arXiv:2303.00633v4 [cs.IT] 2 May 2024

方差-不变性-协方差正则化的信息论视角

Ravid Shwartz-Ziv*纽约大学

Randall Balestriero Meta AI、FAIR

Kenji Kawaguchi新加坡国立大学

TimG. J. Rudner纽约大学

Yann LeCun
纽约大学与 Meta AI、FAIR

摘要

方差-不变性-协方差正则化（VICReg）是一种自监督学习（SSL）方法，已在各种任务中展现出良好的效果。然而，VICReg背后的基本机制尚未得到探索。本文从信息论视角阐述VICReg目标函数。我们首先为确定性网络推导信息论量，以替代不切实际的随机网络假设。随后，我们将VICReg目标函数的优化与互信息优化联系起来，揭示其背后的假设，并便于与其他自监督学习算法进行建设性比较；此外，我们还为VICReg推导出一个泛化界，揭示其在下游任务中的内在优势。基于这些结果，我们提出了一系列基于信息论原理推导出的自监督学习方法，其表现优于现有的自监督学习技术。

1 引言

自监督学习（SSL）是一种很有前景的方法，它通过优化输入与自生成信号之间的代理目标函数来提取有意义的表征。例如，方差-不变性-协方差正则化（VICReg） [7], 是一种广泛使用的自监督学习算法，它采用去相关机制，通过应用方差正则化和协方差正则化来避免学习到平凡解。
一旦代理目标函数得到优化，预训练模型便可作为特征提取器，用于各种下游监督任务，例如图像分类、目标检测、实例分割或姿态估计 [15, 16, 49, 68]。尽管自监督学习方法展现出了令人鼓舞的结果，但解释其有效性的理论基础仍在持续研究中 [5, 43]。

信息论已被证明是增进我们对深度神经网络（DNNs）理解的有效工具，对表征学习应用 [3] 和理论探索 [61, 73]都产生了重大影响。然而，将信息论原理应用于自监督学习时，往往依赖不切实际的假设，导致许多现有的信息论自监督学习方法用途有限。其中一个假设是，待优化的深度神经网络具有随机性——但实际应用中的绝大多数深度神经网络都不满足这一假设。关于这一主题的全面综述，请参阅Shwartz-Ziv和LeCun的研究 [63]。

*联系邮箱: ravid.shwartz.ziv@nyu.edu。

第37届神经信息处理系统大会（NeurIPS 2023）。

本文从信息论视角考察方差-不变性-协方差正则化（VICReg），这是一种为确定性深度神经网络开发的自监督学习方法。我们提出了一种方法，通过将随机性从网络转移到输入数据，来应对确定性网络中的互信息估计难题；这一假设更为合理。这种转变使我们能够对确定性网络进行信息论分析。为建立VICReg目标函数与信息最大化之间的联系，我们找出必要的假设并通过实证进行验证。基于这一分析，我们从信息论视角阐述不同自监督学习算法之间的差异，并提出一类新的自监督学习插件方法。这类新方法利用现有的信息估计器，在多个基准测试任务上取得了最先进的预测性能。最后，我们推导出一个泛化界，将信息优化和VICReg目标函数与下游任务性能联系起来，凸显了VICReg的优势。

我们的主要贡献总结如下：

1. 我们提出了一种新方法，通过依据数据分布假设（第 3 节）将随机性从网络转移到输入上，从信息论视角研究确定性深度神经网络。2. 我们建立了 VICReg目标函数与信息论优化之间的联系，并借此阐明该目标函数背后的假设，并将其与其他自监督学习方法进行比较（第4 节）。3. 基于我们的分析，我们提出了一系列信息论自监督学习方法，并取得了最先进的性能（第 5 节）。4. 我们推导出一个泛化界，直接关联 VICReg 与下游任务的泛化能力，进一步凸显其相较于其他自监督学习方法的实际优势（第 6 节）。

2 背景与预备知识

我们首先介绍分析所需的技术背景。

2.1 连续分段仿射（CPA）映射 .

分段多项式中蕴含着一类丰富的函数：样条算子。简而言之，给定定义域 $\mathbb{R}^D$ 的一个划分 Ω，k 阶样条是在每个区域 $\omega \in \Omega$ 上由 k 阶多项式定义的映射，并且在整个定义域上满足关于 0 至 $k-1$ 阶导数的连续性约束。由于我们将重点讨论仿射样条（$k=1$），为清晰起见这里只定义这一情形。一个 K 维仿射样条 f 通过 $f(\boldsymbol{z}) = \sum_{\omega\in\Omega}(\boldsymbol{A}_\omega \boldsymbol{z} + \boldsymbol{b}_\omega)\mathbb{1}_{\{\boldsymbol{z}\in\omega\}}$ 生成其输出，其中输入为$\boldsymbol{z} \in \mathbb{R}^D$，而 $\boldsymbol{A}_\omega \in \mathbb{R}^{K\times D}$，$\boldsymbol{b}_\omega \in \mathbb{R}^K$，$\forall \omega \in \Omega$ 分别是各区域的斜率和偏移量参数；其关键约束是整个映射在定义域 $f \in \mathcal{C}^0(\mathbb{R}^D)$ 上连续。

深度神经网络作为连续分段仿射映射。深度神经网络（DNN）是一个（非线性）算子 f_Θ，其参数为 Θ，可将输入 $\boldsymbol{x} \in \mathbb{R}^D$ 映射为预测 $\boldsymbol{y} \in \mathbb{R}^K$。DNN 算子的精确定义见 Goodfellow 等人 [27]。为避免符号过于繁杂，除非有助于澄清，否则我们将省略 Θ。在我们的分析中，我们仅假设 DNN 中的非线性函数是连续分段仿射映射——泄漏型 ReLU、绝对值和最大池化算子均属于此类。于是，整个输入到输出的映射成为一个连续分段仿射样条，其隐式划分 Ω 是网络权重和架构的函数 [52, 6]。对于光滑非线性函数，我们的结果可通过一阶泰勒近似论证得到。

2.2 自监督学习。

自监督学习是一组从无标签数据中学习表征函数的技术，这些表征函数随后可以适配于各种下游任务。监督学习依赖有标签数据，而自监督学习则利用自生成信号构造代理目标。自监督学习面临的挑战是在没有标签的情况下学习有用的表征。其目标是避免模型将所有输入映射为常数输出的平凡解 [11, 36]。为此，自监督学习采用了多种策略。对比方法（如 SimCLR 及其 InfoNCE 准则）通过区分正例和负样本来学习表征 [16, 55]。相比之下，非对比方法应用正则化技术来防止坍塌 [14, 17, 28]。

方差-不变性-协方差正则化（VICReg）。一种广泛使用的用于训练联合嵌入架构的自监督学习方法 [7]。其损失目标由三个项组成：不变性损失、方差损失和协方差损失：

- 不变性损失： 不变性损失定义为嵌入两两之间的均方欧几里得距离，可确保原始输入与增强输入的表征保持一致。
- 正则化： 正则化项由两个损失项组成：**方差损失**——一种铰链损失，用于维持嵌入中每个变量在一个批次内的标准差；以及**协方差损失**，它会惩罚嵌入协方差矩阵的非对角系数，以促进特征之间的去相关。

VICReg生成两批嵌入，$\boldsymbol{Z} = [f(\boldsymbol{x}_1)\ldots, f(\boldsymbol{x}_B)]$和 $\boldsymbol{Z}' = [f(\boldsymbol{x}'_1)\ldots, f(\boldsymbol{x}'_B)]$,，每批大小为（$B \times K$）。这里，$\boldsymbol{x}_i$和 $\boldsymbol{x}'_i$是样本 I_i的两种不同随机增强。协方差矩阵 $\boldsymbol{C} \in \mathbb{R}^{K\times K}$ 由 $[\boldsymbol{Z}, \boldsymbol{Z}']$得到。因此，VICReg损失可表示如下：

$$\mathcal{L} = \underbrace{\frac{1}{K}\sum_{k=1}^{K}\left(\alpha \max\left(0, \gamma - \sqrt{\boldsymbol{C}_{k,k} + \epsilon}\right) + \beta \sum_{k'\neq k} (\boldsymbol{C}_{k,k'})^2\right)}_{\text{Regularization}} + \underbrace{\eta \|\boldsymbol{Z} - \boldsymbol{Z}'\|_F^2 / N}_{\text{Invariance}}. \tag{1}$$

2.3 深度神经网络与信息论

近年来，信息论方法通过开发和应用信息论估计量及学习原理，在推动深度学习发展和深度神经网络训练方面发挥了重要作用 [3,9, 34, 57, 65, 66, 70, 73]。然而，确定性深度神经网络的信息论目标通常面临一个共同问题：输入与深度神经网络表征之间的互信息是无穷大的。这会导致适定性不良的优化问题。

人们提出了多种策略来应对这一挑战。其中一种方法是使用带有变分界的随机深度神经网络，将确定性网络的输出作为条件分布的参数 [44, 62]。另一种方法由 Dubois 等人提出 [22],，假设两个视图之间数据增强的随机性是网络中随机性的主要来源。然而，这些方法假设随机性来自深度神经网络，这与通常的实践相悖。其他研究则假设输入是随机的，但没有对网络的表征分布特性作出任何假设。相反，这些研究依赖一般性的下界来分析目标 [72, 76]。

3 深度神经网络中的自监督学习：信息论视角

为了分析确定性网络中的信息，我们首先需要确立自监督学习的信息论视角（第3.1节）。随后，我们利用数据分布假设（第3.2节）来证明它适用于确定性自监督学习网络。

3.1 信息论视角下的自监督学习

我们的讨论从多视图信息最大化原理开始，该原理旨在最大化视图 X' 与第二个表征 Z 之间的互信息 $I(Z; X')$。正如 Federici 等人 [24], 所示，我们可以通过采用以下下界来优化这一信息量：

$$I(Z, X') = H(Z) - H(Z|X') \geq H(Z) + \mathbb{E}_{x'}[\log q(z|x')]. \tag{2}$$

这里，$H(Z)$ 表示 Z 的熵。在监督学习中，标签 Y 保持固定，因此熵项 $H(Y)$ 是一个常数。因此，优化仅关注对数损失，$\mathbb{E}_{x'}[\log q(z|x')]$,，它可以是交叉熵或平方损失。

然而，对于联合嵌入网络，可能会出现退化解，即所有输出都“坍塌”到一个不理想的值 [16]。考察公式 (2)后，我们发现熵并非固定不变，而是可以优化的。因此，仅最小化对数损失可能会导致表征坍塌为平凡解，必须对其进行正则化。

3.2 理解数据分布假设

此前我们提到，简单的分析可能会认为确定性深度神经网络中的信息是无限的。为了解决这一问题，我们探究以下假设是否能够为神经网络输出提供一个易于处理的分布：数据集是支撑集互不重叠的高斯混合。与假设神经网络本身是随机的相比，这一假设的限制性较小，因为它关注的是数据的生成过程，而非模型和训练过程。关于假设网络具有随机性的局限性，以及随机网络与随机输入之间的比较，详见附录 N。在第 4.2 节中，我们验证了这一假设适用于真实世界数据集。

所谓的流形假设允许我们将任意点视为一个高斯随机变量，其低秩协方差矩阵与数据流形的切空间对齐 [25],这使我们能够考察潜在表示相对于观测均值的条件性，即 $X|\boldsymbol{x}^* \sim \mathcal{N}(\boldsymbol{x};$ $\boldsymbol{x}^*,\ \Sigma_{\boldsymbol{x}^*})$。在此，$\Sigma_{\boldsymbol{x}^*}$ 的特征向量与 $\boldsymbol{x}^*$ 处数据流形的切空间对齐，而该切空间会随 $\boldsymbol{x}^*$ 在空间中的位置变化。在这种设定下，一个数据集被视为由一组不同的点组成 $\boldsymbol{x}_n^*,\ n = 1$ ..... N，而完整的数据分布则表示为一系列具有低秩协方差的高斯密度之和，定义如下：

$$X \sim \sum_{n=1}^{N} \mathcal{N}(\boldsymbol{x}_n^*, \Sigma_{\boldsymbol{x}_n^*})^{\mathbb{I}_{\{T=n\}}} \quad \text{with} \quad T \sim \mathrm{Cat}(N). \tag{3}$$

这里，T 是一个均匀的类别随机变量。为简单起见，我们假设 $\mathcal{N}$（$\boldsymbol{x}_i^*$，$\Sigma_{\boldsymbol{x}_i^*}$）的有效支撑集 互不重叠（关于这一假设的经验验证，请参见第 4.2 节）。有效支撑集定义为 $\{x \in \mathbb{R}^D :$ p（x） $> \epsilon\}$。于是，我们可以如下近似密度函数：

$$p(\boldsymbol{x}) \approx \mathcal{N}\left(\boldsymbol{x}; \boldsymbol{x}_{n(\boldsymbol{x})}^*, \Sigma_{\boldsymbol{x}_{n(\boldsymbol{x})}^*}\right)/N, \tag{4}$$

其中，$\mathcal{N}(\boldsymbol{x};.,.)$ 是 $\boldsymbol{x}$ 处的高斯分布密度，且 $n(\boldsymbol{x}) = \arg\min_n (\boldsymbol{x} - \boldsymbol{x}_n^*)^T \Sigma_{\boldsymbol{x}_n^*} (\boldsymbol{x} - \boldsymbol{x}_n^*$).

3.3 深度神经网络变换下的数据分布

考虑一个仿射样条算子 f，如第 2.1 节所示，它将一个维度为 D 的空间映射到一个维度为 K 的空间，其中 $K \geq D$。该映射的像或张成空间表示如下：

$$\mathrm{Im}(f) \triangleq \{f(\boldsymbol{x}) : \boldsymbol{x} \in \mathbb{R}^D\} = \bigcup\nolimits_{\omega \in \Omega} \mathrm{Aff}(\omega; \boldsymbol{A}_\omega, \boldsymbol{b}_\omega) \tag{5}$$

在该方程中，$\mathrm{Aff}(\omega; \boldsymbol{A}_\omega, \boldsymbol{b}_\omega) = \{\boldsymbol{A}_\omega \boldsymbol{x} + \boldsymbol{b}_\omega :\ \boldsymbol{x} \in \omega\}$ 表示由区域 ω 的逐区域参数 $\boldsymbol{A}_\omega, \boldsymbol{b}_\omega$定义的仿射变换。$\Omega$ 表示包含 $\boldsymbol{x}$的输入空间划分。为了实际计算逐区域仿射映射，我们将 $\boldsymbol{A}_\omega$ 设为网络在对应输入 x处的雅可比矩阵，并将 b 定义为 $f(x) - \boldsymbol{A}_\omega x$。因此，基于 $\boldsymbol{A}_\omega$ 引起的坐标变换和 $\boldsymbol{b}_\omega$引起的平移，由各输入空间划分区域 Ω 上的仿射变换构成深度神经网络映射 $\omega \in$ 。

当输入空间关联着一个密度分布时，该密度会经由映射 f进行变换。计算 $f(X)$的密度通常难以处理。然而，在 第3.2节的不相交支撑假设下，我们可以通过增加原型的数量 N，任意提升密度的表征能力。因此，每个高斯分布的支撑集都包含在其均值所在的区域 ω 内，从而得到以下定理：

定理 1。 给定 公式 (4)中的设定，无条件 DNN 输出密度 Z可以近似为以下仿射变换后的分布的混合分布 $\boldsymbol{x}|\boldsymbol{x}_n^*{}_{\boldsymbol{x}}$：

$$Z \sim \sum_{n=1}^{N} \mathcal{N}\Big(\boldsymbol{A}_{\omega(\boldsymbol{x}_n^*)} \boldsymbol{x}_n^* + \boldsymbol{b}_{\omega(\boldsymbol{x}_n^*)}, \boldsymbol{A}_{\omega(\boldsymbol{x}_n^*)}^T \Sigma_{\boldsymbol{x}_n^*} \boldsymbol{A}_{\omega(\boldsymbol{x}_n^*)}\Big)^{(\ )^{\mathbb{1}_{\{T=n\}}}},$$

其中 $\omega(\boldsymbol{x}_n^*) = \omega \in \Omega \iff \boldsymbol{x}_n^* \in \omega$ 是原型 $\boldsymbol{x}_n^*$ 所在的划分区域。证明。 参见附录 B。 □

换言之，定理1 表明，当输入噪声较小时，我们可以将条件输出密度简化为单个高斯分布：($Z'|X' = x_n) \sim \mathcal{N}(\mu(x_n)\ \Sigma(x_n))$，其中 $\mu(x_n) = \boldsymbol{A}_{\omega(\boldsymbol{x}_n)} \boldsymbol{x}_n + \boldsymbol{b}_{\omega(\boldsymbol{x}_n)}$ 且 $\Sigma(x_n) = \boldsymbol{A}_{\omega(\boldsymbol{x}_n)}^T \Sigma_{\boldsymbol{x}_n} \boldsymbol{A}_{\omega(\boldsymbol{x}_n)}$。

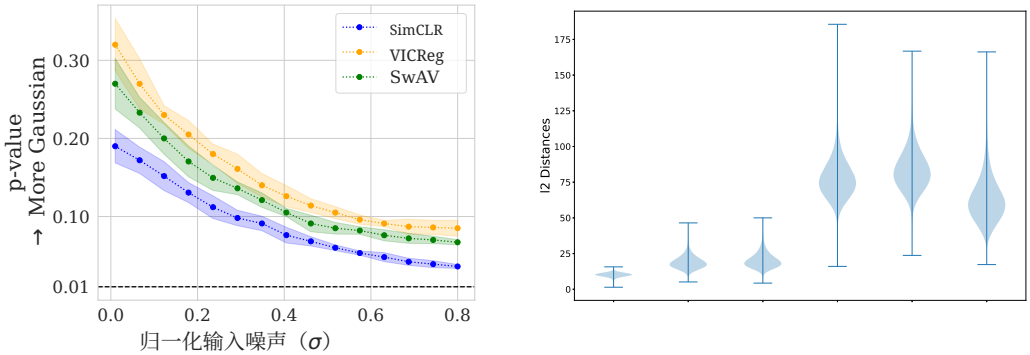

图 1：**左：在输入噪声较小时，自监督学习训练的网络输出更接近高斯分布**。在不同输入噪声水平下，于 ImageNet 上训练的不同自监督学习模型的正态性检验p值。虚线表示能够以 99% 的置信度拒绝零假设（高斯分布）的临界点。**右：每个点周围的高斯分布彼此不重叠**。这些图展示了不同数据集的原始图像之间的 $l2$距离。可以看出，对于更复杂的真实世界数据集，这些距离最大。

4 信息优化与 VICReg 优化目标

基于我们之前的讨论，我们使用数据分布假设将确定性网络中的条件输出建模为高斯混合。这使我们能够将自监督学习训练目标表述为最大化互信息， $I(Z;\ X')$ 和 $I(Z';X)$。

然而，一般而言，这种互信息难以处理。因此，我们将利用对网络表征的推导，基于期望损失获得一个可处理的变分近似，并对其进行优化。

计算期望损失需要对输出中的随机性进行边缘化。我们可以采用高斯观测模型进行最大似然估计。要计算关于 x' 样本的期望损失，我们必须对 Z'中的随机性进行边缘化。这个过程意味着条件解码器服从高斯分布：$(Z|X'=x_n)\ \sim \mathcal{N}(\mu(x_n)\ I+\Sigma(x_n)$。

然而，计算关于 Z 样本的期望对数损失颇具挑战。因此，我们将关注一个下界——关于 Z' 样本的期望对数损失。利用詹森不等式，我们可以得到以下下界：

$$\mathbb{E}_{x'}\left[\log q(z|x')\right] \geq \mathbb{E}_{z'|x'}\left[\log q(z|z')\right] = \frac{1}{2}(d\log 2\pi - (z-\mu(x'))^2 - \operatorname{Tr}\log\Sigma(x')). \tag{6}$$

对 Z取期望，可得

$$\mathbb{E}_{z|x}\left[\mathbb{E}_{z'|x'}\left[\log q(z|z')\right]\right] = \frac{1}{2}(d\log 2\pi - (\mu(x)-\mu(x'))^2 - \log\left(|\Sigma(x)|\cdot|\Sigma(x')|\right)). \tag{7}$$

综合以上各项，可得

$$I(Z;X') \geq H(Z) + \frac{d}{2}\log 2\pi - \frac{1}{2}\mathbb{E}_{x,x'}[(\mu(x)-\mu(x'))^2 + \log(|\Sigma(x)|\cdot|\Sigma(x')|)]. \tag{8}$$

完整推导见附录 A。为在实践中优化这一目标函数，我们使用经验数据分布近似 $p(x,\ x')$

$$L(x_1\ldots x_N, x'_1\ldots x'_N) \approx \frac{1}{N}\sum_{i=1}^{N} \underbrace{H(Z) - \log\left(|\Sigma(x_i)|\cdot|\Sigma(x'_i)|\right)}_{\text{Regularizer}} \underbrace{- \frac{1}{2}\left(\mu(x_i)-\mu(x'_i)\right)^2}_{\text{Invariance}}. \tag{9}$$

4.1 方差-不变性-协方差正则化：一种信息论视角 e

接下来，我们将 VICReg 与基于信息论的目标函数联系起来。公式 (9)中的“不变性项”会使来自同一张像的增强结果彼此靠近，这与 VICReg目标函数中使用的项相同。然而，正则化项的计算面临重大挑战。熵估计是信息论中的一个成熟问题，通常使用高斯混合密度来表示表征。然而，高斯混合的微分熵没有闭式解 [56]。

一种直接近似熵的方法是获取分布的前两个矩，从而得到熵的一个上界。然而，最小化上界并不一定能优化原始目标。尽管已有报告表明，最小化上界 [47, 54], 取得了成功，但这种方法可能会导致训练过程不稳定。

令 Σ_Z 表示 Z的协方差矩阵。我们利用前两个矩来近似希望最大化的熵。由于不变性项与原始 VICReg目标函数的形式相同，因此我们只考察正则化项。因此，得到如下近似

$$\bar{L}(x_1 \dots x_N, x'_1 \dots x'_N) \approx \sum_{i=1}^{N} \log \frac{|\Sigma_Z(x_1 \dots x_N)|}{|\Sigma(x_i)| \cdot |\Sigma(x'_i)|}. \tag{10}$$

定理 2。假设 $\Sigma(x_i)$ 和 $\Sigma(x'_i)$ 的特征值，以及高斯均值 $\mu(x_i)$ 与 $\mu(x'_i)$ 之间的差值均有界，则最大化问题的解为

$$\max_{\Sigma_Z} \left\{ \sum_{i=1}^{N} \log \frac{|\Sigma_Z(x_1 \dots x_N)|}{|\Sigma(x_i)| \cdot |\Sigma(x'_i)|} \right\} \tag{11}$$

将 Σ_Z 设为对角矩阵。

证明。参见附录 J。 □

根据定理2，我们可以通过对协方差矩阵进行对角化并增大其对角元素，来最大化公式（10）。这一目标可通过最小化 Σ_Z的非对角元素来实现——这就是 VICReg 的协方差准则——同时最大化其对角元素的对数之和。尽管这种方法简单高效，但也存在一个缺点：对角值可能趋近于零，从而导致对数计算不稳定。解决这一问题的一种方法是使用上界，直接计算对角元素之和，由此得到 VICReg 的方差项。这便建立了我们的信息论目标与 VICReg 三个关键分量之间的联系。

4.2 关于数据分布假设的实证验证

为了验证我们的理论，我们测试了随着输入噪声减小，条件输出密度 $P(Z|X)$ 是否会变为高斯分布。我们使用在 CIFAR-10、CIFAR-100 和 ImageNet 数据集上分别以 SimCLR 或 VICReg 目标函数训练的 ResNet-50 模型。我们从测试数据集中为每张像抽取 512 个高斯样本，并考察每个样本在深度神经网络的倒数第二层是否仍呈高斯分布。我们采用 D’Agostino-Pearson 检验来判断这一假设是否成立 [19]。

图 1 （左）展示了 p-值随归一化标准差的变化。在低噪声水平下，使用 VICReg 时，我们以 85% 的概率拒绝网络的条件输出密度为非高斯分布这一假设。然而，随着输入噪声增加，网络输出偏离高斯分布。

接下来，我们验证了关于模型的数据分布具有不重叠的有效支撑集这一假设。我们计算了七个数据集中的图像之间成对距离的分布：MNIST l_2 、CIFAR10、CIFAR[42], 、CIFAR100 [41], Flowers102 [53], Food 101 [12], 和 FGVAircaft [46]。图 1（右）显示，即使对于原始像素，成对距离也远大于零。这意味着我们可以在每个点周围使用一个较小的高斯分布而不会发生重叠，从而验证了这一假设符合实际情况。

5 通过信息最大化实现自监督学习模型

公式（8）的实际应用涉及若干关键的“设计选择”。我们首先比较现有的自监督学习模型如何实现该公式，考察所用的估计量，并讨论其假设所带来的影响。随后，我们介绍用于自监督学习的新方法，这些方法采用了信息论领域中的先进估计量，性能优于当前方法。

5.1 VICReg 与 SimCLR

为了评估它们关于信息最大化的基本假设和策略，我们将 VICReg 与对比式自监督学习方法（如 SimCLR）以及非对比式方法（如 BYOL 和 SimSiam）进行比较。

使用 SimCLR 进行对比学习。 在他们的研究中，Lee 等人 [44] 利用冯·米塞斯-费舍尔分布，将 SimCLR 目标与有关表征的信息变分界联系起来。我们将对确定性网络中信息的分析应用于他们的研究，并比较 SimCLR 与 VICReg 之间的主要差异：

(i) **条件分布：** SimCLR 假设编码器服从冯·米塞斯-费舍尔分布，而 VICReg 假设其服从高斯分布。(ii) **熵估计：** SimCLR 根据输入样本的有限求和对其进行近似。相比之下，VICReg 仅根据 Z 的二阶矩估计熵。开发融合这两种区别的自监督学习方法，是未来一个值得探索的方向。

实证比较。 我们在 CIFAR-10 上训练了用于 VICReg、SimCLR 和 BYOL 的 ResNet-18，并使用 成对距离 熵估计器直接比较它们的熵。（更多详情请参见 附录 K。）该估计器并未被任何方法直接优化，而是用于独立验证。结果（图 2）显示，所有方法的熵在训练过程中都不断增加，其中 SimCLR 的熵最低，而 VICReg 的熵最高。

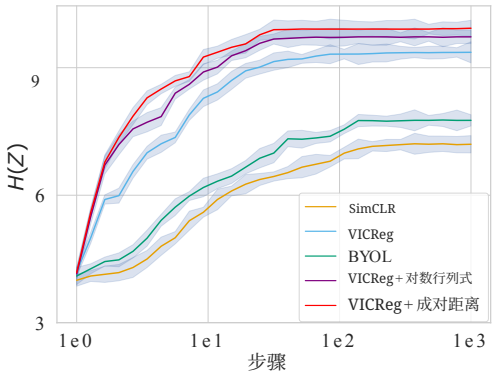

图 2：**VICReg 在训练过程中具有更高的熵。** 不同自监督学习方法在训练过程中的熵。实验使用 ResNet-18 在 CIFAR-10 上进行。误差棒表示 5 次试验的一个标准误差。

5.2 可选熵估计器族

接下来，我们建议将当前自监督学习方法中的不变性项与能够优化熵的插件方法相结合。

熵估计器。VICReg目标函数通过经验协方差矩阵的对角项来近似其对数行列式。正如第 4.1 节所述，这种方法存在一些缺点。一种替代方案是采用不同的熵估计器。对数行列式熵估计器 [75] 是其中一种选择，它能提供更紧的上界。该估计器采用带缩放噪声的 α 阶微分熵；先前研究表明，它是高维特征的紧致估计器，并且对随机噪声具有鲁棒性。然而，由于该估计器提供的是熵的上界，最大化此上界并不能保证优化原始目标。为此，我们还引入了一种基于单个混合成分成对距离的下界估计器 [39]。在这一类估计器中，每个估计器都采用一个用于确定分量密度之间成对距离的函数。这些估计器计算效率高，通常也易于优化。有关其他熵估计器，请参见附录 F。除 VICReg 外，这些方法还可作为多种自监督学习算法的即插即用估计器。除 VICReg 外，我们还进行了实验，将这些估计器与 BYOL 算法结合使用。

表1：**所提出的熵估计器优于以往方法**。使用 ResNet-18、ConvNetX 和 VIT 作为骨干网络进行线性评估时，CIFAR-10、CIFAR-100 和 Tiny-ImageNet 上的 top-1 准确率。误差棒表示三次试验的一个标准误差。

方法	CIFAR-10	Tiny-ImageNet		CIFAR-100	
	ResNet-18	ConvNetX	VIT	ConvNetX	VIT
SimCLR	89.72 ± 0.05	50.86 ± 0.13	51.16 ± 0.13	67.21 ± 0.24	67.31 ± 0.18
Barlow Twins	88.81 ± 0.10	51.34 ± 0.10	51.40 ± 0.16	68.54 ± 0.15	68.02 ± 0.12
SwAV	89.12 ± 0.13	50.76 ± 0.14	51.54 ± 0.20	68.93 ± 0.14	67.89 ± 0.21
MoCo	89.46 ± 0.08	52.36 ± 0.21	53.06 ± 0.21	70.32 ± 0.15	69.89 ± 0.14
VICReg	89.32 ± 0.09	51.02 ± 0.26	52.12 ± 0.25	70.09 ± 0.20	70.12 ± 0.17
BYOL	89.21 ± 0.11	52.24 ± 0.17	53.44 ± 0.20	70.01 ± 0.27	69.59 ± 0.22
VICReg + 成对距离 (**我们的方法**)	$\mathbf{90.37 \pm 0.09}$	52.61 ± 0.15	53.70 ± 0.13	71.10 ± 0.16	70.50 ± 0.19
VICReg + 对数行列式（**我们的方法**）	90.27 ± 0.08	52.91 ± 0.17	$\mathbf{54.89 \pm 0.20}$	71.23 ± 0.18	70.61 ± 0.17
BYOL + 成对距离（**我们的方法**）	90.19 ± 0.14	$\mathbf{53.47 \pm 0.22}$	54.33 ± 0.21	$\mathbf{71.39 \pm 0.25}$	$\mathbf{71.09 \pm 0.24}$
BYOL + 对数行列式（**我们的方法**）	90.11 ± 0.16	53.19 ± 0.25	54.67 ± 0.27	71.20 ± 0.21	70.79 ± 0.26

实验设置。实验在三个图像数据集上进行：CIFAR-10、CIFAR-100 [40], 和 Tiny-ImageNet [20]。CIFAR-10 使用了 ResNet-18 [31] 。相比之下，CIFAR-100 和 Tiny-ImageNet 使用了 ConvNeXt [45]和视觉变换器 [21] 。作为对比，我们考察了以下自监督学习方法：VICReg、SimCLR、BYOL、SwAV [14], 、Barlow Twins [74],和 MoCo [33]。通过线性评估来评估表征质量。不同方法的详细说明见附录 H。

结果。 如 表 1所示，所提出的熵估计器优于原有的自监督学习方法。与最初的实现相比，使用更精确的熵估计器能够提升 VICReg 和 BYOL 的性能。值得注意的是，作为下界的成对距离估计器取得了更好的结果，这与理论上优先最大化真实熵下界的观点相一致。我们的研究结果表明，在本框架的指导下，恰当地选择熵估计器有助于提升性能。

6 下游任务的泛化界

在前文中，我们将信息论原理与 VICReg 目标函数联系起来。现在，我们旨在通过一个泛化界，将这一联系扩展到下游泛化。这一联系进一步表明，VICReg 的泛化与信息最大化和隐式正则化相一致。

记号。 考虑输入点 x、输出 $y \in \mathbb{R}^r$、大小为n 的有标签训练数据 $S = ((x_i, y_i))_{i=1}^n$， 以及大小为 m的无标签训练数据 $\bar{S} = ((x_i^+, x_i^{++}))_{i=1}^m$， 其中 x_i^+ 和 x_i^{++} 共享相同的（未知）标签。利用无标签训练数据，我们定义不变性损失

$$I_{\bar{S}}(f_\theta) = \frac{1}{m} \sum_{i=1}^{m} \|f_\theta(x_i^+) - f_\theta(x_i^{++})\|,$$

其中 f_θ 是基于无标签数据 S训练得到的表征。我们定义有标签损失 $\ell_{x,\, y}(w) = \|W f_\theta(x) - y\|$ 其中 $w = \mathrm{vec}[W] \in \mathbb{R}^{dr}$ 是矩阵 $W \in \mathbb{R}^{r \times d}$的向量化。令$w_S = \mathrm{vec}[W_S]$ 为最小范数解，使其 $W_S =$最小化 $_{W'} \|W'\|_F$，满足

$$W' \in \arg\min_W \frac{1}{n} \sum_{i=1}^{n} \|W f_\theta(x_i) - y_i\|^2.$$

我们还定义了表示矩阵

$$Z_S = [f(x_1), \ldots, f(x_n)] \in \mathbb{R}^{d \times n} \quad \text{and} \quad Z_{\bar{S}} = [f(x_1^+), \ldots, f(x_m^+)] \in \mathbb{R}^{d \times m},$$

以及投影矩阵

$$\mathbf{P}_{Z_S} = I - {Z_S}^\top (Z_S {Z_S}^\top)^\dagger Z_S \quad \text{and} \quad \mathbf{P}_{Z_{\bar{S}}} = I - {Z_{\bar{S}}}^\top (Z_{\bar{S}} {Z_{\bar{S}}}^\top)^\dagger Z_{\bar{S}}.$$

我们定义标签矩阵 $Y_S = [y_1, \ldots, y_n]^\top \in \mathbb{R}^{n \times r}$和未知标签矩阵 $Y_{\bar{S}} = [y_1^+, \ldots, y_m^+]^\top \in \mathbb{R}^{m \times r}$，其中 y_i^+是 x_i^+的未知标签。设 $\mathcal{F}$为 f_θ的假设空间。对于给定的假设空间 $\mathcal{F}$，我们定义归一化 Rademacher 复杂度

$$\tilde{\mathcal{R}}_m(\mathcal{F}) = \frac{1}{\sqrt{m}} \mathbb{E}_{\bar{S},\xi} \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^{m} \xi_i \| f(x_i^+) - f(x_i^{++}) \| \right],$$

其中 ξ_1、.....、ξ_m是 $\{-1, 1\}$中的独立均匀随机变量。其归一化方式使得$\tilde{\mathcal{R}}_m(\mathcal{F}) = O(1)$，当 $m \to \infty$时。

6.1 方差-不变性-协方差正则化的泛化界

现在我们将证明，VICReg目标函数能够提升有监督下游任务上的泛化能力。更具体地说，在控制协方差$Z_{\bar{S}} {Z_{\bar{S}}}^\top$和表征复杂度 $\tilde{\mathcal{R}}_m(\mathcal{F})$ 的同时，最小化无标签不变性损失，可以最小化期望的有标签损失：

定理 3.（非正式版本）。对于任意 $\delta > 0$，其概率至少为 $1 - \delta$，

$$\mathbb{E}_{x,y}[\ell_{x,y}(w_S)] \leq I_{\bar{S}}(f_\theta) + \frac{2}{\sqrt{m}} \|\mathbf{P}_{Z_{\bar{S}}} Y_{\bar{S}}\|_F + \frac{1}{\sqrt{n}} \|\mathbf{P}_{Z_S} Y_S\|_F + \frac{2\tilde{\mathcal{R}}_m(\mathcal{F})}{\sqrt{m}} + \mathcal{Q}_{m,n}, \tag{12}$$

其中 $\mathcal{Q}_{m,n} = O(G\sqrt{\ln(1/\delta)/m + \ln(1/\delta)/n}) \to 0$当 $m, n \to \infty$时。在 $\mathcal{Q}_{m,n}$中，衰减速率为 $1/\sqrt{m}$ 的项的 G 取值取决于 f_θ 和 w 的假设空间，而衰减速率为 $1/\sqrt{n}$ 的项则与任何假设空间无关。

证明。 定理 3 的完整版本及其证明见 附录 I。 □

项 $\|\mathbf{P}_{Z_{\bar{S}}} Y_{\bar{S}}\|_F$定理 3中包含不可观测的标签矩阵 $Y_{\bar{S}}$。然而，我们可以通过使用 $\|\mathbf{P}_{Z_{\bar{S}}} Y_{\bar{S}}\|_F \leq \|\mathbf{P}_{Z_{\bar{S}}}\|_F \|Y_{\bar{S}}\|_F$并最小化 $\|\mathbf{P}_{Z_{\bar{S}}}\|_F$来最小化这一项。因子$\|\mathbf{P}_{Z_{\bar{S}}}\|_F$在协方差 $Z_{\bar{S}} {Z_{\bar{S}}}^\top$的秩最大时达到最小值。可以通过增大对角元素、同时减小非对角元素来实现这一点，VICReg 正是这样做的。

项 $\|\mathbf{P}_{Z_S} Y_S\|_F$ 仅包含可观测变量，我们可以直接使用训练数据测量该项的值。此外，当协方差 $Z_S {Z_S}^\top$ 的秩最大时，项 $\|\mathbf{P}_{Z_S} Y_S\|_F$ 也达到最小。由于协方差 $Z_S {Z_S}^\top$ 和 $Z_{\bar{S}} {Z_{\bar{S}}}^\top$ 会通过集中不等式彼此靠拢，误差量级为 $O(\sqrt{(\ln(1/\delta))/n} + \tilde{\mathcal{R}}_m(\mathcal{F})\sqrt{(\ln(1/\delta))/m})$，~~我们~~也可以通过增大~~$Z_{\bar{S}} {Z_{\bar{S}}}^\top$~~的对角元素、同时减小其非对角元素，来最小化 $\|\mathbf{P}_{Z_S} Y_S\|_F$ 的上界，VICReg 正是采用了这种做法。

因此，VICReg 可以理解为一种通过最小化不变性损失，同时控制协方差以最小化3 定理中的泛化界的方法，具体而言，是最小化 与标签无关的上界 $\|\mathbf{P}_{Z_{\bar{S}}} Y_{\bar{S}}\|_F$和 $\|\mathbf{P}_{Z_S} Y_S\|_F$。如果我们知道无标签数据的标签部分信息 $Y_{\bar{S}}$，就可以利用这些信息直接最小化 $\|\mathbf{P}_{Z_{\bar{S}}} Y_{\bar{S}}\|_F$和 $\|\mathbf{P}_{Z_S} Y_S\|_F$。

6.2 泛化界的比较

SimCLR 的泛化界 [59] 要求有标签类别的数量趋于无穷，才能消除泛化差距；而定理3中的 VICReg 泛化界则不要求标签类别的数量趋于无穷，也能使泛化差距趋于零。这反映出，与 SimCLR 不同，VICReg 不使用负样本对，因此也不使用基于这样一种隐含假设的损失函数，即负样本对的标签不同。另一个区别是，随着 n增大，我们的 VICReg 泛化界会改善，而 SimCLR 以往的泛化界 [59] 不依赖于 n。这是因为 Saunshi 等人 [59] 在设定中假设可以针对每个类别部分访问真实分布，因此有标签数据量的重要性被消除 n ，而我们的研究并未作此假设。

因此，定理 3 中的泛化界通过 $G\sqrt{\ln(1/\delta)/m + \sqrt{\ln(1/\delta)/n}}$，为 VICReg 中 m 与 n 的影响之比提供了新的见解。最后，定理 3 也阐明了 VICReg 相较于标准监督训练的优势。也就是说，在标准训练中，通过拉德马赫复杂度得到的泛化界要求假设空间 $\tilde{}$ $\tilde{}$、$\mathcal{R}_n(\mathcal{W})/\sqrt{n}$ 和 $\mathcal{R}_n(\mathcal{F})/\sqrt{n}$ 的复杂度与有标签数据量 n 相关，而不是与无标签数据量 m 相关。因此，定理 3 表明，采用自监督学习（SSL）后，我们可以用关于 m 的假设空间复杂度替代关于 n 的复杂度。由于无标签数据点的数量通常远多于有标签数据点的数量，这阐明了自监督学习的优势。我们的泛化界与近期的信息瓶颈泛化界 [38] 不同，因为我们的证明和泛化界都不依赖于信息瓶颈。

6.3 通过最大化互信息理解定理 2

定理 3 结合上一节的结果表明，为了提升下游任务的泛化能力，在 SSL 中通过最小化不变性损失 $I_{\bar{S}}(f_\theta)$ 来最大化互信息 $I(Z;X')$，同时控制协方差 $Z_{\bar{S}}{Z_{\bar{S}}}^\top$，是有益的。项 $2\tilde{\mathcal{R}}_m(\mathcal{F})/\sqrt{m}$ 体现了控制表征复杂度 f_θ的重要性。为了进一步理解这一项，我们考虑将 $\mathcal{F}$ 的参数空间离散化，使其具有有限的 $|\mathcal{F}| < \infty$。于是，根据 Massart有限类引理，我们有 $\tilde{\mathcal{R}}_m(\mathcal{F}) \leq C\sqrt{\ln|\mathcal{F}|}$，其中 $C > 0$为某个常数。此外，Shwartz-Ziv [61]表明，我们可以用 $2^{I(Z;X)}$近似 $\ln|\mathcal{F}|$。因此，在定理 3 中，项 $I_{\bar{S}}(f_\theta) + \frac{2}{\sqrt{m}}\|\mathrm{P}_{Z_{\bar{S}}}Y_{\bar{S}}\|_F + \frac{1}{\sqrt{n}}\|\mathrm{P}_{Z_S}Y_S\|_F$对应于 $I(Z;X')$，我们希望最大化该项，同时压缩项 $2\mathcal{R}_m(\mathcal{F})/\sqrt{m}$，后者对应于 $\mathcal{H}(Z;X)$ [23, 64, 67]。

~尽管我们可以显式地对信息添加正则化以控制 $2\mathcal{R}_m(\mathcal{F})/\sqrt{m}$，但~有可能 $I(Z;X|X')$ 和 $2\mathcal{R}_m(\mathcal{F})/\sqrt{m}$ 会通过设计选择所产生的隐式偏差而受到隐式正则化 [29, 69, 30]。因此，定理3 将VICReg的信息论理解与下游泛化的概率保证联系起来。

7 局限性

在本文中，我们提出了以信息最大化为前提的新颖自监督学习方法。尽管我们的方法在一些数据集上展现出优异性能，但计算约束阻碍了我们在更大数据集上对其进行测试。此外，我们的研究依赖于某些假设；尽管进行了严格的验证工作，这些假设可能并不普遍成立。虽然我们努力进行细致的测试与验证，但必须注意的是，某些假设可能不适用于所有情境或条件。在解读本研究结果时，应当考虑这些局限性。

8 结论

我们通过信息论视角分析了用于自监督学习的方差-不变性-协方差正则化。通过将信息论分析所需的随机性转移到输入分布中，我们展示了如何从信息论原理推导出VICReg目标函数；利用这一视角，我们强调了VI-CReg目标函数中隐含的假设，推导出了适用于下游任务的VICReg泛化界，并将其与信息最大化联系起来。

基于这些发现，我们提出了一种受VICReg启发的新型自监督学习目标。我们的概率保证表明，在部分标签信息的情形下，可以通过使协方差矩阵与部分可观测标签矩阵对齐，进一步改进VICReg。这也为未来研究开辟了若干方向，包括设计更优的信息论量估计量，以及探究不同自监督学习方法是否适用于特定的数据特征。

参考文献

[1] Alessandro Achille、Giovanni Paolini 和 Stefano Soatto。深度神经网络中的信息在哪里？*arXiv*预印本 *arXiv:1905.12213*，2019年。[2] N.A. Ahmed 和 D.V. Gokhale。多元分布的熵表达式及其估计量。*IEEE*信息论汇刊，35(3):688–692，1989年。doi: 10.1109/ 18.30996。[3] Alexander A Alemi、Ian Fischer、Joshua V Dillon 和 Kevin Murphy。深度变分信息瓶颈。*arXiv*预印本 *arXiv:1612.00410*，2016年。[4] Rana Ali Amjad 和 Bernhard Claus Geiger。运用信息瓶颈原理学习用于基于神经网络分类的表征。*IEEE*模式分析与机器智能汇刊，2019年。[5] Sanjeev Arora、Hrishikesh Khandeparkar、Mikhail Khodak、Orestis Plevrakis 和 Nikunj Saunshi。对比无监督表征学习的理论分析。*arXiv*预印本 *arXiv:1902.09229*，2019年。[6] Randall Balestriero 和 Richard Baraniuk。深度网络的样条理论。载于*ICML*论文集，第80卷，第374–383页，2018年7月。[7] Adrien Bardes、Jean Ponce 和 Yann LeCun。VICReg：用于自监督学习的方差-不变性-协方差正则化。*arXiv*预印本 *arXiv:2105.04906*，2021年。[8] Peter L Bartlett 和 Shahar Mendelson。Rademacher复杂度与高斯复杂度：风险界与结构性结果。机器学习研究期刊，3(11月):463–482，2002年。[9] Mohamed Ish mael Belghazi、Aristide Baratin、Sai Rajeswar、Sherjil Ozair、Yoshua Bengio、Aaron Courville 和 R Devon Hjelm。MINE：互信息神经估计。*arXiv*预印本 *arXiv:1801.04062*，2018年。[10] Itamar Ben-Ari 和 Ravid Shwartz-Ziv。用于视频异常检测的注意力卷积LSTM图像修复网络。*arXiv*预印本 *arXiv:1811.10228*，2018年。[11] Ido Ben-Shaul、Ravid Shwartz-Ziv、Tomer Galanti、Shai Dekel 和 Yann LeCun。自监督学习的逆向工程。*arXiv*预印本 *arXiv:2305.15614*，2023年。[12] Lukas Bossard、Matthieu Guillaumin 和 Luc Van Gool。Food-101——利用随机森林挖掘判别性组件。载于计算机视觉–*ECCV 2014*：第*13*届欧洲计算机视觉会议，瑞士苏黎世，*2014*年*9*月*6*日至*12*日，论文集，第*VI*部分第*13*卷，第446–461页。Springer，2014年。[13] Brendon J Brewer。使用嵌套采样计算熵。熵，19(8):422，2017年。[14] Mathilde Caron、Ishan Misra、Julien Mairal、Priya Goyal、Piotr Bojanowski 和 Armand Joulin。通过对比聚类分配进行视觉特征的无监督学习。神经信息处理系统进展，33:9912–9924，2020年。[15] Mathilde Caron、Hugo Touvron、Ishan Misra、Herv´e J´egou、Julien Mairal、Piotr Bojanowski 和 Armand Joulin。自监督视觉变换器的新兴特性。载于*IEEE/CVF*国际计算机视觉会议论文集，第9650–9660页，2021年。[16] Ting Chen、Simon Kornblith、Mohammad Norouzi 和 Geoffrey Hinton。用于视觉表示对比学习的简单框架。载于国际机器学习会议，第1597–1607页。PMLR，2020年。[17] Xinlei Chen 和 Kaiming He。探索简单的孪生表示学习。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第15750–15758页，2021年。[18] Corinna Cortes、Mehryar Mohri 和 Afshin Rostamizadeh。基于中心化对齐的核学习算法。机器学习研究期刊，13(1):795–828，2012年。

[19]Ralph B. D'Agostino。正态性的综合检验：针对中等规模和大规模样本。*Biometrika*，58(2):341–348，1971年。ISSN 00063444。URL http://www.jstor.org/stable/2334522。[20] Jia Deng、Wei Dong、Richard Socher、Li-Jia Li、Kai Li 和 Li Fei-Fei。ImageNet：一个大规模分层图像数据库。载于*2009*年*IEEE*计算机视觉与模式识别会议，第248–255页。IEEE，2009年。[21] Alexey Dosovitskiy、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、Xiaohua Zhai、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Sylvain Gelly等。图像胜过16×16个词：用于大规模图像识别的Transformer。*arXiv*预印本 *arXiv:2010.11929*，2020年。[22] Yann Dubois、Benjamin Bloem-Reddy、Karen Ullrich 和 Chris J Maddison。用于无损预测的有损压缩。神经信息处理系统进展，34，2021年。[23] Marco Federici、Anjan Dutta、Patrick Forr´e、Nate Kushman 和 Zeynep Akata。通过多视图信息瓶颈学习鲁棒表征。载于国际学习表征会议，2019年。[24] Marco Federici、Anjan Dutta、Patrick Forr´e、Nate Kushman 和 Zeynep Akata。通过多视图信息瓶颈学习鲁棒表征。*arXiv*预印本 *arXiv:2002.07017*，2020年。[25] Charles Fefferman、Sanjoy Mitter 和 Hariharan Narayanan。检验流形假设。美国数学会杂志，29(4):983–1049，2016年。[26] Ziv Goldfeld、Ewout van den Berg、Kristjan Greenewald、Igor Melnyk、Nam Nguyen、Brian Kingsbury 和 Yury Polyanskiy。估计深度神经网络中的信息流。*arXiv*预印本 *arXiv:1810.05728*，2018年。[27] I. Goodfellow、Y. Bengio 和 A. Courville。深度学习，第1卷。麻省理工学院出版社，2016年。[28] Jean-Bastien Grill、Florian Strub、Florent Altch´e、Corentin Tallec、Pierre Richemond、Elena Buchatskaya、Carl Doersch、Bernardo Avila Pires、Zhaohan Guo、Mohammad Gheshlaghi Azar等。自举潜变量：一种新的自监督学习方法。神经信息处理系统进展，33:21271–21284，2020年。[29] Suriya Gunasekar、Blake E Woodworth、Srinadh Bhojanapalli、Behnam Neyshabur 和 Nati Srebro。矩阵分解中的隐式正则化。载于神经信息处理系统进展，第6151–6159页，2017年。[30] Suriya Gunasekar、Jason D Lee、Daniel Soudry 和 Nati Srebro。线性卷积网络中梯度下降的隐式偏差。载于神经信息处理系统进展，第9461–9471页，2018年。[31] Kaiming He、张祥雨、任少卿和孙剑。用于图像识别的深度残差学习。*CoRR*，abs/1512.03385，2015年。[32] Kaiming He、张祥雨、任少卿和孙剑。用于图像识别的深度残差学习。载于*IEEE*计算机视觉与模式识别会议论文集，第770–778页，2016年。[33] Kaiming He、范浩奇、吴育昕、Saining Xie 和 Ross Girshick。用于无监督视觉表征学习的动量对比。载于*IEEE/CVF*计算机视觉与模式识别会议，第9729–9738页，2020年。[34] R Devon Hjelm、Alex Fedorov、Samuel Lavoie-Marchildon、Karan Grewal、Phil Bachman、Adam Trischler 和 Yoshua Bengio。通过互信息估计与最大化学习深层表征。*arXiv*预印本 *arXiv:1808.06670*，2018年。[35] Marco Huber、Tim Bailey、Hugh Durrant-Whyte 和 Uwe Hanebeck。关于高斯混合随机向量的熵近似。载于*IEEE*智能系统多传感器融合与集成国际会议，第181–188页，2008年9月。doi: 10.1109/MFI.2008。4648062。

[36] Li Jing、Pascal Vincent、Yann LeCun 和 Yuandong Tian。理解对比式自监督学习中的维度坍塌。载于国际学习表征会-议，2022年。URL https://openreview.net/forum?id=YevsQ05DEN7。[37] Kenji Kawaguchi、Zhun Deng、Kyle Luh 和 Jiaoyang Huang。鲁棒性通过数据依赖的泛化界蕴含泛化。载于国际机器学习会议（*ICML*），2022年。[38] Kenji Kawaguchi、Zhun Deng、Xu Ji 和 Jiaoyang Huang。信息瓶颈如何帮助深度学习？载于国际机器学习会议（*ICML*），2023年。[39] Artemy Kolchinsky 和 Brendan D Tracey。利用成对距离估计混合熵。熵，19(7):361，2017年。[40] Alex Krizhevsky。从微小图像中学习多层特征。2009年。[41] Alex Krizhevsky 和 Geoffrey Hinton。从微小图像中学习多层特征。技术报告0，多伦多大学，多伦多，安大略省，2009年。[42] 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。[43] Jason D Lee、Qi Lei、Nikunj Saunshi 和 Jiacheng Zhuo。预测你已知的内容有助于学习：可证明的自监督学习。神经信息处理系统进展，34，2021年。[44] Kuang-Huei Lee、Anurag Arnab、Sergio Guadarrama、John Canny 和 Ian Fischer。压缩视觉表征。神经信息处理系统进展，34，2021年。[45] Zhuang Liu、Hanzi Mao、Chao-Yuan Wu、Christoph Feichtenhofer、Trevor Darrell 和 Saining Xie。面向2020年代的卷积网络。载于*IEEE/CVF*计算机视觉与模式识别会议，第11976–11986页，2022年。[46] Subhransu Maji、Esa Rahtu、Juho Kannala、Matthew Blaschko 和 Andrea Vedaldi。飞机的细粒度视觉分类。*arXiv*预印本 *arXiv:1306.5151*，2013年。[47] Julieta Martinez、Jashan Shewakramani、Ting Wei Liu、Ioan Andrei Bˆarsan、Wenyuan Zeng 和 Raquel Urtasun。置换、量化与微调：神经网络的高效压缩。载于*IEEE/CVF*计算机视觉与模式识别会议，第15699–15708页，2021年。[48] David McAllester 和 Karl Stratos。互信息度量的形式化局限。载于国际人工智能与统计会议，第875–884页。PMLR，2020年。[49] Ishan Misra 和 Laurens van der Maaten。预文本不变表征的自监督学习。载于*IEEE/CVF*计算机视觉与模式识别会议，第6707–6717页，2020年。[50] Neeraj Misra、Harshinder Singh 和 Eugene Demchuk。多元正态分布熵的估计。多元分析杂志，92(2):324–342，2005年。ISSN 0047-259X。doi: https://doi.org/10.1016/j.jmva.2003.10.003。[51] Mehryar Mohri、Afshin Rostamizadeh 和 Ameet Talwalkar。机器学习基础。麻省理工学院出版社，2012年。[52] Guido F Montufar、Razvan Pascanu、Kyunghyun Cho 和 Yoshua Bengio。深度神经网络线性区域的数量。载于*Proc. NeurIPS*，第2924–2932页，2014年。[53] Maria-Elena Nilsback 和 Andrew Zisserman。大类别数量下的自动花卉分类。载于*2008*年第六届印度计算机视觉、图形学与图像处理会议，第722–729页。IEEE，2008年。

[54] Sebastian Nowozin、Botond Cseke 和 Ryota Tomioka。f-gan：使用变分散度最小化训练生成式神经采样器。载于神经信息处理系统进展，第 271–279 页，2016年。[55] Aaron van den Oord、Yazhe Li 和 Oriol Vinyals。基于对比预测编码的表征学习。*arXiv* 预印本 *arXiv:1807.03748*，2018年。[56] Liam Paninski。熵与互信息的估计。神经计算，15(6)：1191–1253，2003年。[57] Zoe Piran、Ravid Shwartz-Ziv 和 Naftali Tishby。双重信息瓶颈。*arXiv* 预印本 *arXiv:2006.04641*，2020年。[58] Ben Poole、Sherjil Ozair、A¨aron van den Oord、Alexander A. Alemi 和 George Tucker。关于互信息的变分界。*CoRR*，abs/1905.06922，2019年。网址：http://arxiv.org/abs/1905.06922。[59] Nikunj Saunshi、Orestis Plevrakis、Sanjeev Arora、Mikhail Khodak 和 Hrishikesh Khande-parkar。对比无监督表征学习的理论分析。载于国际机器学习会议，第 5628–5637 页。PMLR，2019年。[60] Shai Shalev-Shwartz 和 Shai Ben-David。理解机器学习：从理论到算法。剑桥大学出版社，2014年。[61] Ravid Shwartz-Ziv。深度神经网络中的信息流。*arXiv* 预印本 *arXiv:2202.06749*，2022年。[62] Ravid Shwartz-Ziv 和 Alexander A Alemi。无限宽神经网络无限集成中的信息。载于近似贝叶斯推断进展研讨会，第 1–17 页。PMLR，2020年。[63] Ravid Shwartz-Ziv 和 Yann LeCun。压缩还是不压缩——自监督学习与信息论：综述。*arXiv* 预印本 *arXiv:2304.09355*，2023年。[64] Ravid Shwartz-Ziv 和 Naftali Tishby。通过信息压缩深度神经网络，（2017）。*arXiv* 预印本 *arXiv:1703.00810*，2017年。[65] 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。[66] Ravid Shwartz-Ziv、Amichai Painsky 和 Naftali Tishby。深度神经网络中的表征压缩与泛化，2018年。[67] Ravid Shwartz-Ziv、Randall Balestriero 和 Yann LeCun。我们在自监督学习中最大化的是什么？*arXiv* 预印本 *arXiv:2207.10081*，2022年。[68] Ravid Shwartz-Ziv、Micah Goldblum、Hossein Souri、Sanyam Kapoor、Chen Zhu、Yann Le-Cun 和 Andrew Gordon Wilson。预训练你的损失：借助信息性先验实现简易贝叶斯迁移学习。*arXiv* 预印本 *arXiv:2205.10279*，2022年。[69] Daniel Soudry、Elad Hoffer、Mor Shpigel Nacson、Suriya Gunasekar 和 Nathan Srebro。梯度下降在可分数据上的隐式偏差。机器学习研究期刊，19(1)：2822–2878，2018年。[70] Thomas Steinke 和 Lydia Zakynthinou。通过条件互信息推理泛化问题。载于学习理论会议，第 3437–3452 页。PMLR，2020年。[71] Michael Tschannen、Josip Djolonga、Paul K Rubenstein、Sylvain Gelly 和 Mario Lucic。关于表征学习中的互信息最大化。*arXiv* 预印本 *arXiv:1907.13625*，2019年。[72] Tongzhou Wang 和 Phillip Isola。通过超球面上的对齐与均匀性理解对比表征学习。载于国际机器学习会议，第 9929–9939 页。PMLR，2020年。

[73] Aolin Xu 和 Maxim Raginsky。学习算法泛化能力的信息论分析。神经信息处理系统进展，第 30 卷，2017年。[74] Jure Zbontar、Li Jing、Ishan Misra、Yann LeCun 和 St´ephane Deny。Barlow Twins：通过冗余减少进行自监督学习。载于国际机器学习会议，第 12310–12320 页。PMLR，2021年。[75] Zhanghao Zhouyin 和 Ding Liu。使用对数行列式熵估计器理解神经网络。*arXiv* 预印本 *arXiv:2105.03705*，2021年。[76] Roland S Zimmermann、Yash Sharma、Steffen Schneider、Matthias Bethge 和 Wieland Brendel。对比学习逆转数据生成过程。载于国际机器学习会议，第 12979–12990 页。PMLR，2021年。

附录

目录

本附录的组织如下：

- 在附录 A中，我们详细推导了公式（8）的下界。
- 在附录 B中，我们给出了关于网络表征分布的定理1 的完整证明。
- 在附录 C中，我们提供了额外的经验验证。具体而言，我们通过实证检验第 4.1 节中的信息最大化问题的最优解是否为对角矩阵。
- 在附录 D中，我们使用期望最大化（EM）展示高斯混合模型（GMM）下的坍塌现象，并说明其与自监督学习（SSL）的关系，以及如何防止该现象。
- 附录 E 提供了有关 SimCLR 方法的更多细节。
- 附录 F 详细介绍了熵估计器及其影响、假设和局限性。
- 在附录 G中，我们给出了全文中所使用的已知引理的证明。
- 在附录 H中，我们详细介绍了第 5.2 节中实验所使用的超参数、数据集和架构。
- 在附录 I中，我们给出了第 6 节中下游任务泛化界的完整证明。
- 在附录 J中，我们完整证明了第 4.1 节关于信息优化与 VICReg目标函数之间联系的定理。
- 附录 K提供了第 5.1 节中进行的不同自监督学习方法之间熵比较实验的详细信息。
- 在附录 L中，我们详细介绍了本研究的可复现性信息。
- 在附录 O中，我们讨论了这项工作的更广泛影响。本节探讨了研究发现超出当前研究范围所带来的影响、意义和潜在应用。

附录 A 关于 $\mathbb{E}_{x'}\left[\log q(z|x')\right]$ 的下界

在本节补充材料中，我们给出关于 $\mathbb{E}_{x'}\left[\log\ q(z|x')\right]$ 下界的完整推导。由于 $Z'|X'$ 是高斯分布，我们可以将其写为 $Z' = \mu(x')\ + L(x')\epsilon$ 其中 $\epsilon \sim \mathcal{N}(0,\ 1)$，且 $L(x')^T L(x')\ = \Sigma(x')$。现在，令 $\Sigma_r = I$，即可得到:

$$
\begin{aligned}
&\mathbb{E}_{x'}\left[\log q(z|x')\right] && (13)\\
\geq&\mathbb{E}_{z'|x'}\left[\log q(z|z')\right] && \\
=&\mathbb{E}_{z'|x'}\left[\frac{d}{2}\log 2\pi - \frac{1}{2}\left(z-z'\right)^T\left(I\right))^{-1}\left(z-z'\right)\right] && (14)\\
=&\frac{d}{2}\log 2\pi - \frac{1}{2}\mathbb{E}_{z'|x'.}\ \left[(z-z')^2\right] && (15)\\
=&\frac{d}{2}\log 2\pi - \frac{1}{2}\mathbb{E}_{\epsilon}\left[\left(z-\mu(x')-L(x')\epsilon\right)^2\right] && (16)\\
=&\frac{d}{2}\log 2\pi - \frac{1}{2}\mathbb{E}_{\epsilon}\left[\left(z-\mu(x')\right)^2 - 2\left(z-\mu(x')*L(x')\epsilon\right) + \left(\left(L(x')\epsilon\right)^T\left(L(x')\epsilon\right)\right)\right] && (17)\\
=&\frac{d}{2}\log 2\pi - \frac{1}{2}\mathbb{E}_{\epsilon}\left[\left(z-\mu(x')\right)^2\right] + \left(z-\mu(x')L(x')\right)\mathbb{E}_{\epsilon}\left[\epsilon\right] - \frac{1}{2}\mathbb{E}_{\epsilon}\left[\epsilon^T L(x')^T L(x')\epsilon\right] && (18)\\
=&\frac{d}{2}\log 2\pi - \frac{1}{2}\left(z-\mu(x')\right)^2 - \frac{1}{2}Tr\log\Sigma(x') && (19)
\end{aligned}
$$

其中 $\mathbb{E}_{x'}\left[\log\ q(z|x')\right]\ =\ \mathbb{E}_{x'}\left[\log\mathbb{E}_{z'|x'}\left[q(z|z')\right]\right]\ \geq\ \mathbb{E}_{z'}\left[\log\ q(z|z')\right]$ 根据詹森不等式，$\mathbb{E}_{\epsilon}[\epsilon] = 0$ 以及 $\mathbb{E}_{\epsilon}\left[\epsilon\left(L(x')^T L(x')\ \epsilon\right.\right] = Tr\log\ \Sigma(x')$ 根据 Hutchinson 估计量。

$$
\begin{aligned}
\mathbb{E}_{z|x}\left[\mathbb{E}_{z'|x'}\left[\log q(z|z')\right]\right] =&\mathbb{E}_{z|x}\left[\frac{d}{2}\log 2\pi - \frac{1}{2}\left(z-\mu(x')\right)^2 - \frac{1}{2}Tr\log\Sigma(x')\right] && (20)\\
=&\frac{d}{2}\log 2\pi - \frac{1}{2}\mathbb{E}_{z|x}\left[\left(z-\mu(x')\right)^2\right] - \frac{1}{2}Tr\log\Sigma(x') && (21)\\
=&\frac{d}{2}\log 2\pi - \frac{1}{2}\mathbb{E}_{\epsilon}\left[\left(\mu(x)+L(x)\epsilon-\mu(x')\right)^2\right] - \frac{1}{2}Tr\log\Sigma(x') && (22)\\
=&\frac{d}{2}\log 2\pi - \frac{1}{2}\mathbb{E}_{\epsilon}\left[\left(\mu(x)-\mu(x')\right)^2\right] + \mathbb{E}_{\epsilon}\left[\left(\mu(x)-\mu(x')\right)L(x)\epsilon\right] && \\
&- \frac{1}{2}\mathbb{E}_{\epsilon}\left[\epsilon^T L(x)^T L(x)\epsilon\right] - \frac{1}{2}Tr\log\Sigma(x') && (23)\\
=&\frac{d}{2}\log 2\pi - \frac{1}{2}\left(\mu(x)-\mu(x')\right)^2 - \frac{1}{2}Tr\log\Sigma(x) - \frac{1}{2}Tr\log\Sigma(x') && (24)\\
=&\frac{d}{2}\log 2\pi - \frac{1}{2}\left(\mu(x)-\mu(x')\right)^2 - \frac{1}{2}\log\left(\left|\Sigma(x)\right|\cdot\left|\Sigma(x')\right|\right) && (25)
\end{aligned}
$$

附录 B 深度网络变换后的数据分布

定理 4. 给定 公式 (4) 中的设定，无条件 DNN 输出密度记为 Z近似于（假设高斯分布在其有效支撑集上截断，且该支撑集包含在 DN 输入空间划分的单个区域 ω 内）仿射变换后的分布的混合$\boldsymbol{x}|\boldsymbol{x}^*_{n(\boldsymbol{x})}$ 例如，对于高斯分布的情形

$$Z \sim \sum_{n=1}^{N} \mathcal{N}\Big(\boldsymbol{A}_{\omega(\boldsymbol{x}^*_n)}\boldsymbol{x}^*_n + \boldsymbol{b}_{\omega(\boldsymbol{x}^*_n)}, \boldsymbol{A}^T_{\omega(\boldsymbol{x}^*_n)}\Sigma_{\boldsymbol{x}^*_n}\boldsymbol{A}_{\omega(\boldsymbol{x}^*_n)}\Big)^{T=n},$$

其中， $\omega(\boldsymbol{x}^*_n) = \omega \in \Omega \iff \boldsymbol{x}^*_n \in \omega$ 是原型 $\boldsymbol{x}^*_n$ 所在的划分区域 .

证明。 我们知道，如果 $\int_\omega p(\boldsymbol{x}|\boldsymbol{x}^*_{n(\boldsymbol{x})})d\boldsymbol{x} \approx 1$ ，那么 f 在 p的有效支撑集内是线性的。因此，来自 p 的任意样本几乎必然位于单个区域 $\omega \in \Omega$ 内，因此，整个映射相对于 p都可以视为线性的。因此，输出分布是基于各区域仿射映射对输入分布进行线性变换的结果。 □

附录 C 额外的实证验证

为通过实证验证定理2，我们实证检验了以下最优解是否

$$\sum_i \log \frac{|\Sigma_Z|}{|\Sigma_{Z|X_i}||\Sigma_{Z'|X'_i}|}$$

是对角矩阵。我们在 CIFAR-10 上使用 ResNet18 训练了 VICReg，并对 Σ_Z进行了随机扰动（采用不同的尺度）。然后，对于每次扰动，我们计算了扰动后的矩阵与对角矩阵之间的平均距离，以及该项的实际值

$$\sum_i \log \frac{|\Sigma_Z||\Sigma_{Z'|X'_i}|}{|\Sigma_{Z|X_i}|}$$

。在图 3中，我们绘制了该项与最优值之差随其到对角矩阵距离的变化情况。可以看到，当我们接近对角矩阵时，便能得到最优解。这一观测为定理2提供了经验验证。

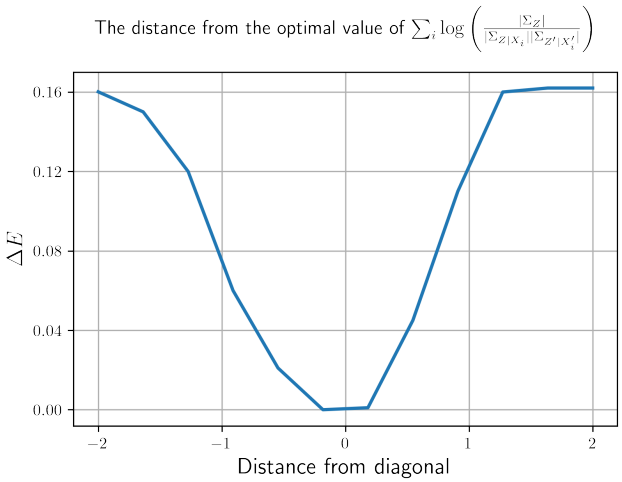

图 3：**该优化问题的最优解是一个对角矩阵**。不同扰动尺度下与对角矩阵的平均距离。实验采用 ResNet-18 网络，在 CIFAR-10 上进行。

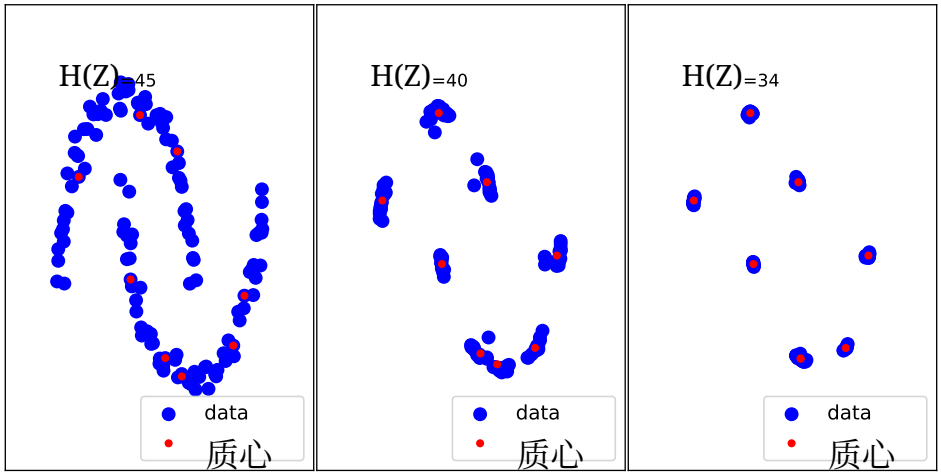

图 4：**在强制数据与质心之间建立类似于 K 均值的一一映射（即使用较小且固定的协方差矩阵）时，GMM 训练过程的演变。我们看到，坍塌并未发生**。左侧——在输入样本固定的情况下，我们观察到没有发生坍塌，并且中心的熵较高。右侧——当我们使输入样本可训练并优化其位置时，所有点都坍塌到同一个点上，导致熵急剧下降。

附录 D EM 与 GMM

让我们考察一个由两个交织月牙构成的玩具数据集，以说明 GMM 下的坍塌现象（图 1，右）。我们首先使用最大似然训练经典 GMM，其中均值根据随机样本初始化，协方差则设为单位矩阵。红点表示训练后的高斯分布均值，蓝点表示数据点。在输入样本固定的情况下，我们观察到没有发生坍塌，且中心的熵较高（图 4，左）。然而，当我们使输入样本可训练并优化其位置时，所有点都会坍塌到同一个点上，导致熵急剧下降（图 4，右）。

为防止坍塌，我们遵循 K 均值算法，强制使用稀疏后验，即使用较小的初始标准差，并且仅学习均值。这会强制建立一一映射，使所有点都趋近于均值而不会发生坍塌，从而得到较高的熵（图 4 ——附录中的中间图）。防止坍塌的另一种方法是对输入和参数使用不同的学习率。在这种设置下，参数的坍塌不会使似然最大化。 图 1 （右）展示了使用不同学习率训练输入和参数时 GMM 的结果。当参数学习率相对于输入学习率足够高时，熵下降得慢得多，并且不会发生坍塌。

附录 E SimCLR

在对比学习中，同一图像的不同增强视图会相互吸引（正样本对），而不同的增强视图则会相互排斥（负样本对）。MoCo [33] 和 SimCLR [16] 是自监督视觉表征学习的近期实例，它们缩小了自监督学习与全监督学习之间的差距。SimCLR 对一张图像应用随机增强以创建两个不同的视图， x 和 y，并使用共享编码器对二者进行编码，得到表征 r_x 和 r_y。 r_x 和 r_y 均经过 $l2$ 范数归一化。SimCLR 版本的 InfoNCE目标函数为：

$$\mathbb{E}_{x,y}\left[-\log\left(\frac{e^{\frac{1}{\eta}r_y^T r_x}}{\sum_{k=1}^{K} e^{\frac{1}{\eta}r_{y_k}^T r_x}}\right)\right],$$

其中， η 是温度项， K 是一个小批量中的视图数量。

附录 F 熵估计器

熵估计是信息论中的经典问题之一，其中高斯混合密度是最常见的表征之一。只要分量数量足够多，高斯混合就能以任意精度逼近任何平滑函数。然而，对于高斯混合，微分熵并不存在闭式解。文献中已有若干近似方法，包括较松的上界和下界 [35]。蒙特卡罗（MC）采样是近似高斯混合熵的一种方法。只要 MC 样本足够多，就能获得精度任意高的无偏熵估计。不幸的是，MC 采样的计算成本很高，通常需要大量样本，尤其是在高维情况下 [13]。VIGCreg 利用经验分布的前两个矩，采用了最直接的熵近似方法之一。尽管如此，先前的研究发现，在许多情况下，这种方法对熵的近似效果并不好 [35]。另一种选择是使用对数行列式函数。已有多种估计量被提出用于实现该函数，包括一致最小方差无偏（UMVU）方法 [2], 和贝叶斯方法 [50]。然而，这些方法往往需要复杂的优化。文献 [75] 中介绍的对数行列式估计器使用了缩放噪声下的微分熵 α 阶熵。他们证明，该方法可应用于高维特征，并且对随机噪声具有鲁棒性。基于泰勒级数展开， [35] 提出了高斯混合随机向量熵的一个下界。他们对每个高斯混合分量的对数进行泰勒级数展开，从而对熵度量进行解析求值。此外，他们还提出了一种拆分高斯密度的技术，以避免出现方差较大的分量，因为这类分量会导致计算成本高昂。Kolchinsky 和 Tracey [39] 提出了一类新颖的混合熵估计量。对于这类估计量中的每一个，都定义了一个分量密度之间的成对距离函数。只要成对距离函数和每个分量分布的熵易于计算，这些估计量的计算效率就很高。此外，该估计量连续且平滑，因此适用于优化问题。另外，他们还提出了熵的下界（使用切尔诺夫距离）和上界（使用 KL 散度）；当分量分布被划分为彼此充分分离的簇时，这两个界均为精确值，

附录 G 已知引理

我们在证明中使用以下众所周知的定理作为引理。为求完整起见，我们将它们列于下文。这些都是经典结果，并非我们的结果。

引理 G.1。（霍夫丁不等式）设 $X_1, ..., X_n$ 是相互独立的随机变量，满足 $a \leq X_i \leq b$ 几乎必然成立。考虑这些随机变量的平均值，$S_n = \frac{1}{n}(X_1 + \cdots + X_n)$。那么，对于所有 $t > 0$,

$$\mathbb{P}_S\left(\mathrm{E}\left[S_n\right] - S_n \geq (b-a)\sqrt{\frac{\ln(1/\delta)}{2n}}\right) \leq \delta,$$

and

$$\mathbb{P}_S\left(S_n - \mathrm{E}\left[S_n\right] \geq (b-a)\sqrt{\frac{\ln(1/\delta)}{2n}}\right) \leq \delta.$$

证明。 利用霍夫丁不等式可得，对于所有 $t > 0$,

$$\mathbb{P}_S\left(\mathrm{E}\left[S_n\right] - S_n \geq t\right) \leq \exp\left(-\frac{2nt^2}{(b-a)^2}\right),$$

and

$$\mathbb{P}_S\left(S_n - \mathrm{E}\left[S_n\right] \geq t\right) \leq \exp\left(-\frac{2nt^2}{(b-a)^2}\right),$$

令 $\delta = \exp\left(-\frac{2nt^2}{(b-a)^2}\right)$，并求解 $t > 0$，

$$
\begin{aligned}
1/\delta &= \exp\left(\frac{2nt^2}{(b-a)^2}\right) \\
\implies \ln(1/\delta) &= \frac{2nt^2}{(b-a)^2} \\
\implies \frac{(b-a)^2 \ln(1/\delta)}{2n} &= t^2 \\
\implies t &= (b-a)\sqrt{\frac{\ln(1/\delta)}{2n}}
\end{aligned}
$$

□

已有研究表明，可以通过拉德马赫复杂度得到泛化界 [8, 51, 60]。以下是对 [51, 定理 31] 的一个简单修改，用于给出非负一般损失函数的单侧界：

引理 G.2。设 $\mathcal{G}$ 是一组陪域为 $[0, M]$ 的函数。那么，对于任意 $\delta > 0$，以至少 $1-\delta$ 的概率，在对 m 个样本 $S = (q_i)_{i=1}^m$ 进行独立同分布抽样后，以下结论对所有 $\psi \in \mathcal{G}$ 均成立：

$$
\mathbb{E}_q[\psi(q)] \leq \frac{1}{m}\sum_{i=1}^m \psi(q_i) + 2\mathcal{R}_m(\mathcal{G}) + M\sqrt{\frac{\ln(1/\delta)}{2m}}, \tag{26}
$$

其中 $\mathcal{R}_m(\mathcal{G}) := \mathbb{E}_{S,\ \xi}[\text{上确界}_{\psi\in\mathcal{G}}\ \frac{1}{m}\sum_{i=1}^m \xi_i\psi(q_i)]$，且 ξ_1、……、ξ_m 是取值于 $\{-1, 1\}$ 的独立均匀随机变量。

证明。设 $S = (q_i)_{i=1}^m$，且 $S' = (q_i')_{i=1}^m$。定义

$$
\varphi(S) = \sup_{\psi\in\mathcal{G}} \mathbb{E}_{x,y}[\psi(q)] - \frac{1}{m}\sum_{i=1}^m \psi(q_i). \tag{27}
$$

为了将 McDiarmid 不等式应用于 $\varphi(S)$，我们计算 $|\varphi(S) - \varphi(S')|$ 的上界，其中 S 和 S' 是仅在任意索引 i_0 处恰好相差一个点的两个测试数据集；即，对于所有 $i \neq i_0$ 和 $S_{i_0} \neq S'_{i_0}$，$S_i = S'_i$。于是，

$$
\varphi(S') - \varphi(S) \leq \sup_{\psi\in\mathcal{G}} \frac{\psi(q_{i_0}) - \psi(q'_{i_0})}{m} \leq \frac{M}{m}. \tag{28}
$$

类似地，$\varphi(S) - \varphi(S') \leq \frac{M}{m}$。因此，根据 McDiarmid 不等式，对于任意 $\delta > 0$，以至少 $1-\delta$ 的概率，

$$
\varphi(S) \leq \mathbb{E}_S[\varphi(S)] + M\sqrt{\frac{\ln(1/\delta)}{2m}}. \tag{29}
$$

此外，

$$
\begin{aligned}
\mathbb{E}_S[\varphi(S)] &= \mathbb{E}_S\left[\sup_{\psi\in\mathcal{G}} \mathbb{E}_{S'}\left[\frac{1}{m}\sum_{i=1}^m \psi(q_i')\right] - \frac{1}{m}\sum_{i=1}^m \psi(q_i)\right] && (30)\\
&\leq \mathbb{E}_{S,S'}\left[\sup_{\psi\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^m (\psi(q_i') - \psi(q_i))\right] && (31)\\
&\leq \mathbb{E}_{\xi,S,S'}\left[\sup_{\psi\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^m \xi_i(\psi(q_i') - \psi(q_i))\right] && (32)\\
&\leq 2\mathbb{E}_{\xi,S}\left[\sup_{\psi\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^m \xi_i\psi(q_i)\right] = 2\mathcal{R}_m(\mathcal{G}), && (33)
\end{aligned}
$$

第一行根据各项的定义得出，第二行运用了 Jensen 不等式和上确界的凸性，第三行则是因为对于每个 $\xi_i \in \{-1, +1\}$，每一项 $\xi_i(\ell(f(x_i')\ y_i') - \ell(f(x_i)\ y_i))$ 的分布与 $(\ell(f(x_i')\ y_i') - \ell(f(x_i)\ y_i))$ 的分布相同，因为 S 和 S' 是从同一分布中独立同分布抽取的。第四行运用了上确界的次可加性。 □

附录 H 最大熵估计器的实现细节

本节将详细介绍第 5.2 节中所进行实验的实施情况。

Set上我们的实验在 CIFAR-10 [41] 上进行。我们使用 ResNet-18 [32] 作为骨干网络 .

训练流程：实验过程分为两个连续阶段：无监督预训练和随后的线性评估。首先进行无监督预训练，在此期间训练编码器网络。完成后，我们进入线性评估阶段，以评估预训练编码器所生成表征的质量。

预训练阶段结束后，我们遵循既有基线方法所采用的微调流程，详见 [14]。

在线性评估阶段，我们首先对线性分类器进行有监督训练。具体而言，我们使用编码器网络生成的表征，同时冻结网络参数，并使用相同的训练数据集。随后，我们使用单独的验证数据集测量训练后的线性分类器的测试准确率。这种方法使我们能够以稳健且系统的方式评估模型性能。
每个模型的训练过程持续 800 个 epoch，批次大小为 512。我们采用随机梯度下降（SGD）优化器，动量设为 0.9，权重衰减设为 $1e^-4$。学习率初始值为 0.5，并按照余弦衰减调度进行调整，同时结合线性预热阶段。

在数据增强过程中，每张输入图像都会生成两个增强版本。具体做法是对每张图像进行随机裁剪，再将其调整回原始分辨率。随后，对图像进行随机水平翻转、颜色抖动、灰度转换、高斯模糊和极化，以进一步增强数据。

在线性评估阶段，线性分类器采用批次大小为 256 进行 100 个 epoch 的训练。此阶段同样使用 SGD 优化器，动量设为 0.9，且不进行权重衰减。学习率按照余弦衰减调度进行调整，初始值为 0.2，最低降至 $2e-4$。

附录 I 下游任务的泛化界

在本附录中，我们给出定理 3 的完整版本，并附上其证明和补充讨论。

I.1 补充符号与细节

首先介绍补充符号和细节。我们用 $x \in \mathcal{X}$ 表示输入，用$y \in \mathcal{Y} \subseteq \mathbb{R}^r$ 表示输出。定义 $p(y) = \mathbb{P}(Y = y)$ 为获得标签 y 的概率，并令 $\hat{p}(y) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{y_i = y\}$ 为 $p(y)$ 的经验估计。令 ζ 为标签范数的上界，使得对所有 $y \in \mathcal{Y}$，均有 $\|y\|_2 \leq \zeta$。定义无标签数据的最小范数解 $W_{\bar{S}}$ 为 $W_{\bar{S}} = \text{minimize}_{W'} \|W'\|_F \text{s.t. } W' \in \arg\min_W \frac{1}{m}\sum_{i=1}^{m} \|Wf_\theta(x_i^+) - g^*(x_i^+)\|^2$。令 κ_S为使用训练后的模型时，每个样本的欧几里得范数损失的数据依赖上界，使得对所有 $(x, y) \in \mathcal{X} \times \mathcal{Y}$，均有$\|W_S f_\theta(x) - y\| \leq \kappa_S$。类似地，令 $\kappa_{\bar{S}}$为每个样本的欧几里得范数损失的数据依赖上界，使得对所有 $(x, y) \in \mathcal{X} \times \mathcal{Y}$，均有 $\|W_{\bar{S}} f_\theta(x) - y\| \leq \kappa_{\bar{S}}$。定义 W_S与 $W_{\bar{S}}$之差为 $c = \|W_S - W_{\bar{S}}\|_2$。令 $\mathcal{W}$为 W的假设空间，使得 $W_{\bar{S}} \in \mathcal{W}$。我们用 $\tilde{\mathcal{R}}_m(\mathcal{W} \circ \mathcal{F}) = \frac{1}{\sqrt{m}} \mathbb{E}_{\bar{S},\xi}[\sup_{W \in \mathcal{W}, f \in \mathcal{F}} \sum_{i=1}^{m} \xi_i \|g^*(x_i^+) - Wf(x_i^+)\|]$ 表示集合 $\{x^+ \mapsto \|g^*(x^+) - Wf(x^+)\| : W \in \mathcal{W}, f \in \overline{\mathcal{F}}\}$的归一化 Rademacher 复杂度。我们用 κ 表示每个样本的欧几里得范数损失的一个上界，使得对所有 $(x, y, W, f) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{W} \times \mathcal{F}$，均有 $\|Wf(x) - y\| \leq \kappa$。

我们采用以下数据生成过程模型，该模型曾用于一篇分析对比学习的论文中 [59, 10]。对于有标签数据，首先从 ρ 分布中抽取 y ，其中 $\mathcal{Y}$，然后从以标签 y为条件的条件分布 $\mathcal{D}_y$ 中抽取 x 。也就是说，我们有联合分布 $\mathcal{D}(x, y) = \mathcal{D}_y(x)\rho(y)$，其中 $((x_i, y_i))_{i=1}^n \sim \mathcal{D}^n$。对于无标签数据，

首先，从分布 ρ中分别抽取每个未知标签 y^+ 和 y^-，然后从条件分布 $\mathcal{D}_{y^+}$ 中分别抽取每个正例 x^+ 和 x^{++}，而负例 x^- 则从 $\mathcal{D}_{y^-}$中抽取。不同于对比学习的分析，我们不需要负样本。设 $\tau_{\bar{S}}$为不变性损失的数据依赖上界，对于训练表示 $\|f_\theta$，对所有$(\bar{x}, x) \sim \mathcal{D}_y^2$和 $y \in \mathcal{Y}$，均有$\bar{x}(-f_\theta)\, x(\| \leq \tau_{\bar{S}})\| \leq \tau_{\bar{S}}$。设 τ为不变性损失的数据无关上界，对于训练表示$\|f$，对所有$(\bar{x}, x) \sim \mathcal{D}_y^2$、$y \in \mathcal{Y}$和 $f \in \mathcal{F}$，均有$\bar{x}(-f)\, x(\| \leq \tau)\| \leq \tau$。为简单起见，我们假设存在一个函数 g^*，使得对所有$(x, y) \in \mathcal{X} \times \mathcal{Y}$，均有$y = g^*(x) \in \mathbb{R}^r$。若不作此假设，则最终结果中会增加标签噪声的平均值；假设标签噪声的均值为零，则当样本量 n 和 m 增加时，该项趋于零。

0.2 定理 3 的证明3

定理 3 的证明。令 $W = W_S$，其中 W_S是如下问题的最小范数解： W_S =最小化$_{W'} \|W'\|_F$，满足 $W' \in \arg\min_W \frac{1}{n}\sum_{i=1}^n \|W f_\theta(x_i) - y_i\|^2$。令 $W^* = W_{\bar{S}}$，其中$W_{\bar{S}}$是如下问题的最小范数解：$W^* = W_{\bar{S}}$ =最小化$_{W'} \|W'\|_F$，满足$W' \in \arg\min_W \frac{1}{m}\sum_{i=1}^m \|W f_\theta(x_i^+) - g^*(x_i^+)\|^2$。由于 $y = g^*(x)$,

$$y = g^*(x) \pm W^* f_\theta(x) = W^* f_\theta(x) + (g^*(x) - W^* f_\theta(x)) = W^* f_\theta(x) + \varphi(x)$$

其中 $\varphi(x) = g^*(x) - W^* f_\theta(x)$。定义 $L_S(w) = \frac{1}{n}\sum_{i=1}^n \|W f_\theta(x_i) - y_i\|$。利用这些,

$$\begin{aligned}
L_S(w) &= \frac{1}{n}\sum_{i=1}^n \|W f_\theta(x_i) - y_i\| \\
&= \frac{1}{n}\sum_{i=1}^n \|W f_\theta(x_i) - W^* f_\theta(x_i) - \varphi(x_i)\| \\
&\geq \frac{1}{n}\sum_{i=1}^n \|W f_\theta(x_i) - W^* f_\theta(x_i)\| - \frac{1}{n}\sum_{i=1}^n \|\varphi(x_i)\| \\
&= \frac{1}{n}\sum_{i=1}^n \|\tilde{W} f_\theta(x_i)\| - \frac{1}{n}\sum_{i=1}^n \|\varphi(x_i)\|
\end{aligned}$$

˜其中 $W = W - W^*$。现考虑用于 $i = 1$的新新样本 $\bar{x}_i \sim \mathcal{D}_{y_i}$，.....， n，以进一步将上述内容改写为:

$$\begin{aligned}
L_S(w) &\geq \frac{1}{n}\sum_{i=1}^n \|\tilde{W} f_\theta(x_i) \pm \tilde{W} f_\theta(\bar{x}_i)\| - \frac{1}{n}\sum_{i=1}^n \|\varphi(x_i)\| \\
&= \frac{1}{n}\sum_{i=1}^n \|\tilde{W} f_\theta(\bar{x}_i) - (\tilde{W} f_\theta(\bar{x}_i) - \tilde{W} f_\theta(x_i))\| - \frac{1}{n}\sum_{i=1}^n \|\varphi(x_i)\| \\
&\geq \frac{1}{n}\sum_{i=1}^n \|\tilde{W} f_\theta(\bar{x}_i)\| - \frac{1}{n}\sum_{i=1}^n \|\tilde{W} f_\theta(\bar{x}_i) - \tilde{W} f_\theta(x_i)\| - \frac{1}{n}\sum_{i=1}^n \|\varphi(x_i)\| \\
&= \frac{1}{n}\sum_{i=1}^n \|\tilde{W} f_\theta(\bar{x}_i)\| - \frac{1}{n}\sum_{i=1}^n \|\tilde{W}(f_\theta(\bar{x}_i) - f_\theta(x_i))\| - \frac{1}{n}\sum_{i=1}^n \|\varphi(x_i)\|
\end{aligned}$$

这意味着

$$\frac{1}{n}\sum_{i=1}^n \|\tilde{W} f_\theta(\bar{x}_i)\| \leq L_S(w) + \frac{1}{n}\sum_{i=1}^n \|\tilde{W}(f_\theta(\bar{x}_i) - f_\theta(x_i))\| + \frac{1}{n}\sum_{i=1}^n \|\varphi(x_i)\|.$$

此外，由于 $y = W^* f_\theta(x) + \varphi(x)$，将 $\bar{y}_i = W^* f_\theta(\bar{x}_i) + \varphi(\bar{x}_i)$ 写作（其中 $\bar{y}_i = y_i$ since $\bar{x}_i \sim \mathcal{D}_{y_i}$ for $i = 1, \ldots\ldots, n$），

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\|\tilde{W} f_\theta(\bar{x}_i)\| &= \frac{1}{n}\sum_{i=1}^{n}\|W f_\theta(\bar{x}_i) - W^* f_\theta(\bar{x}_i)\| \\ &= \frac{1}{n}\sum_{i=1}^{n}\|W f_\theta(\bar{x}_i) - \bar{y}_i + \varphi(\bar{x}_i)\| \\ &\geq \frac{1}{n}\sum_{i=1}^{n}\|W f_\theta(\bar{x}_i) - \bar{y}_i\| - \frac{1}{n}\sum_{i=1}^{n}\|\varphi(\bar{x}_i)\|\end{aligned}$$

综合以上结果，我们得到

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\|W f_\theta(\bar{x}_i) - \bar{y}_i\| \leq L_S(w) + \frac{1}{n}\sum_{i=1}^{n}\|\tilde{W}(f_\theta(\bar{x}_i) - f_\theta(x_i))\| \\ + \frac{1}{n}\sum_{i=1}^{n}\|\varphi(x_i)\| + \frac{1}{n}\sum_{i=1}^{n}\|\varphi(\bar{x}_i)\|.\end{aligned} \tag{34}$$

为了界定方程 34左侧的上界，现在我们分析以下随机变量：

$$\mathbb{E}_{X,Y}[\|W_S f_\theta(X) - Y\|] - \frac{1}{n}\sum_{i=1}^{n}\|W_S f_\theta(\bar{x}_i) - \bar{y}_i\|, \tag{35}$$

其中 $\bar{y}_i = y_i$ 由于 $\bar{x}_i \sim \mathcal{D}_{y_i}$ 对于 $i = 1, \ldots\ldots, n$。重要的是，这意味着由于 W_S 取决于 y_i，W_S 取决于 $\bar{y}_i$。因此，随机变量集合 $\|W_S f_\theta(\bar{x}_1) - \bar{y}_1\|, \ldots\ldots, \|W_S f_\theta(n_n) - \bar{y}_n\|$ 并非相互独立。相应地，我们无法应用标准集中不等式来界定方程 35。学习理论中的一种标准方法是，先将方程 35 的上界界定为 $\mathbb{E}_x, {}_y\|W_S f_\theta(x) - y\| - \frac{1}{n}\sum_{i=1}^{n}\|W_S f_\theta(\bar{x}_i) - \bar{y}_i\| \leq \sup_{W\in\mathcal{W}} \mathbb{E}_x, {}_y\|W f_\theta(x) - y\| - \frac{1}{n}\sum_{i=1}^{n}\|W f_\theta(\bar{x}_i) - \bar{y}_i\|$，其中假设空间 $\mathcal{W}$（与 S 独立），并注意到右侧现在包含相互独立的随机变量集合 $\|W f_\theta(\bar{x}_1) - \bar{y}_1\|, \ldots\ldots, \|W f_\theta(n_n) - \bar{y}_n\|$，因此我们可以对其使用标准集中不等式。这一推理引出了假设空间 $\mathcal{W}$的拉德马赫复杂度。然而，假设空间 $\mathcal{W}$ 的复杂度可能非常大，从而导致界限较松。在此证明中，我们将展示，通过采用一种截然不同的方法，利用条件期望处理相关随机变量 $\|W_S f_\theta(\bar{x}_1) - \bar{y}_1\|, \ldots\ldots, \|W_S f_\theta(n_n) - \bar{y}_n\|$，可以避免依赖假设空间 $\mathcal{W}$。直观而言，我们利用了这样一个事实：对于这些相关随机变量，在给定每个 $y \in \mathcal{Y}$ 的条件下，它们具有条件独立结构。

我们首先将期望损失写成条件期望损失之和：

$$\begin{aligned}\mathbb{E}_{X,Y}[\|W_S f_\theta(X) - Y\|] &= \sum_{y\in\mathcal{Y}} \mathbb{E}_{X,Y}[\|W_S f_\theta(X) - Y\| \mid Y = y]\mathbb{P}(Y = y) \\ &= \sum_{y\in\mathcal{Y}} \mathbb{E}_{X_y}[\|W_S f_\theta(X_y) - y\|]\mathbb{P}(Y = y),\end{aligned}$$

其中，X_y 是条件分布对应的随机变量，且条件为 $Y = y$。据此，我们将公式35 分解为两项：

$$\begin{aligned}&\mathbb{E}_{X,Y}[\|W_S f_\theta(X) - Y\|] - \frac{1}{n}\sum_{i=1}^{n}\|W_S f_\theta(\bar{x}_i) - \bar{y}_i\| \\ &= \left(\sum_{y\in\mathcal{Y}} \mathbb{E}_{X_y}[\|W_S f_\theta(X_y) - y\|]\frac{|\mathcal{I}_y|}{n} - \frac{1}{n}\sum_{i=1}^{n}\|W_S f_\theta(\bar{x}_i) - \bar{y}_i\|\right) \\ &\quad + \sum_{y\in\mathcal{Y}} \mathbb{E}_{X_y}[\|W_S f_\theta(X_y) - y\|]\left(\mathbb{P}(Y = y) - \frac{|\mathcal{I}_y|}{n}\right),\end{aligned} \tag{36}$$

其中

$$\mathcal{I}_y = \{i \in [n] : y_i = y\}.$$

等式右侧的第一项 36 可进一步利用以下方法化简

$$\frac{1}{n}\sum_{i=1}^{n} \|W_S f_\theta(\bar{x}_i) - \bar{y}_i\| = \frac{1}{n}\sum_{y\in\mathcal{Y}}\sum_{i\in\mathcal{I}_y} \|W_S f_\theta(\bar{x}_i) - y\|,$$

as

$$\sum_{y\in\mathcal{Y}} \mathbb{E}_{X_y}[\|W_S f_\theta(X_y) - y\|]\frac{|\mathcal{I}_y|}{n} - \frac{1}{n}\sum_{i=1}^{n} \|W_S f_\theta(\bar{x}_i) - \bar{y}_i\|$$

$$= \frac{1}{n}\sum_{y\in\tilde{\mathcal{Y}}} |\mathcal{I}_y| \left(\mathbb{E}_{X_y}[\|W_S f_\theta(X_y) - y\|] - \frac{1}{|\mathcal{I}_y|}\sum_{i\in\mathcal{I}_y} \|W_S f_\theta(\bar{x}_i) - y\| \right),$$

˜其中 $\mathcal{Y} = \{y \in \mathcal{Y} : \ |\mathcal{I}_y| \neq 0\}$。将这些代入等式等式36可得

$$\mathbb{E}_{X,Y}[\|W_S f_\theta(X) - Y\|] - \frac{1}{n}\sum_{i=1}^{n} \|W_S f_\theta(\bar{x}_i) - \bar{y}_i\| \tag{37}$$

$$= \frac{1}{n}\sum_{y\in\tilde{\mathcal{Y}}} |\mathcal{I}_y| \left(\mathbb{E}_{X_y}[\|W_S f_\theta(X_y) - y\|] - \frac{1}{|\mathcal{I}_y|}\sum_{i\in\mathcal{I}_y} \|W_S f_\theta(\bar{x}_i) - y\| \right)$$

$$+ \sum_{y\in\mathcal{Y}} \mathbb{E}_{X_y}[\|W_S f_\theta(X_y) - y\|] \left(\mathbb{P}(Y = y) - \frac{|\mathcal{I}_y|}{n} \right)$$

需要注意的是，尽管等式 37 右侧的 $\|W_S f_\theta(\bar{x}_1) - \bar{y}_1\|, \ldots, \|W_S f_\theta(\bar{x}_n) - \bar{y}_n\|$ 是相关随机变量，但 $\|W_S f_\theta(\bar{x}_1) - y\|, \ldots, \|W_S f_\theta(\bar{x}_n) - y\|$ 是独立随机变量，因为 W_S 和 $\bar{x}_i$ 相互独立，且此处 y 为固定值。因此，根据霍夫丁不等式˜（引理 G.1），并对 $y \in \mathcal{Y}$ 使用并集界，我们可以得到，以至少 $1 - \delta$ 的概率，以下结论对所有 $y \in \mathcal{Y}$ 均成立：

$$\mathbb{E}_{X_y}[\|W_S f_\theta(X_y) - y\|] - \frac{1}{|\mathcal{I}_y|}\sum_{i\in\mathcal{I}_y} \|W_S f_\theta(\bar{x}_i) - y\| \leq \kappa_S \sqrt{\frac{\ln(|\tilde{\mathcal{Y}}|/\delta)}{2|\mathcal{I}_y|}}.$$

这意味着其概率至少为 $1 - \delta$,

$$\frac{1}{n}\sum_{y\in\tilde{\mathcal{Y}}} |\mathcal{I}_y| \left(\mathbb{E}_{X_y}[\|W_S f_\theta(X_y) - y\|] - \frac{1}{|\mathcal{I}_y|}\sum_{i\in\mathcal{I}_y} \|W_S f_\theta(\bar{x}_i) - y\| \right)$$

$$\leq \frac{\kappa_S}{n}\sum_{y\in\tilde{\mathcal{Y}}} |\mathcal{I}_y| \sqrt{\frac{\ln(|\tilde{\mathcal{Y}}|/\delta)}{2|\mathcal{I}_y|}}$$

$$= \kappa_S \left(\sum_{y\in\tilde{\mathcal{Y}}} \sqrt{\frac{|\mathcal{I}_y|}{n}} \right) \sqrt{\frac{\ln(|\tilde{\mathcal{Y}}|/\delta)}{2n}}.$$

将此界代入方程 37，我们可以得到，以至少 $1 - \delta$ 的概率，

$$\mathbb{E}_{X,Y}[\|W_S f_\theta(X) - Y\|] - \frac{1}{n}\sum_{i=1}^{n} \|W_S f_\theta(\bar{x}_i) - \bar{y}_i\| \tag{38}$$

$$\leq \kappa_S \left(\sum_{y\in\tilde{\mathcal{Y}}} \sqrt{\hat{p}(y)} \right) \sqrt{\frac{\ln(|\tilde{\mathcal{Y}}|/\delta)}{2n}} + \sum_{y\in\mathcal{Y}} \mathbb{E}_{X_y}[\|W_S f_\theta(X_y) - y\|] \left(\mathbb{P}(Y = y) - \frac{|\mathcal{I}_y|}{n} \right)$$

其中

$$\hat{p}(y) = \frac{|\mathcal{I}_y|}{n}.$$

此外，对于等式 38右侧的第二项，根据 [37], 中的引理 1，我们有，以至少 $1-\delta$的概率，

$$\begin{aligned}
&\sum_{y\in\mathcal{Y}} \mathbb{E}_{X_y}[\|W_S f_\theta(X_y) - y\|] \left(\mathbb{P}(Y=y) - \frac{|\mathcal{I}_y|}{n}\right) \\
&\leq \left(\sum_{y\in\mathcal{Y}} \sqrt{p(y)}\mathbb{E}_{X_y}[\|W_S f_\theta(X_y) - y\|\right) \sqrt{\frac{2\ln(|\mathcal{Y}|/\delta)}{2n}} \\
&\leq \kappa_S \left(\sum_{y\in\mathcal{Y}} \sqrt{p(y)}\right) \sqrt{\frac{2\ln(|\mathcal{Y}|/\delta)}{2n}}
\end{aligned}$$

其中 $p(y) = \mathbb{P}(Y=y)$。将此界代入等式 38 并使用并集界，我们有，以至少 $1-\delta$的概率，

$$\begin{aligned}
&\mathbb{E}_{X,Y}[\|W_S f_\theta(X) - Y\|] - \frac{1}{n}\sum_{i=1}^n \|W_S f_\theta(\bar{x}_i) - \bar{y}_i\| \qquad (39)\\
&\leq \kappa_S \left(\sum_{y\in\tilde{\mathcal{Y}}} \sqrt{\hat{p}(y)}\right) \sqrt{\frac{\ln(2|\tilde{\mathcal{Y}}|/\delta)}{2n}} + \kappa_S \left(\sum_{y\in\mathcal{Y}} \sqrt{p(y)}\right) \sqrt{\frac{2\ln(2|\mathcal{Y}|/\delta)}{2n}} \\
&\leq \left(\sum_{y\in\mathcal{Y}} \sqrt{\hat{p}(y)}\right) \kappa_S \sqrt{\frac{2\ln(2|\mathcal{Y}|/\delta)}{2n}} + \left(\sum_{y\in\mathcal{Y}} \sqrt{p(y)}\right) \kappa_S \sqrt{\frac{2\ln(2|\mathcal{Y}|/\delta)}{2n}} \\
&\leq \kappa_S \sqrt{\frac{2\ln(2|\mathcal{Y}|/\delta)}{2n}} \sum_{y\in\mathcal{Y}} \left(\sqrt{\hat{p}(y)} + \sqrt{p(y)}\right)
\end{aligned}$$

结合等式 34 和等式 39 可得，以至少 $1-\delta$的概率，

$$\begin{aligned}
&\mathbb{E}_{X,Y}[\|W_S f_\theta(X) - Y\|] \qquad (40)\\
&\leq \frac{1}{n}\sum_{i=1}^n \|W_S f_\theta(\bar{x}_i) - \bar{y}_i\| + \kappa_S \sqrt{\frac{2\ln(2|\mathcal{Y}|/\delta)}{2n}} \sum_{y\in\mathcal{Y}} \left(\sqrt{\hat{p}(y)} + \sqrt{p(y)}\right) \\
&\leq L_S(w_S) + \frac{1}{n}\sum_{i=1}^n \|\tilde{W}(f_\theta(\bar{x}_i) - f_\theta(x_i))\| \\
&\quad + \frac{1}{n}\sum_{i=1}^n \|\varphi(x_i)\| + \frac{1}{n}\sum_{i=1}^n \|\varphi(\bar{x}_i)\| + \kappa_S \sqrt{\frac{2\ln(2|\mathcal{Y}|/\delta)}{2n}} \sum_{y\in\mathcal{Y}} \left(\sqrt{\hat{p}(y)} + \sqrt{p(y)}\right).
\end{aligned}$$

现在我们将分析等式 $\frac{1}{n}\sum_{i=1}^n \|\varphi(x_i)\| + \frac{1}{n}\sum_{i=1}^n \|\varphi(\bar{x}_i)\|$ 右侧的项，该项位于等式 40中。由于 $W^* = W_{\bar{S}}$,

$$\frac{1}{n}\sum_{i=1}^n \|\varphi(x_i)\| = \frac{1}{n}\sum_{i=1}^n \|g^*(x_i) - W_{\bar{S}} f_\theta(x_i)\|.$$

利用霍夫丁不等式（引理 G.1），我们可得，对于任意 $\delta > 0$，其成立的概率至少为$1-\delta$,

$$\frac{1}{n}\sum_{i=1}^n \|\varphi(x_i)\| \leq \frac{1}{n}\sum_{i=1}^n \|g^*(x_i) - W_{\bar{S}} f_\theta(x_i)\| \leq \mathbb{E}_{x^+}[\|g^*(x^+) - W_{\bar{S}} f_\theta(x^+)\|] + \kappa_{\bar{S}} \sqrt{\frac{\ln(1/\delta)}{2n}}.$$

此外，利用 [51, 定理 3.1] 及损失函数 $x^+ \mapsto \|g^*(x^+) - Wf(x^+)\|$ (即引理 G.2)，我们可得，对于任意 $\delta > 0$，其成立的概率至少为 $1-\delta$，

$$\mathbb{E}_{x^+}[\|g^*(x^+) - W_{\bar{S}} f_\theta(x^+)\|] \le \frac{1}{m}\sum_{i=1}^m \|g^*(x_i^+) - W_{\bar{S}} f_\theta(x_i^+)\| + \frac{2\tilde{\mathcal{R}}_m(\mathcal{W}\circ\mathcal{F})}{\sqrt{m}} + \kappa\sqrt{\frac{\ln(1/\delta)}{2m}} \tag{41}$$

其中 $\tilde{\mathcal{R}}_m(\mathcal{W}\circ\mathcal{F}) = \frac{1}{\sqrt{m}}\mathbb{E}_{\bar{S},\xi}[\sup_{W\in\mathcal{W}, f\in\mathcal{F}} \sum_{i=1}^m \xi_i \|g^*(x_i^+) - Wf(x_i^+)\|]$ 是集合 $\{x^+ \mapsto \|g^*(x^+) - Wf(x^+)\| : W \in \mathcal{W}, f \in \mathcal{F}\}$ 的归一化 Rademacher 复杂度（其归一化方式为˜使得 $\mathcal{R}_m(\mathcal{F}) = O(1)$，当 $m \to \infty$ 取常见值时 $\mathcal{F}$），而 $\xi_1, \ldots, \xi_m$ 是取值于 $\{-1, 1\}$中的独立均匀随机变量。利用并集界可得，对于任意$\delta > 0$，以至少 $1-\delta$的概率，

$$\frac{1}{n}\sum_{i=1}^n \|\varphi(x_i)\| \le \frac{1}{m}\sum_{i=1}^m \|g^*(x_i^+) - W_{\bar{S}} f_\theta(x_i^+)\| + \frac{2\tilde{\mathcal{R}}_m(\mathcal{W}\circ\mathcal{F})}{\sqrt{m}} + \kappa\sqrt{\frac{\ln(2/\delta)}{2m}} + \kappa_{\bar{S}}\sqrt{\frac{\ln(2/\delta)}{2n}}$$

类似地，对于任意 $\delta > 0$，以至少 $1-\delta$的概率，

$$\frac{1}{n}\sum_{i=1}^n \|\varphi(\bar{x}_i)\| \le \frac{1}{m}\sum_{i=1}^m \|g^*(x_i^+) - W_{\bar{S}} f_\theta(x_i^+)\| + \frac{2\tilde{\mathcal{R}}_m(\mathcal{W}\circ\mathcal{F})}{\sqrt{m}} + \kappa\sqrt{\frac{\ln(2/\delta)}{2m}} + \kappa_{\bar{S}}\sqrt{\frac{\ln(2/\delta)}{2n}}.$$

因此，通过应用并集界可知，对任意 $\delta > 0$，其概率至少为 $1-\delta$，

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^n \|\varphi(x_i)\| + \frac{1}{n}\sum_{i=1}^n \|\varphi(\bar{x}_i)\| \\ &\le \frac{2}{m}\sum_{i=1}^m \|g^*(x_i^+) - W_{\bar{S}} f_\theta(x_i^+)\| + \frac{4\mathcal{R}_m(\mathcal{W}\circ\mathcal{F})}{\sqrt{m}} + 2\kappa\sqrt{\frac{\ln(4/\delta)}{2m}} + 2\kappa_{\bar{S}}\sqrt{\frac{\ln(4/\delta)}{2n}} \end{aligned} \tag{42}$$

为了分析等式右侧的第一项 42，回顾一下

$$W_{\bar{S}} = \underset{W'}{\text{minimize}} \|W'\|_F \text{ s.t. } W' \in \arg\min_W \frac{1}{m}\sum_{i=1}^m \|Wf_\theta(x_i^+) - g^*(x_i^+)\|^2. \tag{43}$$

这里，由于 $Wf_\theta(x_i^+) \in \mathbb{R}^r$，我们有

$$Wf_\theta(x_i^+) = \text{vec}[Wf_\theta(x_i^+)] = [f_\theta(x_i^+)^\top \otimes I_r]\,\text{vec}[W] \in \mathbb{R}^r,$$

其中 $I_r \in \mathbb{R}^{r\times r}$ 是单位矩阵，而 $[f_\theta(x_i^+)^\top \otimes I_r] \in \mathbb{R}^{r\times dr}$ 是这两个矩阵的克罗内克积，并且 $\text{vec}[W] \in \mathbb{R}^{dr}$ 是矩阵 $W \in \mathbb{R}^{r\times d}$的向量化。因此，定义 $A_i = [f_\theta(x_i^+)^\top \otimes I_r] \in \mathbb{R}^{r\times dr}$ 并采用 $w = \text{vec}[W]$ 及其逆映射$W = \text{vec}^{-1}[w]$ (即在固定顺序下，从 $\mathbb{R}^{r\times d}$ 到 $\mathbb{R}^{dr}$ 的向量化的逆映射)，我们可以将等式 43改写为

$$W_{\bar{S}} = \text{vec}^{-1}[w_{\bar{S}}] \quad \text{where} \quad w_{\bar{S}} = \underset{w'}{\text{minimize}} \|w'\|_F \text{ s.t. } w' \in \arg\min_w \sum_{i=1}^m \|g_i - A_i w\|^2,$$

其中 $g_i = g^*(x_i^+) \in \mathbb{R}^r$。由于函数 $w \mapsto \sum_{i=1}^m \|g_i - A_i w\|^2$ 是凸函数，该函数的极小值点的充分必要条件为

$$0 = \nabla_w \sum_{i=1}^m \|g_i - A_i w\|^2 = 2\sum_{i=1}^m A_i^\top (g_i - A_i w) \in \mathbb{R}^{dr}$$

这意味着

$$\sum_{i=1}^m A_i^\top A_i w = \sum_{i=1}^m A_i^\top g_i.$$

换句话说，

$$A^\top A w = A^\top g \quad \text{where } A = \begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_m \end{bmatrix} \in \mathbb{R}^{mr \times dr} \text{ and } g = \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_m \end{bmatrix} \in \mathbb{R}^{mr}$$

因此，

$$w' \in \arg\min_w \sum_{i=1}^m \|g_i - A_i w\|^2 = \{(A^\top A)^\dagger A^\top g + v : v \in \text{Null}(A)\}$$

其中，$(A^\top A)^\dagger$ 是矩阵 $A^\top A$ 的摩尔–彭若斯逆，Null(A)是矩阵 A的零空间。因此，最小范数解为

$$\text{vec}[W_{\bar{S}}] = w_{\bar{S}} = (A^\top A)^\dagger A^\top g.$$

因此，利用这个 $W_{\bar{S}}$，可得

$$\begin{aligned}
\frac{1}{m}\sum_{i=1}^m \|g^*(x_i^+) - W_{\bar{S}} f_\theta(x_i^+)\| &= \frac{1}{m}\sum_{i=1}^m \sqrt{\sum_{k=1}^r ((g_i - A_i w_{\bar{S}})_k)^2} \\
&\le \sqrt{\frac{1}{m}\sum_{i=1}^m \sum_{k=1}^r ((g_i - A_i w_{\bar{S}})_k)^2} \\
&= \frac{1}{\sqrt{m}} \sqrt{\sum_{i=1}^m \sum_{k=1}^r ((g_i - A_i w_{\bar{S}})_k)^2} \\
&= \frac{1}{\sqrt{m}} \|g - A w_{\bar{S}}\|_2 \\
&= \frac{1}{\sqrt{m}} \|g - A(A^\top A)^\dagger A^\top g\|_2 = \frac{1}{\sqrt{m}} \|(I - A(A^\top A)^\dagger A^\top) g\|_2
\end{aligned}$$

其中，不等式是由詹森不等式和平方根函数的凹性得出的。因此，可得

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \|\varphi(x_i)\| + \frac{1}{n}\sum_{i=1}^n \|\varphi(\bar{x}_i)\| \qquad (44) \\
&\le \frac{2}{\sqrt{m}} \|(I - A(A^\top A)^\dagger A^\top) g\|_2 + \frac{4\mathcal{R}_m(\mathcal{W} \circ \mathcal{F})}{\sqrt{m}} + 2\kappa \sqrt{\frac{\ln(4/\delta)}{2m}} + 2\kappa_{\bar{S}} \sqrt{\frac{\ln(4/\delta)}{2n}}
\end{aligned}$$

结合方程40和方程44 并使用并集界，可得

$$\begin{aligned}
&\mathbb{E}_{X,Y}[\|W_S f_\theta(X) - Y\|] \qquad (45) \\
&\le L_S(w_S) + \frac{1}{n}\sum_{i=1}^n \|\tilde{W}(f_\theta(\bar{x}_i) - f_\theta(x_i))\| + \frac{2}{\sqrt{m}} \|\mathbf{P}_A g\|_2 \\
&\quad + \frac{4\mathcal{R}_m(\mathcal{W} \circ \mathcal{F})}{\sqrt{m}} + 2\kappa \sqrt{\frac{\ln(8/\delta)}{2m}} + 2\kappa_{\bar{S}} \sqrt{\frac{\ln(8/\delta)}{2n}} \\
&\quad + \kappa_S \sqrt{\frac{2\ln(4|\mathcal{Y}|/\delta)}{2n}} \sum_{y \in \mathcal{Y}} \left(\sqrt{\hat{p}(y)} + \sqrt{p(y)} \right).
\end{aligned}$$

其中 $\tilde{W} = W_S - W^*$ 和$\mathbf{P}_A = I - A(A^\top A)^\dagger A^\top$。

现在，我们将分析公式 45右侧的第二项：

$$\frac{1}{n}\sum_{i=1}^n \|\tilde{W}(f_\theta(\bar{x}_i) - f_\theta(x_i))\| \le \|\tilde{W}\|_2 \left(\frac{1}{n}\sum_{i=1}^n \|f_\theta(\bar{x}_i) - f_\theta(x_i)\| \right), \qquad (46)$$

˜ ˜其中， $\|W\|_2$ 是 W的谱范数。由于 $\bar{x}_i$ 与 x_i 具有相同的标签，正如 $\bar{x}_i \sim \mathcal{D}_{y_i}$ （且 $x_i \sim \mathcal{D}_{y_i}$），并且 f_θ 是使用无标签数据 S 训练的，因此利用霍夫丁不等式（引理G.1）可知，以至少 $1-\delta$的概率，

$$\frac{1}{n}\sum_{i=1}^{n}\|f_\theta(\bar{x}_i)-f_\theta(x_i)\| \le \mathbb{E}_{y\sim\rho}\mathbb{E}_{\bar{x},x\sim\mathcal{D}_y^2}[\|f_\theta(\bar{x})-f_\theta(x)\|]+\tau_{\bar{S}}\sqrt{\frac{\ln(1/\delta)}{2n}}. \tag{47}$$

此外，利用 [51, 定理 3。1] 对于损失函数$(x,\bar{x}) \mapsto \|f_\theta(\bar{x})-f_\theta(x)\|$ （即引理G.2），我们可以得到，以至少 $1-\delta$的概率，

$$\mathbb{E}_{y\sim\rho}\mathbb{E}_{\bar{x},x\sim\mathcal{D}_y^2}[\|f_\theta(\bar{x})-f_\theta(x)\|] \le \frac{1}{m}\sum_{i=1}^{m}\|f_\theta(x_i^+)-f_\theta(x_i^{++})\|+\frac{2\tilde{\mathcal{R}}_m(\mathcal{F})}{\sqrt{m}}+\tau\sqrt{\frac{\ln(1/\delta)}{2m}} \tag{48}$$

其中 $\tilde{\mathcal{R}}_m(\mathcal{F}) = \frac{1}{\sqrt{m}}\mathbb{E}_{\bar{S},\xi}[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\xi_i\|f(x_i^+) - f(x_i^{++})\|]$ 是集合 $\{(x^+,x^{++}) \mapsto \|f(x^+) - f(x^{++})\|$ 的归一化 Rademacher 复杂度： $f\in\mathcal{F}\}$ （其归一化方式使得，对于 $\mathcal{F}$ 的典型取值，当 $m\to\infty$时，$\mathcal{R}_m(\mathcal{F}) = O(1)$），而 $\xi_1,\dots,\xi_m$ 是取值于 $\{-1,1\}$ 的独立均匀随机变量。因此，应用并集界可得，对于任意 $\delta>0$，以至少 $1-\delta$ 的概率，

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\|\tilde{W}(f_\theta(\bar{x}_i)-f_\theta(x_i))\| \\ &\le \|\tilde{W}\|_2\left(\frac{1}{m}\sum_{i=1}^{m}\|f_\theta(x_i^+)-f_\theta(x_i^{++})\|+\frac{2\tilde{\mathcal{R}}_m(\mathcal{F})}{\sqrt{m}}+\tau\sqrt{\frac{\ln(2/\delta)}{2m}}++\tau_{\bar{S}}\sqrt{\frac{\ln(2/\delta)}{2n}}\right).\end{aligned} \tag{49}$$

结合方程45 和方程49，利用并集界可知，其概率至少为 $1-\delta$，

$$\begin{aligned}&\mathbb{E}_{X,Y}[\|W_S f_\theta(X)-Y\|] \\ &\le L_S(w_S)+\|\tilde{W}\|_2\left(\frac{1}{m}\sum_{i=1}^{m}\|f_\theta(x_i^+)-f_\theta(x_i^{++})\|+\frac{2\tilde{\mathcal{R}}_m(\mathcal{F})}{\sqrt{m}}+\tau\sqrt{\frac{\ln(4/\delta)}{2m}}+\tau_{\bar{S}}\sqrt{\frac{\ln(4/\delta)}{2n}}\right) \\ &\quad+\frac{2}{\sqrt{m}}\|\mathbf{P}_A g\|_2+\frac{4\mathcal{R}_m(\mathcal{W}\circ\mathcal{F})}{\sqrt{m}}+2\kappa\sqrt{\frac{\ln(16/\delta)}{2m}}+2\kappa_{\bar{S}}\sqrt{\frac{\ln(16/\delta)}{2n}} \\ &\quad+\kappa_S\sqrt{\frac{2\ln(8|\mathcal{Y}|/\delta)}{2n}}\sum_{y\in\mathcal{Y}}\left(\sqrt{\hat{p}(y)}+\sqrt{p(y)}\right) \\ &= L_S(w_S)+\|\tilde{W}\|_2\left(\frac{1}{m}\sum_{i=1}^{m}\|f_\theta(x_i^+)-f_\theta(x_i^{++})\|\right)+\frac{2}{\sqrt{m}}\|\mathbf{P}_A g\|_2+Q_{m,n}\end{aligned} \tag{50}$$

其中

$$\begin{aligned}Q_{m,n} = \|\tilde{W}\|_2&\left(\frac{2\tilde{\mathcal{R}}_m(\mathcal{F})}{\sqrt{m}}+\tau\sqrt{\frac{\ln(3/\delta)}{2m}}+\tau_{\bar{S}}\sqrt{\frac{\ln(3/\delta)}{2n}}\right) \\ &+\kappa_S\sqrt{\frac{2\ln(6|\mathcal{Y}|/\delta)}{2n}}\sum_{y\in\mathcal{Y}}\left(\sqrt{\hat{p}(y)}+\sqrt{p(y)}\right) \\ &+\frac{4\mathcal{R}_m(\mathcal{W}\circ\mathcal{F})}{\sqrt{m}}+2\kappa\sqrt{\frac{\ln(4/\delta)}{2m}}+2\kappa_{\bar{S}}\sqrt{\frac{\ln(4/\delta)}{2n}}.\end{aligned}$$

定义 $Z_{\bar{S}} = [f(x_1^+)\dots,f(x_m^+)] \in \mathbb{R}^{d\times m}$。于是，我们有 $A = [Z_{\bar{S}}{}^\top \otimes I_r]$。因此，

$$\mathbf{P}_A = I-[Z_{\bar{S}}{}^\top\otimes I_r][Z_{\bar{S}}Z_{\bar{S}}{}^\top\otimes I_r]^\dagger[Z_{\bar{S}}\otimes I_r] = I-[Z_{\bar{S}}{}^\top(Z_{\bar{S}}Z_{\bar{S}}{}^\top)^\dagger Z_{\bar{S}}\otimes I_r] = [\mathbf{P}$$

$_{Z_{\bar{S}}}\otimes I_r]$其中 $\mathbf{P}_{Z_{\bar{S}}} = I_m - Z_{\bar{S}}{}^\top(Z_{\bar{S}}Z_{\bar{S}}{}^\top)^\dagger Z_{\bar{S}} \in \mathbb{R}^{m\times m}$。通过定义 $Y_{\bar{S}} = [g^*(x_1^+)\dots,\ g^*(x_m^+)$ $]^\top \in\mathbb{R}^{m\times r}$，由于 $g = \mathrm{vec}[Y_{\bar{S}}^\top]$，

$$\|\mathbf{P}_A g\|_2 = \|[\mathbf{P}_{Z_{\bar{S}}}\otimes I_r]\,\mathrm{vec}[Y_{\bar{S}}^\top]\|_2 = \|\,\mathrm{vec}[Y_{\bar{S}}^\top\mathbf{P}_{Z_{\bar{S}}}]\|_2 = \|\mathbf{P}_{Z_{\bar{S}}}Y_{\bar{S}}\|_F \tag{51}$$

另一方面，回顾可知， W_S 是最小范数解，因为

$$W_S = \underset{W'}{\text{minimize}} \|W'\|_F \text{ s.t. } W' \in \arg\min_W \frac{1}{n}\sum_{i=1}^{n} \|W f_\theta(x_i) - y_i\|^2.$$

通过求解可得

$$W_S = Y^\top {Z_S}^\top (Z_S {Z_S}^\top)^\dagger,$$

其中 $Z_S = [f(x_1)\ldots, f(x_n)] \in \mathbb{R}^{d\times n}$ 以及 $Y_S = [y_1, \ldots, y_n]^\top \in \mathbb{R}^{n\times r}$。于是，

$$\begin{aligned} L_S(w_S) = \frac{1}{n}\sum_{i=1}^{n} \|W_S f_\theta(x_i) - y_i\| &= \frac{1}{n}\sum_{i=1}^{n} \sqrt{\sum_{k=1}^{r}((W_S f_\theta(x_i) - y_i)_k)^2} \\ &\le \sqrt{\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{r}((W_S f_\theta(x_i) - y_i)_k)^2} \\ &= \frac{1}{\sqrt{n}}\|W_S Z_S - Y^\top\|_F \\ &= \frac{1}{\sqrt{n}}\|Y^\top({Z_S}^\top(Z_S {Z_S}^\top)^\dagger Z_S - I)\|_F \\ &= \frac{1}{\sqrt{n}}\|(I - {Z_S}^\top(Z_S {Z_S}^\top)^\dagger Z_S)Y\|_F \end{aligned}$$

因此，

$$L_S(w_S) = \frac{1}{\sqrt{n}}\|\mathbf{P}_{Z_S} Y\|_F \tag{52}$$

其中 $\mathbf{P}_{Z_S} = I - {Z_S}^\top(Z_S {Z_S}^\top)^\dagger Z_S$。

结合方程50–方程 52 并利用 $1 \le \sqrt{2}$，我们有以至少$1 - \delta$的概率，

$$\mathbb{E}_{X,Y}[\|W_S f_\theta(X) - Y\|] \le c I_{\bar{S}}(f_\theta) + \frac{2}{\sqrt{m}}\|\mathbf{P}_{Z_{\bar{S}}} Y_{\bar{S}}\|_F + \frac{1}{\sqrt{n}}\|\mathbf{P}_{Z_S} Y_S\|_F + Q_{m,n}, \tag{53}$$

其中

$$\begin{aligned} Q_{m,n} = c\left(\frac{2\tilde{\mathcal{R}}_m(\mathcal{F})}{\sqrt{m}} + \tau\sqrt{\frac{\ln(3/\delta)}{2m}} + \tau_{\bar{S}}\sqrt{\frac{\ln(3/\delta)}{2n}}\right) \\ + \kappa_S\sqrt{\frac{2\ln(6|\mathcal{Y}|/\delta)}{2n}}\sum_{y\in\mathcal{Y}}\left(\sqrt{\hat{p}(y)} + \sqrt{p(y)}\right) \\ + \frac{4\mathcal{R}_m(\mathcal{W}\circ\mathcal{F})}{\sqrt{m}} + 2\kappa\sqrt{\frac{\ln(4/\delta)}{2m}} + 2\kappa_{\bar{S}}\sqrt{\frac{\ln(4/\delta)}{2n}}. \end{aligned}$$

□

Now,

附录 J 信息优化与VICReg目标函数

假设1。 $\Sigma(x_j)$ 的特征值处于某个范围 $a \le \lambda(\Sigma(x_j)) \le b$。假设2。高斯分布的均值之间的差异有界

$$M = \max_{i,j} \|\mu(X_i) - \mu(X_j)\|^2$$

引理 J.1. 每个 $\mu(X_j)\mu(X_j)^T$ 的最大特征值至多为 M。

证明。 项 $\mu(X_j)\mu(X_j)^T$ 是均值向量 $\mu(X_j)$ 的外积，它是一个对称矩阵。对称矩阵的特征值等于原矩阵奇异值的平方。由于向量的奇异值等于其绝对值，因此 $\mu(X_j)\mu(X_j)^T$ 的最大特征值等于 $\mu(X_j)$ 的最大绝对值的平方。根据第二个假设，该值至多为 M。

□

引理 J.2. $-\mu_Z\mu_Z^T$ 的最大特征值为非正数，且其绝对值至多为M。

证明。 $-\mu_Z\mu_Z^T$ 是总体均值向量 μ_Z的负外积，因此是一个对称矩阵。其特征值均为非正数，且等于 μ_Z的奇异值的负平方。由于向量的奇异值等于其绝对值， $-\mu_Z\mu_Z^T$ 的最大特征值的绝对值等于 μ_Z最大绝对值的平方；根据第二个假设，该值也不超过 M。

□

引理 J.3. Σ_Z 的特征值之和有界

$$\sum_i \lambda_i(Z) \leq (b+M) \times K$$

证明。 给定一个高斯混合模型，其中每个分量 $Z|x_j$ 的均值为 $\mu(X_j)$，协方差矩阵为 $\Sigma(x_j)$，则该混合分布可表示为:

$$Z = \sum_j p_j Z|x_j$$

其中， p_j 为混合系数。该混合分布的协方差矩阵 Σ_Z为:

$$\Sigma_Z = \sum_j p_j \left(\Sigma(x_j) + \mu(X_j)\mu(X_j)^T\right) - \mu_Z\mu_Z^T$$

其中， μ_Z 为混合分布的均值。

根据引理 1.1、1.2以及假设 1和2，$(\Sigma(x_j)$, $\mu(X_j)\mu(X_j)^T$和 $\mu_Z\mu_Z^T$. 的最大特征值分别不超过 b、M和 M。因此，根据两个对称矩阵之和的韦尔不等式， Σ_Z 的最大特征值不超过 $b+M$。

$$\lambda_{max}(\Sigma_Z) \leq \frac{1}{K}\sum_{i=1}^{K}(max(\lambda(\Sigma(X_i))) + M) \leq b + M$$

这意味着我们可以用以下方式界定 Σ_Z 的特征值之和:

$$\sum_i \lambda_i(\Sigma_Z) \leq (b+M) \times K$$

□

引理 J.4. 设 Σ_Z 是一个大小为 $N \times N$的半正定矩阵。考虑如下优化问题:

$$\begin{aligned} &\textit{maximize } \log\det(\Sigma_Z) \\ &\textit{such that:} \\ &\sum_{i=1}^{N} \lambda_i(\Sigma_Z) \leq c \\ &\Sigma_Z \succeq 0 \end{aligned}$$

其中 $\lambda_i(\Sigma_Z)$ 表示 Σ_Z 的第 i 个特征值， c 是一个常数。该问题的解是一个对角元素相等的对角矩阵。

证明。 矩阵的行列式等于其特征值的乘积，因此目标函数 $\log\det(\Sigma_Z)$ 可改写为 $\sum_{i=1}^{N}\log(\lambda_i(\Sigma_Z))$。于是，我们要在特征值之和不超过 c 且 Σ_Z 为半正定的约束条件下，使该和最大化。

对凹函数 $\log(x)$ 应用詹森不等式，并使用权重 $1/N$，可得$\frac{1}{N}\sum_{i=1}^{N}\log(\lambda_i(\Sigma_Z)) \le \log(\frac{1}{N}\sum_{i=1}^{N}\lambda_i(\Sigma_Z))$。当且仅当所有 $\lambda_i(\Sigma_Z)$ 都相等时，等号成立。
令 $\lambda_i(\Sigma_Z) = x$ 对所有 i 均成立，可知约束 $\sum_{i=1}^{N}\lambda_i(\Sigma_Z) \le c$ 变为 $Nx \le c$，从而得到该约束下的最优特征值 $x = c/N$。
由于 Σ_Z 是半正定的，因此可以通过正交变换将其对角化，同时保持其特征值之和与行列式不变。因此，该问题的解是一个对角矩阵，其所有对角元素都等于 c/N。

证明至此完成。 □

定理 J.5. 给定一个高斯混合模型，其中每个分量 $Z|X_i$ 的协方差矩阵为$\Sigma(X_i)$，在上述假设下，优化问题的解为

$$\textit{maximize} \sum_i \log \frac{|\Sigma_Z|}{|\Sigma(X_i)|}$$

是一个对角元素相等的对角矩阵 Σ_Z。

证明。 目标函数可以分解如下:

$$\begin{aligned}\sum_i \log \frac{|\Sigma_Z|}{|\Sigma(X_i)|} &= \sum_i \left(\log|\Sigma_Z| - \log|\Sigma(X_i)|\right)\\ &= K\log|\Sigma_Z| - \sum_i \log\left|\Sigma_(X_i)\right|,\end{aligned}$$

其中，K 是高斯混合模型中的分量数。
在这个优化问题中，我们要优化的是 Σ_Z。项 $\sum_i \log|\Sigma(X_i)|$ 相对于 Σ_Z是常数，因此我们可以专注于最大化 $K\log|\Sigma_Z|$。
由于矩阵的行列式等于其特征值的乘积，$\log|\Sigma_Z|$ 是 Σ_Z的特征值的对数之和。因此，最大化 $\log|\Sigma_Z|$ 相当于最大化 Σ_Z的特征值的对数之和。

根据引理 1.4，当特征值之和受到约束时，最大化半正定矩阵 Σ_Z 的特征值对数之和这一问题的解是一个对角元素相等的对角矩阵。

根据引理 1.3，我们知道 Σ_Z 的特征值之和受 $(b+M)\times K$ 的约束。因此，在这些约束条件下最大化 $K\log|\Sigma_Z|$ 时，解将是对角元素相等的对角矩阵。至此，定理得证。 □

附录 K 熵比较——实验细节

我们使用 ResNet-18 [32] 作为骨干网络。每个模型均以 512 的批次大小训练 800 个周期。我们使用动量为 0.9、权重衰减为 $1e^{-4}$的 SGD 优化器。初始学习率为 0.5。该学习率遵循带线性预热调度的余弦衰减。对于数据增强，每个输入图像会生成两个增强版本。在此过程中，每张图像均以随机尺寸进行裁剪，并调整至原始分辨率，随后随机应用水平镜像、颜色抖动、灰度转换、高斯模糊和色调反转。对于熵估计，我们采用与 [39], 相同的方法，该方法假设表征周围为具有恒定方差的高斯混合分布，并利用表征之间的距离计算熵的下界。

$$H(Z) := H(Z|C) - \sum_i c_i \ln \sum_j c_j e^{-D(p_i||p_j)}. \tag{54}$$

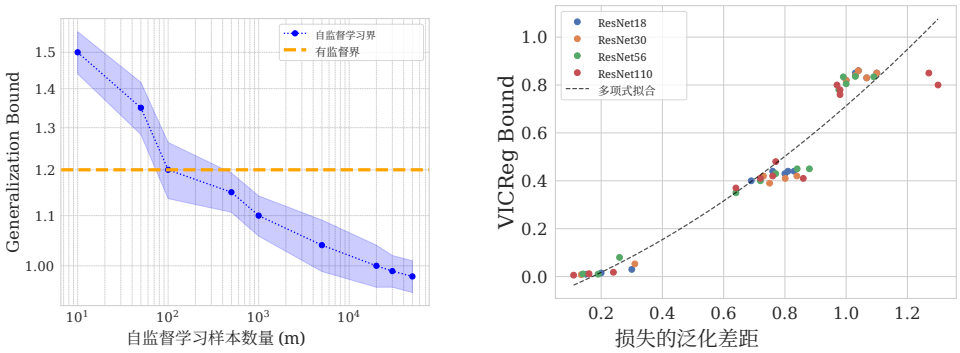

图5：**我们的泛化界能更准确地预测损失中的泛化差距。**（左）我们的自监督学习 VICReg 泛化界优于最先进的监督学习泛化界。（右）VICReg 的泛化差距与我们的泛化界之间具有很强的相关性。皮尔逊相关系数为 0.9633。实验在 CIFAR-10 上进行。

其中， p_i 和 p_j 是第 i 个和第 j 个样本的表征的分布， $D(p_i||p_j)$表示这些分布之间的散度。此外， c_i 表示分量 $i(c_i \geq 0, \sum_i c_i = 1)$ 的权重， C 是一个离散随机变量，其中 $P(C = i) = c_i$ 。

附录 L 可复现性声明

我们研究中的所有方法均基于现有方法及其开源实现。我们为预训练和下游实验提供了详细的实现设置。

附录 M 关于泛化差距的实验

本节进一步给出关于我们的泛化与泛化差距之间联系的实证结果。

附录 N 关于假设网络随机性的局限性

首先，我们简要概述在深度神经网络中估计互信息（MI）所面临的挑战。这一背景将为详细讨论我们方法的优势奠定基础

在深度神经网络中估计互信息会带来各种实际挑战。例如，具有严格单调非线性函数的确定性深度神经网络会产生无穷大或恒定的互信息，从而使优化变得复杂 [4]，或导致优化问题定义不适定。一种解决方案是假设表征服从某个未知分布，并直接估计互信息 [9, 58]。然而，这在理论上具有挑战性，并且需要较大的样本量 [56, 48]，还容易产生显著的估计误差 [71]。

为解决这些问题，先前的研究提出了关于神经网络随机性的各种假设。例如， [1] 认为深度神经网络的参数中存在噪声，而 [1] 假设表征中存在噪声。如果这些假设能够优于确定性深度神经网络，或产生与实际训练得到的表征相似的结果，那么它们就是合理的。

我们的实证结果表明，与向网络本身注入噪声相比，输入噪声的性能显著更好，且其表征与确定性网络的表征更为相似。我们使用 CIFAR-100 和 Tiny-ImageNet，结合 ConvNeXt 与 VICReg，将我们在输入层引入噪声的模型与在网络表征中添加噪声的模型进行了比较 [26], 。下表汇总的结果显示，与我们的输入噪声模型相比，在表征中注入噪声会导致性能大幅下降。

表 2：不同噪声水平下，噪声网络与噪声输入在不同数据集上的比较

β （噪声水平）	Tiny-ImageNet		CIFAR100	
	噪声网络	噪声输入（我们的方法）	噪声网络	噪声输入（我们的方法）
$\beta = 0$（无噪声）	53.1	53.1	70.1	70.1
$\beta = 0.05$	51.7	53.0	69.7	70.0
$\beta = 0.1$	50.2	52.8	68.8	69.6
$\beta = 0.2$	48.1	52.3	67.1	68.9

我们的方法侧重于输入噪声，因此训练期间无需显式注入噪声——这与假定网络噪声固有存在的方法不同。

最后，我们验证了输入噪声这一假设能够与确定性网络实现更好的对齐。随后 [14], 我们使用中心化核对齐（CKA） [18]比较确定性深度神经网络与我们的噪声模型（输入噪声），以及表征中含有噪声的深度神经网络 [26]。结果如下表所示，证实了相比于假设深度神经网络参数中存在噪声，假设输入噪声能够与确定性深度神经网络的表征实现更紧密的对齐。

表 3：不同噪声水平下各种方法的性能比较

噪声水平/方法	确定性网络	噪声网络	噪声输入（我们的方法）
$\beta = 0.05$	0.97	0.82	0.93
$\beta = 0.1$	0.97	0.69	0.85
$\beta = 0.2$	0.97	0.54	0.77
$\beta = 0.3$	0.97	0.32	0.69

总之，我们的输入噪声模型能够更有效地捕捉确定性深度神经网络的行为，并且比假设网络固有随机性更加稳健。

附录 O 更广泛的影响

在机器学习领域，自监督学习算法已广泛应用于各种实际场景，从图像识别到自然语言处理，涵盖范围十分广泛。随着近年来数据的激增，这些算法的实用性呈指数级增长，并经常作为众多现实应用中的关键分量。本文提出了一类新的自监督学习算法，在性能方面展现出优越性。鉴于自监督学习算法的广泛应用，本文提出的改进有望显著提升依赖这些技术的系统的有效性。因此，这些改进算法的影响可能十分深远，有望影响当前采用机器学习的众多应用和行业。与此同时，这种潜在价值也伴随着一种可能性：所提出的方法可能做出无法使现实世界从业者受益的虚假承诺，甚至在部署于敏感应用时造成危害。基于这一原因，出于透明度和可复现性的考虑，我们公开了数值结果和实现细节。与所有新的先进方法一样，我们的改进也可能提升被用于恶意目的的模型。