深度学习算法泛化能力的理解：基于核化 R´ényi 熵的视角

Yuxin Dong[1] , **Tieliang Gong**[1*] ， **Hong Chen**[2] 和**Chen Li**[1]

[1]西安交通大学计算机科学与技术学院，中国西安 710049[2]华中农业大学理学院，中国武汉 430070 adidasgtl@gmail.com, dongyuxin@stu.xjtu.edu.cn, chenh@mail.hzau.edu.cn, cli@xjtu.edu.cn

arXiv:2305.01143v1 [stat.ML] 2 May 2023

摘要

近年来，信息论分析已成为理解深度神经网络泛化行为的一种流行框架。它能够直接分析随机梯度下降／朗之万下降（SGD/SGLD）学习算法，而无需满足 Lipschitz 连续性或凸性条件等强假设。然而，该框架下现有的泛化误差界仍远非最优；由于高维信息量难以处理，要大幅改进这些上界颇具挑战。为解决这一问题，我们首先利用希尔伯特空间中的算子表示，提出一种新颖的信息论度量：核化 R´ényi 熵。它继承了香农熵的性质，可通过简单的随机抽样有效计算，同时不受输入维度影响。随后，我们基于核化 R´ényi 熵建立了 SGD/SGLD 的泛化误差界，其中互信息量可以直接计算，从而能够评估每个中间步骤的紧致性。我们表明，我们的信息论上界取决于沿迭代点计算的随机梯度统计量，并且严格优于当前最先进（SOTA）的结果。大规模实证研究[1]也支持了这些理论发现。

1 引言

现代深度神经网络（深度神经网络）能够记忆全部训练数据，同时还能很好地泛化到未见数据，因而取得了惊人的成功。传统统计学习理论中的泛化界限无法解释这一经验观察，因为它们将泛化归因于假设空间受限的复杂度，而这类复杂度通常对尺度敏感 [Zhang 等,2021]。相反，近期研究发现，算法选择会显著影响深度神经网络的泛化行为 [Hardt 等, 2016; Bartlett 等, 2017], ，这引发了广泛的研究兴趣，促使人们探究不同学习算法的理论性质 [Pensia 等, 2018; Neu等, 2021; Wang 等, 2021; Li 和 Liu, 2022]。

随机梯度下降（SGD）已成为现代深度神经网络训练的主力方法。尽管其简单易用，SGD 也能在复杂的非凸优化问题中实现高效率 [Bottou 等, 2018]。这推动了针对深度学习算法可证明泛化界的大量研究。第一类研究采用一致稳定性的概念，始于 [Hardt等, 2016]对期望收敛的研究，随后许多工作利用类似思路进行了探索[Bassily 等, 2020；Lei 等，2021b；Yang 等，2021b；Yang 等，2021a]。另一类研究将深度神经网络的泛化与信息论分析联系起来[Xu 和 Raginsky, 2017], 并展现出分析含噪迭代学习算法的巨大潜力： [Pensia等, 2018] 首次研究了随机梯度朗之万动力学（SGLD，即一种在每次迭代中注入高斯噪声的 SGD 变体）的泛化能力，后续研究进一步改进了其结果 [Negrea 等，2019；Wang 等, 2021]; [Neu 等, 2021] 随后通过引入辅助权重过程中的虚拟噪声，为 SGD 建立了信息论界，这些界后来在 [Wang 和 Mao, 2021]中得到了进一步收紧。除了稳定性和信息论视角，研究者还从 PAC-贝叶斯 [Neyshabur 等，2018；Yang等, 2019] 以及模型压缩 [Arora 等，2018；Zhou 等, 2018] 等角度开展泛化分析。

尽管当前对深度学习算法泛化能力的理解与解释已取得令人瞩目的成果，但这些界仍然受到限制，因为它们严重依赖于关于假设空间的强假设或维度，应用于大规模深度神经网络时很容易变得无意义。例如，基于一致稳定性的泛化界通常假设经验风险函数满足 Lipschitz 连续性和光滑性 [Hardt *etal.*，2016；Lei *et al.*，2021b] ，或假设凸性和 Polyak-Lojasiewicz（PL）条件等全局最优假设 [Lei *et al.*，2021a；Li 和 Liu, 2022] 以保证收敛，而这些条件在实践中很难满足。相比之下，信息论泛化结果并不

*通讯作者。[1]证明见 https://github.com/Gamepiaynmo/KRE

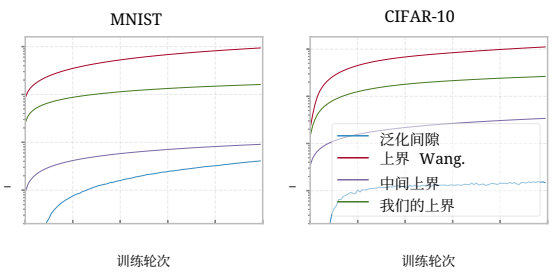

图 1：使用多层感知机和卷积神经网络模型的随机梯度朗之万动力学的泛化误差。我们在第 5 节中提供更详细的分析。

依赖于关于风险函数的强假设，但依赖于假设空间的维度，这通常会导致被严重高估的上界。如图 1 所示，真实泛化误差与当前最优的信息论泛化界 [Wang *etal.*, 2021]之间存在一个从 10^2 到 10^3 的差距。此外，高维信息量的难处理性为进一步收紧这些界带来了额外障碍，因为与中间信息量实际值相比，无论在数值上还是理论上，评估其紧致性都不切实际，而这些中间信息量被用于证明过程中。

本文通过采用一种替代性信息度量——核化 R´enyi 熵，为含噪迭代学习算法建立了可计算的信息论界。这种新的信息量继承了原始香农定义的优雅性质，同时能够直接根据给定样本计算，且不依赖输入维度。随后，我们利用核化 R´enyi 熵对这些学习算法的期望泛化误差进行界定，其中界中的每个关键信息量都可以在训练过程中通过简单的随机抽样直接获取并可视化。基于可视化结果，我们通过严格收紧界，改进了以往的信息论界。例如，如图 1 所示，上述工作 [Wang 等人 用梯度的方差对式（1）中的关键互信息量给出了 2021] 上界，但该估计严重偏高，比实际值（中间界）松了 10 10^2 倍。为缩小这一差距，我们将梯度向量不同维度之间的协方差纳入考量；这一方法同样适用于 [Pensia 等人（2018 年）以及 Wang 和 Mao 的工作 2021], ，并在多个深度学习基准上展现出显著改进。总而言之，本文的主要贡献包括：

• 我们基于希尔伯特空间中的算子表示提出了核化的 R´熵。与经典的香农熵不同，我们的信息量无论维度如何都可以直接计算，同时仍兼容现有的信息论泛化框架。

• 我们基于核化的 R´enyi 熵概念，为随机梯度下降和随机梯度朗之万动力学建立了互信息泛化界，并将其应用于合成数据和真实世界的学习任务进行可视化。可视化结果表明，以往信息论泛化结果仍有多项潜在的改进空间。

• 通过考虑梯度向量不同维度之间的相关性，我们基于一项观察得到了改进的上界。实证研究随后表明，与此前的 SOTA 结果相比，我们的上界紧致程度提高了 5 倍。

2 预备知识

给定随机变量 X，我们将相应的样本空间记为 $\mathcal{X}$，将样本记为小写字母 $\mathbf{x}$，并将概率密度函数（概率密度函数）记为 p_X。我们用 $\|\cdot\|$ 表示向量的欧几里得范数或矩阵的弗罗贝尼乌斯范数，用 I_d 表示 d 维单位矩阵。

2.1 问题设定

设 $\mathcal{Z}$ 为我们关注的实例空间，$\mathcal{W}$ 为假设空间。设 $S = \{Z_i\}_{i=1}^n$ 为包含 n 个独立同分布样本的数据集，这些样本取值于 $\mathcal{Z}$；$W \in \mathcal{W} \subset \mathbb{R}^d$ 为学习算法 $\mathcal{A}$ 根据某个条件分布 $P_{W|S}$ 得到的输出，该条件分布将 $\mathcal{Z}^n$ 映射到 $\mathcal{W}$。设 ℓ：$\mathcal{W} \times \mathcal{Z} \to \mathbb{R}$ 为损失函数。我们的目标是寻找一个参数 $w \in \mathbb{R}^d$，使总体风险 L最小，其定义如下：

$$L(w) \triangleq \mathbb{E}_Z[\ell(w, Z)].$$

由于数据分布通常未知，我们转而最小化经验风险

$$L_S(w) \triangleq \frac{1}{n}\sum_{i=1}^{n} \ell(w, Z_i).$$

对于由 $P_{W|S}$ 表征的学习算法 $\mathcal{A}$，相应的泛化误差定义为 $L(w)$ 与 $L_S(w)$ 之间的期望差值，即

$$\mathrm{gen}(W, S) \triangleq \mathbb{E}_{W,S}[L(W) - L_S(W)].$$

我们始终假设，对于任意 Z，$\ell(w, z)$ 关于 w 几乎处处可微，并且对于任意 $w \in \mathcal{W}$，$\ell(w, Z)$ 是 R-次高斯的。在这些假设下，[Xu 和 Raginsky, 2017] 证明，任意学习算法 $\mathcal{A}$ 的泛化误差有如下上界

$$|\mathrm{gen}(W, S)| \leq \sqrt{\frac{2R^2 I(S; W)}{n}}, \tag{1}$$

其中，$I(S; W)$ 表示输入数据集 S 与输出参数向量 W 之间的互信息。由于现代深度神经网络具有高维特性，这一量通常无法计算，这也给推导更紧的泛化界带来了额外障碍。

2.2 R´enyi 熵及其推广

回顾一下，对于给定的连续随机变量 X，其概率密度函数 p_X 上定义的 R´enyi α 阶熵 $H_\alpha(X)$ 为：

$$H_\alpha(X) \triangleq \frac{1}{1-\alpha}\log \int_{\mathcal{X}} p^\alpha(\mathbf{x})\, \mathrm{d}\mathbf{x}, \tag{2}$$

其中，极限情况下， $\alpha \to 1$ 可还原为香农熵。要精确计算这一信息量，需要了解底层数据分布，而在实际应用中，这种分布通常是未知的。为缓解这一问题， [Giraldo等人， 2014]提出了一种新颖的熵度量方法，通过有限的数据点利用希尔伯特空间表示。具体而言，该方法在由数据点投影到再生核希尔伯特空间（RKHS）后构造的归一化厄米矩阵的特征谱方面，与量子 Rényi 熵相似。本文沿用这一希尔伯特空间表示框架，并对相关再生核施加了轻微限制：

假设 1。 设 $\kappa(\mathbf{x},\mathbf{x}') = \langle \phi(\mathbf{x}),\ \phi(\mathbf{x}') \rangle$ 为一个再生核，其中 $\phi\colon\ \mathcal{X} \mapsto \mathcal{H}$ 是相应的特征映射。假设 κ 满足

- 归一化： $\kappa(\mathbf{x},\mathbf{x}) = 1$ 对于任意 $\mathbf{x} \in \mathcal{X}$；
- 平移不变： $\kappa(\mathbf{x},\mathbf{x}') = f(\|\mathbf{x}-\mathbf{x}'\|)$ 对于某个函数 $f\colon\ \mathbb{R}^+ \mapsto \mathbb{R}^+$；
- L_2 可积： $\forall \mathbf{x} \in \mathcal{X},\ \int_{\mathcal{X}} \kappa^2(\mathbf{x},\mathbf{x}')\,\mathrm{d}\mathbf{x}' < \infty$。

给定随机变量 $X \in \mathcal{X}$，定义线性算子 $G_X\colon \mathcal{H} \mapsto \mathcal{H}$ 为 $G_X f \triangleq \mathbb{E}_X[\phi(\mathbf{x})\langle \phi(\mathbf{x}),\ f\rangle]$。可以验证，当核 κ 归一化时，$\mathrm{tr}(G_X) = 1$，因此 G_X 的特征值构成一个概率分布，它是 X分布的自然密度估计量。

3 核化 Rényi 熵：一种替代性信息度量

本节通过扩展 [Giraldo 等人的工作，将其从有限样本情形推广至无限样本情形，从而引入核化 Rényi 熵，使我们能够直接分析基于概率密度函数的熵量，而不受实际抽样过程的影响。我们的定义通过令$\alpha \to 1$继承了原始香农熵的优雅性质，同时仍可通过简单的随机抽样直接求得。

命题 1. 给定如上定义在随机变量 $X \in \mathcal{X}$ 上的线性算子 G_X，且该随机变量的概率密度函数为 p_X，则有

$$\lim_{\alpha\to 1} \frac{1}{1-\alpha} \log \mathrm{tr}(G_X^{\alpha}) = -\mathrm{tr}(G_X \log G_X) = -\iint_{\mathcal{X}^2} p_X(\mathbf{x}) \log p_X(\mathbf{x}')\kappa^2(\mathbf{x},\mathbf{x}')\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{x}'.$$

命题 1 直接导出以下对 $\alpha \to 1$ 阶核化 Rényi 熵的定义：

定义 1. 给定连续随机变量 X 及其概率密度函数 p_X，则 X 的 $\alpha \to$ 阶核化 ***Rényi*** 熵1 定义为

$$S_1(X) \triangleq -C_\kappa \iint_{\mathcal{X}^2} p_X(\mathbf{x}) \log p_X(\mathbf{x}')\kappa^2(\mathbf{x},\mathbf{x}')\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{x}'.$$

其中 $C_\kappa = 1/\int_{\mathcal{X}} \kappa^2(\mathbf{0},\mathbf{x})\,\mathrm{d}\mathbf{x} > 0$ 是归一化因子，使核函数的平方积分为1。

与对于高维分布而言难以处理的经典香农定义相比，定义无论维度如何，都可以直接计算。为此，可以从 p_X 中随机采样 m 个数据点 $\{\mathbf{x}_i\}_{i=1}^m$，并记 $G_X\colon\ \mathcal{H} \mapsto \mathcal{H}$ 作为G_X的经验版本，即 $\hat{G}_X f \triangleq \frac{1}{m}\sum_{i=1}^m \phi(\mathbf{x}_i)\langle \phi(\mathbf{x}_i),\ f\rangle$。可以验证，$\hat{G}_X$ 是 G_X的无偏估计，这进一步得到核化 Rényi 熵的以下有限样本近似：

命题 2。 设 $\{\mathbf{x}_i\}_{i=1}^m$ 为从 X中采样得到的独立同分布数据点，设 $K \in \mathbb{R}^{m\times m}$ 为由 $K_{ij} = \frac{1}{m}\kappa(\mathbf{x}_i,\mathbf{x}_j)$构造的核矩阵。那么，以 $1-\delta$的置信度，

$$|S_1(X) - \hat{S}_1(X)| \le \frac{9C_\kappa\sqrt{2\log\frac{2}{\delta}}}{\sqrt[3]{m}}, \tag{3}$$

其中 $\hat{S}_1(X) = -C_\kappa \mathrm{tr}(K \log K)$。

请注意，上述集中结果仅与样本数量有关，而与维度无关，因此在高维情形下可以直接计算核化熵。这一性质对于分析现代深度神经网络的行为大有裨益，因为这类网络通常包含数千甚至数百万个参数。还可以注意到，只需将 $\kappa = \kappa_X \otimes \kappa_Y$用作联合分布 $P_{X,Y}$的核函数，即可轻松将定义 1 扩展为多变量联合熵。基于这些设定，可以相应地推导出核化 Rényi 散度和互信息：

定义 2. 给定定义在 $\mathcal{X}$ 上的概率测度 P、 Q 及其概率密度函数 p、 q， P 与 Q 之间的核化 ***Rényi*** 散度定义为：

$$D_1(P \,\|\, Q) \triangleq C_\kappa \iint_{\mathcal{X}^2} p(\mathbf{x}) \log \frac{p(\mathbf{x}')}{q(\mathbf{x}')}\kappa^2(\mathbf{x},\mathbf{x}')\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{x}'.$$

定义 3. 给定归一化核 κ_X、 κ_Y，连续随机变量 X、 Y 及其概率密度函数 p_X、 p_Y， X 与 Y 之间的核化 ***Rényi*** 互信息定义为：

$$I_1(X;Y) \triangleq C_{\kappa_X} C_{\kappa_Y} \iiiint_{\mathcal{Y}^2\times\mathcal{X}^2} p_{X,Y}(\mathbf{x},\mathbf{y}) \cdot \log \frac{p_{X,Y}(\mathbf{x}',\mathbf{y}')}{p_X(\mathbf{x}')p_Y(\mathbf{y}')}\kappa_X^2(\mathbf{x},\mathbf{x}')\kappa_Y^2(\mathbf{y},\mathbf{y}')\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{x}'\,\mathrm{d}\mathbf{y}\,\mathrm{d}\mathbf{y}'.$$

香农熵与核化 Rényi 熵的主要区别在于核函数 κ。具体而言，当 κ 为狄拉克 δ 函数时，我们的定义可以还原为原始的香农熵。为刻画两者之间的差异，我们引入差异函数

$$u_X^{\kappa}(\mathbf{x}) \triangleq C_\kappa \int_{\mathcal{X}} [\log p_X(\mathbf{x}) - \log p_X(\mathbf{x}')]\kappa^2(\mathbf{x},\mathbf{x}')\,\mathrm{d}\mathbf{x}',$$

及其期望形式

$$E_X^{\kappa}(p) \triangleq \left| \int_{\mathcal{X}} p(\mathbf{x}) u_X(\mathbf{x})\,\mathrm{d}\mathbf{x} \right|.$$

为方便起见，我们将**期望差异** $E_X^{\kappa}(p_X)$记为 $E_X^{\kappa\prime}$ E_X^{κ} p_X $C_\kappa \int_{\mathcal{X}} p_X$ κ^2 ， ′ E_X^{κ} ， p_X ≜，并将(ˆ)记为，其中ˆ (x) (x) (xx) dx。

命题 3。令 $\kappa(\mathbf{x}、\mathbf{x}') = \mathbb{1}_{\|\mathbf{x}-\mathbf{x}'\|<c}$。假设概率密度函数 $p_X(\cdot)$ 满足：

•连续性：$\forall \mathbf{x} \in \mathcal{X}$，$\lim_{\mathbf{x}'\to\mathbf{x}} p_X(\mathbf{x}') = p_X(\mathbf{x})$；

•正性：$\forall \mathbf{x} \in \mathcal{X}$，$\lim_{\mathbf{x}'\to\mathbf{x}} p_X(\mathbf{x}') > 0$；

那么我们有 $\lim_{c\to 0} E_X^\kappa \to 0$ 和 $\lim_{c\to 0} E_X^\kappa{}' \to 0$。

命题 3 表明，当归一化核函数 $C_\kappa \cdot \kappa^2$ 具有非常尖锐的峰值，即 c 较小时，期望差异项 E_X^κ 和 $E_X^\kappa{}'$ 都趋于 0。正如我们将在命题 4 中所示，令 $c \to 0$ 对应于核化 R´ényi 熵恢复原始香农定义的情形。当 X 是连续随机变量时，上述连续性假设很容易满足。当 X 在某个区间 $[a,b]$ 内截断，使得对于任意 $x \in [a,b]$（例如，随机采样的图像数据总是被 $[0,255]$截断），$p_X(x) > 0$时，正性假设也自然成立；或者当 X 的分布具有足够重的尾部，使得对于任意有限的 $\mathbf{x} \in \mathcal{X}$，$p_X(\mathbf{x}) > 0$（例如，高斯分布广泛用于模型参数初始化）时，正性假设也自然成立。这些都是现代深度神经网络中的常见情形。

命题 4. 设 X、$X' \in \mathcal{X}$、$Y \in \mathcal{Y}$、$Z \in \mathcal{Z}$ 为连续随机变量，其概率测度分别为 P_X、$P_{X'}$、P_Y 和 P_Z。那么

1. $H(X) \le S_1(X) \le H(X) + E_X^\kappa{}'$.
2. $D_1(P_X \parallel P_{X'}) \ge -E_X^\kappa$.
3. $I_1(X;Y) = D_1(P_{X,Y} \parallel P_X \otimes P_Y) \ge 0$.
4. $I_1(X;Y) = S_1(X) + S_1(Y) - S_1(X,Y)$.
5. $I_1(X;Y|Z) = I_1(X;Y,Z) - I_1(X;Z)$.
6. *Let X,Y,Z form Markov chain $X \to Y \to Z$, then $I_1(X;Y) \ge I_1(X;Z)$ and $I_1(Y;Z) \ge I_1(X;Z)$.*

命题 4 表明，核化 R´ényi 熵继承了原始香农熵的基本性质，从而确保其与现有信息论分析框架兼容。性质 1 验证了当 $c \to 0$ 时，命题 3 中的核化 R´ényi 熵可还原为原始香农熵。结合以下性质，这一结论也适用于散度和互信息量。性质 2 指出，尽管核化 R´ényi 散度不一定为正，但在恰当选择核 κ 时，其值不会显著小于 0。性质 4 和性质 5 表明，可通过估计多个单独熵量的值来估计互信息量。性质6 是数据处理不等式在核化 R´ényi 熵中的对应版本。

接下来，我们将使用高斯核构造核化的 R´ényi 信息量，以界定期望泛化误差，即

$$\kappa(\mathbf{x},\mathbf{x}') = \exp(-\|\mathbf{x}-\mathbf{x}'\|_2^2/2\sigma_\kappa^2),$$

其中，σ_κ 表示核宽度。需要注意的是，σ_κ的选择存在一定的权衡：较小的 σ_κ 意味着 $E_X^\kappa \approx 0$，并且会退化为香农熵。然而，这会导致归一化因子 C_κ 较大，从而产生较大的估计误差，如命题 2 所示。在实践中，我们通常按照所有成对数据点之间欧几里得距离的前 10% 至 20% 来选择 σ_κ，这一做法是根据 Yu 等人，2019]的建议。

4 核化 R´ényi 熵的泛化界

本节介绍了基于核化 R´enyi 熵的迭代含噪学习算法的信息论泛化界。首先，我们证明，式（1）中关于期望泛化误差的互信息界同样适用于我们的核化版本：

定理 1. 假设对于每个 $w \in \mathcal{W}$，$\ell(w,Z)$ 关于 Z 是 R-次高斯的，则

$$|\mathbb{E}_{S,W}[L(W) - L_S(W)]| \le \sqrt{\frac{2R^2 \hat{I}_1(S;W)}{n}}, \quad \textit{and}$$

$$\mathbb{E}_{S,W}[L(W) - L_S(W)]^2 \le \frac{4R^2(\hat{I}_1(S;W) + \log 3)}{n},$$

其中 $\hat{} I_1(S;W) = I_1(S;W) + E_{S,W}^\kappa$。

注释 1. 定理 *1* 从核化 *R´ényi* 熵的角度阐释了信息论泛化 *[Xu 和 Raginsky, 2017]*。如命题 *3* 所示，$E_{S,W}^\kappa$ 是与 κ 相关联的 S 和 W 之间联合分布的期望差异，当 $\sigma_\kappa \to 0$ 时，该差异为零。值得注意的是，定理 *1* 用同一个量 $I_1(S;W)$同时给出了 $\mathrm{gen}(W,S)$ 的期望和方差的上界，因此还可通过集中不等式（例如马尔可夫不等式和切比雪夫不等式）得到泛化误差的高概率界。

接下来，我们将针对经验风险最小化中的小批次迭代式含噪学习算法，应用我们的泛化结果。假设算法 $\mathcal{A}$ 在 T 步数内完成，并令 $W_0 \in \mathcal{W}$ 为初始参数向量。在第 t 步，随机选取一个与当前参数向量独立的数据点批次 $B_t \subset S$，并将其用于计算梯度下降的方向：

$$g(w,B_t) \triangleq \frac{1}{|B_t|}\sum_{z\in B_t} \nabla_w \ell(w,z).$$

那么，更新规则可以形式化表示为

$$W_t = W_{t-1} - \eta_t g(W_{t-1},B_t) + \xi_t, \tag{4}$$

其中，W_t 表示第 t步的参数向量，η_t 是学习率，$\xi_t \in \mathcal{W}$ 是一个与 W_{t-1} 和 B_t相互独立的随机向量。显然，$W_0 \to W_1 \to \cdots \to W_T$构成一个马尔可夫链。

4.1 随机梯度朗之万动力学

SGLD 算法是经典 SGD 算法的一种变体，如式（4）所示，它在每次梯度更新时注入随机噪声。常见的选择是各向同性高斯噪声，即 $\xi_t \sim N(0,\sigma_t^2 I_d)$，因为在固定方差 σ_t^2 的条件下，它具有最大熵，从而得到最紧的上界。 [Pensia 等人， 2018] 推导出了 SGLD 的以下信息论泛化界：

引理1。设 W_T 为随机梯度朗之万动力学算法经过 T 次更新后得到的参数向量，则

$$I(W_T;S) \le \sum_{t=1}^{T} \frac{d}{2}\log\left(\frac{\eta_t^2 L}{d\sigma_t^2} + 1\right), \tag{5}$$

其中 $L =$最大值$_{w\in\mathcal{W},\ z\in\mathcal{Z}} \|g(w,z)\|_2^2$。

注意，引理1中的常数 L 的上界为 Lipschitz 常数的平方，前提是 $\ell(w,z)$ 关于 w 具有 Lipschitz 连续性，且该上界依赖于维度。随后，[Wang 等人 2021]通过移除 Lipschitz 假设改进了这一结果：

引理 2. 在与引理*1*相同的条件下：

$$I(W_T;S) \le \sum_{t=1}^{T} \frac{\eta_t^2 V_t}{2\sigma_t^2}, \tag{6}$$

其中，V_t 是第 t步的梯度方差，定义如下：

$$V_t \triangleq \mathbb{E}_{W_{t-1},B_t}\left[\|g(W_{t-1},B_t) - \mathbb{E}_{B_t}[g(W_{t-1},B_t)]\|_2^2\right].$$

乍看之下，上述上界似乎与维度 d无关。然而，梯度方差 V_t实际上取决于 d，因为它是随机梯度向量各维度所产生的方差之和。主要原因在于，他们使用各向同性高斯分布来对随机梯度的熵给出上界；根据我们的实证结果（见图 2 和图 3），这种做法严重高估了熵。为了解决这一问题，我们考虑利用梯度不同维度之间的相关性，从而为随机梯度朗之万动力学得到更紧的上界：

定理 2. 在与引理*1*相同的条件下：

$$I_1(W_T;S) \le \sum_{t=1}^{T} I_1(W_t;B_t|W_{t-1}) \tag{7}$$

$$\le \sum_{t=1}^{T}\left(\frac{1}{2}\log\left|\frac{\eta_t^2}{\sigma_t^2}\mathbb{V}_t + I\right| + E^{\kappa}_{W_t|W_{t-1}}\right), \tag{8}$$

$$I(W_T;S) \le \sum_{t=1}^{T}\frac{1}{2}\log\left|\frac{\eta_t^2}{\sigma_t^2}\mathbb{V}_t + I\right|, \tag{9}$$

其中，$\mathbb{V}_t$ = 协方差$[g(W_{t-1}, B_t)]$ 是梯度协方差矩阵，$|\cdot|$ 表示矩阵行列式。

注 2。 定理 *2* 表明，核化的 *Rényi* 互信息 $I_1(W_T;S)$ 由梯度协方差矩阵的行列式给出上界，其中包含梯度向量不同维度之间的完整相关性。极限情形 $\sigma_\kappa \to 0$ 给出了式 *(9)* 中香农互信息 $I(W_T;S)$ 的一个上界。注意，式 *(7)* 中的核化 *Rényi* 信息量可直接根据 B_t 计算，从而使我们能够验证这些中间上界的紧致性。结合定理 *1*，可以得到随机梯度朗之万动力学的期望泛化误差上界。

以下命题表明，我们的上界严格优于引理1和引理2中的上界。

命题 5. 给定如上所定义的 $\mathbb{V}_t$、V_t 和 L，令 $\{c_i\}_{i=1}^r$ 为 $\{n\}$ 的一个不相交划分，即对于任意 $i \ne j$，$c_1 \cup \cdots \cup c_r = \{n\}$ 且 $c_i \cap c_j = \Phi$。令 $\mathbb{V}_t^i$ 为 $\mathbb{V}_t$ 的子矩阵，其列和行的索引均由 c_i 指定，并定义

$$\theta_c(\mathbb{V}) = \frac{1}{2}\log\left|\frac{\eta_t^2}{\sigma_t^2}\mathbb{V} + I\right|, \quad \theta_v(V) = \frac{d}{2}\log\left(\frac{\eta_t^2 V}{d\sigma_t^2} + 1\right),$$

$$\textit{then} \qquad \theta_c(\mathbb{V}_t) \le \sum_{i=1}^{r}\theta_c(\mathbb{V}_t^i) \le \theta_v(V_t) \le \theta_v(L).$$

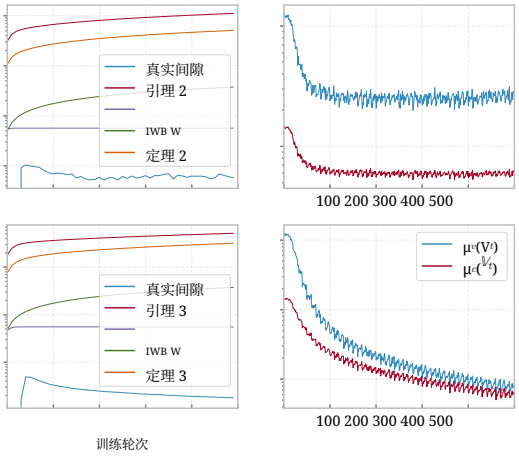

图 2：合成数据上的泛化界比较。

备注 3. 量 θ_c 和 θ_v 分别对应定理 *2* 和引理 *2* 中的上界。当模型规模 d 较大时，由于内存有限，无法计算整个协方差矩阵 $\mathbb{V}_t$。命题 *5* 为 $\theta_c(\mathbb{V}_t)$ 提出了一个替代上界：按照相关性将参数向量划分为不同的组（例如，将模型的每一层作为一组），对每组计算 $\theta_c(\mathbb{V}_t^i)$，然后将结果相加，即可用少得多的内存完成计算。在极端情况下，若 W 中的每个参数各自构成一组，计算 θ_c 所需的内存不会超过 θ_v，同时该上界仍严格优于后者。

4.2 随机梯度下降

与随机梯度朗之万动力学不同，随机梯度下降算法在每次更新中都不涉及随机噪声，即 $\xi_t = 0$。这实际上给信息论泛化分析带来了额外的困难，因为用于推导随机梯度朗之万动力学上界的策略不再适用。为了解决这一问题，[Neu 等人提出引入一个辅助权重过程 W_t，通过手动加入虚拟噪声，然后建立这两个过程（W_t 和 W_t）之间的差异联系。让

$$\tilde{W}_0 = W_0, \quad \text{and}$$
$$\tilde{W}_t = \tilde{W}_{t-1} - \eta_t g(W_{t-1},B_t) + \tilde{\xi}_t, \quad \text{for } t>0,$$

其中 $\tilde{\xi}_t \sim N(0,\sigma_t^2 I)$ 是随机高斯向量。显然，我们有 $\tilde{W}_t = W_t + \Delta_t$，其中 $\Delta_t = \sum_{i=1}^{t}\tilde{\xi}_i$。Wang 和 Mao 的近期研究 [2021] 为随机梯度下降建立了如下信息论泛化界：

引理 3。 假设 $L(W_T) \le \mathbb{E}_{\Delta_t}[L(W_T+\Delta_t)]$，且 ℓ二阶可微。则对于任意 σ_1、$\cdots$、$\sigma_T > 0$，我们有

$$\text{gen}(W_T;S) \le \frac{1}{2}\sum_{t=1}^{T}\sigma_t^2\mathbb{E}_{W_T}[\mathbb{H}(W_T)] + |\text{gen}(\tilde{W}_T;S)|,$$

$$\textit{and} \qquad I\left(\tilde{W}_T;S\right) \le \sum_{t=1}^{T}\frac{d}{2}\log\left(\frac{\eta_t^2 V_t}{d\sigma_t^2}+1\right), \tag{10}$$

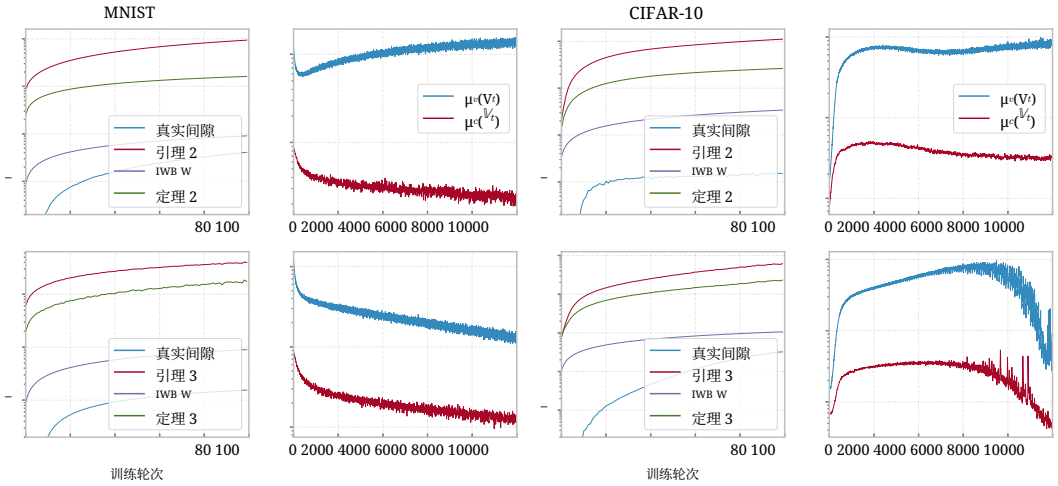

图 3：MNIST 和 CIFAR-10 上随机梯度朗之万动力学与随机梯度下降的信息论泛化界的可视化与比较。

其中 $\mathbb{H}(W_T) = \mathbb{E}_Z[\mathrm{tr}(H_{W_T}(Z))]$，且 $H_{W_T}(Z)$ 是关于 W_T 的损失函数 $\ell(W_T, Z)$ 的 *Hessian* 矩阵。

同样，上述互信息界依赖于 d。我们采用定理 2 中探讨的相同策略来缓解这一问题：

定理 3. 假设 κ 满足假设 *1*，并且满足引理 *3* 的相同条件：

$$I_1\left(\tilde{W}_T; S\right) \leq \sum_{t=1}^{T}\left(\frac{1}{2}\log\left|\frac{\eta_t^2}{\sigma_t^2}\mathbb{V}_t + I\right| + E^{\kappa}_{\tilde{W}_t|\tilde{W}_{t-1}}\right),$$

$$I\left(\tilde{W}_T; S\right) \leq \sum_{t=1}^{T}\frac{1}{2}\log\left|\frac{\eta_t^2}{\sigma_t^2}\mathbb{V}_t + I\right|. \tag{11}$$

注释 4. 定理 *3* 在核化 *Rényi* 熵框架下，为随机梯度下降算法建立了信息论泛化界，其中式 *(11)* 对应于极限情形 $\sigma_\kappa \to 0$。我们强调，我们的结果通过梯度协方差给出了泛化误差的上界；如命题 *5* 所示，由于$\theta_c(\mathbb{V}_t) \leq \theta_v(V_t)$，该上界严格优于引理 *3* 中的结果。请注意，定理 *2* 和定理 *3* 中的上界还可以通过考虑随机梯度的高阶矩进一步改进，但这会带来显著的计算负担，因为 s 阶矩张量的大小为 d^s，因此这种潜在改进在实践中的意义不大。

5 实证研究

在本节中，我们将对前几节推导出的 SGLD/SGD 可计算泛化界进行可视化，并验证先前结果以及我们在定理 2 和定理 3 中提出的改进结果的紧致性。为简单起见，我们对学习率 $\eta_t = \eta$ 和高斯噪声 $\sigma_t = \sigma$ 均采用常数值。在计算过程中，我们忽略期望差异 E^{κ}_X 项，因为通过选取适当的 σ_κ 值，这些项趋于 0，而且实际上无法计算。我们不采用动量、权重衰减和批次归一化等高级调优技术。为了计算定理 1 中的 R，我们收集每个训练轮次中各批次的损失值，并令 $R = \frac{1}{2}[\max_t \ell(W_{t-1}, \mathcal{B}_t) - \min_t \ell(W_{t-1}, B_t)]$。为了计算上述互信息上界中的 $\mathbb{V}_t$ 和 V_t，我们使用 BackPACK PyTorch 库 [Dangel 等，2020] 以获取每个数据输入批次中 $\mathbb{V}_t$ 和 V_t 的经验估计值。为了计算定理 3 中的 $\mathbb{H}(W_T)$，我们使用 PyHessian 库获取 Hessian 矩阵。每项实验重复进行 100 次，以获取 W_t 和 B_t 的独立同分布（i.i.d.）样本，随后利用这些样本构建公式 (3) 中的核矩阵 K，以计算我们信息论上界中的核化 Rényi 互信息。

5.1 合成数据

我们的第一个实验采用了一个简单的线性回归问题

$$y = \mathbf{w}^\top \mathbf{x} + \varepsilon,$$

其中，$\mathbf{x}$ 是 10 维输入向量，y 是回归目标，$\mathbf{w}$ 是线性系数，ε 是某个均值为零的随机噪声。我们训练了一个包含一个宽度为 10 的隐藏层的多层感知机。对于 100 个相互独立的训练过程，我们使用相同的策略，为每个过程生成一个大小为$n = 100$的独立训练数据集。左侧的图 2 展示了现有理论泛化界与真实泛化差距之间的比较，其中我们用相应的关键信息量表示信息论界：$I_1(W_T; S)$记为 IWS，而 $I_1(W_t; B_t|W_{t-1})$的求和记为 IWB|W。请注意，尽管式 (7) 是在随机梯度朗之万动力学算法的背景下推导出的，但它同样适用于随机梯度下降算法，因为该不等式唯一的前提是马尔可夫链关系 $S \to \{B_t\}_{t=1}^T \to \{W_t\}_{t=1}^T$。

如图所示，IWS 始终小于 IWB|W，并且二者始终处于真实间隙曲线与 SGLD 的式（9）上界（或 SGD 的式（11）上界）之间，这表明我们的近似能够成功反映这些信息论量 $I(W_t; S)$ 和 $I(W_t; B_t|W_{t-1})$ 的实际行为。我们可以从可视化结果中获得一些重要的见解：

1）IWS 与真实泛化间隙之间的差距表明，损失函数$\ell(w, Z)$ 关于 Z 的次高斯常数 R 被高估了。假设训练良好的深度神经网络的损失低于随机初始化是合理的，因此预计该常数 R 会随着训练过程的进行而减小。这一观察结果可用于进一步收紧式（1）和定理 1 中的上界。

2) IWS 很快达到峰值，随后随着训练过程的推进而下降（更多细节请参见附录）；其表现与真实间隙曲线相符。相比之下，IWB|W 持续增长，因为每个 $I(W_t; B_t|W_{t-1})$ 始终严格为正。这一观察表明，尽管模型总能从新的批次中学到一些知识（即 $I(W_t; B_t|W_{t-1}) \geq 0$），但 W 包含的数据集 S的总信息量（即 $I(W_t; S)$）很快就达到上限：网络实际上正在遗忘此前学到的信息。目前的信息论泛化界研究尚未捕捉到这种“遗忘”现象，因此，随着训练轮次越来越多，IWS 与 IWB|W 之间的差距逐渐扩大。

3) IWB|W 与我们改进后的上界之间仍然存在的差距表明，即使考虑了噪声梯度的完整相关性，当前的上界仍远非最优。近期研究也支持这一观察：[Gurbuzbalaban等人，2021；Camuto 等人， 2021], 指出，随机梯度下降生成的随机梯度向量具有重尾特性，而若假设其服从高斯分布，则会显著高估其熵。另一种推测是，深度神经网络中存在某些隐式的自正则化机制[Mahoney 和 Martin，2019；Martin 和 Mahoney，2021], 导致权重所包含的信息远低于其理论容量。

图 2 的右侧直观地比较了定理 2 中的上界 θ_c（梯度协方差矩阵）与引理 2 中的上界 θ_v（梯度方差）。可以看出，θ_v（V_t）始终大于 θ_c（$\mathbb{V}_t$），尤其是在训练过程开始时。这一观察验证了我们在命题 5 中的主张。

5.2 真实世界数据

随后，我们在真实世界数据集上可视化可计算的泛化界，以展示核化 R´enyi 熵的可扩展性。遵循实验设置中的彩在线

[Wang 等人， 2021], 我们在 MNIST 上训练了一个隐藏层更宽的多层感知机，并在 CIFAR-10 上训练了一个 4 层卷积神经网络。在 100 次独立的训练过程中，每次都从整个数据集中均匀抽取一部分数据对作为训练数据集，以模拟 S 的随机性。详细的实验设置见附录。

同样，不同泛化界之间的比较结果如图 3 所示。可以看出，IWB|W 的曲线仍始终落在相邻界之间的正确区间内，并且完美地反映了梯度协方差上界的上升趋势。式（6）中的梯度方差界相较于 IWB|W 仍被严重高估。作为比较，我们在式（9）（或式（11））中收紧后的上界涵盖了 SGLD 的 IWB|W 曲线与引理 2（或 SGD 的引理

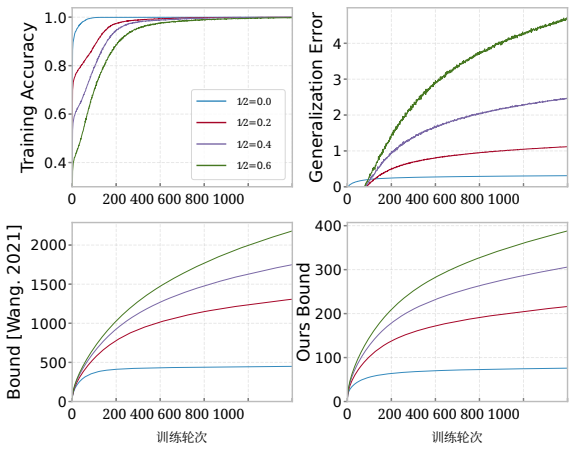

图 4：MNIST 上的随机标签实验。

3（适用于随机梯度下降））。此外，可以看出，在随机梯度朗之万动力学算法的训练过程中，θ_c 的曲线很快停止上升，并在后续训练轮次中开始下降，而 θ_v 的曲线在整个训练过程中始终持续上升。这一观察结果进一步验证了我们改进后的泛化界的紧致性。

接下来，我们进行随机标签实验，以展示改进后的泛化界在不同标签噪声水平下的紧致性。我们沿用上述相同的实验设置，同时以超参数 ρ 指定的概率，将训练标签随机替换为带噪标签。如图 4 所示，标签噪声水平越高，泛化误差也越高。Wang 和 Mao 的上界 [、2021] 以及我们的上界，都成功反映了训练过程中真实泛化间隙的变化趋势；不过，根据引理 3，我们的上界比他们的上界紧 5 倍。更多关于 CIFAR-10 和不同模型规模的实验结果，请参阅附录。

6 结论

在这项工作中，我们解决了香农信息量在实践中难以估计这一常见问题。这给信息论泛化分析带来了额外障碍，因为无法评估以往泛化界 $I(W; S)$、 $I(W_t; B_t|W_{t-1})$ 中所用任何中间信息量的紧致程度 [Wang 等, 2021; Wang and Mao，2021]，，导致相应的上界（引理 2 和 3）被严重高估。为了解决这一问题，我们提出了一种名为核化 R´ényi 熵的替代熵度量，它无论维度如何都可以直接估计，并且仍兼容现有的泛化分析框架。我们成功地将其用于推导并可视化含噪迭代学习算法的信息论泛化界，揭示了多个有待进一步探索的潜在方向。随后，我们基于其中一项发现证明了随机梯度朗之万动力学和随机梯度下降的更紧泛化界，表明这些界在多个深度学习基准上显著优于现有研究成果。

致谢

本研究得到国家重点研发计划（2021ZD0110700）、国家自然科学基金委员会（62106191、12071166、62192781、61721002）、挪威研究理事会（RCN，项目编号309439）、教育部创新研究团队（IRT 17R86）、中国工程科技知识中心项目以及中国工程院项目（一带一路培训的中国慕课线上线下混合式教育服务系统）的资助。

参考文献

[Arora 等人，2018] Sanjeev Arora、Rong Ge、Behnam Neyshabur 和 Yi Zhang。通过压缩方法为深度神经网络提供更强的泛化界。载于《国际机器学习会议》，第 254–263 页。PMLR，2018。[Bartlett 等人，2017] Peter L Bartlett、Dylan J Foster 和 Matus J Telgarsky。神经网络的谱归一化间隔界。载于《神经信息处理系统进展》，第 30 卷，2017年。[Bassily 等人，2020] Raef Bassily、Vitaly Feldman、Crist´obal Guzm´an 和 Kunal Talwar。随机梯度下降在非光滑凸损失上的稳定性。载于《神经信息处理系统进展》，33：4381–4391，2020。[Bottou 等人，2018] L´eon Bottou、Frank E Curtis 和 Jorge Nocedal。大规模机器学习的优化方法。《SIAM评论》，60(2)：223–311，2018。[Camuto 等人，2021] Alexander Camuto、Xiaoyu Wang、Lingjiong Zhu、Chris Holmes、Mert Gurbuzbalaban 和 Umut Simsekli。高斯噪声注入中的非对称重尾与隐式偏差。载于《国际机器学习会议》，第 1249–1260 页。PMLR，2021。[Dangel 等人，2020] Felix Dangel、Frederik Kunstner 和 Philipp Hennig。BackPACK：在反向传播中纳入更多内容。载于《国际学习表征会议》，2020。[Giraldo 等人，2014] Luis Gonzalo Sanchez Giraldo、Murali Rao 和 Jose C Principe。利用无限可分核从数据中度量熵。《IEEE信息论汇刊》，61(1)：535–548，2014。[Gurbuzbalaban 等人，2021] Mert Gurbuzbalaban、Umut Simsekli 和 Lingjiong Zhu。随机梯度下降中的重尾现象。载于《国际机器学习会议》，第 3964–3975 页。PMLR，2021。[Hardt 等人，2016] Moritz Hardt、Ben Recht 和 Yoram Singer。训练更快，泛化更好：随机梯度下降的稳定性。载于《国际机器学习会议》，第 1225–1234 页。PMLR，2016。[Harutyunyan 等人，2021] Hrayr Harutyunyan、Maxim Ra- ginsky、Greg Ver Steeg 和 Aram Galstyan。黑箱学习算法的信息论泛化界。载于《神经信息处理系统进展》，34：24670–24682，2021。

[Lei 等人，2021a] Yunwen Lei、Ting Hu 和 Ke Tang。凸损失函数下多遍随机梯度下降的泛化性能。机器学习研究杂志，22：1–41，2021。[Lei等人，2021b] Yunwen Lei、Zhenhuan Yang、Tianbao Yang 和 Yiming Ying。极小极大问题中随机梯度方法的稳定性与泛化。载于国际机器学习会议，第 6175–6186 页。PMLR，2021。[Li 和 Liu，2022] Shaojie Li 和 Yong Liu。重尾条件下非凸随机梯度下降的高概率保证。载于国际机器学习会议，第 12931–12963 页。PMLR，2022。

[Mahoney 和 Martin，2019] Michael Mahoney 和 Charles Martin。神经网络模型中的传统型与重尾自正则化。载于国际机器学习会议，第 4284–4293 页。PMLR，2019。

[Martin 和 Mahoney，2021] Charles H Martin 和 Michael W Mahoney。深度神经网络中的隐式自正则化：来自随机矩阵理论的证据及其对学习的启示。机器学习研究期刊，22(165):1–73，2021。

[Negrea 等，2019] Jeffrey Negrea、Mahdi Haghifam、Gintare Karolina Dziugaite、Ashish Khisti 和 Daniel M Roy。基于数据依赖估计的随机梯度朗之万动力学信息论泛化界。神经信息处理系统进展，32，2019。

[Neu 等，2021] Gergely Neu、Gintare Karolina Dziugaite、Mahdi Haghifam 和 Daniel M Roy。随机梯度下降的信息论泛化界。载于学习理论会议，第 3526–3545 页。PMLR，2021。

[Neyshabur 等, 2018] Behnam Neyshabur、Srinadh Bho- janapalli 和 Nathan Srebro。针对神经网络谱归一化间隔界的 PAC-贝叶斯方法。发表于国际学习表征会议，2018。

[Pensia 等, 2018] Ankit Pensia、Varun Jog 和 Po-Ling Loh。含噪迭代算法的泛化误差界。发表于*2018 IEEE* 信息论国际研讨会（*ISIT*），第 546–550 页。IEEE，2018。

[Wang 和 Mao, 2021] Ziqiao Wang 和 Yongyi Mao。关于使用随机梯度下降训练的模型的泛化：信息论界及其启示。发表于国际学习表征会议，2021。

[Wang 等, 2021] Hao Wang、Yizhe Huang、Rui Gao 和 Flavio Calmon。利用高斯信道的性质分析随机梯度朗之万动力学的泛化能力。神经信息处理系统进展，34：24222–24234，2021。

[Xu 和 Raginsky，2017] 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，30，2017年。

[Yang 等， 2019] Jun Yang、Shengyang Sun 和 Daniel M Roy。通过移位拉德马赫过程获得快速率 PAC-贝叶斯泛化界。神经信息处理系统进展，32，2019。[Yang等，2021a] Zhenhuan Yang、Yunwen Lei、Siwei Lyu 和 Yiming Ying。具有非光滑损失的成对学习中随机梯度下降的稳定性与差分隐私。载于国际人工智能与统计学会议，第 2026–2034 页。PMLR，2021。[Yang等，2021b] Zhenhuan Yang、Yunwen Lei、Puyu Wang、Tianbao Yang 和 Yiming Ying。用于成对学习的简单随机梯度下降和在线梯度下降算法。神经信息处理系统进展，34：20160–20171，2021。[Yu 等， 2019] Shujian Yu、Luis Gonzalo Sanchez Giraldo、Robert Jenssen 和 Jose C Principe。基于矩阵的 R´enyi α阶熵泛函的多变量扩展。*IEEE* 模式分析与机器智能汇刊，42(11)：2960–2966，2019。[Zhang等， 2021] Chiyuan Zhang、Samy Bengio、Moritz Hardt、Benjamin Recht 和 Oriol Vinyals。理解深度学习（仍然）需要重新思考泛化。*ACM* 通讯，64(3)：107–115，2021。[Zhou等， 2018] Wenda Zhou、Victor Veitch、Morgane Austern、Ryan P Adams 和 Peter Orbanz。ImageNet 规模下非平凡的泛化界：一种 PAC-贝叶斯压缩方法。载于神经信息处理系统进展，2018。

《理解深度学习算法泛化能力：基于核化 Rényi 熵的视角》附录

A 第 3 节的证明

A.1 命题 1 的证明

命题 1(重述)。 对于定义在具有概率密度函数 p_X 的随机变量 $X \in \mathcal{X}$ 上、如上所述的线性算子 G_X，我们有

$$\lim_{\alpha\to 1}\frac{1}{1-\alpha}\log \mathrm{tr}(G_X^\alpha) = -\mathrm{tr}(G_X \log G_X) = -\iint_{\mathcal{X}^2} p_X(\mathbf{x}) \log p_X(\mathbf{x}')\kappa^2(\mathbf{x},\mathbf{x}')\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{x}'.$$

证明。 设 $\{\psi_i\}_{i=1}^{N_\mathcal{H}}$ 是 $\mathcal{H}$ 的完备正交基。根据洛必达法则:

$$\begin{aligned}
\lim_{\alpha\to 1}\frac{1}{1-\alpha}\log \mathrm{tr}(G_X^\alpha) &= \lim_{\alpha\to 1} -\frac{\frac{\partial}{\partial\alpha}\mathrm{tr}(G_X^\alpha)}{\mathrm{tr}(G_X^\alpha)} = -\mathrm{tr}(G_X\log G_X)\\
&= -\sum_{i=1}^{N_\mathcal{H}} \langle G_X\psi_i, \log G_X\psi_i\rangle\\
&= -\sum_{i=1}^{N_\mathcal{H}} \int_\mathcal{X} p_X(\mathbf{x})\langle\psi_i,\phi(\mathbf{x})\rangle\langle\phi(\mathbf{x}),\log G_X\psi_i\rangle\,\mathrm{d}\mathbf{x}\\
&= -\int_\mathcal{X} p_X(\mathbf{x})\langle\log G_X\phi(\mathbf{x}), \sum_{i=1}^{N_\mathcal{H}}\langle\psi_i,\phi(\mathbf{x})\rangle\psi_i\rangle\,\mathrm{d}\mathbf{x}\\
&= -\int_\mathcal{X} p_X(\mathbf{x})\langle\log G_X\phi(\mathbf{x}),\phi(\mathbf{x})\rangle\,\mathrm{d}\mathbf{x}\\
&= -\int_\mathcal{X} p_X(\mathbf{x})\langle\int_\mathcal{X}\log p_X(\mathbf{x}')\phi(\mathbf{x}')\langle\phi(\mathbf{x}'),\phi(\mathbf{x})\rangle\,\mathrm{d}\mathbf{x}',\phi(\mathbf{x})\rangle\,\mathrm{d}\mathbf{x}\\
&= -\iint_{\mathcal{X}^2} p_X(\mathbf{x})\log p_X(\mathbf{x}')\langle\phi(\mathbf{x}'),\phi(\mathbf{x})\rangle\langle\phi(\mathbf{x}'),\phi(\mathbf{x})\rangle\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{x}'\\
&= -\iint_{\mathcal{X}^2} p_X(\mathbf{x})\log p_X(\mathbf{x}')\kappa^2(\mathbf{x},\mathbf{x}')\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{x}'.
\end{aligned}$$

□

A.2 命题 2 的证明

命题 2 （重述）。 设 $\{\mathbf{x}_i\}_{i=1}^m$ 是从 X中抽取的独立同分布数据点，令 $K\in\mathbb{R}^{m\times m}$ 为由 $K_{ij}=\frac{1}{m}\kappa(\mathbf{x}_i,\mathbf{x}j)$构造的核矩阵。那么，以 $1-\delta$的置信度，

$$|S_1(X)-\hat{S}_1(X)| \le \frac{9C_\kappa\sqrt{2\log\frac{2}{\delta}}}{\sqrt[3]{m}},$$

其中 $\hat{S}_1(X)=-C_\kappa\mathrm{tr}(K\log K)$。

证明。 设 λ_i 和 $\mu_i, i\in[1,m]$ 分别为 G_X 和 $\hat{G}_X$ 的特征值。参照 [Giraldo等人对定理 6.2 的证明，同时取 $\varphi(x)=|x|$，我们有，以 $1-\delta$的概率，

$$\sum_{i=1}^m|\lambda_i-\mu_i| \le C\sqrt{\frac{2\log\frac{2}{\delta}}{m}},$$

其中 $C=\max_{\mathbf{x}\in\mathcal{X}}\kappa(\mathbf{x},\mathbf{x})=1$。那么对于任意 $s>0$，我们有

$$\left|\mathrm{tr}(G_X\log G_X)-\mathrm{tr}(\hat{G}_X\log\hat{G}_X)\right| = \left|\sum_{i=1}^m\lambda_i\log\lambda_i-\sum_{i=1}^m\mu_i\log\mu_i\right| \le \sum_{i=1}^m|\lambda_i\log\lambda_i-\mu_i\log\mu_i|$$

$$\leq \sum_{i=1}^{m} \max(-|\lambda_i-\mu_i|\log|\lambda_i-\mu_i|, -(1-|\lambda_i-\mu_i|)\log(1-|\lambda_i-\mu_i|)) \tag{12}$$

$$\leq \sum_{i=1}^{m} -\sqrt{\frac{2\log\frac{2}{\delta}}{m^3}}\log\sqrt{\frac{2\log\frac{2}{\delta}}{m^3}} = \sqrt{\frac{2\log\frac{2}{\delta}}{m}}\log\sqrt{\frac{m^3}{2\log\frac{2}{\delta}}}$$

$$\leq s\sqrt{\frac{2\log\frac{2}{\delta}}{m}}\left(\frac{m^3}{2\log\frac{2}{\delta}}\right)^{\frac{1}{2s}} \leq sm^{\frac{3}{2s}-\frac{1}{2}}\sqrt{2\log\frac{2}{\delta}}, \tag{13}$$

其中，当 $|\lambda_1-\mu_1| = \cdots = |\lambda_n-\mu_n| = \frac{1}{m}\sqrt[2]{\log_m \frac{2}{\delta}}$ 时，(12) 取得最大值；(13) 则由以下事实得出：对于任意 $t>0$，$\log x \leq x^t/t$。令 $s=9$，我们有

$$\begin{aligned}|S_1(X)-\hat{S}_1(X)| &= C_\kappa|\mathrm{tr}(G_X\log G_X)-\mathrm{tr}(K\log K)| \\ &= C_\kappa\left|\mathrm{tr}(G_X\log G_X)-\mathrm{tr}(\hat{G}_X\log\hat{G}_X)\right| \leq \frac{9C_\kappa\sqrt{2\log\frac{2}{\delta}}}{\sqrt[3]{m}}.\end{aligned}$$

□

A.3 命题 3 的证明

命题 3（重述）。设 $\kappa(\mathbf{x},\mathbf{x}') = \mathbb{1}_{\|\mathbf{x}-\mathbf{x}'\|<c}$。假设概率密度函数 $p_X(\cdot)$ 满足：

- 连续性：$\forall \mathbf{x}\in\mathcal{X}$，$\lim_{\mathbf{x}'\to\mathbf{x}} p_X(\mathbf{x}') = p_X(\mathbf{x})$；
- 正性：$\forall \mathbf{x}\in\mathcal{X}$，$\lim_{\mathbf{x}'\to\mathbf{x}} p_X(\mathbf{x}') > 0$;

则有 $\lim_{c\to 0} E_X^\kappa \to 0$ 和 $\lim_{c\to 0} {E_X^\kappa}' \to 0$。

证明。对于定义在 $\mathcal{X}$ 上的任意概率密度函数 $p(\cdot)$:

$$\begin{aligned}\lim_{c\to 0} E_X^\kappa(p) &= \lim_{c\to 0} C_\kappa \iint_{\mathcal{X}^2} p(\mathbf{x})[\log p_X(\mathbf{x}) - \log p_X(\mathbf{x}')]\kappa^2(\mathbf{x},\mathbf{x}')\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{x}' \\ &= \lim_{c\to 0} C_\kappa \iint_{\mathcal{X}^2} p(\mathbf{x})[\log p_X(\mathbf{x}) - \log p_X(\mathbf{x}+\mathbf{x}')]\kappa^2(0,\mathbf{x}')\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{x}' \\ &= C_\kappa \int_{\mathcal{X}} \left\{\lim_{c\to 0}\int_{\|\mathbf{x}-\mathbf{x}'\|<c} p(\mathbf{x})[\log p_X(\mathbf{x}) - \log p_X(\mathbf{x}+\mathbf{x}')]\,\mathrm{d}\mathbf{x}\right\}\mathrm{d}\mathbf{x}' \\ &= C_\kappa \int_{\mathcal{X}} 0\,\mathrm{d}\mathbf{x}' = 0.\end{aligned}$$

□

A.4 命题 4 的证明

命题 4（重述）。设 X、$X'\in\mathcal{X}$、$Y\in\mathcal{Y}$、$Z\in\mathcal{Z}$ 是连续随机变量，其概率测度分别为 P_X、$P_{X'}$、P_Y 和 P_Z。则

1. $H(X) \leq S_1(X) \leq H(X) + {E_X^\kappa}'$。2. $D_1(P_X \parallel P_{X'}) \geq -E_X^\kappa$。3. $I_1(X;Y) = D_1(P_{X,Y} \parallel P_X \otimes P_Y) \geq 0$。4. $I_1(X;Y) = S_1(X) + S_1(Y) - S_1(X,Y)$。5. $I_1(X;Y|Z) = I_1(X;Y,Z) - I_1(X;Z)$。6. 设 X、Y、Z 构成马尔可夫链 $X\to Y\to Z$，则 $I_1(X;Y) \geq I_1(X;Z)$，且 $I_1(Y;Z) \geq I_1(X;Z)$。

性质 *1* 的证明。

$$\begin{aligned}S_1(X) - H(X) &= \int_{\mathcal{X}} p_X(\mathbf{x})\log p_X(\mathbf{x})\,\mathrm{d}\mathbf{x} - C_\kappa\iint_{\mathcal{X}^2} p_X(\mathbf{x})\log p_X(\mathbf{x}')\kappa^2(\mathbf{x},\mathbf{x}')\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{x}' \\ &= \int_{\mathcal{X}} p_X(\mathbf{x})\log p_X(\mathbf{x})\,\mathrm{d}\mathbf{x} - \int_{\mathcal{X}} p_X(\mathbf{x})\left(C_\kappa\int_{\mathcal{X}}\log p_X(\mathbf{x}')\kappa^2(\mathbf{x},\mathbf{x}')\,\mathrm{d}\mathbf{x}'\right)\mathrm{d}\mathbf{x}\end{aligned}$$

$$
\begin{aligned}
&\geq \int_{\mathcal{X}} p_X(\mathbf{x}) \log p_X(\mathbf{x}) \, \mathrm{d}\mathbf{x} - \int_{\mathcal{X}} p_X(\mathbf{x}) \left(\log C_\kappa \int_{\mathcal{X}} p_X(\mathbf{x}') \kappa^2(\mathbf{x}, \mathbf{x}') \, \mathrm{d}\mathbf{x}' \right) \mathrm{d}\mathbf{x} \\
&= \int_{\mathcal{X}} p_X(\mathbf{x}) \log \frac{p_X(\mathbf{x})}{C_\kappa \int_{\mathcal{X}} p_X(\mathbf{x}') \kappa^2(\mathbf{x}, \mathbf{x}') \, \mathrm{d}\mathbf{x}'} \, \mathrm{d}\mathbf{x} \\
&= \mathrm{KL}(P_X \parallel Q_X) \geq 0,
\end{aligned} \tag{14}
$$

其中，式（14）由詹森不等式得出，且 Q_X 是概率密度函数为 $q_X(\mathbf{x}) = C_\kappa \int_{\mathcal{X}} p_X(\mathbf{x}')\kappa^2(\mathbf{x}, \mathbf{x}') \, \mathrm{d}\mathbf{x}'$ 的概率测度。同时，

$$
\begin{aligned}
S_1(X) &= -C_\kappa \iint_{\mathcal{X}^2} p_X(\mathbf{x}) \log p_X(\mathbf{x}') \kappa^2(\mathbf{x}, \mathbf{x}') \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{x}' \\
&= -\int_{\mathcal{X}} p_X(\mathbf{x}) \left(C_\kappa \int_{\mathcal{X}} \log p_X(\mathbf{x}') \kappa^2(\mathbf{x}, \mathbf{x}') \, \mathrm{d}\mathbf{x}' \right) \mathrm{d}\mathbf{x} \\
&= -\int_{\mathcal{X}} p_X(\mathbf{x}) \log p_X(\mathbf{x}) \, \mathrm{d}\mathbf{x} - \int_{\mathcal{X}} p_X(\mathbf{x}) \left[C_\kappa \int_{\mathcal{X}} (\log p_X(\mathbf{x}') - \log p_X(\mathbf{x})) \kappa^2(\mathbf{x}, \mathbf{x}') \, \mathrm{d}\mathbf{x}' \right] \mathrm{d}\mathbf{x} \\
&\leq H(X) + E_X^{\kappa\prime}.
\end{aligned}
$$

□

性质 *2* 的证明。令 t p 和 q 分别为 X 和 X' 的概率密度函数。接着，考虑关于 $q(\mathbf{x}$的如下泛函):

$$
J(q) = C_\kappa \iint_{\mathcal{X}^2} p(\mathbf{x}) \log \frac{p(\mathbf{x}')}{q(\mathbf{x}')} \kappa^2(\mathbf{x}, \mathbf{x}') \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{x}' - \eta_0 \left(\int_{\mathcal{X}} q(\mathbf{x}) \, \mathrm{d}\mathbf{x} - 1 \right),
$$

其中 η_0 是拉格朗日乘子，用于确保 $q(\mathbf{x})$ 是一个概率分布。散度 $D_1(P \parallel Q)$ 在泛函导数等于零时达到极值:

$$
\frac{\partial J}{\partial q} = -C_\kappa \int_{\mathcal{X}} \frac{p(\mathbf{x})}{q(\mathbf{x}')} \kappa^2(\mathbf{x}, \mathbf{x}') \, \mathrm{d}\mathbf{x} - \eta_0 = -\frac{C_\kappa \int_{\mathcal{X}} p(\mathbf{x}) \kappa^2(\mathbf{x}, \mathbf{x}') \, \mathrm{d}\mathbf{x}}{q(\mathbf{x}')} - \eta_0 = 0,
$$

化器 $\hat{q}(\mathbf{x}')$ 满足 $\hat{q}(\mathbf{x}') \propto C_\kappa \int_{\mathcal{X}} p(\mathbf{x})\kappa^2(\mathbf{x}, \mathbf{x}') \, \mathrm{d}\mathbf{x}$。结合 $\int_{\mathcal{X}} \hat{q}(\mathbf{x}') \, \mathrm{d}\mathbf{x}' = 1$，我们有

这表明最小 e

$$
\hat{q}(\mathbf{x}') = C_\kappa \int_{\mathcal{X}} p(\mathbf{x}) \kappa^2(\mathbf{x}, \mathbf{x}') \, \mathrm{d}\mathbf{x}.
$$

因此，

$$
\begin{aligned}
D_1(P \parallel Q) &\geq C_\kappa \iint_{\mathcal{X}^2} p(\mathbf{x}) \log \frac{p(\mathbf{x}')}{\hat{q}(\mathbf{x}')} \kappa^2(\mathbf{x}, \mathbf{x}') \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{x}' \\
&= \int_{\mathcal{X}} \left(C_\kappa \int_{\mathcal{X}} p(\mathbf{x}) \kappa^2(\mathbf{x}, \mathbf{x}') \, \mathrm{d}\mathbf{x} \right) \log \frac{p(\mathbf{x}')}{\hat{q}(\mathbf{x}')} \, \mathrm{d}\mathbf{x}' \\
&\geq -\int_{\mathcal{X}} \hat{q}(\mathbf{x}') \left(C_\kappa \int_{\mathcal{X}} (\log p(\mathbf{x}') - \log p(\mathbf{x})) \kappa^2(\mathbf{x}, \mathbf{x}') \, \mathrm{d}\mathbf{x} \right) \mathrm{d}\mathbf{x}' \\
&\geq -E_X^\kappa.
\end{aligned} \tag{15}
$$

其中，式（15）由詹森不等式得出。 □

性质 *3* 的证明。第一个等式直接由核化散度（定义 2）和互信息（定义 3）的定义得出。将 $n \to \infty$ 代入 [Giraldo 等人 的命题 4.1，即可得出 $I_1(X;Y)$ 的正性，2014]。

□

性质 *4* 的证明。请注意

$$
\begin{aligned}
S_1(X) &= -C_{\kappa_X} \iint_{\mathcal{X}^2} p_X(\mathbf{x}) \log p_X(\mathbf{x}') \kappa_X^2(\mathbf{x}, \mathbf{x}') \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{x}' \\
&= -C_{\kappa_X} \iiint_{\mathcal{Y} \times \mathcal{X}^2} p_{X,Y}(\mathbf{x}, \mathbf{y}) \log p_X(\mathbf{x}') \kappa_X^2(\mathbf{x}, \mathbf{x}') \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{x}' \, \mathrm{d}\mathbf{y} \\
&= -C_{\kappa_X} C_{\kappa_Y} \iiiint_{\mathcal{Y}^2 \times \mathcal{X}^2} p_{X,Y}(\mathbf{x}, \mathbf{y}) \log p_X(\mathbf{x}') \kappa_X^2(\mathbf{x}, \mathbf{x}') \kappa_Y^2(\mathbf{y}, \mathbf{y}') \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{x}' \, \mathrm{d}\mathbf{y} \, \mathrm{d}\mathbf{y}'.
\end{aligned}
$$

同理，我们有 $S_1(Y) = -C_{\kappa_X} C_{\kappa_Y} \iiiint_{\mathcal{Y}^2 \times \mathcal{X}^2} p_{X,\ Y}(\mathbf{x},\ \mathbf{y}) \log\ p_Y(\mathbf{y}')\kappa_X^2(\mathbf{x},\ \mathbf{x}')\kappa_Y^2(\mathbf{y},\ \mathbf{y}') \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{x}' \mathrm{d}\mathbf{y} \, \mathrm{d}\mathbf{y}'$。结合以上表达式即可完成证明。

□

性质*5*的证明。根据核化互信息的定义，我们可以推导出核化条件互信息的表达式如下：

$$I_1(X;Y|Z) = C_{\kappa_X} C_{\kappa_Y} C_{\kappa_Z} \iint_{\mathcal{Z}^2} \iint_{\mathcal{Y}^2} \iint_{\mathcal{X}^2} p_{X,Y,Z}(x,y,z) \log \frac{p_{X,Y|Z}(x',y'|z')}{p_{X|Z}(x'|z') p_{Y|Z}(y'|z')} \tag{16}$$

$$\cdot \kappa_X^2(x,x') \kappa_Y^2(y,y') \kappa_Z^2(z,z') \,\mathrm{d}x \,\mathrm{d}x' \,\mathrm{d}y \,\mathrm{d}y' \,\mathrm{d}z \,\mathrm{d}z'. \tag{17}$$

根据条件核化互信息和无条件核化互信息的定义，性质5的证明可直接得出。

□

性质*6*的证明。 根据性质5，我们有

$$I_1(X;Y,Z) = I_1(X;Y|Z) + I_1(X;Z) = I_1(X;Z|Y) + I_1(X;Y).$$

根据马尔可夫条件，给定 Y时， X 和 Z 条件独立，即$p_{X,\ Z|Y}(x,\ \ z|y) = p_{X|Y}(x|y) p_{Z|Y}(z|y)$。根据条件互信息的定义，我们有 $I_1(X;Z|Y) = 0$。于是，由于 $I_1(X;Y|Z)$的正性，性质6的第一个不等式得证。类似地，可以证明性质6的第二部分。

□

B 第4节的证明

B.1 定理1的证明

引理4。 设 P、 Q 是定义在同一可测空间上的概率测度，其中 P 相对于 Q绝对连续。则

$$D_1(P \parallel Q) + E_P^\kappa \geq \mathbb{E}_P[X] - \log \mathbb{E}_Q[e^X].$$

其中， X 是任意随机变量，且 e^X 是 Q-可积的，并且 $\mathbb{E}_P[X]$ 存在。

证明。 定义 Q^X 为一个概率测度，使得

$$Q^X(\Omega) = \int_\Omega \frac{e^X}{\mathbb{E}_Q[e^X]} \,\mathrm{d}Q,$$

则 Q 关于 Q^X绝对连续。注意到

$$\begin{aligned}
D_1(P \parallel Q) + E_P^\kappa &= D_1(P \parallel Q^X) + E_P^\kappa + C_\kappa \iint_{\mathcal{X}^2} p_X(x) \log \frac{e^{x'}}{\mathbb{E}_Q[e^X]} \kappa^2(x,x') \,\mathrm{d}x \,\mathrm{d}x' \\
&\geq C_\kappa \iint_{\mathcal{X}^2} p_X(x) \log e^{x'} \kappa^2(x,x') \,\mathrm{d}x \,\mathrm{d}x' - C_\kappa \iint_{\mathcal{X}^2} p_X(x) \log \mathbb{E}_Q[e^X] \kappa^2(x,x') \,\mathrm{d}x \,\mathrm{d}x' \\
&= \int_{\mathcal{X}} p_X(x) \left(C_\kappa \int_{\mathcal{X}} x' \kappa^2(x,x') \,\mathrm{d}x' \right) \mathrm{d}x - \log \mathbb{E}_Q[e^X] \int_{\mathcal{X}} p_X(x) \left(C_\kappa \int_{\mathcal{X}} \kappa^2(x,x') \,\mathrm{d}x' \right) \mathrm{d}x \\
&= \int_{\mathcal{X}} p_X(x) x \,\mathrm{d}x - \log \mathbb{E}_Q[e^X] \int_{\mathcal{X}} p_X(x) \,\mathrm{d}x = \mathbb{E}_P[X] - \log \mathbb{E}_Q[e^X].
\end{aligned}$$

□

引理5。 （*[Harutyunyan*等人， *2021]*中的引理 *2*）设 X 是一个 R-次高斯零均值随机变量，则$\forall \lambda \in \left[0, \frac{1}{4R^2}\right)$:

$$\mathbb{E}\left[e^{\lambda X^2}\right] \leq 1 + 8\lambda R^2.$$

引理6。 （*[Harutyunyan*等人， *2021]*中的引理 *3*）设 X 和 Y 为独立随机变量。设 f 为可测函数，使得 f（x,Y） 是 R-次高斯，且对所有 $x \in \mathcal{X}$， $\mathbb{E}_Y[f(x,Y)] = 0$，则 f（X,Y） 也是 R-次高斯。

定理 引理 1（重述）。假设对于每个 $w \in \mathcal{W}$， ℓ（w,Z）相对于 Z 是 R-次高斯的， n

$$|\mathbb{E}_{S,W}[L(W) - L_S(W)]| \leq \sqrt{\frac{2R^2 \hat{I}_1(S;W)}{n}}, \quad \textit{and}$$

$$\mathbb{E}_{S,W}[L(W) - L_S(W)]^2 \leq \frac{4R^2(\hat{I}_1(S;W) + \log 3)}{n},$$

其中 $\hat{} \ I_1(S;W) = I_1(S;W) + E_{S,W}^\kappa$.

证明。令 $f(w,s)=L(w)-L_s(w)$，并令 W' 和 S' 分别为 W 和 S 以及 $\lambda\in[0,\infty)$ 的独立副本，则

$$\begin{aligned} I_1(W;S)+E^{\kappa}_{W,S} &= D_1(P_{W,S}\parallel P_W\otimes P_S)+E^{\kappa}_{W,S}\\ &\geq \mathbb{E}_{W,S}[\lambda f(W,S)]-\log\mathbb{E}_{W',S'}\left[e^{\lambda f(W',S')}\right]\\ &= \mathbb{E}_{W,S}[\lambda f(W,S)]-\log\mathbb{E}_{W,S'}\left[e^{\lambda f(W,S')}\right] \end{aligned} \tag{18}$$

根据引理4以及 W、W'和 S'相互独立这一事实。注意，对于每个 $w\in\mathcal{W}$，$f(w,S)$是$R/\sqrt{n}$-次高斯的，因为$L_S(w)$是 n 个独立同分布的 R-次高斯随机变量的平均值。此外，对于任意固定的 w，$\mathbb{E}_S[f(w,S)]=0$。于是根据引理6，$f(W,S)$是$R/\sqrt{n}$-次高斯的。因此，

$$\log\mathbb{E}_{W,S'}\left[e^{\lambda f(W,S')-\lambda\mathbb{E}_{W,S'}[f(W,S')]}\right]\leq\frac{\lambda^2R^2}{2n},$$
$$\log\mathbb{E}_{W,S'}\left[e^{\lambda f(W,S')}\right]\leq\frac{\lambda^2R^2}{2n}.$$

代入式（18），得到

$$I_1(W;S)+E^{\kappa}_{W,S}\geq\lambda\mathbb{E}_{W,S}[f(W,S)]-\frac{\lambda^2R^2}{2}\geq\frac{n}{2R^2}\mathbb{E}^2_{W,S}[f(W,S)].$$

第一部分的证明至此完成。对于第二部分，令 $\tilde{f}(w,s)=(L(w)-L_s(w))^2$，并令$\lambda\in\left[0,\frac{1}{4R^2}\right)$，则

$$\begin{aligned} I_1(W;S)+E^{\kappa}_{W,S} &\geq \mathbb{E}_{W,S}\left[\lambda\tilde{f}(W,S)\right]-\log\mathbb{E}_{W',S'}\left[e^{\lambda\tilde{f}(W',S')}\right]\\ &\geq \mathbb{E}_{W,S}\left[\lambda(L(W)-L_S(W))^2\right]-\log(1+8\lambda R^2)\\ &\geq \frac{1}{4R^2}\mathbb{E}_{W,S}\left[(L(W)-L_S(W))^2\right]-\log 3, \end{aligned}$$

根据引理4、5并取 $\lambda=\frac{1}{4R^2}$。第二部分的证明至此完成。 □

B.2 定理2的证明

引理 7。设随机变量 $X\sim N(0,\Sigma)$ 和 X' 为满足 $\mathbb{E}[X']=0$ 和协方差$[X']=\Sigma$的任意连续随机变量，则 $S_1(X')\leq S_1(X)+E^{\kappa}_{X'}$。此外，若 κ 为核宽度为 σ_κ的高斯分布，则 $S_1(X)=\frac{d}{2}\log(2\pi e)+\frac{1}{2}\log|\Sigma|+\frac{\sigma_\kappa^2}{4}\operatorname{tr}[\Sigma^{-1}]$。

证明。设 $p(\cdot)$ 和 $q(\cdot)$ 分别为 X 和 X' 的概率密度函数。注意到 $p_\kappa=C_\kappa\kappa^2(0,\cdot)$ 的积分为 1，因此可以将其视为一个概率分布，其协方差矩阵为 $\frac{1}{2}\sigma_\kappa^2I$。于是有

$$\begin{aligned} S_1(X)-S_1(X') &= C_\kappa\iint_{\mathcal{X}^2}\left[q(\mathbf{x})\left(\log p(\mathbf{x}')+\log\frac{q(\mathbf{x}')}{p(\mathbf{x}')}\right)-p(\mathbf{x})\log p(\mathbf{x}')\right]\kappa^2(\mathbf{x},\mathbf{x}')\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{x}'\\ &= C_\kappa\iint_{\mathcal{X}^2}[q(\mathbf{x})-p(\mathbf{x})]\log p(\mathbf{x}')\kappa^2(\mathbf{x},\mathbf{x}')\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{x}'+C_\kappa\iint_{\mathcal{X}^2}q(\mathbf{x})\log\frac{q(\mathbf{x}')}{p(\mathbf{x}')}\kappa^2(\mathbf{x},\mathbf{x}')\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{x}'\\ &= C_\kappa\iint_{\mathcal{X}^2}[q(\mathbf{x})-p(\mathbf{x})]\left(-\frac{d}{2}\log(2\pi)-\frac{1}{2}\log|\Sigma|-\frac{1}{2}\mathbf{x}'^{\top}\Sigma^{-1}\mathbf{x}'\right)\kappa^2(\mathbf{x},\mathbf{x}')\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{x}'+D_1(Q\parallel P)\\ &\geq -\left(\frac{d}{2}\log(2\pi)+\frac{1}{2}\log|\Sigma|\right)\int_{\mathcal{X}}[q(\mathbf{x})-p(\mathbf{x})]\left(C_\kappa\int_{\mathcal{X}}\kappa^2(\mathbf{x},\mathbf{x}')\,\mathrm{d}\mathbf{x}'\right)\mathrm{d}\mathbf{x}\\ &\quad-\frac{1}{2}\operatorname{tr}\left\{\left[\int_{\mathcal{X}}[q(\mathbf{x})-p(\mathbf{x})]\left(C_\kappa\int_{\mathcal{X}}\mathbf{x}'\mathbf{x}'^{\top}\kappa^2(\mathbf{x},\mathbf{x}')\,\mathrm{d}\mathbf{x}'\right)\mathrm{d}\mathbf{x}\right]\Sigma^{-1}\right\}-E^{\kappa}_{X'}\\ &= -\left(\frac{d}{2}\log(2\pi)+\frac{1}{2}\log|\Sigma|\right)\int_{\mathcal{X}}[q(\mathbf{x})-p(\mathbf{x})]\,\mathrm{d}\mathbf{x}\\ &\quad-\frac{1}{2}\operatorname{tr}\left\{\left[\int_{\mathcal{X}}[q(\mathbf{x})-p(\mathbf{x})]\left(\mathbf{x}\mathbf{x}^{\top}+\frac{1}{2}\sigma_\kappa^2I\right)\mathrm{d}\mathbf{x}\right]\Sigma^{-1}\right\}-E^{\kappa}_{X'}\\ &= -\frac{1}{2}\operatorname{tr}\left[\left(\int_{\mathcal{X}}[q(\mathbf{x})-p(\mathbf{x})]\mathbf{x}\mathbf{x}^{\top}\,\mathrm{d}\mathbf{x}\right)\Sigma^{-1}\right]-E^{\kappa}_{X'}=-\frac{1}{2}\operatorname{tr}\left[(\Sigma-\Sigma)\Sigma^{-1}\right]-E^{\kappa}_{X'}=-E^{\kappa}_{X'}. \end{aligned}$$

考虑核 κ 为高斯核的情况，即 $\kappa^2(\mathbf{x},\ \mathbf{x}') = \exp\left(-\|\mathbf{x}-\mathbf{x}'\|_2^2/\sigma_\kappa^2\right)$，且 $C_\kappa = (\pi\sigma_\kappa^2)^{-d/2}$，我们有

$$
\begin{aligned}
S_1(X) &= \frac{1}{\sqrt{(2\pi)^d|\Sigma|}}\frac{1}{\sqrt{(\pi\sigma_\kappa^2)^d}}\iint_{\mathcal{X}^2}\exp\left(-\frac{1}{2}\mathbf{x}^\top\Sigma^{-1}\mathbf{x}\right)\left[\frac{d}{2}\log(2\pi)+\frac{1}{2}\log|\Sigma|+\frac{1}{2}\mathbf{x}'^\top\Sigma^{-1}\mathbf{x}'\right]\exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|_2^2}{\sigma_\kappa^2}\right)\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{x}' \\
&= \frac{d}{2}\log(2\pi)+\frac{1}{2}\log|\Sigma|+\frac{1}{\sqrt{(2\pi)^d|\Sigma|}}\int_{\mathcal{X}}\frac{1}{2}\exp\left(-\frac{1}{2}\mathbf{x}^\top\Sigma^{-1}\mathbf{x}\right)\left(\frac{1}{\sqrt{(\pi\sigma_\kappa^2)^d}}\int_{\mathcal{X}}\mathbf{x}'^\top\Sigma^{-1}\mathbf{x}'\exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|_2^2}{\sigma_\kappa^2}\right)\mathrm{d}\mathbf{x}'\right)\mathrm{d}\mathbf{x} \\
&= \frac{d}{2}\log(2\pi)+\frac{1}{2}\log|\Sigma|+\frac{1}{\sqrt{(2\pi)^d|\Sigma|}}\int_{\mathcal{X}}\frac{1}{2}\exp\left(-\frac{1}{2}\mathbf{x}^\top\Sigma^{-1}\mathbf{x}\right)\mathrm{tr}\left[\left(\mathbf{x}\mathbf{x}^\top+\frac{1}{2}\sigma_\kappa^2 I\right)\Sigma^{-1}\right]\mathrm{d}\mathbf{x} \\
&= \frac{d}{2}\log(2\pi)+\frac{1}{2}\log|\Sigma|+\frac{1}{2}\mathrm{tr}\left[\left(\Sigma+\frac{1}{2}\sigma_\kappa^2 I\right)\Sigma^{-1}\right] = \frac{d}{2}\log(2\pi e)+\frac{1}{2}\log|\Sigma|+\frac{\sigma_\kappa^2}{4}\mathrm{tr}[\Sigma^{-1}].
\end{aligned}
$$

□

引理 8. 设 X、Y、Δ和 $\xi\sim N(0、\sigma^2 I)$为独立随机变量。设 $f\colon\ \mathcal{W}\times\mathcal{Z}^b\to\mathcal{W}$为行列式函数，且 $\Omega(\cdot)=\mathbb{E}_X[f(\cdot,X)]$。则

$$
I_1(f(Y+\Delta,X)+\xi;X|Y)\le\frac{1}{2}\log\left|\frac{1}{\sigma^2}\mathrm{Cov}[g(Y,\Delta,X)]+I\right|+E^\kappa_{f(Y+\Delta,X)-\Omega(Y+\Delta)+\xi|Y,\Delta}.
$$

证明。令 $g(Y,\Delta,X) = f(Y+\Delta,X) - \Omega(Y+\Delta)$，则

$$
\begin{aligned}
&I_1(f(Y+\Delta,X)+\xi;X|Y=\mathbf{y},\Delta=\delta) && (19)\\
&= I_1(g(Y,\Delta,X)+\xi;X|Y=\mathbf{y},\Delta=\delta) \\
&= S_1(g(Y,\Delta,X)+\xi|Y=\mathbf{y},\Delta=\delta) - S_1(g(Y,\Delta,X)+\xi|X,Y=\mathbf{y},\Delta=\delta) \\
&= S_1(g(Y,\Delta,X)+\xi|Y=\mathbf{y},\Delta=\delta) - S_1(\xi) \\
&= S_1(g(Y,\Delta,X)+\xi|Y=\mathbf{y},\Delta=\delta) - \frac{d}{2}\log(2\pi e\sigma^2) - \frac{d\sigma_\kappa^2}{4\sigma^2} \\
&\le \frac{d}{2}\log(2\pi e)+\frac{1}{2}\log\left|\mathrm{Cov}[g(Y,\Delta,X)|Y=\mathbf{y},\Delta=\delta]+\sigma^2 I\right| - \frac{d}{2}\log(2\pi e\sigma^2)+E^\kappa_{g(Y,\Delta,X)+\xi|Y,\Delta} \\
&\quad + \frac{\sigma_\kappa^2}{4}\mathrm{tr}\left[\left(\mathrm{Cov}[g(Y,\Delta,X)|Y=\mathbf{y},\Delta=\delta]+\sigma^2 I\right)^{-1}\right] - \frac{d\sigma_\kappa^2}{4\sigma^2} && (20)\\
&\le \frac{1}{2}\log\left|\frac{1}{\sigma^2}\mathrm{Cov}[g(Y,\Delta,X)|Y=\mathbf{y},\Delta=\delta]+I\right|+E^\kappa_{g(Y,\Delta,X)+\xi|Y,\Delta} && (21)
\end{aligned}
$$

其中，式 (20) 根据引理 7 并注意到以下事实可得：

$$
\begin{aligned}
\mathrm{Cov}[g(Y,\Delta,X)+\xi|Y=\mathbf{y},\Delta=\delta] &= \mathrm{Cov}[g(Y,\Delta,X)|Y=\mathbf{y},\Delta=\delta]+\mathrm{Cov}[\xi] \\
&= \mathrm{Cov}[g(Y,\Delta,X)|Y=\mathbf{y},\Delta=\delta]+\sigma^2 I,
\end{aligned}
$$

而式 (21) 根据以下事实可得：

$$
\mathrm{tr}\left[\left(\mathrm{Cov}[g(Y,\Delta,X)|Y=\mathbf{y},\Delta=\delta]+\sigma^2 I\right)^{-1}\right]\le\mathrm{tr}\left[\left(\sigma^2 I\right)^{-1}\right].
$$

由此得到以下上界：

$$
\begin{aligned}
I_1(f(Y+\Delta,\ X)+\xi;\ X|Y) &\le I_1(f(Y+\Delta,\ X)+\xi,\ \Delta;\ X|Y) && (22)\\
&= I_1(f(Y+\Delta,X)+\xi,\Delta;X|Y) - I_1(\Delta;X|Y) && (23)\\
&= I_1(f(Y+\Delta,\ X)+\xi;\ X|Y,\ \Delta) && (24)\\
&= \mathbb{E}_{Y,\Delta}[I_1(f(Y+\Delta,X)+\xi;X|Y=\mathbf{y},\Delta=\delta)] \\
&\le \mathbb{E}_{Y,\Delta}\left[\frac{1}{2}\log\left|\frac{1}{\sigma^2}\mathrm{Cov}[g(Y,\Delta,X)|Y=\mathbf{y},\Delta=\delta]+I\right|\right]+E^\kappa_{g(Y,\Delta,X)+\xi|Y,\Delta} && (25)\\
&\le \frac{1}{2}\log\left|\frac{1}{\sigma^2}\mathbb{E}_{Y,\Delta}[\mathrm{Cov}[g(Y,\Delta,X)|Y=\mathbf{y},\Delta=\delta]]+I\right|+E^\kappa_{g(Y,\Delta,X)+\xi|Y,\Delta} && (26)\\
&= \frac{1}{2}\log\left|\frac{1}{\sigma^2}\mathrm{Cov}[g(Y,\Delta,X)]+I\right|+E^\kappa_{g(Y,\Delta,X)+\xi|Y,\Delta}, && (27)
\end{aligned}
$$

其中，(23) 是因为 Δ 和 X 相互独立，(25) 是通过应用 (21) 得到的，(26) 是由詹森不等式和对数行列式函数的凹性得到的，而 (27) 则由全方差公式以及注意到以下事实得到

$$\mathrm{Cov}[\mathbb{E}_X[g(Y,\Delta,X)|Y=\mathbf{y},\Delta=\delta]]=\mathrm{Cov}[\mathbb{E}_X[f(Y,\Delta,X)-\mathbb{E}_X[f(Y,\Delta,X)]|Y=\mathbf{y},\Delta=\delta]]=0.$$

□

定理 2 （重述）。 在与引理*1*相同的条件下：

$$I_1(W_T;S)\le\sum_{t=1}^{T}I_1(W_t;B_t|W_{t-1})\le\sum_{t=1}^{T}\left(\frac{1}{2}\log\left|\frac{\eta_t^2}{\sigma_t^2}\mathbb{V}_t+I\right|+E^{\kappa}_{W_t|W_{t-1}}\right),$$

$$I(W_T;S)\le\sum_{t=1}^{T}\frac{1}{2}\log\left|\frac{\eta_t^2}{\sigma_t^2}\mathbb{V}_t+I\right|,$$

其中 $\mathbb{V}_t=\mathrm{Cov}[g(W_{t-1},B_t)]$ 是梯度协方差 矩阵，$|\cdot|$ 表示矩阵行列式。

证明。 注意到 $S\to(B_1,\cdots,B_T)\to(W_1,\cdots,W_T)$ 构成一个马尔可夫链，因此

$$\begin{aligned}I_1(W_T;S)&\le I_1(W_T;B_1,\cdots,B_T)\le I_1(W_0,W_1,\cdots,W_T;B_1,\cdots,B_T)\\&=I_1(W_0;B_1,\cdots,B_T)+I_1(W_1;B_1,\cdots,B_T|W_0)+I_1(W_2;B_1,\cdots,B_T|W_0,W_1)\\&\quad+\cdots+I_1(W_T;B_1,\cdots,B_T|W_0,\cdots,W_T).\end{aligned}$$

对于每个 $t\in[1,T]$，我们有

$$\begin{aligned}I_1(W_t;B_1,\cdots,B_t|W_0,\cdots,W_{t-1})&=S_1(W_t|W_0,\cdots,W_{t-1})-S_1(W_t|B_1,\cdots,B_t,W_0,\cdots,W_{t-1})\\&=S_1(W_t|W_{t-1})-S_1(W_t|B_t,W_{t-1})\\&=I_1(W_t;B_t|W_{t-1}).\end{aligned}$$

设引理 8 中的 $X=B_t,Y=W_{t-1},\Delta=0,\xi=\xi_t$ 和 $f(W_{t-1},B_t)=-\eta_t g(W_{t-1},B_t)$，则有

$$\begin{aligned}I_1(W_t;B_t|W_{t-1})&=I_1(W_t-W_{t-1};B_t|W_{t-1})=I_1(-\eta_t g(W_{t-1},B_t)+\xi_t;B_t|W_{t-1})\\&\le\frac{1}{2}\log\left|\frac{\eta_t^2}{\sigma_t^2}\mathrm{Cov}[g(W_{t-1},B_t)]+I\right|+E^{\kappa}_{W_t|W_{t-1}},\end{aligned}$$

至此，证明完成。 □

B.3 命题5的证明

引理9。 设 V是一个按以下方式分块的 $n\times n$ 阶对称正定矩阵

$$V=\begin{bmatrix}A & C^\top\\ C & B\end{bmatrix},$$

*w*其中 A、B 分别是大小为 $n_1\times n_1$ 和 $n_2\times n_2$ 的对称矩阵。则 $|V|\le|A||B|$.

证明。 注意到

$$V=D\begin{bmatrix}A & 0\\ 0 & B-CA^{-1}C^\top\end{bmatrix}D^\top,\quad\text{where } D=\begin{bmatrix}I_{n_1} & 0\\ CA^{-1} & I_{n_2}\end{bmatrix},$$

于是我们有 $|V|=|D||A||B-CA^{-1}C^\top||D^\top|=|A||B-CA^{-1}C^\top|$。设 $D^\dagger$ 是 D的伪逆，则对任意长度为 n的列向量 $\mathbf{x}$：

$$\mathbf{x}^\top\begin{bmatrix}A & 0\\ 0 & B-CA^{-1}C^\top\end{bmatrix}\mathbf{x}=(\mathbf{x}^\top D^\dagger)V(\mathbf{x}^\top D^\dagger)^\top\ge 0,$$

因此 $B-CA^{-1}C^\top$ 是半正定的。类似地，我们可以证明 $CA^{-1}C^\top$ 是半正定的。设 λ_i、μ_i、ν_i、$i\in\{1,\cdots,n_2\}$ 分别是 $B-CA^{-1}C^\top$、B和 $CA^{-1}C^\top$ 的特征值，且按降序排列，则根据韦尔不等式，对所有 $i\in\{1,\cdots,n_2\}$，我们有 $\lambda_i\le\mu_i-\nu_{n_2}\le\mu_i$，这意味着 $|B-CA^{-1}C^\top|\le|B|$。结合上述结果即可完成证明：$|V|=|A||B-CA^{-1}C^\top|\le|A||B|$。 □

命题 5 (重述)。 给定如上所定义的 $\mathbb{V}_t$、V_t 和 L，设 $\{c_i\}_{i=1}^r$ 是 $\{n\}$ 的一个划分，即 $c_1\cup\cdots\cup c_r=\{n\}$ 且对于任意 $1\le i<j\le r$ 有 $c_i\cap c_j=\Phi$。设 $\mathbb{V}_t^i$ 是 $\mathbb{V}_t$ 的子矩阵，其列和行由 c_i 索引，并定义

$$\theta_c(\mathbb{V})=\frac{1}{2}\log\left|\frac{\eta_t^2}{\sigma_t^2}\mathbb{V}+I\right|,\quad\theta_v(V)=\frac{d}{2}\log\left(\frac{\eta_t^2V}{d\sigma_t^2}+1\right),$$

$$\textit{then}\qquad\theta_c(\mathbb{V}_t)\le\sum_{i=1}^{r}\theta_c(\mathbb{V}_t^i)\le\theta_v(V_t)\le\theta_v(L).$$

表 1：用于训练深度学习模型的超参数。

超参数	合成数据	MNIST	CIFAR-10
学习率（η）	0.001	0.01	0.01
训练数据集大小（n）	100	5000	5000
训练轮次	50	100	100
批量大小	10	50	50
步数（T）	500	10000	10000
噪声方差（σ_t^2）	10^{-3}	10^{-5}	10^{-5}

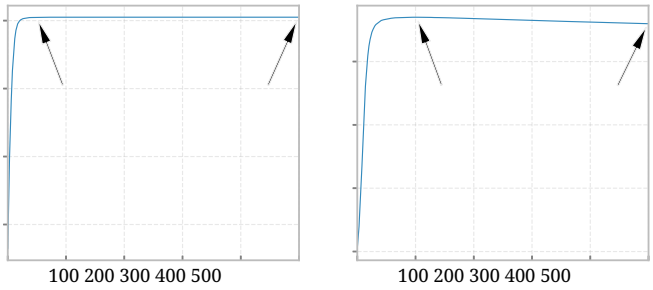

图 5： $I_1(W_T;\, S)$ 在合成数据上的表现。

证明。 注意到 $V_t = \mathrm{tr}[\mathbb{V}_t]$。由于协方差矩阵始终是对称半正定的，我们将 $\mathbb{V}_t$ 的特征值记为 λ_1、 $\cdots$、 $\lambda_d \geq 0$，则

$$\log\left|\frac{\eta_t^2}{\sigma_t^2}\mathbb{V}_t + I\right| = \log\left[\prod_{i=1}^{d}\left(\frac{\eta_t^2\lambda_i}{\sigma_t^2}+1\right)\right] \leq \log\left[\frac{1}{d}\sum_{i=1}^{d}\left(\frac{\eta_t^2\lambda_i}{\sigma_t^2}+1\right)\right]^d = d\log\left[\frac{\eta_t^2}{d\sigma_t^2}\sum_{i=1}^{d}\lambda_i + 1\right] = d\log\left[\frac{\eta_t^2 V_t}{d\sigma_t^2}+1\right],$$

其中，唯一的不等式依据是几何平均数总是小于算术平均数。令 $V_t^i = \mathrm{tr}[\mathbb{V}_t^i]$，则采用相同的策略，可以证明对于所有 $i \in \{1, \cdots, r\}$:

$$\theta_c(\mathbb{V}_t^i) \leq \theta_v(V_t^i),$$

那么根据詹森不等式，我们有:

$$\sum_{i=1}^{r}\theta_c(\mathbb{V}_t^i) \leq \sum_{i=1}^{r}\theta_v(V_t^i) = \sum_{i=1}^{r} d\log\left[\frac{\eta_t^2 V_t^i}{d\sigma_t^2}+1\right] \leq d\log\left[\frac{\eta_t^2}{d\sigma_t^2}\sum_{i=1}^{r}V_t^i + 1\right] = d\log\left[\frac{\eta_t^2}{d\sigma_t^2}V_t + 1\right] = \theta_v(V_t).$$

接下来，通过递归应用引理9，我们可以证明

$$\theta_c(\mathbb{V}_t) = \frac{1}{2}\log\left|\frac{\eta_t^2}{\sigma_t^2}\mathbb{V}_t + I\right| \leq \frac{1}{2}\log\prod_{i=1}^{r}\left|\frac{\eta_t^2}{\sigma_t^2}\mathbb{V}_t^i + I\right| = \frac{1}{2}\sum_{i=1}^{r}\log\left|\frac{\eta_t^2}{\sigma_t^2}\mathbb{V}_t^i + I\right| = \sum_{i=1}^{r}\theta_c(\mathbb{V}_t^i).$$

为了证明命题 5 的最后一个不等式，注意到

$$\begin{aligned} V_t &= \mathbb{E}_{B_t}[\|g(W_{t-1}, B_t) - \mathbb{E}_{B_t}[g(W_{t-1}, B_t)]\|_2^2] = \mathbb{E}_{B_t}[\|g(W_{t-1}, B_t)\|_2^2] - \|\mathbb{E}_{B_t}[g(W_{t-1}, B_t)]\|_2^2 \\ &\leq \mathbb{E}_{B_t}[\|g(W_{t-1}, B_t)\|_2^2] \leq \max_{w\in\mathcal{W}, z\in\mathcal{Z}}\|g(w,z)\|_2^2 = L, \end{aligned}$$

由对数函数的单调性即可完成证明。 □

B.4 定理 3 的证明

引理 10. 互信息 $I_1(\tilde{W}_T;\, S) \leq \sum_{t=1}^{T} I_1(-\eta_t g(W_{t-1}, B_t) + \tilde{\xi}_t;\, B_t|\tilde{W}_{t-1})$。

证明。

$$
\begin{aligned}
I_1(\tilde{W}_T;S) &= I_1(\tilde{W}_{T-1}-\eta_T g(W_{T-1},B_T)+\tilde{\xi}_T;S)\\
&\le I_1(\tilde{W}_{T-1},-\eta_T g(W_{T-1},B_T)+\tilde{\xi}_T;S) \qquad (28)\\
&= I_1(\tilde{W}_{T-1};S)+I_1(-\eta_T g(W_{T-1},B_T)+\tilde{\xi}_T;S|\tilde{W}_{T-1}) \qquad (29)\\
&\le I_1(\tilde{W}_{T-2};S)+I_1(-\eta_{T-1} g(W_{T-2},B_{T-1})+\tilde{\xi}_{T-1};S|\tilde{W}_{T-2})\\
&\quad + I_1(-\eta_T g(W_{T-1},B_T)+\tilde{\xi}_T;S|\tilde{W}_{T-1})\\
&\le \cdots\\
&\le I_1(\tilde{W}_0;S)+\sum_{t=1}^{T} I_1(-\eta_t g(W_{t-1},B_t)+\tilde{\xi}_t;S|\tilde{W}_{t-1}) \qquad (30)\\
&= \sum_{t=1}^{T} I_1(-\eta_t g(W_{t-1},B_t)+\tilde{\xi}_t;S|\tilde{W}_{t-1}) \qquad (31)\\
&\le \sum_{t=1}^{T} I_1(-\eta_t g(W_{t-1},B_t)+\tilde{\xi}_t;B_t|\tilde{W}_{t-1}) \qquad (32)\\
&= \sum_{t=1}^{T} I_1(\tilde{W}_t-\tilde{W}_{t-1};B_t|\tilde{W}_{t-1}) = \sum_{t=1}^{T} I_1(\tilde{W}_t;B_t|\tilde{W}_{t-1}),
\end{aligned}
$$

其中，注意到 $Z \to (X,\ Y) \to f(X,\ Y)$ 构成马尔可夫链，再应用命题~4中的性质5，即可得到式(28)；式(29)由命题4中的性质5得出；重复上述步骤可得式(30)；注意到 W_0~ ~和 S 相互独立，可得式(31)；根据 $S \to B_t \to -\eta_t g(w,\ B_t) + \xi_t|w = W_{t-1}$ 构成马尔可夫链这一事实，可得式(32)。 □

定理 3 （重述）。在与引理 *3* 相同的条件下：

$$
I_1\Big(\tilde{W}_T;S\Big) \le \sum_{t=1}^{T}\left(\frac{1}{2}\log\left|\frac{\eta_t^2}{\sigma_t^2}\mathbb{V}_t + I\right| + E^{\kappa}_{\tilde{W}_t|\tilde{W}_{t-1}}\right),
$$

$$
I\Big(\tilde{W}_T;S\Big) \le \sum_{t=1}^{T}\frac{1}{2}\log\left|\frac{\eta_t^2}{\sigma_t^2}\mathbb{V}_t + I\right|.
$$

证明。应用引理 10，可得

$$
I_1(\tilde{W}_T;S) \le \sum_{t=1}^{T}\left(\frac{1}{2}\log\left|\frac{\eta_t^2}{\sigma_t^2}\mathrm{Cov}[g(W_{t-1},B_t)] + I\right| + E^{\kappa}_{\tilde{W}_t|\tilde{W}_{t-1}}\right),
$$

通过应用引理 8，并令 $X = B_t$、$Y = W_{t-1}$、$\Delta = -\Delta_{t-1}$、$\xi = \xi_t$ 和 $f(W_{t-1},\ B_t) = -\eta_t g(W_{t-1},\ B_t)$，可得 ~ ~。 □

C 实验详情

深度学习模型在配备英特尔至强 CPU（2.10GHz，48 核）、256GB 内存和 4 块 Nvidia Tesla V100 GPU（32GB）的设备上进行训练。对于 MNIST 数据集，我们训练了一个包含一个大小为 128 的隐藏层的多层感知机。对于 CIFAR10 数据集，我们训练了一个包含 4 个大小为 3×3 的卷积层（32、32、48、48）的卷积神经网络，后接两个大小为 48 的全连接层。上述所有层均使用 ReLU 作为激活函数。训练所用的超参数列于表 1。

如图 5 所示，IWS 的曲线在经过几个训练轮次后很快停止上升并开始下降。这与真实泛化误差的行为一致，支持了我们的主张：$I_1(W_T;\ S)$ 的估计成功反映了 $I(W_T;\ S)$ 的行为。

在图 6 中，我们在 CIFAR-10 上进行随机标签实验；在图 7 中，我们测试了多层感知机在 MNIST 数据集上、隐藏层维度不同时的泛化行为。可以看出，我们的上界仍然比引理 3 的上界紧 5 倍以上。

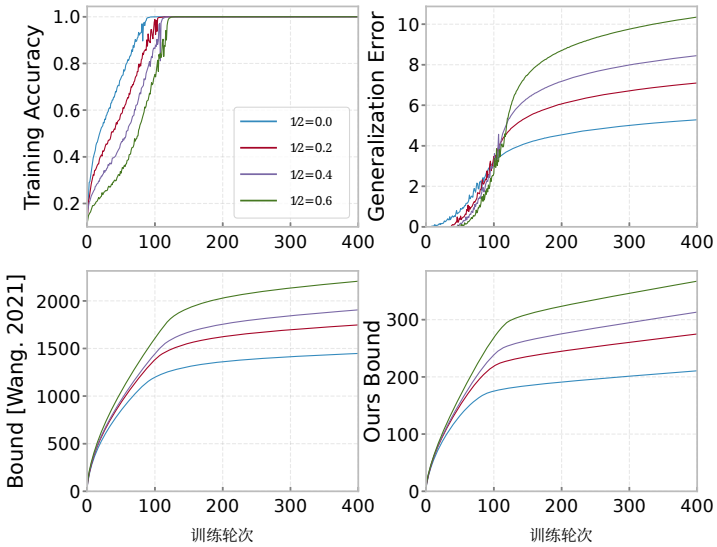

图 6：在 CIFAR-10 上进行的随机标签实验。

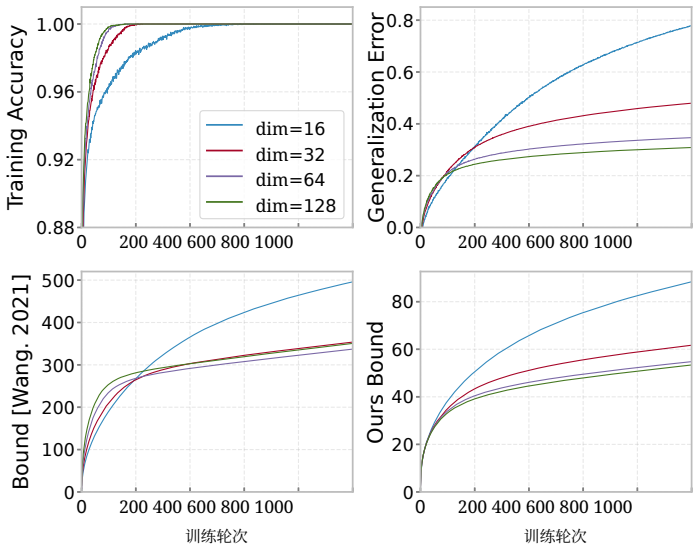

图 7：隐藏层维度不同时多层感知机在 MNIST 上的泛化。