计算机科学
COMPUTER SCIENCE

深度泛化机制的再思考:过参数化与高维噪声扰动下的一致收敛界重构

李鹏奇, 丁立中, 张春晖, 傅稼润

引用本文

李鹏奇, 丁立中, 张春晖, 傅稼润. 深度泛化机制的再思考:过参数化与高维噪声扰动下的一致收敛界重构[J]. 计算机科学, 2026, 53(4): 33-39.

LI Pengqi, DING Lizhong, ZHANG Chunhui, FU Jiarun. Rethinking Deep Generalization Mechanisms:Establishment of Uniform Convergence Bounds Under Overparameterization and High-dimensional Noise Perturbations [J]. Computer Science, 2026, 53(4): 33-39.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.250600129

深度泛化机制的再思考:过参数化与高维噪声扰动下的一致收敛界重构

李鹏奇　丁立中　张春晖　傅稼润
北京理工大学计算机学院　北京 100081
(pengqi.li@bit.edu.cn)

摘　要　深度神经网络在具备强大的表达能力的同时展现出优异的泛化性能,这与统计学习理论中“模型复杂度损害泛化”的经典论断存在本质冲突,导致传统框架下的深度泛化机制分析陷入困境。经典一致收敛界理论具有依赖参数空间维度、忽略算法隐式偏置等局限,难以直接适配深度网络核心特性。针对这一理论裂隙,构建了融合深度模型关键特征的新型统计学习理论框架,重构了一致收敛理论对深度模型泛化机制的解释范式。通过构建保留深度网络过参数化结构与高维噪声扰动特征的代理线性模型,首次推导出有效的一致收敛界,揭示了高维特征空间中噪声扰动对泛化性能的良性作用机制,突破了传统低维学习理论框架的局限性;基于深度泛化机制构造了数据规模敏感的规范化训练过程,揭示一致收敛界与泛化误差随样本复杂度增长呈现同步衰减的规律,证实了一致收敛理论对深度模型泛化机制的解释能力。基于理论与实验双重证据,突破了一致收敛泛化界的适配瓶颈,重新打开了一致收敛理论分析深度模型泛化性这扇即将被关闭的大门。

关键词:泛化误差;一致收敛界;修剪假设空间;高维概率;泛化机制
中图分类号　TP391

Rethinking Deep Generalization Mechanisms:Establishment of Uniform Convergence Bounds Under Overparameterization and High-dimensional Noise Perturbations

LI Pengqi,DING Lizhong,ZHANG Chunhui and FU Jiarun
School of Computer Science,Beijing Institute of Technology,Beijing 100081,China

Abstract　Deep neural networks demonstrate both powerful expressive capabilities and exceptional generalization performance, which fundamentally conflicts with the classical statistical learning tenet that “model complexity harms generalization”, rendering the analysis of deep generalization mechanisms under traditional frameworks intractable. Classic uniform convergence theory, constrained by its reliance on parameter space dimensionality and neglect of algorithmic implicit bias, fails to directly align with the core characteristics of deep networks. To address this theoretical gap, this paper constructs a novel statistical learning framework that integrates key features of deep models, thereby redefining the explanatory paradigm of uniform convergence theory for deep generalization mechanisms. It derives the first effective uniform convergence bound for deep networks by introducing a surrogate linear model that preserves overparameterization and high-dimensional noise-perturbation features, which reveals a benign role of high-dimensional noise in improving generalization beyond classical low-dimensional theory. Building on this deep generalization mechanism, it further proposes a scale-sensitive regularized training scheme and shows that the bound and the generalization error decay with increasing sample complexity. Supported by both theoretical and empirical evidence, this work breaks through the adaptability bottleneck of uniform convergence bounds and reopens the door for uniform convergence theory to analyze the generalization of deep models.

Keywords　Generalization error, Uniform convergence bound, Pruned hypothesis space, High-dimensional probability, Generalization mechanism

1　引言

深度神经网络[1-3](Deep Neural Networks,DNNs)在具备强大的非线性表征能力的同时具有卓越的泛化性,其区别于传统机器学习模型的关键特征在于参数量显著超越训练样本规模且权重中存在高维噪声扰动。然而,深度神经网络的泛化行为与统计学习理论所揭示的“复杂的模型往往泛化能力较差”规律[4-5]产生冲突,这意味着传统框架下的深度泛化

到稿日期:2025-06-20　返修日期:2026-01-10
基金项目:国家重点研发计划(2022YFB2703100);国家自然科学基金(62376028,U22A2099);国家自然科学基金优秀青年科学基金(海外)
This work was supported by the National Key Research and Development Program of China(2022YFB2703100),National Natural Science Foundation of China(62376028,U22A2099) and Excellent Young Scientists Fund(Overseas) of the National Natural Science Foundation of China.
通信作者:丁立中(lizhong.ding@outlook.com)

机制分析陷入困境。经典的泛化理论主要建立在一致收敛界基础上，结合算法的隐式偏置衍生出不同复杂度度量的算法-数据双依赖泛化界，包括 PAC-Bayes 界[6-7]、覆盖数界[8]、拉德马赫复杂度界[9-10]、神经正切核界[11]、一致稳定性界[12-14]等。在此背景下，构建适配深度网络关键特征的新型统计学习理论框架并证明有效的一致收敛界，不仅能够揭示深度网络的核心泛化机制，更能为建立理论驱动的可靠模型架构设计提供关键理论依据与范式基础。

然而，一致收敛界在适配深度网络关键特征上存在理论裂隙，通过一致收敛界揭示深度网络泛化机制的目标也尚未实现[15]，这引发了学术界对传统一致收敛泛化界理论在深度学习场景下适用性的根本性质疑。文献[16]通过研究高维参数空间中神经网络的泛化行为，指出传统一致收敛界过于依赖参数空间维度，进而无法预测过参数化模型的泛化性能；文献[17]在特定过参数化线性模型上证明了，充分拟合训练数据的分类器在包含其输出的任何假设空间上一致收敛界均失效；文献[18]发现“双下降”现象，即过参数化的深度网络的测试误差最终会随网络复杂度的增加而再次下降，故认为传统的一致收敛理论对深度网络的泛化机制缺乏解释能力。其他工作转而从其他视角建立与泛化误差的联系，例如信息瓶颈[19]、网络压缩[20]、因果推断[21]、差分隐私[22]等。

为突破深度神经网络泛化理论的适配性瓶颈，构建面向深度网络关键特征的新型统计学习理论框架，本文通过建立与深度模型对齐的过参数化高维噪声扰动代理模型，首次推导出包含算法隐式偏置的紧一致收敛界，证明并验证了泛化界与噪声维度呈$O(1/\sqrt{D})$的依赖关系，为揭示深度学习的泛化机制提供了超越传统低维理论的新视角；实验分析不同数据规模对训练动态的影响，结合“双下降”现象规范训练过程，揭示一致收敛界与泛化误差具有相似的数据依赖性，证实一致收敛理论对深度模型泛化机制的解释能力。通过理论与实证的双重验证，重新确立了一致收敛理论在深度学习泛化分析中的核心地位。

2 相关工作

2.1 一致收敛泛化界

泛化误差(Aeneralization Error)直接表征模型从训练数据到未知数据的泛化能力，其数值越小，意味着算法在分布外样本上的预测性能与训练集的拟合程度越一致。在统计学习理论框架下，泛化误差的上界分析往往通过建立泛化界(Generalization Bound)进行处理，其中假设空间复杂度的构造涵盖了多种学习理论工具，包括 PAC-Bayes 框架[6-7]、Rademacher 复杂度[9-10]、神经正切核理论[11]、一致稳定性[12-14]等。虽然这些方法在代数技巧上存在差异，但其核心均可视为一致收敛理论的不同变体。近些年，深度泛化理论已从传统的权重范数与容量度量拓展至图结构学习场景，针对谱 GNN[23]、等变 GNN[24]及对抗鲁棒情形建立了节点度无关的谱范数 PAC-Bayes 界与分布外风险上界，揭示了扩散矩阵谱半径、权重谱范数与扰动幅度共同决定模型样本复杂度与迁移误差。

然而，后续研究[25-26]通过理论分析与实证研究表明，直接在整个假设空间上应用一致收敛存在固有缺陷。这促使深度网络泛化分析转向结合具体优化算法的动力学特征，以算法依赖与数据依赖的方式重构假设空间。尽管如此，当前泛化界研究仍呈现出显著的局限性分化现象：一方面，部分数值较小的泛化界仅适用于特定网络变体(如随机网络[6]、压缩网络[20]或随机-压缩-再训练网络[23])；另一方面，多数泛化界往往过松且难以有效刻画泛化误差随数据复杂度[7]、数据规模[15]与网络深度[27]的变化规律，因而无法解释深度神经网络在过参数化状态下的卓越泛化表现。这些工作揭示出统计学习范式在兼容深度网络特性时面临着紧性与机理可解释性的双重约束难题。

2.2 深度模型的线性视角

将深度模型[1-3]的通用近似能力[5,28]与传统机器学习模型[29-31]的泛化能力结合，协同增强了模型的非线性表征优势与统计可解释性，是现代深度学习研究的热点课题。一致收敛在传统机器学习模型上取得了显著的成功，故在深度特征赋能的线性模型上证明一致收敛界，有望突破深度神经网络的泛化理论瓶颈。文献[32]构建了基于单层 Softmax 注意力机制的对偶模型 $f(\boldsymbol{x})=\boldsymbol{W}\phi(\boldsymbol{x})$，证明了注意力层在上下文学习过程中的推理过程与其对偶模型的参数更新过程形成严格对应关系。这里 $\phi(\cdot)$为特征映射，因此对于高维随机特征 $\phi(\boldsymbol{x})$来说，该对偶模型可以看作过参数化的线性分类器，研究线性模型可为理解 Transformer 类模型的隐式逻辑提供关键理论映射。此外，神经正切核理论[33-35]揭示了当深度神经网络的宽度趋向无穷大时，其非线性表征能力会渐近退化为高维线性模型，揭示了无限宽神经网络与过参数化线性分类器的动力学内在一致性。在深度统计学习理论框架下重新审视过参数化线性分类器，是揭示深度神经网络泛化机理的有效方法。

3 泛化机制再思考的动机

3.1 统计学习视角

设假设空间$\mathcal{H}$为从输入空间 χ 到实数域 $\mathbb{R}$ 的假设函数类，$\mathcal{D}$为定义在 $\chi\times\{-1,+1\}$上的联合概率分布。对于任一损失函数$\mathcal{L}$与任一假设函数 $h\in\mathcal{H}$，定义期望损失$\mathcal{L}_{\mathcal{D}}(h):=\mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\mathcal{L}(h(\boldsymbol{x}),y)]$；给定大小为 m 的样本集 $S\sim\mathcal{D}^m$，定义相应的经验损失$\hat{\mathcal{L}}_S(h):=\frac{1}{m}\sum_{i=1}^{m}[\mathcal{L}(h(\boldsymbol{x}_i),y_i)]$。设$\mathcal{A}$为学习算法，记 $h_S:=\mathcal{A}(S)$为该算法在任一样本集 S 上输出的假设函数。

对于给定的置信参数 $\delta\in(0,1)$，学习算法的泛化误差(Generalization Error)是训练集 S 上习得的假设函数h_S 的经验损失$\hat{\mathcal{L}}_S(h_S)$与关于分布$\mathcal{D}$的期望损失$\mathcal{L}_{\mathcal{D}}(h_S)$之差的概率上界：学习算法$\mathcal{A}$关于损失$\mathcal{L}$的泛化误差满足$\mathbb{P}_{S\sim\mathcal{D}^m}[\mathcal{L}_{\mathcal{D}}(h_S)-\hat{\mathcal{L}}_S(h_S)\leqslant\epsilon_{\text{gen}}(m,\delta)]\geqslant1-\delta$的最小$\epsilon_{\text{gen}}(m,\delta)$。泛化误差量化了机器学习模型在未知数据分布上的预测能力与训练集经验性能之间的统计偏离程度，是解释模型性能的核心依据。为了从理论上研究泛化误差的行为，最通用的方法是构造双侧一致收敛界(Uniform Convergence Bound)，对假设空间中的所有假设函数的预测性能施加一致的约束，即关于损失函数

$\mathcal{L}$的一致收敛界是满足$\mathbb{P}_{S\sim\mathcal{D}^m}[\sup_{h\in\mathcal{H}}|\mathcal{L}_{\mathcal{D}}(h)-\hat{\mathcal{L}}_S(h)|\leqslant\epsilon_{\text{unif}}(m,\delta)]\geqslant 1-\delta$的最小$\epsilon_{\text{unif}}(m,\delta)$。

3.2 对一致收敛界的质疑

适配深度网络过参数化与高维噪声扰动的特征一直是一致收敛理论面临的困境。文献[17]认为其无法顺利应用的原因在于高概率存在破坏一致收敛的样本-假设对$\{(S,h):S\sim\mathcal{D}^m,h\in\mathcal{H}\}$，而泛化误差的定义中，因为$\{(S,h_S):S\sim\mathcal{D}^m\}$之间具有样本-假设严格映射关系而排除了这种情形。为了构造最紧的一致收敛界，选择最小的假设空间，即仅包含算法$\mathcal{A}$在分布$\mathcal{D}$下可能选取的剪枝假设空间：

$$\mathcal{H}(\mathcal{D},\mathcal{A},m):=\{h_S\in\mathcal{H}:S\sim\mathcal{D}^m,h_S=\mathcal{A}(S)\}$$

经证明，对该假设空间的进一步剪枝操作都将破坏其闭包性，致使泛化误差界失效，由此建立的算法依赖一致收敛界可表述为：学习算法$\mathcal{A}$关于损失$\mathcal{L}$的最紧的算法依赖一致收敛界为满足关系式$\mathbb{P}_{S\sim\mathcal{D}^m}[\sup_{h\in\mathcal{H}(\mathcal{D},\mathcal{A},m)}|\mathcal{L}_{\mathcal{D}}(h)-\hat{\mathcal{L}}_S(h)|\leqslant\epsilon_{\text{unif-alg}}(m,\delta)]\geqslant 1-\delta$的最小$\epsilon_{\text{unif-alg}}(m,\delta)$。

4 再思考：过参数化与高维噪声扰动下的一致收敛界重构

为了重新建立经典一致收敛理论与现代深度网络之间的桥梁，思考一致收敛框架在深度模型泛化分析中的新范式，本章建立与深度模型对齐的过参数化高维噪声扰动代理模型，首次确立了充分考虑算法隐式偏置的算法-数据双依赖一致收敛界，并经验地揭示了训练过程中噪声维度与样本容量的依赖关系，为揭示深度学习的泛化机制提供了超越传统低维理论的新视角。

4.1 理论思考

1）设任一输入向量$\boldsymbol{x}\in\mathbb{R}^{K+D}$分解为$(\boldsymbol{x}_1,\boldsymbol{x}_2)$，其中$\boldsymbol{x}_1\in\mathbb{R}^K$为低维信号分量，$\boldsymbol{x}_2\in\mathbb{R}^D$为高维噪声分量($K\ll m\ll D$)。定义类别中心$\boldsymbol{u}\in\mathbb{R}^K$由向量控制，满足$\|\boldsymbol{u}\|_2=1/\sqrt{m}$。数据分布$\mathcal{D}$满足以下性质：1）标签$y$等概率取$+1$和$-1$；2）信号分量$\boldsymbol{x}_1$服从确定性分布$\boldsymbol{x}_1=k_1 y\boldsymbol{u}$；3）噪声分量$\boldsymbol{x}_2$服从球面高斯分布$\boldsymbol{x}_2\sim\mathcal{N}(0,(k_2^2/D)\boldsymbol{I})$。

由于两个类别的信号分量均值在K维信号空间中沿相反方向分离，且噪声分量对两类影响相同，故该分布在$\mathbb{R}^{K+D}$中是线性可分的。学习算法$\mathcal{A}$采用原点初始化的线性分类器，权重向量$\boldsymbol{w}=(\boldsymbol{w}_1,\boldsymbol{w}_2)$，决策函数$h(\boldsymbol{x})=\boldsymbol{w}_1\boldsymbol{x}_1+\boldsymbol{w}_2\boldsymbol{x}_2$。给定样本集$S$，$\mathcal{A}$执行学习率为1的梯度下降，目标函数为$\min_{\boldsymbol{w}}\sum_{i=1}^m -y_i h(\boldsymbol{x}_i)$，学习到的信号权重与分类边界对齐：$\boldsymbol{w}_1=k_1 m\boldsymbol{u}$，噪声权重服从球面高斯分布：

$$\boldsymbol{w}_2=\sum_{i=1}^m y^{(i)}\boldsymbol{x}_2^{(i)}\sim\mathcal{N}(0,(k_2^2 m/D)\boldsymbol{I})$$

该设置与深度模型的对齐体现在3个维度：1）参数空间的噪声扰动，模拟实际训练梯度噪声的统计特性；2）深度网络范数的经验观察，$\boldsymbol{w}=\Theta(\sqrt{m})$，与深度网络参数演化过程吻合；3）适度过参数化($m\ll D$)。特别地，文献[15]的设置在这里是特例($k_1=2,k_2^2=32$)。本节主要证明如下定理。

定理1 对于任意给定的$\epsilon>0$和$\delta>0$，当噪声维度D满足下界条件$D=\Omega(\max(m\ln\frac{m}{\delta},m\ln\frac{l}{\epsilon}))$时，对于0-1损失函数$\mathcal{L}$，其泛化误差满足$\epsilon_{\text{gen}}(m,\delta)\leqslant\epsilon$；在此基础上当参数满足$k_1^2>k_2^2$时，算法依赖的一致收敛界满足$\epsilon_{\text{unif-gen}}(m,\delta)\leqslant\epsilon$。

证明

1）对于$\epsilon_{\text{gen}}(m,\delta)\leqslant\epsilon$：只需证明对于$1-\delta$的大概率样本$S$，有$\hat{\mathcal{L}}_S(h_S)=0$且$\mathcal{L}_D(h_S)\leqslant\epsilon$。

(1) $\hat{\mathcal{L}}_S(h_S)=0$

对于$\hat{\mathcal{L}}_S(h_S)=\sum_{i=1}^m\mathcal{L}(h_S(\boldsymbol{x}^{(i)}),y^{(i)})$，只需证明对于大概率的样本$S$，$\forall i\in[m]$，有$y^{(i)}h_S(\boldsymbol{x}^{(i)})\geqslant 0$。做分解：

$$y^{(i)}h(\boldsymbol{x}^{(i)})=y^{(i)}\boldsymbol{w}_1\cdot\boldsymbol{x}_1^{(i)}+y^{(i)}\cdot y^{(i)}\|\boldsymbol{x}_2^{(i)}\|^2+y^{(i)}\cdot\boldsymbol{x}_2^{(i)}\cdot\sum_{j\neq i}y^{(j)}\boldsymbol{x}_2^{(j)}$$

其中，$\|\cdot\|$为ℓ_2范数。

第一项$y^{(i)}\boldsymbol{w}_1\cdot\boldsymbol{x}_1^{(i)}$：由数据分布$\mathcal{D}$的性质，有：

$$y^{(i)}\boldsymbol{w}_1\cdot\boldsymbol{x}_1^{(i)}=k_1 y^{(i)}m\boldsymbol{u}\cdot k_1 y^{(i)}\boldsymbol{u}=k_1^2$$

第二项$\|\boldsymbol{x}_2^{(i)}\|^2$：由数据分布$\boldsymbol{x}_2\sim\mathcal{N}(0,(k_2^2/D)\boldsymbol{I})$，有：

$$(\sqrt{D}/k_2^2)\boldsymbol{x}_2\sim N(0,\boldsymbol{I})$$

由等周不等式得到高斯集中性(文献[36]的定理5.2.2)，即对任一$\boldsymbol{x}\sim\mathcal{N}(0,\boldsymbol{I})$与Lipschitz连续函数$f:\mathbb{R}^D\to\mathbb{R}$，有：

$$\|f(\boldsymbol{x})-\mathbb{E}f(\boldsymbol{x})\|_{\psi_2}\leqslant C\|f\|_{\text{Lip}}$$

其中，$\|\cdot\|_{\psi_2}$指次高斯范数[7]，$\|\cdot\|_{\text{Lip}}$为Lipschitz范数，C是常数。特别地，取$f(\boldsymbol{x})=\|\boldsymbol{x}\|_2$，则：

$$\|\|\sqrt{D}\boldsymbol{x}_2/k_2\|_2-\sqrt{D}\|_{\psi_2}\leqslant C$$

由范数的线性性质，有：

$$\|\|\boldsymbol{x}_2\|_2-k_2\|_{\psi_2}\leqslant Ck_2/\sqrt{D}$$

由次高斯范数的性质(文献[36]的命题2.5.2)，对于$\forall\delta>0$，有：

$$\mathbb{P}[|\|\boldsymbol{x}_2\|_2-k_2|>\delta]\leqslant 2\exp\{-C'D\delta^2\}$$

其中，C'为与k_2有关的常数。由于

$$|z-1|\geqslant\delta\Rightarrow|z^2-1|\geqslant\max\{\delta,\delta^2\}$$

故

$$\mathbb{P}[|\|\boldsymbol{x}_2\|_2^2/k_2^2-1|>\delta]\geqslant\mathbb{P}[|\|\boldsymbol{x}_2\|_2/k_2-1|>\delta]$$

进而有：

$$\mathbb{P}[|\|\boldsymbol{x}_2\|_2^2-k_2^2|>\delta']\leqslant 2\exp\{-C'D\delta'^2\}$$

令$\delta/3m=2\exp\{-C'D\delta'^2\}$，即$\delta'=\sqrt{\frac{\ln(6m/\delta)}{DC'}}$，则对于$\forall i\in[m]$，有：

$$\mathbb{P}[|\|\boldsymbol{x}_2^{(i)}\|_2^2-k_2^2|>\sqrt{\frac{\ln(6m/\delta)}{DC'}}]\leqslant\delta/3m$$

根据概率的次可数可加性，推出：

$$\mathbb{P}[\forall i\in[m]:|\|\boldsymbol{x}_2^{(i)}\|_2^2-k_2^2|<\sqrt{\frac{\ln(6m/\delta)}{DC'}}]\geqslant 1-\frac{\delta}{3}$$

因此，当$D=\Omega(\ln(m/\delta))$时，$\|\boldsymbol{x}_2^{(i)}\|^2$对大概率$(1-\delta/3)$的样本$S$以$1/\sqrt{D}$的收敛速度趋于$k_2^2$。

第三项$y^{(i)}\cdot\boldsymbol{x}_2^{(i)}\cdot\sum_{j\neq i}y^{(j)}\boldsymbol{x}_2^{(j)}$：该项为相互独立的随机变量的内积，因此

$$\mathbb{E}[y^{(i)}\cdot\boldsymbol{x}_2^{(i)}\cdot\sum_{j\neq i}y^{(j)}\boldsymbol{x}_2^{(j)}]=\mathbb{E}[y^{(i)}\cdot\boldsymbol{x}_2^{(i)}]\cdot\mathbb{E}[\sum_{j\neq i}y^{(j)}\boldsymbol{x}_2^{(j)}]=0$$

由文献[8]的引理E.1，有：

$$\mathbb{P}[\forall i\in[m]:|y^{(i)}\cdot \boldsymbol{x}_2^{(i)}\cdot\sum_{j\neq i}y^{(j)}\boldsymbol{x}_2^{(j)}|\leqslant C''\frac{m}{\sqrt{D}}\sqrt{\ln\frac{6m}{\delta}}]\geqslant 1-\frac{\delta}{3} \tag{1}$$

其中，C''为常数。因此 $D=\Omega(m\ln(\frac{m}{\delta}))$，当 $D=\Omega(m\ln(\frac{m}{\delta}))$ 时，对于大概率$(1-\delta/3)$的样本 S，$y^{(i)}\cdot \boldsymbol{x}_2^{(i)}\cdot\sum_{j\neq i}y^{(j)}\boldsymbol{x}_2^{(j)}$ 以 $1/\sqrt{D}$的收敛速度趋于 0。

总之，对于 $1-2\delta/3$ 的样本 S，有：

$$\forall i\in[m]:|y^{(i)}h(\boldsymbol{x}^{(i)})-(k_1^2+k_2^2)|\leqslant O\left(\frac{m}{\sqrt{D}}\sqrt{\ln\frac{m}{\delta}}\right)$$

当 $D=\Omega(m\ln(\frac{m}{\delta}))$时，$\hat{\mathcal{L}}_S(h_S)$以 $1/\sqrt{D}$的收敛速趋于 0。

(2) $\mathcal{L}_D(h_S)\leqslant\epsilon$

对于$\mathcal{L}_D(h_S)=E_{(\boldsymbol{z},y)\sim\mathcal{D}}[\mathcal{L}(h_S(\boldsymbol{z}),y)]$，只需证明对于大概率的样本 S，有$\mathbb{P}_{(\boldsymbol{z},y)\sim\mathcal{D}}[yh_S(\boldsymbol{z})\geqslant 0]\leqslant\epsilon$即可。对测试点$(\boldsymbol{z},y)$进行分解：

$$yh_S(\boldsymbol{z})=y\boldsymbol{w}_1\cdot\boldsymbol{z}_1+y\cdot\boldsymbol{z}_2\cdot\sum_j y^{(j)}\boldsymbol{x}_2^{(j)}$$

第一项：

$$y^{(i)}\boldsymbol{w}_1\cdot\boldsymbol{x}_1^{(i)}=k_1y^{(i)}m\boldsymbol{u}\cdot k_1y^{(i)}\boldsymbol{u}=k_1^2$$

第二项：

$$yh(\boldsymbol{z})=y\boldsymbol{w}_1\cdot\boldsymbol{z}_1+y\cdot\boldsymbol{z}_2\cdot\sum_j y^{(j)}\boldsymbol{x}_2^{(j)}$$

本质上与 $y^{(i)}\cdot \boldsymbol{x}_2^{(i)}\cdot\sum_{j\neq i}y^{(j)}\boldsymbol{x}_2^{(j)}$ 结构相同，因此有类似的高概率界：

$$|y\cdot\boldsymbol{z}_2\cdot\sum_j y^{(j)}\boldsymbol{z}_2^{(j)}|=O\left(\sqrt{\frac{m}{D}\ln\frac{1}{\epsilon}}\right) \tag{2}$$

因此在满足过参数化条件 $D=\Omega\left(m\ln\frac{1}{\epsilon}\right)$时，$yh_S(\boldsymbol{z})$以 $O(1/\sqrt{D})$的收敛速度趋于 k_1^2，进而有$\mathbb{P}_{(\boldsymbol{z},y)\sim\mathcal{D}}[yh_S(\boldsymbol{z})\geqslant 0]\leqslant\epsilon$。

2) 其次证明$\epsilon_{\text{unif-gen}}(m,\delta)\leqslant\epsilon$。

对于任一样本 S，构造相应的逆噪声样本：

$$S'=\{((\boldsymbol{x}_1,-\boldsymbol{x}_2),y):((\boldsymbol{x}_1,\boldsymbol{x}_2),y)\in S\}$$

基于之前的分析，只需先证明 h_S 能正确分类 S'，再证明几乎不可能采到$|\mathcal{L}_D(h_S)-\hat{\mathcal{L}}_{S'}(h_S)|>\epsilon$的样本。

(1) $\hat{\mathcal{L}}_{S'}(h_S)=0$

设$\boldsymbol{x}_{\text{neg}}^{(i)}=(\boldsymbol{x}_1^{(i)},-\boldsymbol{x}_2^{(i)})\in S'$，做分解：$y^{(i)}h(\boldsymbol{x}_{\text{neg}}^{(i)})=y^{(i)}\boldsymbol{w}_1\cdot\boldsymbol{x}_1^{(i)}-y^{(i)}\cdot y^{(i)}\parallel\boldsymbol{x}_2^{(i)}\parallel^2-y^{(i)}\cdot\boldsymbol{x}_2^{(i)}\cdot\sum_{j\neq i}y^{(j)}\boldsymbol{x}_2^{(j)}$，与泛化误差的证明过程相同。得到满足过参数化条件 $D=\Omega(m\ln(m/\delta))$时，对于 $1-2\delta/3$ 的样本 S，有：

$$\forall i\in[m]:|y^{(i)}h(\boldsymbol{x}_{\text{neg}}^{(i)})-(k_1^2-k_2^2)|\leqslant O\left(\frac{m}{\sqrt{D}}\sqrt{\ln\frac{m}{\delta}}\right) \tag{3}$$

因此，当参数满足约束 $k_1^2>k_2^2$ 时，$\hat{\mathcal{L}}_{S'}(h_S)=0$ 对大概率的样本 S 成立。

(2)低概率采到$|\mathcal{L}_D(h_S)-\hat{\mathcal{L}}_{S'}(h_S)|>\epsilon$的样本。设 S_δ 满足$\mathbb{P}_{S\sim D}[S\in S]\geqslant 1-\delta$，则根据概率的性质，有：

$$\begin{aligned}&\mathbb{P}_{S,S'\sim\mathcal{D}^m}[|\mathcal{L}_{\mathcal{D}}(h_S)-\hat{\mathcal{L}}_{S'}(h_S)|>\epsilon:S,S'\in S_\delta]\\&=1-\mathbb{P}_{S,S'\sim\mathcal{D}^m}[|\mathcal{L}_{\mathcal{D}}(h_S)-\mathcal{L}_{S'}(h_S)|\leqslant\epsilon:S,S'\in S_\delta]\\&\leqslant 1-\mathbb{P}_{S,S'\sim\mathcal{D}^m}[(\mathcal{L}_{\mathcal{D}}(h_S)<\mathcal{L})\cap(\hat{\mathcal{L}}_{S'}(h_S)=0):S,S'\in S_\delta]\\&=\mathbb{P}_{S,S'\sim\mathcal{D}^m}[(\mathcal{L}_{\mathcal{D}}(h_S)\geqslant\epsilon)\cap(\hat{\mathcal{L}}_{S'}(h_S)\neq 0):S,S'\in S_\delta]\\&\leqslant\mathbb{P}_{S,S'\sim\mathcal{D}^m}[\mathcal{L}_{\mathcal{D}}(h_S)\geqslant\epsilon:S\in S_\delta]+\mathbb{P}_{S,S'\sim\mathcal{D}^m}[\hat{\mathcal{L}}_{S'}(h_S)\neq 0:S,S'\in S_\delta]\\&\leqslant\frac{\delta}{3}+\frac{2\delta}{3}=\delta\end{aligned}$$

即几乎不可能采到$|\mathcal{L}_{\mathcal{D}}(h_S)-\hat{\mathcal{L}}_{S'}(h_S)|>\epsilon$的样本。其中：

$$\mathbb{P}_{S,S'\sim\mathcal{D}^m}[\mathcal{L}_{\mathcal{D}}(h_S)\geqslant\epsilon:S\in S_\delta]\leqslant\delta/3$$

由之前的分析得到。

噪声维度满足下界条件 $D=\Omega\left(\max\left(m\ln\frac{m}{\delta},m\ln\frac{l}{\epsilon}\right)\right)$，且参数 $k_1^2>k_2^2$ 时，对于 0-1 损失函数$\mathcal{L}$，其泛化误差与一致收敛界分别满足$\epsilon_{\text{gen}}(m,\delta)\leqslant\epsilon$与$\epsilon_{\text{unif-gen}}(m,\delta)\leqslant\epsilon$。

证毕。

特别地，当 $k_1=2,k_2^2=32$ 时[15]，不满足 $k_1^2>k_2^2$，文献[17]由此质疑一致收敛理论的普适性，得出了如下结论。

引理 1(文献[17]的定理 1，取间隔 $\gamma=0$) 对于任意给定的 $\gamma=0$ 和 $\delta>0$，当噪声维度 D 满足下界条件 $D=\Omega\left(\max\left(m\ln\frac{m}{\delta},m\ln\frac{l}{\epsilon}\right)\right)$时，对于 0-1 损失函数$\mathcal{L}$，其泛化误差与泛化界分别满足$\epsilon_{\text{gen}}(m,\delta)\leqslant\epsilon$与$\epsilon_{\text{unif-gen}}(m,\delta)\geqslant 1-\epsilon$。

然而，参数 k_1,k_2 的选择并不影响线性模型过参数化与高维噪声扰动的特征，因此文献[17]的结论并不具有普适性。定理 1 证明了与深度模型对齐的过参数化高维噪声扰动代理模型上存在一致收敛界，且该一致界的收敛率与高维噪声维度呈$O(1/\sqrt{D})$依赖。

4.2 经验思考

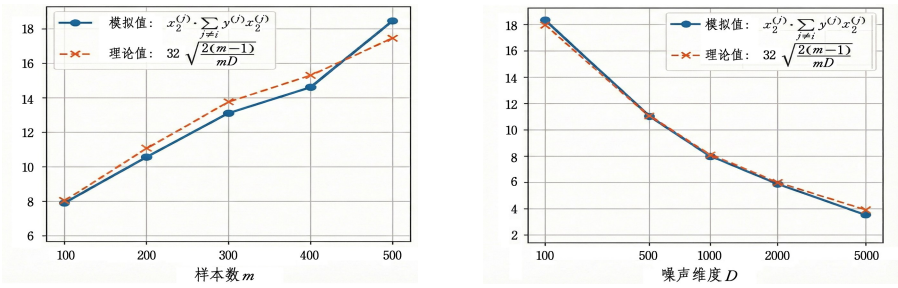
本节对定理 1 的关键步骤及主要结论进行经验验证，进而为揭示一致收敛界的深度模型应用提供经验证据。

基于蒙特卡罗模拟(200 次)的数值实验设置信号维数 $K=10$，初始化权重$\boldsymbol{w}_1=\boldsymbol{0},\boldsymbol{w}_2=\boldsymbol{0}$，通过学习率 $\eta=1$ 的梯度下降优化目标函数$\min_{\boldsymbol{w}}\sum_{i=1}^{m}-y_ih(\boldsymbol{x}_i)$，每组实验重复 5 次以降低随机性，最终结果取均值。

1)首先固定噪声维数 $D=1000$，依次取训练集样本规模 $m\in\{10,100,200,300,400,500\}$，其次固定样本规模 $m=100$，依次取噪声维度 $D\in\{10,100,1000,2000,3000,4000,5000\}$。如图 1 所示，计算模拟式(1)的结果并与理论值比较，验证了训练损失$\hat{\mathcal{L}}_S(h_S)$关于样本规模与高维噪声维度呈$O(\sqrt{m/D})$依赖。

2)与式(1)类似，在图 2 中计算式(2)的模拟结果并与理论值比较，验证了期望损失$\mathcal{L}_D(h_S)$关于样本规模与高维噪声维度呈$O(\sqrt{m/D})$依赖。

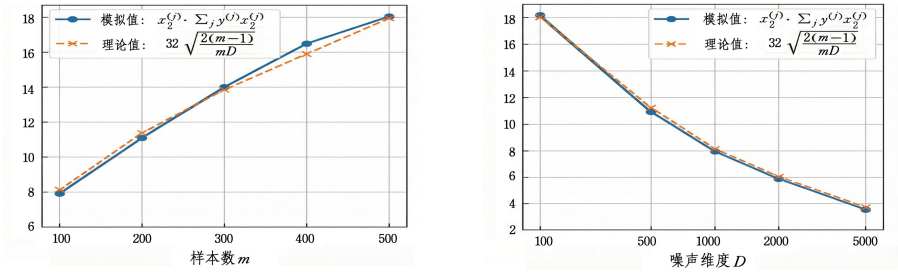
3)数据规模 $m=100$，噪声维数 $D\in[10^2,10^5]$对数均匀采样，构造负样本$\boldsymbol{x}_{\text{neg}}^{(i)}=(\boldsymbol{x}_1^{(i)},-\boldsymbol{x}_2^{(i)})$并计算相应的平均间隔 $\frac{1}{m}\sum_{i=1}^{m}y^{(i)}h(\boldsymbol{x}_{\text{neg}}^{(i)})$。在图 3 中采用 3 组参数组合$(k_1,k_2^2)\in\{(2,32),(4,16),(5,9)\}$，分别表征满足不同参数约束($k_1^2>k_2^2$，$k_1^2=k_2^2$及 $k_1^2<k_2^2$)的情况，基于定理 1 画出理论均值与 95% 置信区间，验证了式(3)的理论预测结果。



(a)样本数的影响($D=1000$)　(b)噪声维度的影响($m=100$)

图 1　样本数量与噪声维度对模拟式(1)的影响

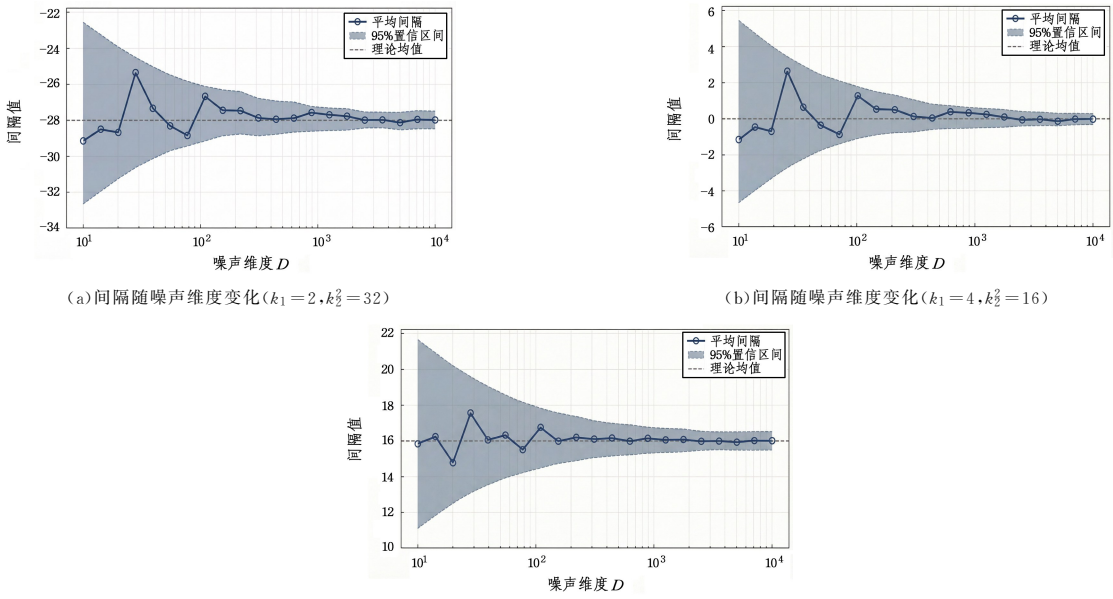
Fig. 1　Effect of sample size and noise dimension on the simulation of Eq. (1)



(a)样本数的影响($D=1000$)　(b)噪声维度的影响($m=100$)

图 2　样本数量与噪声维度对模拟式(2)的影响

Fig. 2　Effect of sample size and noise dimension on the simulation of Eq. (2)



(a)间隔随噪声维度变化($k_1=2,k_2^2=32$)　(b)间隔随噪声维度变化($k_1=4,k_2^2=16$)

(c)间隔随噪声维度变化($k_1=5,k_2^2=9$)

图 3　负样本上的平均间隔在不同设置下的上界估计

Fig. 3　Upper bound estimation of the margin on negative samples under different parameter settings

基于理论与经验的双重思考说明了定理 1 的正确性，验证了间隔上界与高维噪声维度的$O(\sqrt{1/D})$依赖性，证明了一致收敛界在适配深度网络关键特征上的普适性。

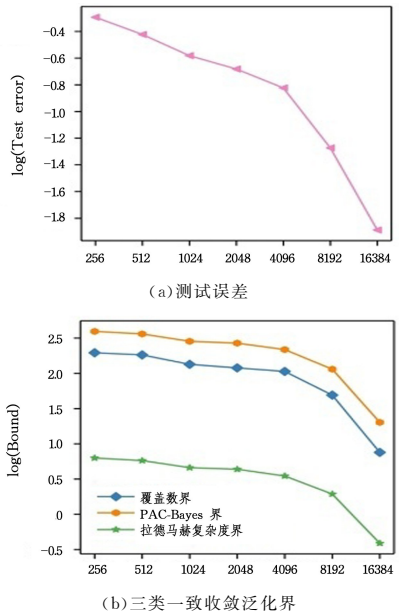
5　一致收敛界揭示深度泛化机制

本章关注一致收敛界的有效性评估问题。第 4 章的泛化性理论分析(定理 1)表明，在过参数化条件与高维噪声扰动下，线性模型的泛化误差收敛速率满足$O(1/\sqrt{D})$，揭示了噪声强度的增加会提升模型的泛化性能。进一步地，实际场景中噪声与样本规模往往存在耦合关系。例如，在大规模数据采集中，样本量的扩张通常伴随标注噪声的累积[37]，因此定理 1 也暗示了过参数化与高维噪声扰动下一致收敛界有随样本量增大而下降的趋势，而这正是泛化界有效的必要条件：能够准确刻画泛化误差的演化规律，特别是在训练集规模扩展时呈现递减特性。

然而，一些研究[8,32,38]在实验过程中发现传统泛化界中理论复杂度项与训练样本量呈正相关关系，这一反常现象导致现有理论框架难以合理解释深度神经网络的泛化行为。本章从实验角度指出设置的系统性偏差，并通过规范化训练过程，在经典的泛化界上初步验证了一致收敛界对深度泛化机

制的解释能力。常见的设置是采用固定间隔误差 0.01(即99%的训练数据被以间隔阈值 γ^* 正确分类)作为模型训练的终止条件,虽具有操作简便性,但忽视了不同数据规模对训练动态的差异性影响:当训练样本量增大时,模型需要更长时间的优化迭代才能达到同等水平的训练误差,而在模型训练过程中,泛化误差通常呈现先降后升的典型现象[18]。因此,采用固定间隔误差阈值作为终止准则的实验设计,实质上是强制不同数据规模的训练过程在参数空间的不同收敛阶段停止。大规模样本对应的训练终止点可能处于泛化误差曲线的上升阶段,该系统性偏差直接导致了实验观测中理论复杂度项与样本量正相关的现象。

本实验采用超球面二分类数据集(内/外半径分别为 1.0/1.2),训练集规模从 256 至 16384 指数增长,固定测试集 8192 样本;构建深度为 2、宽度 1000 的 ReLU 网络,使用 Xavier 初始化,通过 SGD 优化器(学习率 0.2,批量大小 1)进行训练;创新性地引入 20%验证集划分与早停机制(容忍 5 个 epoch 验证误差无改善),在验证误差最低点保存模型参数;计算覆盖数界[8]、PAC-Bayes 界[38]、Rademacher 复杂度界[39]3 种泛化界,严格控制权重初始化一致性,最终通过标准化早停策略确保不同数据规模的模型在各自最优泛化状态接受评估。实验结果如图 4 所示。



(a)测试误差

(b)三类一致收敛泛化界

图 4 规范化训练下,测试误差与三类一致收敛泛化界随训练样本数的变化

Fig. 4 Test error and uniform convergence generalization bounds versus training set size under the normalized training procedure

尽管本研究主要针对 3 类典型理论界展开分析,但所得结论对深度学习泛化理论体系具有普遍解释力——一般的复杂度项均建立于神经网络参数的范数度量基础之上(包括但不限于权重矩阵的谱范数、相对初始位置的 ℓ_2 距离)。

实验结果揭示一致收敛界与泛化误差随样本复杂度增长呈现同步衰减的规律,证明了一致收敛理论对深度模型泛化性的解释能力。

结束语 本文针对一致收敛理论的深度网络适配性瓶颈,构建了融合深度学习关键特征的统计学习模型,取得了以下突破:通过设计与深度模型对齐的过参数化高维噪声扰动代理模型,首次推导出包含算法隐式偏置的数据-算法双依赖一致收敛界,证明并验证了泛化界对高维噪声维度的 $O(1/\sqrt{D})$ 关联,突破了传统低维统计理论对深度模型泛化机制的解释局限;其次,基于数据规模对训练动态的系统性分析,揭示了一致收敛界与泛化误差在数据依赖性上的内在关联,验证了一致收敛理论对深度学习的泛化解释能力。此外,将本文的数据-算法双依赖视角迁移至马尔可夫决策过程,在强化学习的策略隐式偏置场景有潜在应用。本研究重新打开了一致收敛泛化界分析深度模型泛化性这一即将被关闭的大门,为构建适配深度网络特性的新一代统计学习理论奠定了重要基础。

参 考 文 献

[1] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(9): 1992-2001.

[2] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017: 5998-6008.

[3] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners[J]. Advances in Neural Information Processing Systems, 2020, 33: 18779-18797.

[4] BELLEC P H, BOUSQUET O, GUEDJ B, et al. Reconciling modern machine learning practice and the bias-variance trade-off [C]// Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS). PMLR, 2023: 2358-2376.

[5] CYBENKO G. Approximation by superpositions of a sigmoidal function[J]. Mathematics of Control, Signals and Systems, 1989, 2(4): 303-314.

[6] ERINGIS D, HOFMANN T, RAKHLIN A. PAC-Bayes generalisation bounds for dynamical systems including stable RNNs [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(11): 11901-11909.

[7] DZIUGAITE G K, ROY D M. Computing non-vacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data[J]. arXiv: 1703.11008, 2017.

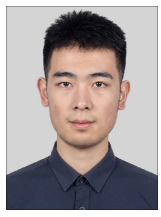
[8] BARTLETT P L, FOSTER D J, TELGARSKY M J. Spectrally-normalized margin bounds for neural networks[C]// Advances in Neural Information Processing Systems. California: NIPS, 2017: 6241-6250.

[9] BARTLETT P L, MENDELSON S. Rademacher and Gaussian ∂complexities: risk bounds and structural results[J]. Journal of Machine Learning Research, 2002, 3(11): 463-482.

[10] SACHS S, OBRIST R, SIMCHI-LEVI D, et al. Generalization guarantees via algorithm-dependent Rademacher complexity [C]// Proceedings of the 36th Annual Conference on Learning Theory. Bangalore: COLT, 2023: 4863-4880.

[11] DU S S, LEE J D, LI H, et al. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks[C]// Proceedings of the 36th International Conference

on Machine Learning. Long Beach, CA: PMLR, 2019: 1617-1626.
[12] HARDT M, RECHT B, SINGER Y. Train faster, generalize better: stability of stochastic gradient descent[C]// Proceedings of the 33rd International Conference on Machine Learning. New York: PMLR, 2016: 1225-1234.
[13] MOU W, WANG L, ZOU D, et al. Generalization bounds of SGLD for non-convex learning: two theoretical viewpoints[J]. arXiv: 1707.05947, 2017.
[14] ATTIA A, KOREN T. Uniform stability for first-order empirical risk minimization[C]// Proceedings of the Thirty-Fifth Conference on Learning Theory. PMLR, 2022: 3313-3332.
[15] JIANG Y, NASSER Y, RAGHAVAN D, et al. NeurIPS 2020 competition: predicting generalization in deep learning[C]// Proceedings of the NeurIPS 2020 Competition and Demonstration Track. Vancouver, BC: NeurIPS, 2020: 170-190.
[16] ADVANI M S, SAXE A M. High-dimensional dynamics of generalization error in neural networks[J]. arXiv: 1710.03667, 2017.
[17] NAGARAJAN V, KOLTER J Z. Uniform convergence may be unable to explain generalization in deep learning[C]// Proceedings of the 33st International Conference on Neural Information Processing Systems. 2019: 11615-11626.
[18] KAWAGUCHI K, ZHANG L, BENGIO Y, et al. How does information bottleneck help deep learning? [C]// International Conference on Machine Learning(ICML). ICML, 2023.
[19] ARORA S, GE R, NEGDY A, et al. Stronger generalization bounds for deep nets via a compression approach[C]// Proceedings of the 35th International Conference on Machine Learning. PMLR, 2018: 3597-3606.
[20] CHATTERJEE S, ZIELINSKI P. On the generalization mystery in deep learning[J]. arXiv: 2203.10036, 2022.
[21] DWORK C, ROTH A. The algorithmic foundations of differential privacy[J]. Foundations and Trends in Theoretical Computer Science, 2014, 9(3/4): 211-407.
[22] VASILEIOU A, JEGELKA S, LEVIE R, et al. Survey on generalization theory for graph neural networks[J]. arXiv: 2503.15650, 2025.
[23] SUN T, LIN J. PAC-Bayesian adversarially robust generalization bounds for graph neural network[J]. arXiv: 2402.04038, 2024.
[24] NEYSHABUR B, TOMIOKA R, SREBRO N. In search of the real inductive bias: on the role of implicit regularization in deep learning[C]// Proceedings of the 3rd International Conference on Learning Representations(ICLR). ICLR, 2015.
[25] ZHANG C Y, BENGIO S, HARDT M, et al. Understanding deep learning requires rethinking generalization[C]// Proceedings of the 5th International Conference on Learning Representations(ICLR). ICLR, 2017.
[26] ZHOU W, SUN Q, POGGIO T, et al. Non-vacuous generalization bounds at the ImageNet scale: a PAC-Bayesian compression approach[C]// Proceedings of the 7th International Conference on Learning Representations(ICLR). New York: ICLR, 2019: 1-14.
[27] NAGARAJAN V, KOLTER J Z. Generalization in deep networks: the role of distance from initialization[J]. arXiv: 1901.01672, 2019.
[28] YUN C, BHOJANAPALLI S, RAWAT A S, et al. Are transformers universal approximators of sequence-to-sequence functions? [J]. arXiv: 1912.10077, 2019.
[29] CORTES C, VAPNIK V. Support-vector networks[J]. Machine Learning, 1995, 20: 273-297.
[30] BISHOP C M, NASRABADI N M. Pattern recognition and machine learning[M]. New York: Springer, 2006.
[31] HOSMER D W, LEMESHOW S, STURDIVANT R X. Applied logistic regression[M]. John Wiley & Sons, 2013.
[32] REN R F, LIU Y. Towards understanding how transformers learn in-context through a representation learning lens[J]. Advances in Neural Information Processing Systems, 2025, 37: 892-933.
[33] JACOT A, GABRIEL F, HONGLER C. Neural tangent kernel: convergence and generalization in neural networks[C]// Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing. 2018.
[34] TAN L, WU S, ZHOU W, et al. Weighted neural tangent kernel: a generalized and improved network-induced kernel[J]. Machine Learning, 2023, 112(8): 2871-2901.
[35] JI Z, TELGARSKY M. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow ReLU networks[J]. arXiv: 1909.12292, 2019.
[36] VERSHYNIN R. High-dimensional probability: an introduction with applications in data science[M]. Cambridge: Cambridge University Press, 2018.
[37] ZHANG X H, LI M, WANG J, et al. A joint training framework for learning under label noise[J]. Journal of Computer Research and Development, 2022, 59(10): 2021-2035.
[38] NEYSHABUR B, BHATIA K, BARTLETT P L, et al. A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks[J]. arXiv: 1707.09564, 2017.
[39] NEYSHABUR B, TOMIOKA R, SREBRO N. Towards understanding the role of over-parametrization in generalization of neural networks[C]// Proceedings of the 35th International Conference on Machine Learning. PMLR, 2018: 3784-3793.



LI Pengqi, born in 2002, Ph.D candidate. His main research interests include deep statistical learning theory, kernel learning, uncertainty in large language models and physical AI.

DING Lizhong, born in 1986, professor. His main research interests include deep statistical learning theory and methods, the emergence mechanisms and reasoning mechanisms of large-scale models, statistical hypothesis testing and deep generative models, and neural-symbolic learning theory and methods.

（责任编辑：何杨）