测能力。在 $Z = f(X) + D$ 中添加高斯噪声可确保互信息[3]有限，因此能够使用蒙特卡洛抽样估计互信息，并对估计误差给出界限。此外，这也将信息瓶颈的信息论视角与潜在空间中发生的几何效应联系起来。事实上，当输入与表征之间的互信息降低时，这意味着以不同数据点的表征为中心的噪声诱导高斯分布重叠得更加明显。因此，仅凭表征来区分同一类别的输入会变得更加困难，这会降低表征与输入之间的互信息，同时使表征与目标之间的互信息保持不变。

如上所述，对于连续输入分布，确定性神经网络和含加性噪声的随机神经网络，其信息平面都呈现几何图景（前一种情况下，几何解释是唯一有效的解释，因为互信息是无限的）。因此，在本研究中，我们考察丢弃层网络中的互信息估计，即随机性由乘性噪声而非加性噪声引入的情形。接下来，我们将探究乘性噪声需要满足哪些条件才能使互信息保持有限，以及由此得到的信息平面是否支持信息瓶颈假设。

3 丢弃网络中的互信息

如上一节所述，若假设输入分布是连续的，那么确定性网络的输入与隐藏表征之间的互信息是无穷大。为解决这一问题，必须引入某种随机性。虽然给激活值添加噪声（Gold- feld 等人，2019年）确实可以计算互信息，但大多数当代神经网络并不采用这种方法。相比之下，带有丢弃的神经网络是实践中最常用的神经网络类别之一，其本质上也是随机的：在神经网络中添加一个丢弃层，相当于将隐藏表示与某种形式的随机噪声相乘。形式上，将随机噪声记为与 D 维度相同的随机变量 $f(X)$，则隐藏表示变为 $Z = f(X) \circ D$，其中 $\circ$ 表示逐元素乘法。最基本的形式中，D 服从伯努利分布（Srivastava 等人，2014年）。这种二元 dropout 得到了广泛应用，直观上可以将其理解为训练期间“关闭”一部分神经元。还有

[2]关于互信息为何为无限值，有多种数学推导。例如，Saxe 等人（2019年，附录 C）对此进行了讨论。[3]至少当 p_x 和 $f(\cdot)$ 满足以下条件时 $f(X)$ 具有有限方差，那么根据加性高斯噪声信道容量的相关结果，互信息的有限性即可得出，参见（Cover 和 Thomas，1991年，式（10.17））。

多种其他 Dropout方案，包括乘性高斯噪声、快速Dropout（Wang 和 Manning，2013年）以及变分 dropout（Kingma 等人，2015年）。信息丢弃（Achille 和 Soatto，2018年）是一种变体，它使用互信息的闭式表达式作为正则化项。为得到这样的闭式形式，丢弃噪声从对数正态分布中采样，并根据激活函数（ReLU 或 Softplus）选择表征的先验分布。推导细节见附录 A.1。

在本节中，我们探讨带有 dropout 的神经网络在输入 X 与表征 Z之间是否确实具有有限的互信息。我们首先通过证明二元 dropout 仍会导致 $I(X;Z) = \infty$，给出一个否定结果；而我们的定理 3.3 表明，具有连续分布的 dropout 能使互信息保持有限。这一事实使我们能够在第 4 节和第 5节中估计此类 dropout 神经网络的互信息。

3.1 二元 dropout

我们首先分析二元 dropout，它会以一定概率使单个神经元"关闭"。更正式地说，每个神经元的输出都会乘以一个相互独立的伯努利随机变量，该变量以预先设定的概率 p 取值为 1。下述定理表明，这种（组合式）随机性不足以阻止 $I(X;Z)$ 变为无穷大。

定理 3.1. 在（*Amjad 和 Geiger，2020年，定理 1*）的设定下，设隐藏层的输出 $f(\cdot)$ 由一个具有 N 个神经元的确定性神经网络参数化；设 $B \in \{0,1\}^N$ 为刻画 *dropout* 模式的一组相互独立的伯努利随机变量；并设 $Z = f_B(X)$ 表示应用随机模式 B 后隐藏层的输出。那么有 $I(X;Z) = \infty$。

在证明中（见附录A.2），我们利用了这样一个事实：Dropout掩码 $b = (1, 1, \ldots\ldots, 1)$ 会导致互信息无穷大。伯努利分布保证 $b = (1, 1, \ldots\ldots, 1)$ 始终具有非零概率，而 $\{0,1\}^N$ 上的其他分布可能不具备这一性质。不过，定理3.1 仍可推广至 $\{0,1\}^N$ 上的任意分布：

定理 3.2。 在（*Amjad 和 Geiger，2020年，定理 1*）的设定下，设隐藏层的输出 $f(\cdot)$ 由一个具有 N 个神经元的确定性神经网络参数化，设 $B \in \{0,1\}^N$ 为表征 *Dropout*模式的二元随机向量，并设 $Z = f_B(X)$ 表示应用随机模式 B 后隐藏层的输出。那么，若 *Dropout*模式几乎必然会阻断网络中的信息流，则以下两种情况必有其一：$I(X;Z) = \infty$，或 $I(X;Z) = 0$。

该定理的证明见附录A.3。

定理3.1和定理3.2 都将丢弃仅应用于部分层的情形作为特例，只需将 B 中对应于未应用丢弃的神经元输出的元素设为 1 即可。如果丢弃仅应用于单个层，则 $f_B(X) = f(X) \circ B'$，其中 B' 表示所考虑层的 Dropout模式，$\circ$ 表示逐元素乘积。

根据定理3.2，对于采用二元丢弃的神经网络，任何有限的互信息估计都是"无限错误"的，因此得到的信息瓶颈不允许作信息论解释。本质上，二元丢弃引入的随机性是组合性的，因此无法补偿输入中存在的"连续"随机性 X。

3.2 带连续噪声的丢弃

正如Srivastava 等人 (2014年)所提出的，丢弃也可以通过均值向量为 $\mu = \mathbf{1}$ 、对角协方差矩阵为 $I\sigma^2$ 且方差固定为 σ^2的连续高斯分布噪声来实现。Achille 和 Soatto(2018年)则提出了对数正态分布的丢弃噪声，其方差取决于输入样本 x （这称为信息丢弃）。本节将连续分布的乘性噪声 D作为高斯分布丢弃和信息丢弃的推广。与从离散分布中采样的二元噪声不同，连续分布噪声使 (Z,X) 的联合分布相对于 Z 和 X 的边缘分布绝对连续，从而使输入 X 与隐藏表征 Z之间的互信息具有有限值。下面的定理指出，即使噪声方差取决于输入，输入与丢弃层隐藏表示之间的互信息也确实是有限的。

定理 3.3。 设 X 在所有维度上均有界，$f(\cdot)$ 由具有利普希茨激活函数的确定性神经网络参数化，且令 $Z = f(X) \circ D(X)$，其中各分量

噪声 $D(X) = (D_1(X) \ldots, D_N(X))$ 在给定 X 的条件下相互独立，且其微分熵和二阶矩本质有界，即对某个 M 及所有 $i = 1, \ldots, N$，$\mathbb{E}[D_i(X)^2] \leq M < \infty$ X-几乎处处成立。于是，若条件期望 $\mathbb{E}[\log(|f(X)|) \mid |f(X)| > 0]$ 的每个元素均有限，则有 $I(X;Z) < \infty$。

定理 3.3（证明见附录 A.4）可用于高斯丢弃，其中 D_i（ $D_i \sim \mathcal{N}\ 1\ \sigma^2$ $D_i\ x\ \sim\ \mathcal{N}\ 0\ \alpha^2\ x$ $x\ =$ ），）；对于信息丢弃，其中（）log（，（））。请注意，对于信息丢弃，我们必须确保（学习得到的）方差 $\alpha^2(x)$ 在上下界之间保持有界；例如，在 Achille 和 Soatto（2018年）的实验中，$\alpha^2(x)$ 被限制为低于 0。7。

要求条件期望 $\mathbb{E}[\log(|f(X)|) \mid |f(X)| > 0]$ 在其每个元素中都是有限的，这一点对证明至关重要。事实上，可以构造一个违反该条件的合成（尽管不切实际）示例：

示例3.4。 令 X' 具有如下概率密度函数

$$p_{x'}(x') = \begin{cases} 2^{-n}, & \text{if } x' \in [2^n, 2^n + 1), n = 1, 2, \ldots \\ 0, & \text{else} \end{cases}$$

显然，$\mathbb{E}[X'] = \infty$。于是，$X = \mathrm{e}^{-X'}$ 有界，因为其字母表是(0,e^{-2}的子集].

现在考虑一个只有一个神经元的单隐藏层神经网络。设从 X到该神经元的权重为1，并假设该神经元使用 *ReLU*激活函数。那么，

$$\mathbb{E}[\log|f(X)|] = \mathbb{E}[\log|X|] = \mathbb{E}[\log|\mathrm{e}^{-X'}|] = \mathbb{E}[-X'] = -\infty \ .$$

可以证明，在此例中，X （以及 $f(X)$）的概率密度函数无界。假设 $f(X)$的概率密度函数 p_f 有界，则定理断言中的条件期望是有限的：假设 $p_f \leq C < \infty$，根据无意识统计学家定律，我们有

$$\begin{aligned} \mathbb{E}_x[\log(|f(X)_i|) \mid |f(X)_i| > 0] &= \int_0^{\|f(X)_i\|_\infty} \log(f) p_f(f) \mathrm{d}f \\ &= \underbrace{\int_0^1 \log(f) p_f(f) \mathrm{d}f}_{I_1} + \underbrace{\int_1^{\|f(X)_i\|_\infty} \log(f) p_f(f) \mathrm{d}f}_{I_2} \ . \end{aligned}$$

显然，I_2 为正且有限。由于 p_f 有界，我们还得到 $I_1 \geq C \int_0^1 \log(f) \mathrm{d}f = Cf(\log(f) - 1)|_0^1 = -C > -\infty$。

然而，对于任意神经网络，p_f 的有界性很难保证。相比之下，p_x 的有界性则更现实，也更容易检验。对于有界的 p_x，我们可以证明（见附录 A.5）期望 $\mathbb{E}[\log(|f(X)|) \mid |f(X)| > 0]$ 对于 ReLU 网络是有限的：

命题 3.5。 考虑一个确定性神经网络函数 $f(\cdot)$，该函数由有限层构成，每层的神经元数量有限，并采用 *ReLU* 激活函数。设 X 是一个连续分布随机变量，其概率密度函数为 p_x，该函数有界（$p_x \leq P < \infty$），且其支撑集有界 $\mathcal{X}$。那么，条件期望 $\mathbb{E}[\log(|f(X)|) \mid |f(X)| > 0]$ 的每个元素都是有限的。

最后请注意，定理 3.3 假设网络在所考虑的丢弃层之前是确定性的。对于前馈网络（例如，没有残差连接的网络），这并不损失一般性：事实上，可以将定理 3.3 应用于带有丢弃的第一个隐藏层表示 $Z^{(1)}$，此时该假设总是成立。于是，对于第 ℓ个隐藏层，无论该层是否也有丢弃，互信息 $I(X; Z^{(\ell)})$ 根据数据处理不等式（Cover 和 Thomas，1991年，定理 2.8.1）都是有限的。因此，定理 3.3 保证第一个连续丢弃层之后的所有隐藏层的互信息都是有限的。

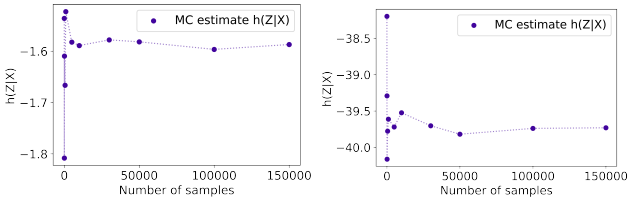
4 连续丢弃条件下互信息的估计

现在，我们考虑如何估计 $I(X;Z)$ 连续分布丢弃网络中的上述量，首先从信息丢弃入手。正如 Achille 和 Soatto (2018年) 所讨论的，采用信息

丢弃的网络使用交叉熵损失 ℓ_{ce}（该损失涉及已知的变分下界 $I(Z;Y) \geq H(Y) - \ell_{ce}$），并使用 $I(X;Z)$ 的变分上界进行正则化。因此，信息平面中显示的各量的估计值会直接用于训练损失，因而易于跟踪，至少对于采用 Softplus 激活函数[4]的网络而言是如此。

对于高斯丢弃，为了估计 $I(X;Z)$ 我们分别近似 $h(Z)$ 和 $h(Z|X)$（伪代码见附录 1中的算法 A.6）。

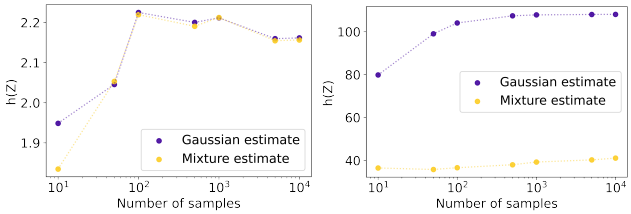
为了估计 $h(Z)$，我们采用蒙特卡罗（MC）估计方法，类似于 Goldfeld 等人（2019年）提出的方法。具体而言，我们将 Z 的分布近似为高斯混合，从中抽取样本 $f(x^{(j)})$ $j = 1, \ldots, |S|$，并在每个样本 $f(x^{(j)})$ 上放置具有对角协方差矩阵的高斯分布，其方差为 $\sigma^2 |f(x^{(j)})_i|^2, i = 1, \ldots, N$。作为合理性检查，我们还计算 $h(Z)$ 的一个上界，该上界由与 Z 具有相同协方差矩阵的高斯分布的熵给出。请注意，估计上界要求样本数量足够大，以确保样本协方差矩阵非奇异，且由此得到的熵估计值有限。



(a) $\sigma = 0.1, n = 2$ (b) $\sigma = 0.1, n = 50$

图 2：互信息的蒙特卡洛估计 $h(Z|X)$不受维度影响，并且随着样本数量增加而趋于稳定。

对于每个固定的 x，条件分布 $p_{z|x}$ 是一个高斯分布 $\mathcal{N}(f(x), \text{diag}(\{\sigma^2 |f(x)_i|^2\}))$。此外，当输入固定时，$Z|X = x$ 的各分量相互独立，因为噪声的各分量相互独立。因此，可以将 $h(Z|X)$ 计算为 $h(Z_i|X)$ 之和，其中 Z_i 是表征向量的第 i 个分量。计算 $h(Z_i|X)$ 需要在输入空间上积分，以计算数学期望 $\mathbb{E}_x[h(Z_i|X = x)]$。这可以通过蒙特卡罗采样来近似。也就是说，我们用 $1/|S| \sum_{=1}^{|S|} h(Z_i|X = x^{(j)})$ 近似 $h(Z_i|X)$，其中



(a) $\sigma = 0.1, n = 1$ (b) $\sigma = 0.1, n = 50$

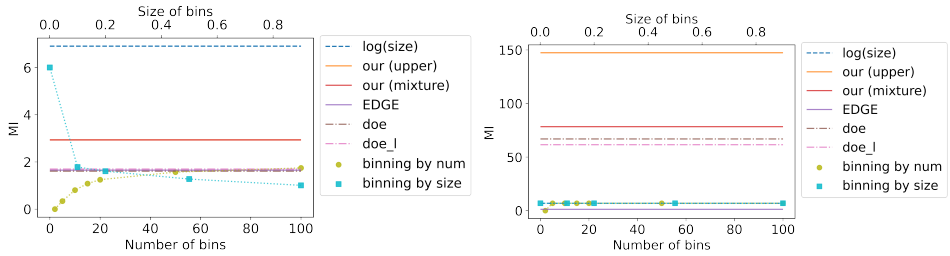
图 3：隐藏表示 Z 的微分熵 $h(Z)$ 估计值。随着 X 的维度增加，与基于高斯混合的蒙特卡洛估计相比，高斯上界变得非常宽松。

$$h(Z_i|X \overset{j}{=} x^{(j)}) = \log(|f(x^{(j)})_i| \sigma \sqrt{2\pi e}).$$

我们考虑一个简单的玩具问题，以验证我们估计互信息的方法：输入 X 由一个 h $= 2X + 0.5$ $\mathcal{N}$ 1 σ^2 $=$ 维标准正态分布生成，并通过函数 $f(X)$ 进行变换，然后施加服从$(,)$分布的高斯丢弃。我们研究样本数量增加时，估计器对 $h(Z|X)$ 的收敛性。对于每个输入数据点，我们生成10 个噪声掩码，从而获得10 个 Z 样本，对应每个 $x^{(j)}$。图2 中的结果表明，对于不同数据维度，随着样本数量增加，估计结果趋于稳定。我们还在图 $h(Z)$ 3中将估计值与上界进行比较。

最后，我们将互信息的估计结果与分箱、EDGE 估计器（Noshad 等人，2019年）以及 McAllester 和 Stratos（2020年）分析的下界进行比较。结果如图 4所示。图中，doe 表示熵差（DoE）估计器，doe l 表示采用逻辑斯谛参数化的 DoE（McAllester 和 Stratos，2020年）。当分箱大小较大时，分箱估计器会低估

[4]事实上，对于 Softplus 激活函数，$I(X;Z)$ 的变分近似具有闭式表达；而对于 ReLU 激活函数，可用的表达式仅适用于最小化，而非计算 $I(X;Z)$ (参见附录A.1)。



(a) $\sigma = 0.1, \quad n = 1$ (b) $\sigma = 0.1, \quad n = 50$

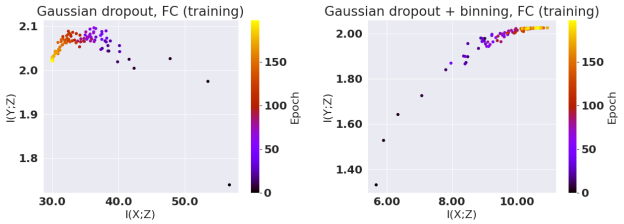
图 4：对于含乘性高斯噪声的玩具示例，不同互信息估计方法的比较。对于低维 X 和 Z，不同的分箱大小会导致分箱估计器得出不同的互信息估计值。对于更高维度，基于分箱的估计值会崩溃。我们的估计值与 McAllester 和 Stratos (2020年) 提出的下界估计非常接近，同时如预期一样仍然更大。

互信息；而分箱大小较小时则会高估互信息（Ross，2014年）。这一点在图中可以清楚地看到：分箱既按大小（上方坐标轴）排列，也按数量（下方坐标轴）排列。此外，对于高维数据，分箱很快就会达到 $\log(|S|)$这一可能的最大值，无法达到更大的互信息值。根据 McAllester 和 Stratos （2020年），基于下界的互信息估计器（例如 MINE（Belghazi 等人，2018年））要想准确逼近互信息的真实值，也需要数量随真实互信息值呈指数增长的数据点；否则，它们总会严重低估互信息。

Furt不同丢弃方差和输入维度下的其他图示见附录 A.6。

5 丢弃网络的信息平面分析

我们使用上一节所述的估计量，对采用高斯丢弃和信息丢弃的网络进行信息瓶颈分析。我们始终只考虑与第一个丢弃层对应的表征 [5]，并以纳特为单位测量互信息，例如使用自然对数。对于估计I（Y； Z），我们对高斯丢弃采用 EDGE 估计器（Noshad 等人，2019年），对信息丢弃采用变分估计器。使用分箱估计器生成的信息平面，则对 I（X； Z）和 I（Y； Z）均使用分箱方法。



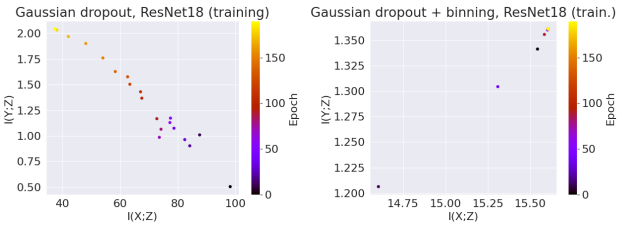
（a）我们的估计器 （b）分箱估计器

图 5：在 MNIST 上训练的采用高斯丢弃的全连接网络的信息平面。与互信息的分箱估计相比，我们的方法显示出压缩现象。

在第一组实验中，我们研究通过所提出的估计器获得的信息平面之间的差异以及通过分箱进行分析。针对 MNIST 数据集的分析采用了 LeNet 网络（LeCun 等人，1998年），其准确率达到 99% ；此外，还采用了一个包含三个隐藏层（$28\times28-512-128-32-10$）和 Softplus 激活函数的简单全连接网络（FC 网络），其准确率达到 97%。我们分析了 LeNet 网络中的信息丢弃和高斯丢弃，以及 FC 网络中仅使用高斯丢弃的情况。两种情况下，丢弃都应用于倒数第二层。图 1 和图 5展示了基于分箱估计器的信息平面与基于我们估计器的信息平面的比较。

[5]这使互信息估计更加高效，因为网络的前一部分是确定性的，从而可以得到 $h(Z|X = x)$ 的解析表达式。不过需要注意的是，由于在更高层中互信息同样保持有限，因此估计也可以扩展到更高层。然而，对于这些层，应使用不同于我们的估计器。

我们还分析了在 CIFAR10 上训练的 ResNet18 的信息平面（见图 6）。我们在输出层之前额外添加了一个包含128 个神经元的瓶颈层和高斯丢弃，其准确率达到 94%。

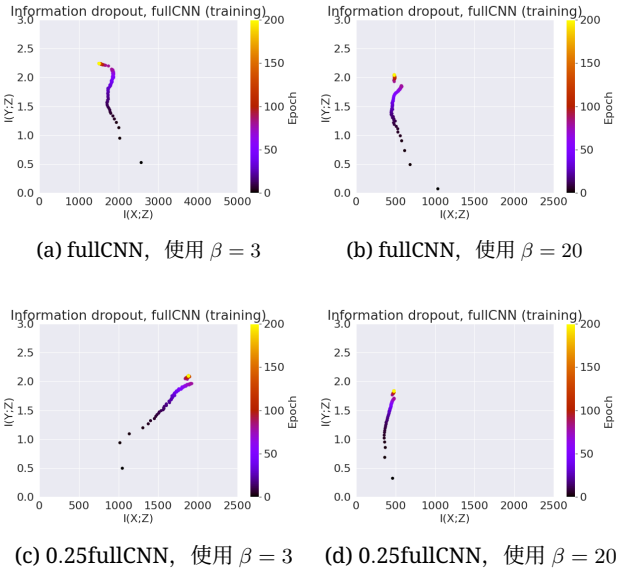


(a) 我们的估计器　　(b) 分箱估计器

图 6：使用高斯丢弃并在 CIFAR10 上训练的 ResNet18 网络的信息平面。与基于分箱的互信息估计器不同，我们的方法清楚地显示了压缩。

有趣的是，对于所有网络和数据集，我们都观察到使用我们的估计器时存在显著的压缩，而使用分箱估计器时则没有压缩（不同分箱大小的情况也如此，见附录 A.8）。这表明，要么丢弃网络中测得的互信息压缩不同于纯粹的几何压缩，要么样本数 $|S|$ 不足以通过分箱可靠地估计 $I(X;Z)$。

在第二组实验中，我们分析了信息丢弃网络中的信息平面，互信息估计方法如前所述。为此，我们使用 Achille 和 Soatto(2018年提供的代码，在 CIFAR10 上训练了一个全卷积神经网络（fullCNN）。训练过程持续了200 个训练轮数，采用带动量的随机梯度下降；与原始设置不同的是，仅在第三个卷积层之后设置一个丢弃层。批量大小设为100，学习率初始设为0.05，并在0.1 后乘以该系数进行衰减，衰减节点为40、80和120训练轮数。网络采用不同的正则化权重进行训练 β，卷积层中的滤波器数量也各不相同。具体而言，完整尺寸的 fullCNN 包含3 层，每层有96 个滤波器，后接4 层，每层有192 个滤波器；而小型网络仅由这些滤波器的25%构成。此外，与原始设置不同，我们允许噪声方差增大至0.95，以观察有限的



(a) fullCNN，使用 $\beta = 3$　　(b) fullCNN，使用 $\beta = 20$

(c) 0.25fullCNN，使用 $\beta = 3$　　(d) 0.25fullCNN，使用 $\beta = 20$

图 7：信息平面显示，根据 β，输入与表征之间的互信息会出现较多（a）、（c）或较少（b）、（d）压缩。较小网络（c）和（d）的 $I(X;Z)$ 值更小。

表征与输入之间的信息更加显著。结果如图 7 所示。可以看出，对 $I(X;Z)$ 进行正则化是有效的（即，较大的 β 会带来较小的 $I(X;Z)$），而过强的正则化（$\beta = 20$）会导致性能变差：测试误差高出 5%，训练误差高出 10%。我们还可以看到，β 较小时压缩更强，而 β 较大时几乎没有压缩。我们推测，只有当允许足够多的信息流经网络时，压缩才会显现出来（这只会在 β 较小时发生）。图 7 (c) 和 (d) 展示了小型 fullCNN 的信息平面。可以看出，较小的网络似乎完全没有压缩（见图 7 (c)），但 $I(X;Z)$ 在整个训练过程中反而持续增加，直到达到图 7 (a) 中的相同水平。这表明，β 决定了信息平面中的信息压缩到何种程度，而训练过程中经过的信息平面曲线取决于神经网络的整体容量。

附加实验的图表可见附录A.8。

6 讨论

信息论压缩是否与泛化能力的提升相关，是与深度神经网络信息平面分析相关的核心问题，也是支持这种分析最主要的理由。然而，只有在互信息有限、因而可以测量的神经网络中，才能检验这种关联。在理论分析中，我们研究了在连续输入分布这一假设下，不同形式的丢弃噪声是否能使互信息保持有限。我们对此给出了肯定的回答：通过证明，对于诱导分布满足特定约束的表征的网络，具有有限微分熵的连续丢弃噪声可以防止 $I(X;Z)$ 变为无穷大。我们还证明，如果输入的概率密度函数有界，那么 ReLU网络中的表征分布就满足这些约束。

基于上述结论，我们提出了一种用于估计高斯丢弃网络中互信息的基于蒙特卡洛的估计方法，并对不同数据集上的高斯丢弃网络和信息丢弃网络进行了信息平面分析。实验表明，分箱估计器的表现与我们的估计器截然不同：我们的估计器在信息平面中大多呈现压缩，而分箱估计器则没有。此外，对于我们的估计器，$I(X;Z)$ 的数值通常比 $I(Y;Z)$ 的数值大几个数量级，尤其是与分箱估计器相比时。假设所提出的估计器足够准确，那么这就使信息论压缩与泛化之间是否存在关联变得值得怀疑。虽然这些初步实验尚不能确切回答这种关联是否存在，但它们揭示了一种具有实际意义的情形，可据此研究这种相关性。

分箱估计器与我们的估计器之间的差异进一步表明，我们使用自己的估计器观察到的信息论压缩要么并非几何压缩，要么是由于样本不足，无法通过分箱估计器获得可靠的估计。这与Goldfeld等人（2019年）的研究形成对比；他们表明，在其带有加性噪声的网络中，信息论压缩与几何压缩存在关联。因此，我们认为，深入研究乘性噪声是否会诱发几何压缩，以及由此产生的压缩是否能提升泛化性能，是未来研究中值得关注的问题。

致谢

The 作者感谢 Michael Kamp、Simon Damm、Ziv Goldfeld 和 Jihao Andreas Lin 提供 r valu了关于本研究的宝贵讨论。

Asja Fischer 感谢德国研究基金会（DFG）在德国卓越战略——EXC 2092 CASA——项目（编号：390781972）下提供的支持。

Know-Center 获得奥地利COMET计划（卓越技术能力中心计划）的资助，该计划由奥地利联邦气候行动、环境、能源、交通、创新和技术部、奥地利联邦数字和经济事务部以及施蒂利亚州共同支持。COMET 由奥地利研究促进署FFG 管理。

参考文献

Alessandro Achille 和 Stefano Soatto。信息丢弃：通过带噪计算学习最优表征。*IEEE*模式分析与机器智能汇刊，40 (12):2897–2905，2018年。2，4，5，8，11

Rana Ali Amjad 和 Bernhard C. Geiger。使用信息瓶颈原理学习神经网络分类的表征。*IEEE*模式分析与机器智能汇刊，42(9):2225–2239，2020年9月。doi: 10.1109/TPAMI.2019.2909031。开放获取： arXiv:1802.09766 [cs.LG]。1，3，4， 12

Mina Basirat、Bernhard C. Geiger 和 Peter M. Roth。信息平面分析的几何视角。*Entropy*，23(6):711，2021年。1

Mohamed Ishmael Belghazi、Aristide Baratin、Sai Rajeshwar、Sherjil Ozair、Yoshua Bengio、Aaron Courville 和 Devon Hjelm。互信息神经估计。载于 ICML 会议录，第 531–540 页。PMLR，2018年。7 Thomas M. Cover 和 Joy A. Thomas。信息论基础。纽约：Wiley，1991年。3, 5, 13 Bernhard C. Geiger。关于神经网络分类器的信息平面分析：综述，2022年12月。开放获取：arXiv:2003.09671 [cs.LG]。1, 2 Ziv Goldfeld、Ewout van den Berg、Kristjan H Greenewald、Igor Melnyk、Nam Nguyen、Brian Kingsbury 和 Yury Polyanskiy。估计深度神经网络中的信息流。载于 ICML 会议录，2019年。1, 3, 6, 9 Durk P Kingma、Tim Salimans 和 Max Welling。变分 dropout 与局部重参数化技巧。神经信息处理系统进展，28，2015年。4 Yann LeCun、L´eon Bottou、Yoshua Bengio 和 Patrick Haffner。应用于文档识别的基于梯度的学习。IEEE会议录，86(11):2278–2324，1998年。7 Stephan Sloth Lorenzen、Christian Igel 和 Mads Nielsen。信息瓶颈：对（量化）神经网络的精确分析。arXiv 预印本 arXiv:2106.12912，2021年。2 David McAllester 和 Karl Stratos。互信息测量的形式化限制。载于 AISTATS 会议录，第 875–884 页。PMLR，2020年。6, 7 Morteza Noshad、Yu Zeng 和 Alfred O Hero。利用依赖图进行可扩展互信息估计。载于 ICASSP 会议录，第 2962–2966 页。IEEE，2019年。6, 7 Brian C. Ross。离散和连续数据集之间的互信息。PLOS ONE，9(2):e87357，2014年。1, 7 Andrew M Saxe、Yamini Bansal、Joel Dapello、Madhu Advani、Artemy Kolchinsky、Brendan D Tracey 和 David D Cox。关于深度学习的信息瓶颈理论。统计力学杂志：理论与实验，2019(12):124020，2019年。2, 3 Ravid Shwartz-Ziv 和 Naftali Tishby。通过信息揭开深度神经网络的黑箱。arXiv 预印本 arXiv:1703.00810，2017年。1 Nitish Srivastava、Geoffrey Hinton、Alex Krizhevsky、Ilya Sutskever 和 Ruslan Salakhutdinov。丢弃：一种防止神经网络过拟合的简易方法。机器学习研究杂志，15(1):1929–1958，2014年。2, 3, 4, 15 Sida Wang 和 Christopher Manning。快速 dropout 训练。载于 ICML 会议录，第 118–126 页。PMLR，2013年。4

A 附录

A.1 信息丢弃

一种使用连续噪声的丢弃方式称为信息丢弃（Achille 和 Soatto，2018年）。这是一种将从对数正态分布中采样的丢弃噪声结合起来的技术$\epsilon \sim p_\epsilon = \log \mathcal{N}(0, \alpha_\theta^2(x))$，其中 $\alpha_\theta(x)$ 是一个依赖于网络参数的可学习参数 θ，并引入正则化项 $KL(p_{z|x_i}||\prod_{i=1}^{|Z|} p_{z_i})$。该正则化项基于用于训练神经网络的信息瓶颈目标：重写信息瓶颈拉格朗日函数并加入解耦项（即，我们希望表征 Z 的每个元素都独立于其他元素），即可得到上述公式。此外，研究者提出使用由激活函数（ReLU 或 Softplus）的选择所定义的先验分布，并采用一种经过实证验证有效性的特定分布。这样的先验分布与选定的丢弃噪声可用于推导 KL 散度的闭式形式，从而便于在训练过程中直接跟踪信息平面值。

下面给出 $I(X;Z)$ 的闭式计算方法，该方法由Achille 和 Soatto (2018年)提出。

$$\begin{aligned} I(X;Z) = KL(p_{x,z}||p_z p_x) &= \int p_{x,z}(x,z) \log\left(\frac{p_{x,z}(x,z)}{p_z(z)p_x(x)}\right) \mathrm{d}x\mathrm{d}z \\ &= \int p_x(x)p_{z|x}(z) \log\left(\frac{p_x(x)p_{z|x}(z)}{p_z(z)p_x(x)}\right) \mathrm{d}x\mathrm{d}z = \int p_x(x) KL(p_{z|x}||p_z)\mathrm{d}x \\ &= \mathbb{E}_x[KL(p_{z|x}||p_z)] \ . \end{aligned}$$

在经验上，我们可以将其近似为 $I(X;Z) = \sum_{j=1}^{|S|} KL(p_{z|x^{(j)}}||p_z)$，其中对大小为 $|S|$ 的 X 样本数据集求和。

首先，我们讨论 ReLU神经网络。在这种情况下，先验分布 p_z 由两部分混合而成：非正规对数均匀分布和位于 0 处的点质量。对于 ReLU 激活值而言，这种先验在经验上是有效的。首先，我们将推导限定在 $f(X) \neq 0$ 的情形（这也意味着 $Z \neq 0$，因为噪声 ϵ 是对数正态分布，不可能为 0）。下文中，当概率密度函数的自变量足以明确其含义时，我们将省略其下标。

$$\begin{aligned} KL(p_{z|x^{(j)}}||p_z) &= KL(p_{\log(z|x^{(j)})}||p_{\log(z)}) && (1) \\ &= \int p(\log(z|x^{(j)})) \log\left(\frac{p(\log(z|x^{(j)}))}{p(\log(z))}\right) \mathrm{d}z \\ &= \int p(\log(\epsilon) + \log(f(x^{(j)}))|x^{(j)}) \log\left(\frac{p(\log(\epsilon) + \log(f(x^{(j)}))|x^{(j)})}{c}\right) \mathrm{d}\epsilon && (2) \\ &= \int p(\log(\epsilon)) \log(p(\log(\epsilon)))\mathrm{d}\epsilon - \int p(\log(\epsilon)) \log(c)\mathrm{d}\epsilon && (3) \\ &= \int p(\log(\epsilon)) \log(p(\log(\epsilon)))\mathrm{d}\epsilon - \log(c) = -h(\log(\epsilon)) - \log(c) && (4) \\ &= -(\log(\alpha(x^{(j)})) + \frac{1}{2}\log(2\pi e)) - \log(c) \ , && (5) \end{aligned}$$

其中，方程 1 成立是因为 KL散度 在严格单调函数（log(·)）下进行参数变换时具有不变性；方程 2 成立是因为 $\log(Z) = \log(\epsilon) + \log(f(X))$，且 $p_{\log(z)} = c$ 对于非正规对数均匀分布成立；方程 3 考虑了$p_{x+const} = p_x$、$\log(f(x))|x^{(j)}$ 是常数这一点，以及$p_{\log(\epsilon)|x^{(j)}} = p_{\log(\epsilon)}$，因为 ϵ 独立于 X；方程 4 利用了 $\int p_{\log(\epsilon)}\mathrm{d}\epsilon = 1$；最后，方程 5 成立是因为 $\log(\epsilon)$ 服从正态分布，其熵可以用闭式计算。

现在我们将 $f(X) = 0$，并且还得到 $Z = 0$。然后 $p_{Z|X} = \delta_0$（点质量或狄拉克δ函数），互信息变为：

$$KL(p_{z|x^{(j)}}||p_z) = \int p_{z|x}(z) \log\left(\frac{p_{z|x}(z)}{p_z(z)}\right) \mathrm{d}z = \int \delta_0 \log\left(\frac{\delta_0}{q\delta_0}\right) \mathrm{d}z = -\log(q) \ , \quad (6)$$

其中， q 是先验分布 p_z中点质量的权重。

结合公式 5 和公式 6 可得到一个可计算的 $I(X;Z)$。如图所示，必须正确组合 $f(X)$ 的非零值和零值，并且还需知道先验分布 p_z的参数：常数 c 和权重 q。这使得它不适用于信息平面分析。

如果网络使用的不是 ReLU 而是 softplus 激活，那么表征分布的先验分布将是标准对数正态分布，而非带有狄拉克 δ 分布的对数均匀分布。在这种情况下，计算非常简单，因为两个对数正态分布之间的 KL 散度可作为相应正态分布之间的 KL 散度来计算：

$$KL(p_{z|x^{(j)}}||p_z) = \frac{1}{2\sigma^2}(\alpha^2(x^{(j)}) + \mu^2) - \frac{\log(\alpha(x^{(j)}))}{\sigma} - \frac{1}{2}\ , \tag{7}$$

其中， $\sigma^2 = 1$和 $\mu = 0$ 是先验分布的已知参数。因此，softplus 激活（公式 7）可以直接计算 $I(X;Z)$。

A.2 定理 3.1 的证明

证明。利用互信息链式法则，我们有

$$\begin{aligned} I(X;Z) = I(X;Z,B) - I(B;X|Z) = I(X;Z|B) + I(B;X) - I(B;X|Z) \\ \geq I(X;Z|B) - H(B) \end{aligned}$$

其中，不等式是由于去掉了 $I(B;X)$，因为 B 和 X 相互独立，并利用了 $I(B;X|Z) \leq HB$这一事实得到的。由于 $B \in \{0,1\}^N$是离散随机变量，立即可得$H(B) \leq N\log 2$。现在注意到

$$I(X;Z|B) = \sum_{b\in\{0,1\}^{\hat{N}}} \mathbb{P}(B=b) I(X;Z|B=b).$$

由于伯努利随机变量相互独立，因此以下情况被赋予了正概率质量： $b = (1,\ 1,\ \ldots\ldots,\ 1)$，即所有神经元均处于激活状态的情况。显然，当 $b = (1,\ 1,\ \ldots\ldots,\ 1)$ 时，可得$Z = f(X)$。因此，根据（Amjad 和 Geiger，2020年，定理1）

$$I(X;Z|B) \geq \mathbb{P}(b = (1,1,\ldots,1)) I(X;f(X)) = \infty$$

且 $I(X;Z) = \infty$。 □

A.3 定理3.2的证明3.2

证明。 如果二元 dropout 使得丢弃掩码 $b = (1,\ 1,\ \ldots\ldots,\ 1)$被赋予非零概率，则该定理的结论可按照定理3.1的证明得到。

现在假设 B使得赋予 $b = (1, 1, \ldots, 1)$ 的概率质量为零。为处理这种情况，我们假设 X 的分布有一部分在某个紧集上具有连续概率密度函数，并且神经网络的激活函数要么是双 Lipschitz 函数，要么是导数严格为正的连续可微函数（遵循 Amjad 和 Geiger（2020，定理 1）的要求）。那么，对于神经网络几乎所有的参数化，根据（Amjad and Geiger，2020，定理 1），我们有 $I(X;\ f(X)) = \infty$。在此设定下， $f_B(X)$仍是一个神经网络，其激活函数要么是双 Lipschitz 函数，要么是导数严格为正的连续可微函数。假设 b 使得网络的输入与所考虑的层并非完全断开，对于这种模式，我们有 $I(X;Z|B = b) = \infty$。否则，显然有 $I(X;Z|B = b) = 0$。对所有模式 b取期望，即可得到定理的结论。 □

A.4 证明定理 3.3

证明。 不失一般性，我们首先只关注表征中非零的维度。具体而言，假设 $Z = (Z_1, \ldots, Z_N)$，且 $B = (B_1, \ldots, B_N)$，其中当 $Z_i = 0$ 时 $B_i = 0$，否则 $B_i = 1$。显然， B 是 Z 的函数，因此 $I(X;Z) = I(X;Z,B) =$

$I(B;\ X)\ +\ I(Z;X|B)$。由于 B 是二元的，我们有 $I(X;\ B)\ \leq\ H(B)\ \leq\ n\log 2$。令 $Z_B\ =(Z_i|i:$ $B_i\ =\ 1)$ 表示 Z 中非零元素构成的子向量，则

$$I(X;Z)\leq n\log 2+\sum_b \mathbb{P}(B=b)I(Z_b;X)$$

其中，若 $B=b, I(Z_b;X)\ =\ I(Z;X|B=b)$ 成立，是因为常数（即取值为 0）的随机变量不会对互信息产生贡献。因此， $I(X;\ Z)$ 有限，当且仅当 $I(Z_b;\ X)\ =\ I(Z;X|B\ =\ b)$ 几乎处处（关于 B）有限。于是，我们现在固定任意一个 $B\ =\ b$，并继续证明 $Z\ =\ Z_b$。

我们将互信息分解为微分熵： $I(X;Z)\ =h(Z)-h(Z|X)$。表征 $h(Z)$ 的微分熵上界为具有相同协方差矩阵 Σ 的高斯随机变量的熵，其中该协方差矩阵与 $Z\ =(Z_1,\ \ldots\ldots,\ Z_N)$ 的分布相同，即不超过 $N/2\log(2\pi)\ +1/2\log(\det(\Sigma))$ $+N/2$ Σ $\Sigma\ \leq$。由 Hadamard 不等式，$\prod_{i=1}^n\sigma_{ii}^2$ σ_{ii}^2 $\sigma_{ii}^2=Var[Z_i]$ 且由于其为半正定矩阵，可得 det( )，其中 为协方差矩阵的对角元素，即 。该方差可以从上方界定。具体而言，由于 X_i 有界，且 $f(\cdot)$ 是 Lipschitz函数的复合，因此 $f(X)_i$ 也有界。回顾 $\mathbb{E}[D_i(x)^2]\ \leq\ M$ 几乎处处（关于 X）成立，可得

$$\begin{aligned}Var[Z_i]&\leq\mathbb{E}_x[f(X)_i^2D_i(X)^2]=\mathbb{E}_x[f(X)_i^2\mathbb{E}_d[D_i(X)^2\mid X]]\\&\leq M\mathbb{E}_x[f(X)_i^2]\leq M\|f(X)_i\|_\infty^2\end{aligned}$$

接下来需要证明 $h(Z|X)\ >\ -\infty$。由于给定 X时 D_i 与 D_j条件独立 X，对于所有 $i\neq j,\ Z$ 的条件微分熵可分解为其各个分量的条件微分熵之和，即 $h(Z|X)=\sum_{i=1}^N h(Z_i|X)$。我们将这个条件熵写成关于 X 的期望，并利用（Cover 和 Thomas，1991年，定理 9.6.4）得到

$$\begin{aligned}h(Z_i|X)&=\mathbb{E}_x[h(Z_i|X=x)]=\mathbb{E}_x[h(D_i(x)|f(x)_i||X=x)]\\&=\mathbb{E}_x[h(D_i(x)|X=x)]+\mathbb{E}_x[\log(|f(X)_i|)]\end{aligned}$$

根据微分熵的变量变换公式。根据定理的断言，这两项均为有限值。第一项有限，因为我们假设$D_i(X)$的微分熵本质有界，即存在一个数 $C<\infty$，使得 $h(D_i(x))\leq C$ X几乎处处成立。第二项有限，因为我们假设条件期望 $\mathbb{E}[\log(|f(X)|)\ |f(X)|>0]$ 对其每个元素均有限，并且由于 $Z_i\neq 0$ 蕴含 $|f(X)_i|>0$。证明至此完成。

□

A.5 命题3.5 的证明

证明。 不失一般性，我们假设 $f(\cdot)$ 的值域维数为 $D\ =\ 1$，即 $f\colon\ \mathcal{X}\ \to\ \mathbb{R}$，其中 $\mathcal{X}\subseteq\mathbb{R}^n$ 是函数定义域。该证明可以直接推广到 $f(\cdot)$ 的多个维度。

由于 $f(\cdot)$ 是使用有限规模的、具有 ReLU 激活函数的神经网络构造的，因此它在函数定义域的有限划分上是分段仿射线性的。由 $\mathcal{X}$ 有界，进而 $|f(\mathcal{X})|$ 有界这一事实，可立即得出 $\mathbb{E}[\log(|f(X)|)\ |f(X)|>0]<\infty$。

为了考察 $\mathbb{E}[\log(|f(X)|)\ \mid\ |f(X)|\ >\ 0]\ >\ -\infty$ 是否成立，将定义域 $\mathcal{X}$ 划分为以下各部分：

1. $\mathcal{X}_0=f^{-1}(\{0\})$ 表示 $f(X)$ 为零的那部分；2. $\{\mathcal{X}_i^c\}_{i=1、\ldots\ldots、\ell}$ 表示 $\mathcal{X}$ 的划分中的各部分，在这些部分上 $f(X)\ =\ c_i$，也就是说，$f(\cdot)$ 为常数；3. $\mathcal{X}^a=\bigcup_{i=1}^m\mathcal{X}_i^a$ 表示划分中所有其他集合 $\{\mathcal{X}_i^a\}_{i=1、\ldots\ldots、m}$ 的并集，在这些集合上 $f(\cdot)$ 不是常数。

~ ~对于最后一个子集，定义函数 $f\colon\ \mathcal{X}^a\ \to\mathbb{R}^n$，其定义方式为 $f(x)\ =$ $(|f(x)|,\ x_2,\ x_3,\ \ldots\ldots,\ x_n)$。注意， $f(\cdot)$ 是分段双射，因此 $W=f(X)$ 的概率密度函数可由

变量替换公式得到：

$$p_{\tilde{w}}(\tilde{w}) = \sum_{x \in \tilde{f}^{-1}(\tilde{w})} \frac{p_x(x)}{|\det(J_{\tilde{f}}(x))|}$$

其中 $J_{\tilde{f}}(x) = \left[\frac{\partial f_i}{\partial x_j}(x)\right]$ 是 $\tilde{f}(\cdot)$ 的雅可比矩阵，其中 $\tilde{f}_1(x) = |f(x)|$ 且 $\tilde{f}_j(x) = x_j$ 对所有 $j \geq 2$ 均成立。由此可知，雅可比矩阵是对角矩阵，其行列式为 $|\frac{\partial f}{\partial x_1}(x)|$。条件随机变量 $W = |f(X)| \mid X \in \mathcal{X}^a$ 的概率密度函数 $p_{w|\mathcal{X}^a}$ 随后可通过对 $p_{\tilde{w}}$ 进行边缘化得到：

$$p_{w|\mathcal{X}^a}(w) = \int p_{\tilde{w}}(w, x_2^n) \mathrm{d}x_2^n = \int \sum_{x \in \tilde{f}^{-1}(w, x_2^n)} \frac{p_x(x)}{|\frac{\partial f}{\partial x_1}(x)|} \mathrm{d}x_2^n \tag{8}$$

其中 $x_2^n = (x_2、\ldots\ldots、x_n)$，并且我们进行了 $(n-1)$ 重积分。

因此，根据勒贝格分解，$W = |f(X)|$ 的分布可以拆分为一个具有概率密度函数 $p_{w|\mathcal{X}^a}$ 的绝对连续分量，以及一个具有有限个质量点的离散分量，对此有 $\mathbb{P}(W = c_i) = \int_{\mathcal{X}_i^c} p_x(x) \mathrm{d}x =: p_x(\mathcal{X}_i^c)$。根据无意识统计学家定律，我们进而得到

$$\begin{aligned}
&\mathbb{E}[\log(|f(X)|) \mid |f(X)| > 0] \\
&= \mathbb{E}[\log(W) \mid W > 0] \\
&= \sum_{i=1}^{\ell} p_x(\mathcal{X}_i^c) \log|c_i| + p_x(\mathcal{X}^a) \int_0^{\infty} \log(w) p_{w|\mathcal{X}^a}(w) \mathrm{d}w \\
&= \sum_{i=1}^{\ell} p_x(\mathcal{X}_i^c) \log|c_i| + p_x(\mathcal{X}^a) \underbrace{\int_0^{\epsilon} \log(w) p_{w|\mathcal{X}^a}(w) \mathrm{d}w}_{I_1} + p_x(\mathcal{X}^a) \underbrace{\int_{\epsilon}^{\infty} \log(w) p_{w|\mathcal{X}^a}(w) \mathrm{d}w}_{I_2}
\end{aligned}$$

其中，最后一行将积分在固定的 $\epsilon \ll 1$ 处分开。显然，第一项求和是有限的，因为对于所有 i 都有 $c_i > 0$。对于其余涉及积分的项，暂且假设 $p_{w|\mathcal{X}^a}(w) \leq C < \infty$。于是，

$$\begin{aligned}
I_1 &= \int_0^{\epsilon} p_w \log(w) dw \geq \int_0^{\epsilon} C \log(w) dw = C(\epsilon \log(\epsilon) - \epsilon) > -\infty \\
I_2 &= \int_{\epsilon}^{\infty} p_w \log(w) dw \geq \int_{\epsilon}^{\infty} p_w \left(1 - \frac{1}{w}\right) dw \geq \int_{\epsilon}^{\infty} p_w \left(1 - \frac{1}{\epsilon}\right) dw \geq 1 - \frac{1}{\epsilon} > -\infty.
\end{aligned}$$

因此，我们还需要证明 $p_{w|\mathcal{X}^a}(w) \leq C$ 对于 $w \in [0, \epsilon]$。为此，我们重新考察方程8，并注意到当满足以下条件时积分是有限的：i）p_x 有界；ii）积分在有界集合上进行；iii）$|\frac{\partial f}{\partial x_1}(x)| \geq \epsilon_1 > 0$。引理的断言确保了条件 i）和 ii）。剩下的工作是证明条件 iii）成立。

注意，与使用 $f(x) = |f(x)|、x_2、x_3、\ldots\ldots、x_n)$ 相比，同一个 $p_{w|\mathcal{X}^a}(w)$ 也可以通过使用分段双射函数 $f(x) = x_1、|f(x)|、x_3、\ldots\ldots、x_n)$ 等方式获得。因此，$p_{w|\mathcal{X}^a}(w) \leq C$，前提是 f 的偏导数在至少一个维度上有下界，即存在一个 i 使得 $|\frac{\partial f}{\partial x_1}(x)| \geq \epsilon_1$。由于我们已经有

$$\|\nabla_x f(x)\|_1 = \sum_{i=1}^{n} \left| \frac{\partial f}{\partial x_i}(x) \right|$$

这等价于要求梯度的 L_1 范数有一个正的下界。事实上，回想一下，f 是分段仿射线性的，且仅有有限个分段，其在 $\mathcal{X}^a$ 上的限制并非常数。因此，在 $\mathcal{X}^a$ 上的限制满足 $\nabla_x f(x) = g_i > 0$，对所有 $x \in \mathcal{X}^a$ 以及某个 $i \in \{1, \ldots, m\}$ 均成立。因此，我们可以找到一个 ϵ_1，使得 $\min_i g_i \geq n \cdot \epsilon_1 > 0$，这意味着存在一个 i，使得对所有 $x \in \mathcal{X}^a$ 都有 $|\frac{\partial f}{\partial x_i}(x)| \geq \epsilon_1$。证明至此完成。□

A.6 高斯丢弃下的互信息估计

在算法1 中，我们说明如何估计 $I(X;Z)$，其中 Z 是高斯丢弃下的一个表征。这是我们在实验中估计互信息时所采用的方法，但在此设置中也可以使用任何其他估计器。

算法 1高斯丢弃下的互信息估计

要求： GMM 均值 σ，无噪声表示 ▷ 用于近似的高斯混合模型中的高斯分布数量；噪声方差；无噪声表征

$reprs \leftarrow []$ ▷ 生成具有对应方差的含噪样本

对所有 nr 无噪声表示 **执行**

  对 $i \leftarrow 1, n$ **do**

    $\epsilon \leftarrow noise_p$

    $reprs \leftarrow reprs + nr * \epsilon$

  结束循环

结束循环

点 ← 无噪声表示[：GMM 均值] ▷ 在有限数量的点上创建高斯混合模型，以便加快计算

$d \leftarrow []$

对所有 p 在 $points$ **执行**

  $d \leftarrow d +$ 高斯分布$(p, \sigma * |p|)$

结束循环

gmm ← 混合模型(d)

$lp \leftarrow []$ ▷ 从高斯混合模型中获取含噪样本的对数概率估计值

对所有 r 表示 **执行**

  $lp \leftarrow lp +$ gmm.对数概率(r)

结束循环

$h(z)$ ←均值(lp)

$h(z|x) \leftarrow 0$ ▷ 使用闭式公式计算条件熵

对于 $i \leftarrow 1,$ dim(表征[0]) **执行** ▷ 对于表征的每个维度

  $h(z|x) \leftarrow h(z|x) +$ mean$(ln(\sqrt{2\pi}e\sigma|$无噪声表示[: i]$|))$ ▷ 使用无噪声表征这里分别处理每个维度

结束循环

$I(x,z) \leftarrow h(z) - h(z|x)$ ▷ 获得互信息的最终估计值

A.7 估计器的评估

图8 展示了在比图 $h(Z)$ 中更高噪声下的上界和估计结果。3更大的噪声会如预期般增大高斯熵上界与基于混合分布的估计之间的差距。

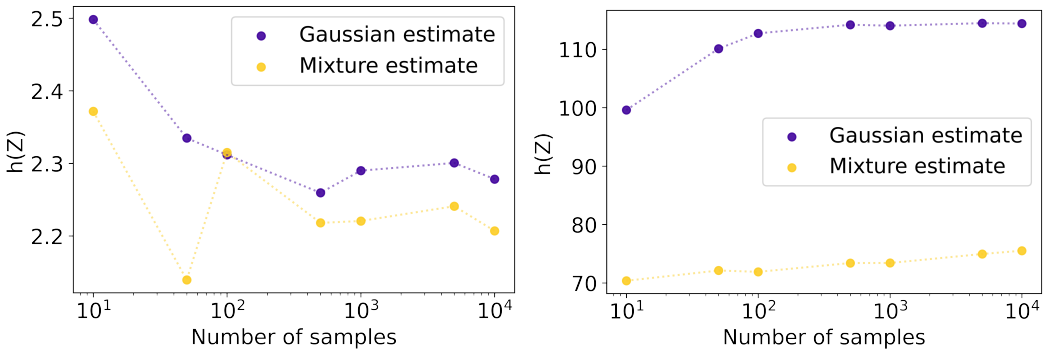
图9 我们可以看到，在更大噪声下，$h(Z|X)$ 的蒙特卡洛估计收敛性。

正如预期，较大的噪声方差会导致较小的互信息值（图 10），而改变维度时观察到的趋势保持不变。

A.8 信息平面分析

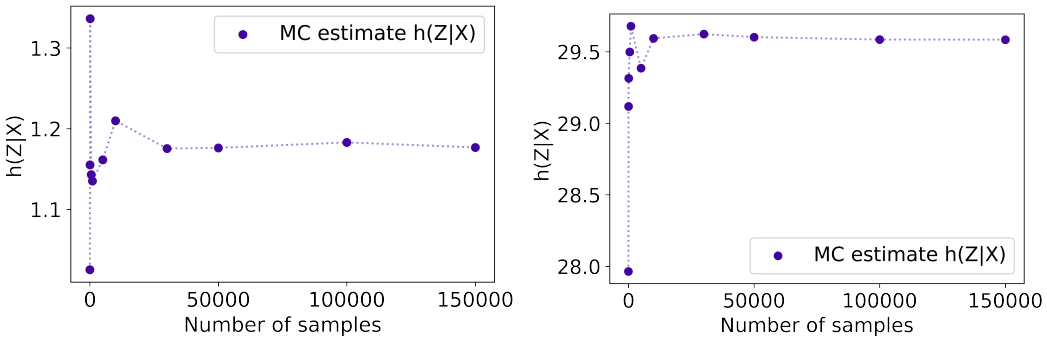
请注意，在实验中，我们分别分析训练样本和测试样本上的信息平面。为了在推理期间获得用于互信息估计的有效隐藏表征样本，我们采用 MC-Dropout，而不是采用关闭丢弃的通常推理方式。根据 Srivastava 等人（2014年），这是获得预测结果的理论上合理方式；而关闭丢弃并进行权重重新缩放所得到的是一种近似方法，可以实现更快的计算。

在图11、图12和图 13 中，我们给出了基于相应数据集（MNIST、MNIST 和 CIFAR10）测试集构建的信息平面。



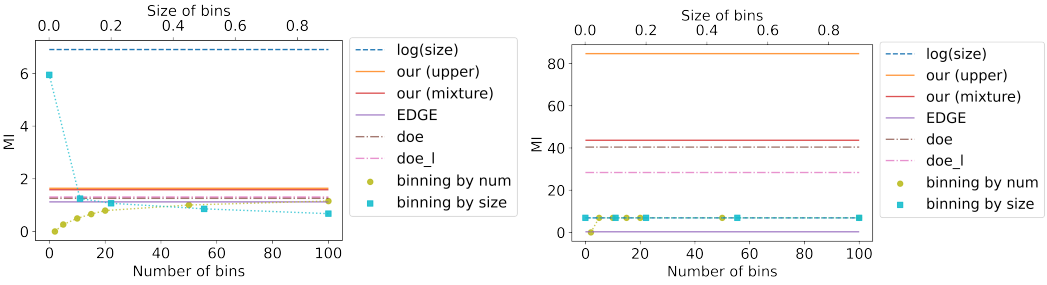
(a) 噪声方差 0.4，维度 1　　(b) 噪声方差 0.4，维度 50

图 8：隐藏表示的熵。可以看出，随着维度增加，与高斯混合估计相比，高斯上界会变得非常宽松。



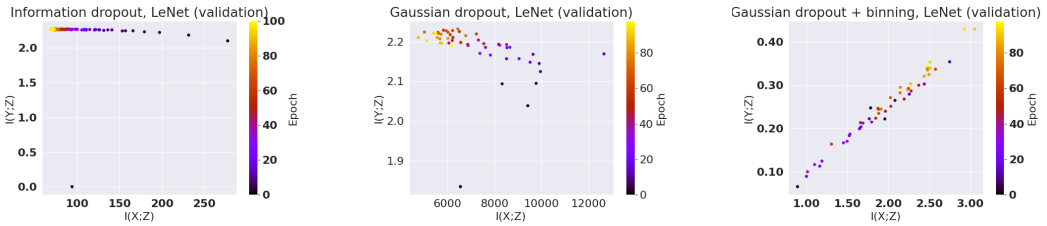
(a) 噪声方差 0.4，维度 2　　(b) 噪声方差 0.4，维度 50

图 9：隐藏表示的条件熵。无论维度如何， $h(Z|X)$ 的蒙特卡洛估计都会随着样本数量的增加而趋于稳定。



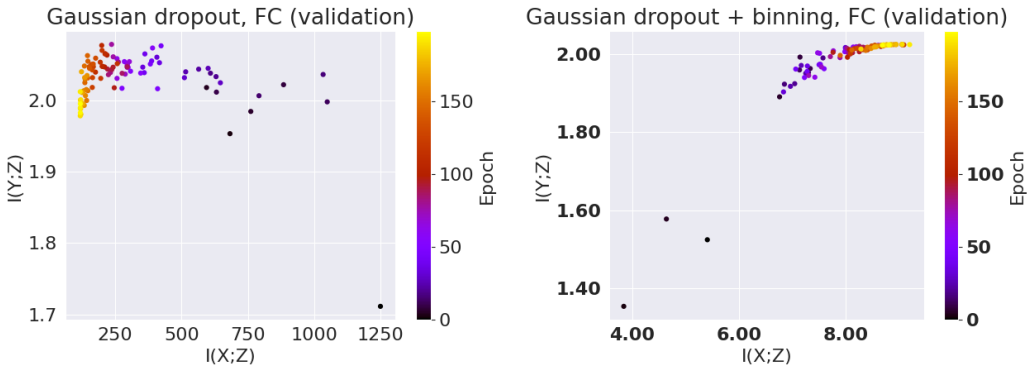
(a) 噪声方差为 0。4，维度为 1　　(b) 噪声方差为 0。4，维度为 50

图 10：针对乘性高斯噪声设置的互信息估计方法比较。



（a）信息丢弃　　（b）高斯丢弃（本文方法）　　（c）高斯丢弃（分箱）

图 11：与基于分箱（离散）互信息估计的信息平面分析相比，基于本文方法的信息瓶颈分析显示，高斯丢弃和信息丢弃均存在压缩现象。

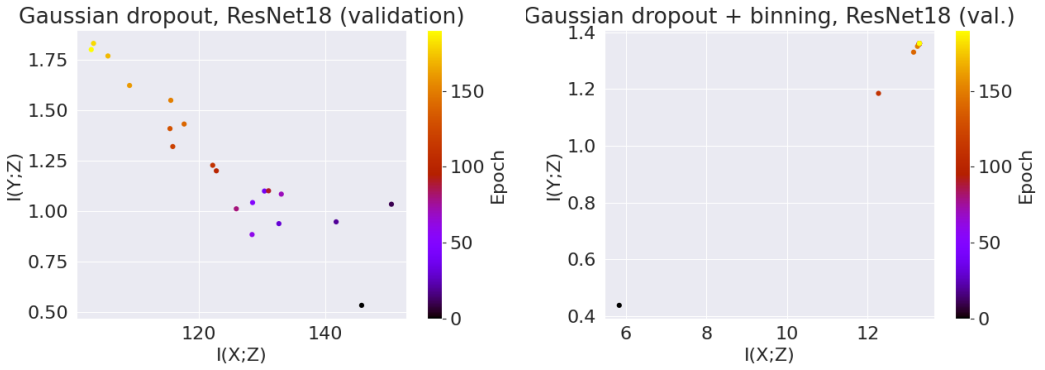


（a）本文估计器　　(b) 分箱估计器

图 12：与基于分箱（离散）互信息估计的信息平面分析相比，基于我们方法的信息平面显示出压缩现象。

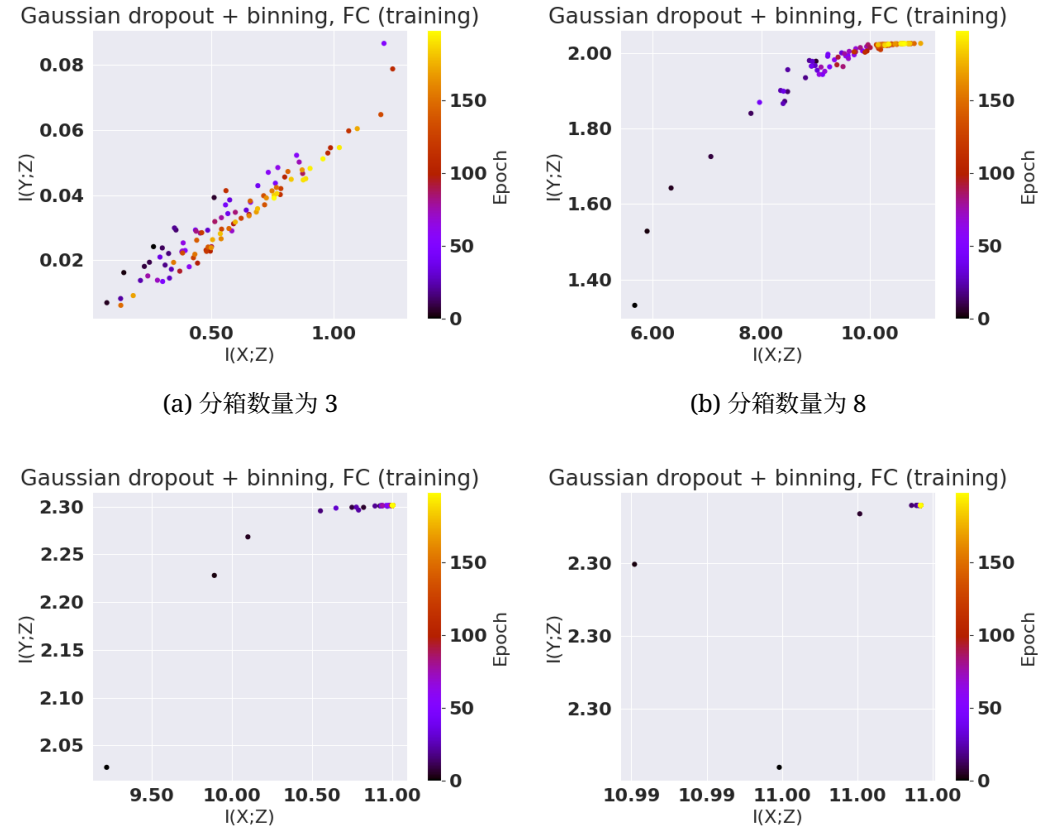
在图 14 中，我们提供了分箱估计器的其他信息平面，展示了用于互信息估计的分箱数量变化时的结果。我们报告了在 MNIST 上使用高斯丢弃方差0.2.训练的全连接网络的结果。

在图 15 中，我们展示了同一个在 MNIST 上训练的全连接网络所得到的信息平面，其中高斯丢弃的方差设为0.4。



(a) 我们的估计器　　(b) 分箱估计器

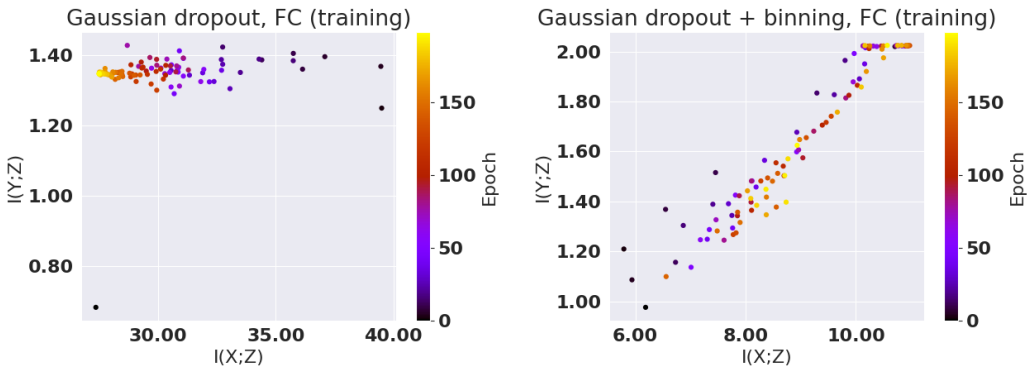
图 13：与信息瓶颈分析中互信息的离散估计相比，我们的方法清晰地展现了压缩现象。



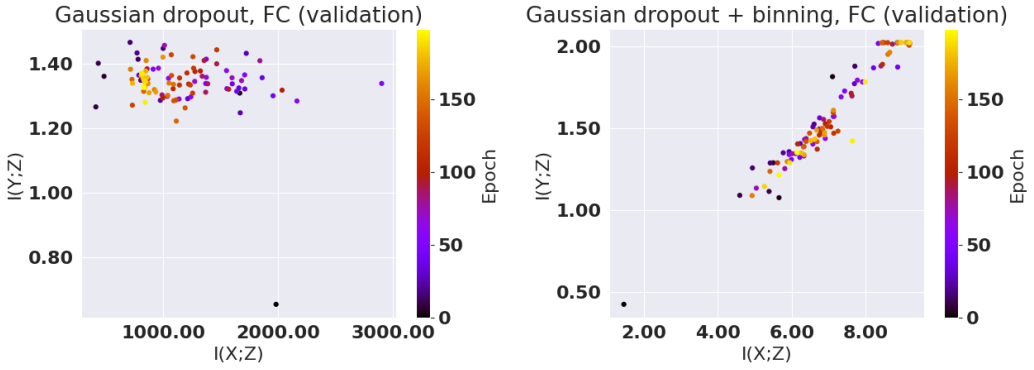
(a) 分箱数量为 3　　(b) 分箱数量为 8

(c) 分箱数量 15　　(d) 分箱数量 30

图 14：用于估计互信息的分箱数量越多，信息瓶颈就越会坍缩为一个点，该点对应于可用样本的数量。所有较少分箱数量的情况都表明，就 $I(X;Z)$ 而言没有压缩。



（a）我们的估计器（训练）　　（b）分箱估计器（训练）



（c）我们的估计器（测试）　　（d）分箱估计器（测试）

图15：与丢弃 0.2 的方差类似，使用我们的估计器进行测量时可以观察到压缩，而使用分箱时则没有压缩。