arXiv:2602.04233v1 [stat.ML] 4 Feb 2026

随着预训练模型规模扩大，可证明的目标样本复杂度改进

Kazuto Fukuchi*[1, 2]、Ryuichiro Hataya[3,4]、和 Kota Matsui[4,5,6]

[1]日本筑波大学[2]日本理化学研究所革新智能统合研究中心[3]日本SB Intuitions Corp.[4]日本京都大学[5]日本滋贺大学[6]日本东京科学大学

2026年2月5日

摘要

预训练模型已成为在广泛的下游任务中高效构建模型不可或缺的工具。关于缩放定律的实证研究凸显了预训练模型的优势，这些研究表明，更大的预训练模型能够显著降低下游学习的样本复杂度。然而，现有的预训练模型理论研究尚无法解释这一现象。本文提出一种受参数高效微调（PEFT）方法启发的新框架——填缝，并据此开展理论研究。此类方法包括基于适配器的微调、低秩适配和部分微调。我们的分析证明，改进的预训练模型能够降低下游任务的样本复杂度，从而为实证观察到的预训练模型规模与下游性能之间的关系（即缩放定律）提供理论依据，而现有研究尚未涵盖这一关系。

1 引言

在不同定义域中利用预训练模型，已成为开发特定应用模型的常见策略。这种方法使得即使在定义域特定数据有限的情况下，也能构建出高精度模型。例如，在医学图像识别领域，研究人员已开发出众多预训练模型，用于疾病诊断和病变区域识别等多种任务（Wen 等人，2021）。近来，能够在统一框架中处理胸部X光和脑部CT图像等不同模态的基础模型也已问世（Azad 等人，2023；Wang 等人，2025）。此外，在药物发现和材料科学等领域，能够处理化学结构的预训练模型和基础模型正成为强大的工具（Xia 等人，2022；Pyzer-Knapp 等人，2025）。

*fukuchi@cs.tsukuba.ac.jp，通讯作者

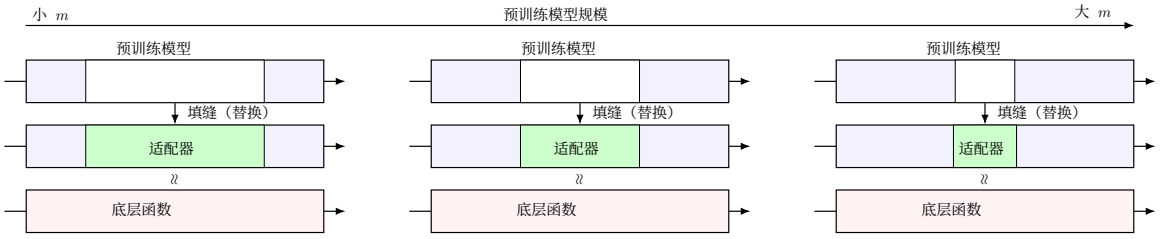

图 1：填缝的概念示意图。蓝色方框表示预训练模型，红色方框表示底层函数。横轴表示源样本量 m，对应于预训练模型的规模。

关于缩放定律 (Henighan 等人，2020；Mikami 等人，2023)的实证研究凸显了利用大型预训练模型的优势。缩放定律最初由 Kaplan 等人 (2020)在大语言模型（LLM）的背景下提出，表明大语言模型的性能会随着模型规模、数据集规模以及训练所用计算量的增加而提升。随后，Henighan 等人(2020)在预训练自回归模型的背景下进一步研究了预训练模型的缩放定律，发现规模更大的预训练模型能够显著降低下游任务微调所需的样本复杂度。Mikami 等人(2023)还在合成到真实迁移学习的背景下证明了预训练模型存在类似的缩放定律，表明预训练模型的性能会随着预训练数据量的增加而提升。本文特别关注预训练模型的数据缩放定律，即下游任务的性能会随着预训练数据量的增加而提升。

预训练模型的有效性也得到了大量理论研究的支持，包括少样本学习方面的研究（Du 等人，2020）以及上下文学习方面的研究（Bai 等人，2023b；Kim 等人，2024）。这些研究大多试图通过建立学习算法误差的上界来证明预训练模型的优势，该上界由源样本量决定 m （用于构建预训练模型）和目标样本量 n （下游任务的样本量）。具体而言，这些分析中得到的误差率通常表示为：

$$\mathbb{E}[E_Q(f_n)] \leq m^{-\alpha} + n^{-\beta}, \tag{1}$$

其中， E_Q（f_n）表示估计的回归函数 f_n 在下游任务的目标分布Q 下的误差，而 α， $\beta > 0$ 是取决于底层函数类的复杂度和学习算法的常数。如果 β 大于从头训练可达到的速率，式 (1)则表明预训练模型有益，前提是 m 足够大，使得 $m^{-\alpha} \leq n^{-\beta}$。

然而，这些理论结果难以解释预训练模型的数据缩放定律。在式（1）的理论分析框架下，预训练模型的数据缩放定律可解释为：源样本量 m （对应预训练模型的规模）增加，从而改善相对于目标样本量 n （下游任务的样本复杂度）的样本复杂度。然而， n 的样本复杂度由 β 在式（1）中刻画，且不随预训练模型的规模变化， m。因此，现有关于式（1）的结果并未充分捕捉预训练模型的数据缩放定律。

为弥补这一不足，本文旨在探讨以下关键研究问题：

预训练模型的哪些特性能够确保样本复杂度得到改善?

随着源样本量增加，目标样本量如何变化?

识别预训练模型的此类特征，有助于更细致地理解它们近期取得成功的原因，尤其是预训练模型的数据缩放定律，并可能促成更高效的预训练模型构建方法。

贡献。 本文的主要贡献在于阐明预训练模型中哪些性质导致了实践中观察到的数据缩放定律。具体而言，我们证明，在对预训练模型的某一假设下，误差率满足

$$\mathbb{E}[E_Q(f_n)] \leq n^{-\beta+o(1)}, \tag{2}$$

对某个常数 $\beta > 0$，其中 o（1） 是一个有限项，它随 m 递减，并在 $m \to \infty$时消失。式 (2) 中的速率表明，关于 n 的样本复杂度会随着源样本量 m 的增加而改善，因为 n 的指数随 m减小。因此，式 (2)中的界为预训练模型的数据缩放定律提供了理论解释。

为了达到式 (2)中的速率，我们引入一个称为可填隙性的新概念，其概念示意见图 1。可填隙性要求预训练模型（图 1中的蓝色方框）具有层次结构，使得底层函数（图 1中的红色方框）可通过插入另一个称为适配器模型 （图 1中的绿色方框）的函数到该层次结构中来近似。我们证明，如果适配器模型的复杂度随着源样本量m 的增加而降低（即适配器模型的深度沿表示 m的横轴在 图 1中收缩），则可以构造一种称为经验填缝的学习算法，使其达到式 (2)中的误差率。该结果表明，降低适配器模型的复杂度是达到式 (2)中速率的充分条件，从而回答了上述研究问题。这也表明，参数高效微调（PEFT）方法在利用预训练模型进行学习时可以是有效的，因为我们可以将利用低复杂度适配器模型理解为PEFT的一种形式。

我们进一步提供了支持理论结果的实证证据。具体而言，我们对卷积神经网络进行微调，以及将视觉能力集成到大语言模型中的实验表明，更大的预训练模型可以通过更浅层的适配器模型适配到下游任务。这些发现进一步支持了我们的理论主张：适配更大的预训练模型能够降低下游任务所需的样本复杂度。

所有缺失的证明均推迟至附录 A。

2 相关工作

领域自适应理论。 领域自适应是迁移学习的一个重要分支，研究这样一种情形：尽管学习任务相同，但训练（源域）数据与测试（目标域）数据来自不同的分布（Redko等人，2020）。本质上，领域自适应是一种迁移学习方法，它利用大量源域数据以及少量明确提供的目标域数据（有标签或无标签）；理论分析通常也基于这一假设展开。为理解域偏移下的泛化性能，研究者已经建立了丰富的理论。由 Ben-David等人(2006，2010) 开创的研究提出了领域自适应的泛化界，将目标误差与三个部分联系起来，即源域误差、域差异，以及一个不可约项（称为联合误差）。一般界可以非正式地表述如下：对于任意一个假设 h，属于一个

给定的类，目标风险 $R_T(h)$ 由源风险 $R_S(h)$ 加上定义域之间的差异度以及表示标签不匹配的常数项所界定。这一结果表明，如果两个定义域非常相似（差异度较小），且共享同一个底层标签函数，那么模型就能在不同定义域之间实现良好的泛化；相反，较大的分布差异或相互冲突的标签定义会导致更大的泛化误差。许多理论成果进一步完善了这些观点。研究者针对不同情形提出了各种差异度度量，例如用于分类和回归的对称差 $\mathcal{H}$-散度（Ben-David等人，2010），并将该理论扩展到多源适应及其他情形（Mansour等人，2008）。其他研究则分析了最大均值差异（Redko等人，2019）或 Wasserstein距离（Shen等人，2018），并为其提供了各自的理论保证。

除基于差异度的理论之外，人们也探索了其他方法。例如，Kpotufe (2017); Ma 等人 (2023); Feng 等人 (2023)研究了利用密度比考虑源域与目标域之间分布偏移的方法的误差上界。此外，Kpotufe 和 Martinet (2021); Pathak 等人 (2022); Galbraith 和 Kpotufe(2023); Fujikawa 等人 (2025)试图利用度量空间中基于超球面的相似性度量来分析误差界。尽管这些分析看似给出了可靠的误差界，但仍可指出若干局限。例如，在基于密度比的方法中，分析往往在真实密度比已知的假设下进行；而实践中该密度比未知，因此未考虑估计误差。此外，在这两种方法中，当源数据分布与目标数据分布的支撑显著不匹配时，相似性度量可能会发散。因此，现有的领域自适应理论不足以作为分析预训练模型误差的工具；需要开发新的工具。

预训练模型的深度学习理论。在理解深度学习模型的统计性能方面已取得大量理论进展（Schmidt-Hieber，2020；Suzuki 和 Nitanda，2021；Kohler 和 Langer，2021；Suh 等人，2022；Nishikawa 等人，2025；Damian 等人，2022，2023）。这一研究方向已进一步扩展到分析预训练模型的统计性质，尤其是确立式（1）中的误差率。例如，Du 等人（2020）在线性特征类和线性结果模型中研究少样本学习问题，证明了式（1）形式的误差率。Bai 等人（2023b）在线性回归模型下研究 Transformer 用于上下文学习的统计性能，同样得到式（1）的形式。Kim 等人（2024）在 Besov 空间中分析 Transformer 的上下文学习，再次表明式（1）形式的误差率。如引言所述，这些结果都具有式（1）的形式，因此未能解释预训练模型中观察到的缩放定律。

大语言模型。 所提出的填缝概念与近期大语言模型（LLM）领域的进展密切相关。尤其是，多模态语言模型的近期训练方法（Li 等人，2023；Bai 等人，2023a；Liu 等人，2023；Dai 等人，2023；Gosthipaty 等人，2024；Liu 等人，2025）以及参数高效微调（PEFT）技术（Houlsby 等人，2019；Guo 等人，2021；Hu 等人，2021）都可以视为与填缝类似。例如，使用预训练视觉特征提取器和大语言模型构建视觉语言模型时，一种典型策略是利用图文对数据集，学习从视觉特征到文本特征的映射，这种映射称为适配器模型。随后，将学得的适配器模型插入视觉特征提取器与大语言模型之间，构成最终的视觉语言模型，这与填缝过程类似。同样，PEFT 方法涉及用复杂度较低的组件替换或增强预训练模型，这也与填缝类似。

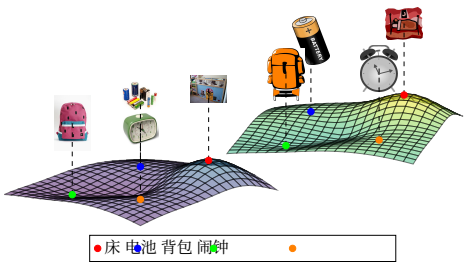

图2：一个说明成功利用预训练模型的示例。

3 通过填缝进行预训练模型学习

记号。设 E $\in$ 3 是概率空间 (Z 3, ν) 中的一个事件。E 的补集记为 E^c，其概率记为 $\mathbb{P}_\nu${E}。对于定义在 (Z 3, ν) 上、取值于可测空间 (X $\mathfrak{X}$) 的随机变量 X，其期望记为 $\mathbb{E}_\nu[X]$。对于 $p \in [1,\infty)$，函数 f : Z $\to \mathbb{R}$ 的L^p-范数记为$\|f\|_{L^p}$(Z) $=(\int$Z $|f|^p d\lambda)^{1/p}$，其中 Z $\subseteq \mathbb{R}^d$，对某个 $d \in \mathbb{N}$ 而言，λ 是勒贝格测度。对于可测空间 (Z 3) 上的测度 ν，f: Z $\to \mathbb{R}$ 的$L^p(\nu)$-范数定义为$\|f\|_{L^p(\nu)}$ $=(\int |f|^p d\nu)^{1/p}$，其中 $p \in [1,\infty)$，且$\|f\|_{L^\infty(\nu)}$ $=$inf{$b > 0$: $\nu(\{z \in$ Z : $|f(z)| \le b\}) = 1$}。

问题设置。我们考虑存在预训练模型时的非参数回归问题。令 $X \in \mathcal{X}$ 和 $Y \in \Omega \subset \mathbb{R}$ 分别表示协变量和结果变量，并令 P 和 Q分别表示源分布和目标分布。在目标分布 Q下，假设 X 和 Y遵循非参数回归模型：

$$Y = f^*(X) + \xi,$$

其中，f^* 是所关注的回归函数，ξ 是与 X独立的零均值噪声项。假设学习器可以访问由从 P中抽取的大小为m的源样本构建的预训练模型 f_{pre}。学习器的目标是结合预训练模型 f_{pre}，利用来自 Q的大小为 n的目标样本估计 f^*。令 f_n 表示估计的回归函数。于是，通过从 f^*出发定义的 $L^2(Q_X)$距离来评估 f_n的准确性：

$$E_Q(f) = \|f - f^*\|^2_{L^2(Q_X)},$$

其中，Q_X 是 Q下 X 的边缘分布。

动机示例。图 2 展示了一个概念性示例，说明如何有效利用预训练模型。示例采用了受 Office-Home 数据集（Venkateswara 等人，2017）启发的图像。图中包含两个流形，分别表示基于两个定义域特征空间，将图像分类为床的概率：真实世界（f^*）和剪贴画（P）。流形上的点对应特征，图像之间用虚线连接。两个流形仅在旋转和平移方面有所不同。利用真实世界图像（源样本）构建预训练模型，以捕捉左侧流形的复杂结构。随后，通过适当的变换将右侧流形与左侧流形对齐，即可准确估计左侧流形上剪贴画图像（目标样本）的 f^*。从头学习复杂的流形结构所需的样本量，大于仅学习旋转和

平移（即线性变换）所需的样本量。因此，利用预训练模型能够在目标样本量较小时实现较高的预测准确率，凸显了使用预训练模型的潜在效用。

填缝。 图中所示的情形 图 2 可用层次结构来描述，并通过可填隙性的概念予以形式化。

定义 1 (可填隙性)。 设 G 是函数类 $g: Z_i \to Z_o$，其中 Z_i 和 Z_o 是任意定义域。给定 $\epsilon > 0$ 和 G，回归器 $f^*: X \to \mathbb{R}$ 若满足以下条件，则称其在 $L^2(Q_X)$- 范数意义下可由 (g_h, g_e) 进行 (ϵ, G)- 填隙，其中 $g_e: X \to Z_i$ 且 $g_h: Z_o \to \mathbb{R}$，

$$\inf_{g_a \in \mathcal{G}} \|f^* - g_h \circ g_a \circ g_e\|_{L^2(Q_X)} \leq \epsilon. \tag{3}$$

为了成功利用预训练模型，我们假设理想回归器 f^* 可由预训练模型 $f_{\text{pre}} = (g_h, g_e)$填隙。在定义 1 中， g_e 表示从图像到特征的映射，对应图 2 中的虚线；图 2 中的流形由这些特征参数化，其中高度对应于 g_h 的输出，空间位置对应于其输入； g_a 表示特定领域变换，分别对应于图 2 左右流形的旋转和平移。式（3）表明，将特定目标变换 g_a 插入预训练模型f_{pre}后，即可（近似）恢复理想回归器 f^*。我们将预训练模型的这一性质称为可填隙性，表示通过填补 g_h 与 g_e之间的间隙来恢复理想回归器 f^* 的情形。我们将 g_h 和 g_e分别称为头部模型和特征提取器模型，并将 g_a 称为适配器模型。

经验填缝。我们提出一种利用预训练模型的学习框架，称为经验填缝。设 (X_1, Y_1) ..., (X_n, Y_n) 是从 Q 中抽取的目标样本。假设理想回归器 f^* 可由预训练模型 $f_{\text{pre}} = (g_h, g_e)$ 以 (ϵ_n, G)-填缝，其中 ϵ_n 是一个常数，可能取决于目标样本量 n，而 G 是给定的函数类。我们通过将微调适配器模型 g_a 插入预训练模型 f_{pre}，为 f_n 设计一种学习算法。给定一个可能取决于 n、由可能的 g_a 构成的类 $\mathsf{G}_n \subseteq \mathsf{G}$，定义 $\mathsf{F}_n = \{g_h \circ g_a \circ g_e : g_a \in \mathsf{G}_n\}$。因此，$\mathsf{F}_n$ 表示将适配器模型 $g_a \in \mathsf{G}_n$ 插入预训练模型 f_{pre} 后可得到的函数集合。估计回归函数 f_n 定义如下

$$f_n = \arg\min_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Y_i)^2.$$

该学习算法等价于最小二乘估计量，但限定在特定的函数类 $\mathcal{F}_n$ 上。

注 *1.* 我们当前对经验填缝的表述涵盖了大多数参数高效微调方法（参数高效微调方法），但无法直接涵盖 *LoRA* 的情形（*Hu* 等，*2021*）。不过，*LoRA* 可以被解释为一种将若干层插入预训练模型的方法。具体而言，令 $g_H \circ \cdots \circ g_1$ 为一个 H 层预训练模型，考虑修改后的版本 $h_{H+1} \circ h_H \circ \cdots \circ h_1 \circ h_0$，其中对于 $i = 1, ..., H$， $h_i(x, y) = (g_i(x)\ y)$，对于 $h_0(x) = (x, x)$，以及 $h_{H+1}(x, y) = x$。于是， h_i 的复合结构等价于 g_i 的复合结构。*LoRA* 可以被视为将适配器层 $h_{1+1/2}, ..., h_{H+1/2}$ 插入这一修改后的预训练模型中。具体而言，通过对 $h_{i+1/2}(x, y) = (x + \Delta W y, x + \Delta W y)$（对于 $i = 1, ..., H$）进行设置，其中 ΔW 为低秩的，即可恢复 *LoRA*。因此，将我们的结果扩展到涵盖插入多个适配器层的情形，将能够涵盖 *LoRA*，这是本工作的一个重要未来方向。

4 经验填缝的误差分析

在本节中，我们利用类 $\mathcal{G}$ 的适当复杂度测度，对与经验填缝相关的误差进行严格分析。在本节中，我们假设 $\mathcal{Z}_o$ 配备了范数 $\|\cdot\|$，并将适配器模型 g_a 的 L^∞-范数定义为：$\mathcal{Z}_i \to \mathcal{Z}_o$ 为 $\|g_a\|_{L^\infty} = \sup_{z\in\mathcal{Z}_i} \|g_a(z)\|$。

假设。 我们在整个分析中作出如下假设：

假设 1。 以下条件成立：

1. 噪声 ξ 是方差代理为 $\sigma^2 < \infty$ 的次高斯噪声，即 $\mathbb{E}[\exp(\lambda\xi)] \le \exp(\lambda^2\sigma^2/2)$ 对所有$\lambda \in \mathbb{R}$成立。
2. Ω 是有界的，即存在常数 $\Delta_\Omega > 0$，使得 $\sup_{x,y\in\Omega} |y - x| \le \Delta_\Omega$。

假设1 在非参数回归与统计学习理论的文献中是标准的。先前工作中也作出了类似假设，例如 Schmidt-Hieber（2020）以及 Suzuki 和 Nitanda（2021），这些工作分析了深度学习与非参数回归方法的理论性质。

复杂度：$\mathcal{G}$。 为刻画经验填缝的误差，我们采用一种同时基于逼近误差与覆盖数的 $\mathcal{G}$ 复杂度测度。给定 $\delta > 0$、类 $\mathcal{G}$，以及定义在 $\mathcal{G}$ 上的范数 $\|\cdot\|$，δ-覆盖数 $N(\delta, \mathcal{G}, \|\cdot\|)$ 是覆盖 $\mathcal{G}$ 所需的、半径为 δ（关于 $\|\cdot\|$）的球的最小个数。下述定义将复杂度这一概念形式化。

定义 2 (β-复杂度)。 设 $\beta > 0$。函数类 G $g : Z_i \to Z_o$ 是 β-复杂的，即对于每个 $J \in \mathbb{N}$，都存在一个类 G_J，使得

$$\ln(N(\delta, \mathcal{G}_J, \|\cdot\|_{L^\infty})) \le cJ\mathrm{polylog}(J, 1/\delta),$$

and

$$\sup_{g^*\in\mathcal{G}} \inf_{g\in\mathcal{G}_J} \|g - g^*\|_{L^\infty} \le c' J^{-\beta},$$

对某些常数 c、$c' > 0$，其中 polylog 表示其参数中的一个多对数因子 *ts.*

定义 2 通过逼近能力与各类复杂度之间的权衡，刻画了$\mathcal{G}$ 类 $\mathcal{G}_J$。这一概念也用于分析稀疏 ReLU 网络等函数类的非参数回归误差界，相关研究见（Schmidt-Hieber，2020；Suzuki 和 Nitanda，2021；Chen 等人，2022）。定义 2 的一个具体应用见第 4.1 节。

赫尔德连续性。 除了 β-复杂度，我们还采用赫尔德连续性来量化预训练模型 f_{pre} 的正则性。

定义 3 (赫尔德连续性)。 设 g 是一个从 X 到 Z 的函数，其中 X 和 Z 分别配备范数 $\|\cdot\|_X$ 和 $\|\cdot\|_Z$。对于 $\alpha \in (0, 1]$，若存在常数 $C > 0$ 使得 g 满足以下条件，则称其为 α-赫尔德连续

$$\|g(x) - g(y)\|_Z \le C\|x - y\|_X^\alpha,$$

对于所有 x、$y \in X$。

误差界。现在，我们给出经验填缝的误差界，条件是$\mathcal{G}$ 具有 β 复杂度，并且理想函数 f^* 可由预训练模型 f_{pre}进行（ϵ_n、$\mathcal{G}$）填缝。

定理 *1.* 设 $\alpha \in (0,1]$ 和 $\beta > 0$。假设 f^* 可由 $f_{\text{pre}} = (g_h, g_e)$ 以 $(n^{-\frac{2\alpha\beta}{2\alpha\beta+1}}, G)$-填缝，其中 G 是某个 β- 复杂类，且 g_h 是 α- 赫尔德连续的。令 f_n 表示通过使用 $G_n = G_J$ 进行经验填缝所得到的估计回归函数，其中 G_J 如定义 *2* 所述，且 $J = \lceil n^{1/(2\alpha\beta+1)} \rceil$。在假设*1*下，我们有

$$\mathbb{E}_{Q^n}[E_Q(f_n)] \leq C n^{-\frac{2\alpha\beta}{2\alpha\beta+1}} \text{polylog}(n),$$

其中 $C > 0$ 是某个常数。

基于 定理 1，我们现在考察预训练算法的有效性。考虑一种预训练算法，它产生预训练模型 $f_{\text{pre},\, m}$，使得 f^* 可由一系列复杂度随 m 增大而降低的适配器模型类 $\mathcal{G}_m$ 进行填缝。具体而言，假设对于每个 m，预训练模型 $f_{\text{pre},\, m} = (g_{h,\, m},\ g_{e,\, m})$ 使得 f^* 可以由一个 $(\beta - \gamma_m)$-复杂类 $\mathcal{G}_m$ 以 $(\epsilon_n, \mathcal{G}_m)$-填缝，其中 γ_m 是一个有界项，且随着 m 增大而趋于零。对于这样的预训练算法，我们得到以下推论。

推论 *1.* 令 $\alpha \in (0,1]$ 和 $\beta > 0$。令 $\gamma_m < \beta$ 为一个依赖于 m 的有界序列。假设 f^* 可由 f_{pre} 按 $(n^{-\frac{2\alpha\beta}{2\alpha\beta+1}+\gamma_m}, G_m)$ 填缝，$_m = (g_{h,m}, g_{e,m})$ 对某个 $(\beta - \gamma_m)$-复杂类 G_m 成立，且 $g_{h,m}$ 是 α-赫尔德连续的。令 f_n 表示通过使用 $G_n = G_J$ 进行经验填缝得到的估计回归函数，其中 G_J 如定义 *2* 所述，且 $J = \lceil n^{1/(2\alpha(\beta-\gamma_m)+1)} \rceil$。在假设*1*下，我们有

$$\mathbb{E}_{Q^n}[E_Q(f_n)] \leq C n^{-\frac{2\alpha\beta}{2\alpha\beta+1}\left(1-\frac{\gamma_m}{\beta(\alpha(\beta-\gamma_m)+1)}\right)} \text{polylog}(n),$$

其中 $C > 0$ 为某个常数。

推论 1 表明，经验填缝可达到 $n^{-\frac{2\alpha\beta}{2\alpha\beta+1}+o(1)}$ 的速率，至多相差一个对数因子，其中 $o(1)$ 是一个有界项，并且随着 m 增大而趋于零，因为 $_{\beta\ \alpha}{}^{\gamma_m}{}_{((\beta-\gamma_m)+1)}$ 会随着 γ_m减小而减小。因此，推论 1 表明，只要 f^* 可由一系列复杂度逐渐降低的类$\mathcal{G}_m$ 填缝，随着源样本量的增加，经验填缝就能降低目标样本复杂度。这一结果为实证研究中观察到的下游任务样本复杂度缩放定律提供了部分理论支持。

4.1 示例：组合空间

现在，我们给出一个具体示例来说明定理 1的应用，并突出展示在回归函数 f^* 由平滑且稀疏的函数复合而成的情形下，填缝的优势（Schmidt-Hieber、2020；Suzuki 和 Nitanda，2021；Kohler 和 Langer，2021）。具体而言，这些研究考虑了以下情形：$f^* \in \mathcal{F}_H = \mathcal{F}_H \circ \cdots \circ \mathcal{F}_1$, 其中

$$\mathcal{F}_H \circ ... \circ \mathcal{F}_1 = \{f_H \circ ... \circ f_1 : f_i \in \mathcal{F}_i, i = 1, ..., H\}, \tag{4}$$

其中，每个 $\mathcal{F}_i$ 都是一类光滑且稀疏的函数，例如稀疏 Hölder 类（Schmidt-Hieber，2020；Kohlerand Langer，2021）或各向异性 Besov 类（Suzuki and Nitanda，2021）。为清晰起见，我们聚焦于 Schmidt-Hieber (2020) 的设定，其中 $\mathcal{F}_i$ 由 β_i-Hölder 光滑函数组成，每个函数都依赖于 t_i 个 d_i 可能变量中的一部分。我们先回顾相关的现有结果，再说明 定理 1 如何应用于这一情境。

无预训练模型时的误差界。 Schmidt-Hieber (2020) 证明，通过在稀疏 ReLU 网络类上进行经验风险最小化得到的估计量 f_n 满足以下误差界：

$$\mathbb{E}_{Q^n}[E_Q(f_n)] \leq C \max_{i=1,\ldots,H} n^{-\frac{2\alpha_i\beta_i}{2\alpha_i\beta_i+t_i}} \operatorname{polylog}(n), \tag{5}$$

对于某个常数 $C>0$，其中 $\alpha_i = \prod_{\ell=i+1}^{H}$ 最大值$\{1, \beta_\ell\}$。这里，α_i 反映了复合结构中各函数之间的（最坏情况下的）Hölder 光滑性$\mathcal{F}_H \circ \cdots \circ \mathcal{F}_i$。经典的非参数回归理论表明，类 $\mathcal{F}_i$ 的极小极大误差率为 $n^{-\frac{2\beta_i}{2\beta_i+t_i}}$。因此，公式（5）中的误差率是各分量类 $\mathcal{F}_i$中最慢的一个，并且每个误差率还会由于复合结构而进一步乘以因子 α_i。

使用预训练模型时的误差界（填缝）。现在考虑存在预训练模型 f_{pre}的情形，并且 f^*可由 $f_{\text{pre}} = (g_{h,\ m}, g_{e,\ m})$ 对某个 $1 < i_e < i_h < H$进行（$1/n$，$\mathcal{F}_{i_h} \circ \cdots \circ \mathcal{F}_{i_e}$）-填缝。类 $\mathcal{F}_{i_h} \circ \cdots \circ \mathcal{F}_{i_e}$的复杂度可以得到有效控制。

定理 2 (Schmidt-Hieber (2020))。 $\mathcal{F}_{i_h} \circ \ldots \circ \mathcal{F}_{i_e}$ 关于 $\min_{i=i_h,\ldots,i_e} \frac{\alpha_i\beta_i}{\alpha_{i_h}t_i}$-复杂，依据 定义 2。

结合 定理 1 与 定理 2，我们得到经验填缝的以下误差界：

$$\mathbb{E}_{Q^n}[E_Q(f_n)] \leq C \max_{i=i_e,\ldots,i_h} n^{-\frac{2\alpha_i\beta_i}{2\alpha_i\beta_i+t_i}} \operatorname{polylog}(n), \tag{6}$$

对于某个常数 $C>0$。比较 公式 (5) 和 公式 (6) 可以得出几点重要见解：

- 两个误差界都对 $n^{-\frac{2\alpha_i\beta_i}{2\alpha_i\beta_i+t_i}}$ 取最大值，但借助预训练模型，索引 i 的取值范围被限制在一个更窄的子集中。这表明，通过经验填缝利用预训练模型，可以实现更低的误差率。
- 公式 公式 (6) 的结果可以理解为：利用预训练模型，学习器可以绕过学习特征提取器部分（$\mathcal{F}_{i_e-1}$，……，$\mathcal{F}_1$）和头部部分（$\mathcal{F}_H$，……，$\mathcal{F}_{i_h+1}$）所涉及的复杂度。此外，预训练模型的复杂度越高（i_e 增加且 i_h减少），误差率的改善幅度就越大。这在一定程度上支持了实证研究中观察到的下游任务样本复杂度缩放定律。
- 特征提取器类$\mathcal{F}_{i_e-1}, \ldots, \mathcal{F}_1$ 对误差界的影响被完全消除了。事实上，即使将这些类替换为任意其他类，只要保持可填隙性，式 (6) 中的误差率仍保持不变。
- 对头部类$\mathcal{F}_H, \ldots, \mathcal{F}_{i_h+1}$ 的依赖仍然存在，但仅通过 Hölder 光滑性参数 α_i体现。

有关此示例的更多详情，请参见附录 B。

5 扩展到分类

在本节s se中，我们将理论结果扩展到二分类问题。

设定。 令 $X \in \mathcal{X}$ 和 $Y \in \{0,1\}$ 分别表示协变量和标签。我们不使用回归设定中的回归函数，而是考虑评分函数 $f^*(x) = \mathbb{P}_Q\{Y=1|X=x\}$。相应地，我们假设学习器可以访问一个预训练模型 f_{pre}

该评分函数 f^*由从 P中抽取的大小为 m的源样本构建而成。给定预训练模型 f_{pre} 以及从 Q中抽取的大小为 n 的目标样本，学习器的目标是构造一个分类器 $h: \mathcal{X} \to \{0,1\}$，使其在目标分布上的分类误差最小，即 $\mathbb{E}_Q[\mathbb{1}\{h(X) \neq Y\}]$。令 h_n 表示得到的估计分类器。

分类器 h 的准确性通过其超额误差来评估。令 h^* 表示使分类误差最小化的贝叶斯最优分类器，其定义为 $h^*(x) = \mathbb{1}\{f^*(x) > 1/2\}$。分类器 h 的超额误差定义为

$$E_Q(h) = \mathbb{E}_Q[\mathbb{1}\{h(X) \neq Y\}] - \mathbb{E}_Q[\mathbb{1}\{h^*(X) \neq Y\}].$$

算法。我们描述基于评分函数f^*估计的插件估计量。该过程首先估计评分函数 f^*，记作 f_n，然后将估计分类器构造为 $h_n(x) = \mathbb{1}\{f_n(x) > 1/2\}$。插件估计量的分类误差可以通过回归误差界加以控制。具体而言，我们有

$$\mathbb{E}_{Q^n}[E_Q(h_n)] \leq 2\sqrt{\mathbb{E}_{Q^n}\left[\|f^* - f_n\|^2_{L^2(Q_X)}\right]}.$$

因此，我们采用平方误差损失进行经验填缝，以估计 f_n。假设理想评分函数 f^* 可由预训练模型 $f_{\text{pre}} = (g_h, g_e)$ 以 $(\epsilon_n, \mathcal{G})$-填缝方式表示。给定一个类$\mathcal{G}_n \subseteq \mathcal{G}$，并令 $\mathcal{F}_n = \{g_h \circ g_a \circ g_e : g_a \in \mathcal{G}_n\}$，估计评分函数定义为

$$f_n = \arg\min_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Y_i)^2.$$

误差分析。我们为定理 1 建立了一个分类情形下的对应版本。

定理*3*。设 $\alpha \in (0, 1]$ 和 $\beta > 0$。假设 f^* 可由某个 β-复杂度函数类 ~~G~~ 通过 $f_{\text{pre}} = (g_h, g_e)$ 以 $(n^{-\frac{2\alpha\beta}{2\alpha\beta+1}}, G)$ 填缝，且 g_h 是 α-赫尔德连续的。令 f_n 表示通过使用 $G_n = G_J$ 进行经验填缝得到的估计评分函数，其中 G_J 如定义 *2* 所述，且 $J = \lceil n^{1/(2\alpha\beta+1)} \rceil$。令 $h_n(x) = \mathbb{1}\{f_n(x) > 1/2\}$。在假设*1*下，我们有

$$\mathbb{E}_{Q^n}[E_Q(h_n)] \leq C n^{-\frac{\alpha\beta}{2\alpha\beta+1}} \operatorname{polylog}(n),$$

其中 $C > 0$ 为某个常数。

6 证明概要

本节概述主要结果的证明，定理 1。定理 1 及其他省略结果的完整证明见补充材料。定理 1的证明依赖于以下误差界，该误差界取决于函数类 $\mathcal{F}_n$。

定理4 (Schmidt-Hieber (2020);Hayakawa 和 Suzuki (2020))。在与 定理 *1* 相同的假设下，我们有

$$\mathbb{E}_{Q^n}\left[\|f_n - f^*\|^2_{L^2(Q_X)}\right] \leq C\left(\inf_{f \in \mathcal{F}_n} \|f - f^*\|^2_{L^2(Q_X)} + \frac{\ln\left(N\left(n^{-1}, \mathcal{F}_n, \|\cdot\|_{L^\infty(Q_X)}\right)\right)}{n}\right),$$

对于某个常数 $C > 0$，其取决于σ_ξ 和 Δ_Ω，在假设*1*中。

为了应用定理 4，有必要根据类$\mathcal{F}_n$ 的 β复杂度来控制$\mathcal{G}_N$的逼近能力和覆盖数。为此，我们建立以下两个命题。

命题 1. 给定函数类 $G = \{g : Z_i \to Z_o\}$，定义 $F = \{g_h \circ g_a \circ g_e : g_a \in G\}$，其中函数 $g_h : Z_o \to \mathbb{R}$ 和 $g_e : X \to Z_i$。假设对于某个 $\alpha \in (0, 1]$，g_h 关于 Z_o 上的范数 $\|\cdot\|$ 是 α-赫尔德连续的。那么，对于 X 上的任意测度 ν 和任意 $\delta > 0$，我们有

$$N(\delta, \mathcal{F}, \|\cdot\|_{L^\infty(\nu)}) \leq N\left(\left(\frac{\delta}{C_\alpha}\right)^{1/\alpha}, \mathcal{G}, \|\cdot\|_{L^\infty}\right),$$

其中 C_α 是 g_h 的赫尔德连续性常数。

命题 2. 给定函数类 $G = \{g : Z_i \to Z_o\}$，定义 $F = \{g_h \circ g_a \circ g_e : g_a \in G\}$，其中函数 $g_h : Z_o \to \mathbb{R}$ 和 $g_e : X \to Z_i$。假设对于某个 $\alpha \in (0, 1]$，g_h 关于 Z_o 上的范数 $\|\cdot\|$ 是 α-赫尔德连续的。对于任意满足 g_a^* 可能不属于 G 的 $f^* = g_h \circ g_a^* \circ g_e$，我们有

$$\inf_{f \in \mathcal{F}} \|f - f^*\|_{L^2(Q_X)} \leq C_\alpha \inf_{g_a \in \mathcal{G}} \|g_a - g_a^*\|_{L^\infty}^\alpha,$$

其中 C_α 是 g_h的赫尔德连续性常数。

将命题 1 和命题 2 分别应用于定理 4中的第一项和第二项，并利用类$\mathcal{G}_J$的 β-复杂度，可得 $\mathbb{E}_{Q^n}\left[\|f_n - f^*\|_{L^2(Q_X)}^2\right] = O(J^{-2\alpha\beta} + J/n)$。按陈述中的方式选择 J，即可得到所需结果。

7 填缝的实证评估

我们进行了两类实验，以支持我们的理论主张。

7.1 CNN 微调

我们首先在 Office-Home 数据集的剪贴画领域上，使用适配器对 ResNet-50（2600 万个参数，He 等人(2016)）和 Wide ResNet-50-2（6800 万个参数，Zagoruyko 和 Komodakis (2016)）进行微调；此前，这些模型已在其他定义域上进行预训练（Venkateswara 等人，2017）。图 3展示了误差率与适配器深度之间的关系，其中适配器由构成多层感知机的线性层组成。由于 ResNet 与 Wide ResNet 的模型架构高度相似，结果表明，较大的模型使用单层适配器即可获得更高的性能，而较小的模型则需要更复杂的适配器。这些发现与我们的理论结果一致，即较大的预训练模型可以通过较简单的适配器进行有效适配。

7.2 将视觉能力集成到大语言模型中

随后，我们通过整合预训练视觉编码器和适配器，将视觉能力融入预训练语言模型。具体而言，我们采用一个 Llama-3 风格 Transformer 模型（1.35 亿个参数，Allal 等人 (2024)），并搭配不同规模的 SigLIP 视觉特征提取器（8600 万和 3.03 亿个参数，Tschannen (2025)），按照 Gosthipaty 等人 (2024)的训练方案对其进行轻微修改后训练，使其成为视觉语言模型。图 4 展示了

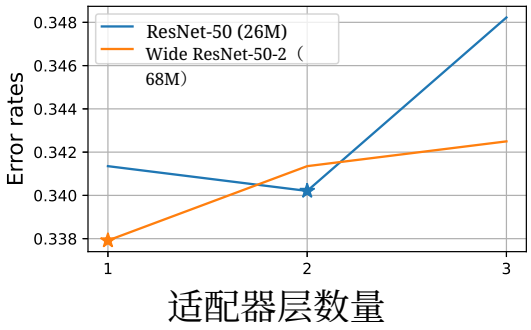

图 3：适配器深度与目标域误差率之间的关系。每个模型的最低误差率以 ⋆ 标出。

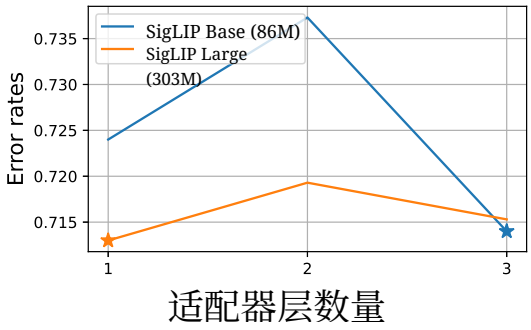

图 4：适配器深度与 MMStar 数据集上的误差率之间的关系（Chen 等人，2024）。

适配器深度与错误率之间的关系。结果再次与我们的理论发现一致。

更多实验细节请参见附录 C。

8 结论

本文从理论上分析了使用预训练模型执行下游任务的优势，并特别关注与预训练模型相关的缩放定律。这些缩放定律表明，更大的预训练模型能够显著降低下游任务所需的样本复杂度。我们通过可填隙性这一全新概念，严格解释了这一现象。实验结果进一步印证了我们的理论发现：具体而言，使用更简单的适配器即可有效地将更大的预训练模型适配到下游任务。这些见解阐明了在下游应用中采用更大预训练模型的益处。

开放问题。本文的重点是确定预训练模型中支撑实践中所观察到的缩放定律的属性，并且我们证明，复杂度递减的适配器模型类所具有的可填隙性，是这些缩放定律成立的充分条件。然而，我们并未提出用于构建具有可填隙性这一属性的预训练模型的具体学习算法。如何设计此类学习算法，仍是一个重要且引人深思的开放问题。

致谢

本研究部分获得日本学术振兴会科学研究費助成事業科研费资助，项目编号 JP23K13011（K.F.）、JP23K28146 和 JP24K20836（K.M），以及日本科学技术振兴机构 BOOST 资助，项目编号 JPMJBY24G2（R.H.）。

参考文献

Allal, L. B., Lozhkov, A., Bakouch, E., von Werra, L. 和 Wolf, T.（2024）。Smollm——速度惊人且功能强大。*Hugging Face* 博客。

Azad, B., Azad, R., Eskandari, S., Bozorgpour, A., Kazerouni, A., Rekik, I. 和 Merhof, D.（2023）。医学影像中的基础模型：全面综述与未来展望。*arXiv* 预印本 *arXiv:2310.18689*。Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C. 和 Zhou, J.（2023a）。Qwen-VL：用于理解、定位、文本阅读等任务的通用视觉语言模型。Bai, Y., Chen, F., Wang, H., Xiong, C. 和 Mei, S.（2023b）。Transformer作为统计学家：通过上下文算法选择实现可证明的上下文学习。载于神经信息处理系统进展，第36卷，第57125–57211页。Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F. 和 Vaughan, J. W.（2010）。跨域学习理论。机器学习，79(1):151–175。Ben-David, S., Blitzer, J., Crammer, K. 和 Pereira, F.（2006）。领域自适应的表征分析。神经信息处理系统进展，19。Chen, L., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Wang, J., Qiao, Y., Lin, D. 和 Zhao, F.（2024）。我们是否采用了正确的方法来评估大型视觉语言模型？载于第三十八届神经信息处理系统年度会议。Chen, M., Jiang, H., Liao, W. 和 Zhao, T.（2022）。使用深度ReLU网络在低维流形上进行非参数回归：函数逼近与统计恢复。信息与推断：*IMA*期刊，11(4):1203–1253。Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P. N. 和 Hoi, S.（2023）。InstructBLIP：通过指令微调迈向通用视觉语言模型。载于神经信息处理系统进展，第36卷，第49250–49267页。Damian, A., Lee, J. 和 Soltanolkotabi, M.（2022）。神经网络可以通过梯度下降学习表征。载于第三十五届学习理论会议论文集，第5413–5452页。PMLR。Damian, A., Nichani, E., Ge, R. 和 Lee, J. D.（2023）。平滑景观增强随机梯度下降的信号：学习单指标模型的最优样本复杂度。神经信息处理系统进展，36:752–784。Du, S. S., Hu, W., Kakade, S. M., Lee, J. D. 和 Lei, Q.（2020）。通过学习表征实现少样本学习：可证明性分析。载于国际学习表征会议。Feng, X., He, X., Wang, C., Wang, C. 和 Zhang, J.（2023）。协变量偏移下基于核的方法的统一分析。神经信息处理系统进展，36:73839–73851。Fujikawa, M., Akimoto, Y., Sakuma, J. 和 Fukuchi, K.（2025）。利用邻域信息分析的力量进行协变量偏移下的分类。载于第二十八届国际人工智能与统计会议论文集，第258卷，第226–234页。PMLR。Galbraith, N. R. 和 Kpotufe, S.（2023）。协变量偏移下的分类树剪枝。*IEEE*信息论汇刊，70(1):456–481。Gosthipaty, A. R., Wiedmann, L., Marafioti, A., Paniego, S., Noyan, M., Cuenca, P. 和 Srivastav, V.（2024）。nanoVLM：使用纯PyTorch训练视觉语言模型的最简仓库。*Hugging Face* 博客。

Guo, D., Rush, A., 和 Kim, Y.（2021）。使用差分剪枝进行参数高效迁移学习。载于 Zong, C., Xia, F., Li, W., 和 Navigli, R.（编），第*59*届计算语言学协会年会暨第*11*届自然语言处理国际联合会议论文集（第*1*卷：长篇论文），第4884–4896页，线上。计算语言学协会。Hayakawa, S. 和 Suzuki, T.（2020）。论深度神经网络学习在稀疏参数空间上的极小极大最优性及其优势。神经网络，123:343–361。He, K., Zhang, X., Ren, S., 和 Sun, J.（2016）。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。Henighan, T., Kaplan, J., Katz, M., Chen, M., Hesse, C., Jackson, J., Jun, H., Brown, T. B., Dhariwal, P., Gray, S., Hallacy, C., Mann, B., Radford, A., Ramesh, A., Ryder, N., Ziegler, D. M., Schulman, J., Amodei, D., 和 McCandlish, S.（2020）。自回归生成建模的缩放定律。Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., Laroussilhe, Q. D., Gesmundo, A., Attariyan, M., 和 Gelly, S.（2019）。面向自然语言处理的参数高效迁移学习。载于第*36*届国际机器学习会议论文集，第2790–2799页。PMLR。Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., 和 Chen, W.（2021）。LoRA：大语言模型的低秩适配。载于国际学习表征会议。Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., 和 Amodei, D.（2020）。神经语言模型的缩放定律。Kim, J., Nakamaki, T., 和 Suzuki, T.（2024）。Transformer是极小极大最优非参数上下文学习器。载于神经信息处理系统进展，第37卷，第106667–106713页。Kohler, M. 和 Langer, S.（2021）。全连接深度神经网络回归估计的收敛速率。统计年刊，49(4):2231–2249。Kpotufe, S.（2017）。Lipschitz密度比、结构化数据与数据驱动调参。载于人工智能与统计，第1320–1328页。PMLR。Kpotufe, S. 和 Martinet, G.（2021）。协变量偏移中的边缘奇异性与标签的益处。统计年刊，49(6):3299–3323。Li, J., Li, D., Savarese, S., 和 Hoi, S.（2023）。BLIP-2：利用冻结图像编码器和大语言模型进行语言-图像预训练引导。载于第*40*届国际机器学习会议论文集，第19730–19742页。PMLR。Liu, H., Li, C., Wu, Q., 和 Lee, Y. J.（2023）。视觉指令微调。载于神经信息处理系统进展，第36卷，第34892–34916页。Liu, Z., Zhu, L., Shi, B., Zhang, Z., Lou, Y., Yang, S., Xi, H., Cao, S., Gu, Y., Li, D. 等（2025）。Nvila：高效的前沿视觉语言模型。载于计算机视觉与模式识别会议论文集，第4122–4134页。Ma, C., Pathak, R., 和 Wainwright, M. J.（2023）。以最优方式应对基于再生核希尔伯特空间的非参数回归中的协变量偏移。统计年刊，51(2):738–761。

Mansour, Y., Mohri, M., 和 Rostamizadeh, A.（2008）。多源领域自适应。神经信息处理系统进展，21。Mikami, H., Fukumizu, K., Murai, S., Suzuki, S., Kikuchi, Y., Suzuki, T., Maeda, S.-i., 和 Hayashi, K.（2023）。合成到真实迁移的缩放定律：预训练的效果有多大？载于 Amini, M.-R., Canu, S., Fischer, A., Guns, T., Kralj Novak, P., 和 Tsoumakas, G.（编），数据库中的机器学习与知识发现，欧洲会议，*ECML PKDD 2022*，格勒诺布尔，法国，*2022*年*9*月*19*日至*23*日，会议论文集，第三部分，第477–492页，查姆。施普林格·自然瑞士。Nishikawa, N., Song, Y., Oko, K., Wu, D., 和 Suzuki, T.（2025）。非线性Transformer能够执行推理时特征学习。载于第四十二届国际机器学习会议。Pathak, R., Ma, C., 和 Wainwright, M.（2022）。一种用于协变量偏移的新型相似性度量及其在非参数回归中的应用。载于国际机器学习会议，第17517–17530页。PMLR。Pyzer-Knapp, E. O., Manica, M., Staar, P., Morin, L., Ruch, P., Laino, T., Smith, J. R., 和 Curioni, A.（2025）。用于材料发现的基础模型——现状与未来方向。*Npj*计算材料，11(1):61。Redko, I., Morvant, E., Habrard, A., Sebban, M., 和 Bennani, Y.（2019）。领域自适应理论进展。爱思唯尔。Redko, I., Morvant, E., Habrard, A., Sebban, M., 和 Bennani, Y.（2020）。领域自适应理论综述：学习界与理论保证。*arXiv*预印本 *arXiv:2004.11829*。Schmidt-Hieber, J.（2020）。使用带ReLU激活函数的深度神经网络进行非参数回归。统计年刊，48(4):1875–1897。Shen, J., Qu, Y., Zhang, W., 和 Yu, Y.（2018）。由Wasserstein距离引导的领域自适应表征学习。载于*AAAI*人工智能会议论文集，第32卷。Suh, N., Zhou, T.-Y., 和 Huo, X.（2022）。通过深度ReLU网络对高维球面上的函数进行逼近与非参数估计。载于第十一届国际学习表征会议。Suzuki, T. 和 Nitanda, A.（2021）。深度学习能够适应各向异性 Besov 空间中模型光滑性的内在维数。载于神经信息处理系统进展，第34卷，第3609–3621页。Curran Associates, Inc. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M. F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B. 等（2025）。Siglip 2：具备更强语义理解、定位能力和密集特征的多语言视觉-语言编码器。*arXiv*预印本 *arXiv:2502.14786*。Venkateswara, H., Eusebio, J., Chakraborty, S., 和 Panchanathan, S.（2017）。用于无监督领域自适应的深度哈希网络。载于*2017*年*IEEE*计算机视觉与模式识别会议（*CVPR*），第5385–5394页。Wang, J., Wang, K., Yu, Y., Lu, Y., Xiao, W., Sun, Z., Liu, F., Zou, Z., Gao, Y., Yang, L. 等（2025）。用于合成医学图像生成及临床应用的自我改进生成式基础模型。自然医学，31(2):609–617。

Wen, Y.、Chen, L.、Deng, Y. 和 Zhou, C.（2021）。重新思考医学影像预训练。《视觉传达与图像表示期刊》，78：103145。
Wu, Y. 和 He, K.（2018）。组归一化。收录于《欧洲计算机视觉会议》（*ECCV*），第 3–19 页。

Xia, J.、Zhu, Y.、Du, Y. 和 Li, S. Z.（2022）。化学预训练模型系统综述。*arXiv* 预印本 *arXiv:2210.16484*。
Zagoruyko, S. 和 Komodakis, N.（2016）。宽残差网络。收录于 *2016*年英国机器视觉会议。英国机器视觉协会。

A 缺失的证明

A.1 定理 定理 1

如第6节所示，定理 1 的证明采用了与Schmidt-Hieber（2020）以及Hayakawa and Suzuki （2020）类似的方法，利用了定理 4。为清晰起见，我们在此给出定理 4的更严谨表述：

定理 *A.1*. 设 $\sigma > 0$ 和 $F > 0$ 为常数。设 X、Y 和 ξ 为定义在概率空间 $(Z, \mathfrak{Z}, \nu)$ 上的随机变量，并设 ν_X 为 X 的边缘分布。给定一个可测函数 $f^*: X \to \mathbb{R}$，满足 $\|f^*\|_{L^\infty(\nu_X)} \leq F$，假设这些变量遵循回归模型 $Y = f^*(X) + \xi$。此外，假设 ξ 与 X 独立，且为具有方差代理参数 σ^2 的次高斯随机变量，即对于所有 $\lambda \in \mathbb{R}$，$\mathbb{E}_\nu\left[e^{\lambda\xi}\right] \leq e^{\lambda^2\sigma^2/2}$。设 $\mathcal{F}$ 为可测函数类 $f: X \to \mathbb{R}$，满足 $\sup_{f\in\mathcal{F}}\|f\|_{L^\infty(\nu_X)} \leq F$。令 $(X_1, Y_1), \ldots, (X_n, Y_n)$ 为 (X, Y) 的 n 独立同分布副本，并定义 $f_n \in \mathcal{F}$ 为达到以下下确界的函数：

$$\inf_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)^2.$$

那么，对于任意满足 $\ln(N(\delta, \mathcal{F}, \|\cdot\|_{L^\infty(\nu_X)})) \geq 1$ 的 $\delta > 0$，我们有

$$\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|^2_{L^2(\nu_X)}\right] \leq C\left(\inf_{f\in\mathcal{F}}\|f - f^*\|^2_{L^2(\nu_X)} + \frac{(F^2+\sigma^2)\ln\left(N\left(\delta, \mathcal{F}, \|\cdot\|_{L^\infty(\nu)}\right)\right)}{n} + (F+\sigma)\delta\right),$$

对某个普适常数 $C > 0$。

对定理 A.1的证明见附录A.2。定理 A.1 刻画了最小二乘估计量 f_n 在类 $\mathcal{F}$ 上的误差界，该界取决于 f^* 被 $\mathcal{F}$ 逼近所产生的逼近误差以及 $\mathcal{F}$的覆盖数。为控制这两项，我们利用 命题 1 和 命题 2。回顾 命题 1 和命题 2的表述：

命题 A.1. 给定函数类 $G = \{g: Z_i \to Z_o\}$，定义 $F = \{g_h \circ g_a \circ g_e: g_a \in G\}$，其中函数 $g_h: Z_o \to \mathbb{R}$ 和 $g_e: X \to Z_i$。假设 g_h 关于 Z_o 上的范数 $\|\cdot\|$ 是 α-赫尔德连续的，其中 $\alpha \in (0, 1]$。那么，对于 X 上的任意测度 ν 以及任意 $\delta > 0$，有

$$N(\delta, \mathcal{F}, \|\cdot\|_{L^\infty(\nu)}) \leq N\left(\left(\frac{\delta}{C_\alpha}\right)^{1/\alpha}, \mathcal{G}, \|\cdot\|_{L^\infty}\right),$$

其中 C_α 是 g_h的赫尔德连续性常数。

命题 A.2。 给定函数类 $G = \{g : Z_i \to Z_o\}$，定义 $F = \{g_h \circ g_a \circ g_e : g_a \in G\}$，其中 $g_h : Z_o \to \mathbb{R}$ 和 $g_e : X \to Z_i$为某些函数。假设 g_h 关于 Z_o 上的某个范数 $\|\cdot\|$ 是 α-赫尔德连续的，其中 $\alpha \in (0, 1]$。对于任意 $f^* = g_h \circ g_a^* \circ g_e$，即使 g_a^* 可能不属于 G，我们有

$$\inf_{f\in\mathcal{F}} \|f - f^*\|_{L^2(Q_X)} \le C_\alpha \inf_{g_a\in\mathcal{G}} \|g_a - g_a^*\|_{L^\infty}^\alpha,$$

其中 C_α 是 g_h的赫尔德连续性常数。

命题 A.1和 命题 A.2 的证明见附录 A.3 和 附录 A.4。基于定理 A.1、命题 A.1和命题 A.2，我们可以证明定理 1:

定理 *1* 的证明定理 *1*。 我们应用定理 A.1 ，取$\mathcal{F} = \mathcal{F}_n$，其中$\mathcal{F}_n = \{g_h \circ g_a \circ g_e : g_a \in \mathcal{G}_n\}$ 对应于由一列类构成的$\mathcal{G}_n$ ，它们属于定义 2 并对应一个 β-复杂类$\mathcal{G}$。由 $\delta = 1/n$，我们有

$$\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|_{L^2(\nu_X)}^2\right] \le C\left(\inf_{f\in\mathcal{F}_n} \|f - f^*\|_{L^2(\nu_X)}^2 + \frac{(F^2+\sigma^2)\ln\left(N\left(n^{-1}, \mathcal{F}_n, \|\cdot\|_{L^\infty(\nu)}\right)\right)}{n} + \frac{(F+\sigma)}{n}\right).$$

由于 $\mathcal{G}_n$ 是由定义 2中对应于 β-复杂类 $\mathcal{G}$的一系列类，根据命题 A.1和A.2，我们有

$$N(n^{-1}, \mathcal{F}_n, \|\cdot\|_{L^\infty(\nu)}) \le N\left(\left(\frac{n^{-1}}{C_\alpha}\right)^{1/\alpha}, \mathcal{G}_n, \|\cdot\|_{L^\infty}\right) \le cJ\text{polylog}(J, (C_\alpha n)^{-1/\alpha}),$$

and

$$\inf_{f\in\mathcal{F}_n} \|f - f^*\|_{L^2(\nu_X)}^2 \le C_\alpha \inf_{g_a\in\mathcal{G}_n} \|g_a - g_a^*\|_{L^\infty}^\alpha \le c' J^{-2\alpha\beta},$$

其中 J 在陈述中定义。因此，我们有

$$\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|_{L^2(\nu_X)}^2\right] \le C\left(J^{-2\alpha\beta} + \frac{J\text{polylog}(J, (C_\alpha n)^{-1/\alpha})}{n} + \frac{1}{n}\right). \tag{7}$$

公式 (7) 在 J 取陈述中定义的值时达到最小值，由此得到所需结论。 □

A.2 定理 A.1 的证明

为清晰简洁起见，我们先引入证明过程中将一直使用的记号。设 $\xi_1, ..., \xi_n$ 为 ξ 的独立同分布副本。对于任意函数 $f : \mathcal{X} \to \mathbb{R}$，将 f 和噪声 ξ_i 的经验 L^2 范数定义为

$$\|f\|_n^2 := \frac{1}{n}\sum_{i=1}^n f^2(X_i), \quad \|\xi\|_n^2 := \frac{1}{n}\sum_{i=1}^n \xi_i^2,$$

以及与噪声变量的经验内积定义为

$$\langle f, \xi \rangle_n := \frac{1}{n} \sum_{i=1}^{n} \xi_i f(X_i).$$

为方便记号，我们用 $\mathbb{V}_\nu[X]$ 表示定义在概率空间 $(\mathcal{Z}\ \mathfrak{Z}, \nu)$ 上的随机变量 X 的方差。

证明分为两步。第一步，我们将 f_n 的期望误差界定为其经验误差。第二步，我们给出 f_n的经验误差界。具体而言，我们证明以下两个引理。

引理 1. 在与 定理 *A.1*相同的假设下，我们有

$$\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|^2_{L^2(\nu_X)}\right] \leq \inf_{\zeta>0}\Bigg((1+\zeta)\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|_n^2\right] + \left(2 + 3\zeta + \frac{1}{2\zeta}\right)\frac{512F^2\ln(N_\delta)}{n} + \left(1 + \zeta + \frac{1}{4\zeta}\right)\frac{9F^2}{n} + 4(2+3\zeta)F\delta\Bigg).$$

其中 $N_\delta = N(\delta, F\ \|\cdot\|_{L^\infty(\nu_X)})$。

引理 2. 在与 定理 *A.1*相同的假设下，我们有

$$\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|_n^2\right] \leq \inf_{\zeta>0}\Bigg((1+\zeta)\inf_{f\in\mathcal{F}}\|f - f^*\|^2_{L^2(\nu_X)} + \left(1 + \frac{\zeta}{2} + \frac{1}{2\zeta}\right)\frac{8\sigma^2\ln(3N_\delta)}{n} + \Big((1+\zeta)\sqrt{2\pi}\sigma + 4F\zeta\Big)\delta\Bigg).$$

其中 $N_\delta = N(\delta, F\ \|\cdot\|_{L^\infty(\nu_X)})$。

结合引理 1 与引理 2 即可立即得到定理 A.1。

定理 *A.1* 的证明。 结合引理 1 与引理 2，并将这些引理中的 ζ 取为通用常数，我们得到

$$\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|^2_{L^2(\nu_X)}\right] \leq C\left(\inf_{f\in\mathcal{F}}\|f - f^*\|^2_{L^2(\nu_X)} + \frac{(\sigma^2 + F^2)\ln(N_\delta)}{n} + (\sigma + F)\delta\right),$$

对某个普适常数 $C > 0$。 □

在接下来的小节中，我们给出引理 1 和引理 2的证明。这两个引理中的界都由类$\mathcal{F}$的覆盖数决定。为方便记号，令 f_k 表示$\mathcal{F}_n$ 关于范数 $\|\cdot\|_{L^\infty(\nu_X)}$的 δ-覆盖中球的球心。定义 $g_k = f_k - f^*$，并令 $K = \arg\min_k \|f_n - f_k\|_{L^\infty(\nu_X)}$。

A.2.1 引理 Lem.1

引理 *1* 的证明。令 $\tau > 0$ 为一个稍后确定的常数。我们有

$$
\begin{aligned}
&\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|^2_{L^2(\nu_X)}\right] \\
=&\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|^2_n\right] + \mathbb{E}_{\nu^n}\left[\|f_n - f^*\|^2_{L^2(\nu_X)} - \|f_n - f^*\|^2_n\right] \\
\leq&\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|^2_n\right] + 8F\delta + \mathbb{E}_{\nu^n}\left[\|g_K\|^2_{L^2(\nu_X)} - \|g_K\|^2_n\right] \\
\leq&\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|^2_n\right] + 8F\delta + \mathbb{E}_{\nu^n}\left[\max\left\{\tau^2, \|g_K\|^2_{L^2(\nu_X)}\right\}\right]^{1/2} \mathbb{E}_{\nu^n}\left[\frac{\left(\|g_K\|^2_{L^2(\nu_X)} - \|g_K\|^2_n\right)^2}{\max\left\{\tau^2, \|g_K\|^2_{L^2(\nu_X)}\right\}}\right]^{1/2}, \qquad (8)
\end{aligned}
$$

其中，第三个不等式由柯西-施瓦茨不等式得出，而在第二个不等式中我们使用了以下结论：

$$
\begin{aligned}
&\|f_n - f^*\|^2_{L^2(\nu_X)} - \|f_n - f^*\|^2_n \\
=&\|g_K\|^2_{L^2(\nu_X)} + \mathbb{E}_\nu[(f_K(X) - f_n(X))(f_K(X) + f_n(X) - 2f^*(X))] \\
&\qquad - \|g_K\|^2_n - \frac{1}{n}\sum_{i=1}^n (f_K(X_i) - f_n(X_i))(f_K(X_i) + f_n(X_i) - 2f^*(X_i)) \\
\leq&\|g_K\|^2 - \|g_K\|^2_n + 8F\delta,
\end{aligned}
$$

几乎处处成立。

对于公式 (8)中的最后一项，我们有

$$
\begin{aligned}
&\mathbb{E}_{\nu^n}\left[\max\left\{\tau^2, \|g_K\|^2_{L^2(\nu_X)}\right\}\right] \\
\leq&\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|^2_{L^2(\nu_X)}\right] + \mathbb{E}_{\nu^n}\left[\|g_K\|^2_{L^2(\nu_X)} - \|f_n - f^*\|^2_{L^2(\nu_X)}\right] + \tau^2 \\
=&\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|^2_{L^2(\nu_X)}\right] + \mathbb{E}_{\nu^n}[\mathbb{E}_\nu[(f_K(X) - f_n(X))(f_K + f_n - 2f^*)]] + \tau^2 \\
\leq&\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|^2_{L^2(\nu_X)}\right] + 4F\delta + \tau^2. \qquad (9)
\end{aligned}
$$

此外，

$$
\mathbb{E}_{\nu^n}\left[\frac{\left(\|g_K\|^2_{L^2(\nu_X)} - \|g_K\|^2_n\right)^2}{\max\left\{\tau^2, \|g_K\|^2_{L^2(\nu_X)}\right\}}\right] = \mathbb{E}_{\nu^n}\left[\left(\frac{1}{n}\sum_{i=1}^n \frac{\left(\|g_K\|^2_{L^2(\nu_X)} - g_K^2(X_i)\right)}{\max\left\{\tau^2, \|g_K\|_{L^2(\nu_X)}\right\}}\right)^2\right].
$$

定义 $Z_{k,\ i} = \frac{(g_k^2\ (X_i) - \|g_k\|^2_{L^2(\nu_X)})}{\text{最大值}\left\{\tau^2, \|g_k\|_{L^2(\nu_X)}\right\}}$。那么 $Z_{k,\ i}$ 是零均值的，并且

$$
\begin{aligned}
\mathbb{E}_\nu\left[Z^2_{k,i}\right] =&\mathbb{E}_\nu\left[\frac{(g_k(X_i) - g_k(X_i'))^2(g_k(X_i) + g_k(X_i'))^2}{\tau^2 \vee \|g_k\|^2_{L^2(\nu_X)}}\right] \\
\leq&8\|g_k\|^2_{L^\infty(\nu_X)} \frac{\mathbb{V}_\nu[g_k(X_i)]}{\max\left\{\tau^2, \|g_k\|^2_{L^2(\nu_X)}\right\}} \leq 8\|g_k\|^2_{L^\infty(\nu_X)} \leq 32F^2,
\end{aligned}
$$

其中 X_i' 是 X_i的独立副本，并且$|Z_{k,\ i}| \le \|g_k\|^2_{L^\infty(\nu_X)}/\tau \le 4F^2/\tau$ 几乎处处成立。根据有界随机变量的伯恩斯坦条件，对于所有 k和 $\lambda \in (0,\ 3\tau/4F^2)$,

$$\mathbb{E}_{\nu^n}\left[\exp\left(\lambda\sum_{i=1}^n Z_{k,i}\right)\right] \le \exp\left(\frac{32n\lambda^2F^2}{2(1-4\lambda F^2/3\tau)}\right).$$

应用 引理 3，对于 $\gamma > 0$,

$$\mathbb{E}_{\nu^n}\left[\max_k\left(\sum_{i=1}^n Z_{k,i}\right)^2\right] \le 512\left(nF^2\ln(N_\delta/\gamma) + \gamma\max\left\{nF^2, \frac{F^4}{9\tau^2}\ln(N_\delta/\gamma)\right\} + \frac{\gamma F^4}{9\tau^2}\right).$$

因此,

$$\begin{aligned}\mathbb{E}_{\nu^n}\left[\frac{\left(\|g_K\|^2_{L^2(\nu_X)} - \|g_K\|^2_n\right)^2}{\max\left\{\tau^2, \|g_K\|^2_{L^2(\nu_X)}\right\}}\right] &\le \frac{1}{n^2}\mathbb{E}_{\nu^n}\left[\max_k\left(\sum_{i=1}^n Z_{k,i}\right)^2\right] \\ &\le \frac{512F^2}{n}\left(\ln(N_\delta/\gamma) + \gamma\max\left\{1, \frac{F^2}{9\tau^2 n}\ln(N_\delta/\gamma)\right\} + \frac{\gamma F^2}{9\tau^2 n}\right). \quad (10)\end{aligned}$$

将 式（9）和 式（10）代入 式（8），可得

$$\begin{aligned}\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|^2_{L^2(\nu_X)}\right] \le\ & \mathbb{E}_{\nu^n}\left[\|f_n - f^*\|^2_n\right] + 8F\delta \\ & + \left(\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|^2_{L^2(\nu_X)}\right] + 4F\delta + \tau^2\right)^{1/2} \\ & \cdot\left(\frac{512F^2}{n}\left(\ln(N_\delta/\gamma) + \gamma\max\left\{1, \frac{F^2}{9\tau^2 n}\ln(N_\delta/\gamma)\right\} + \frac{\gamma F^2}{9\tau^2 n}\right)\right)^{1/2}. \quad (11)\end{aligned}$$

现在，对于 a，b，$c > 0$，若 $x^2 \le a + b\sqrt{x^2 + c^2}$，则 $x^2 \le a$或

$$x^4 - 2x^2\left(a + \frac{b^2}{2}\right) + \left(a^2 - b^2c^2\right) \le 0,$$

这意味着

$$x^2 \le a + \frac{b^2}{2} + b\sqrt{a + \frac{b^2}{4} + c^2}. \quad (12)$$

因此，由式 (11)可得

$$\begin{aligned}\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|^2_{L^2(\nu_X)}\right] \le\ & \mathbb{E}_{\nu^n}\left[\|f_n - f^*\|^2_n\right] + 8F\delta \\ & + \frac{512F^2}{n}\left(\ln(N_\delta/\gamma) + \gamma\max\left\{1, \frac{F^2}{9\tau^2 n}\ln(N_\delta/\gamma)\right\} + \frac{\gamma F^2}{9\tau^2 n}\right) \\ & + \left(\frac{512F^2}{n}\left(\ln(N_\delta/\gamma) + \gamma\max\left\{1, \frac{F^2}{9\tau^2 n}\ln(N_\delta/\gamma)\right\} + \frac{\gamma F^2}{9\tau^2 n}\right)\right)^{1/2}\end{aligned}$$

$$\cdot\left(\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|_n^2\right] + \tau^2 + \frac{512F^2}{n}\left(\ln(N_\delta/\gamma) + \gamma\max\left\{1, \frac{F^2}{9\tau^2 n}\ln(N_\delta/\gamma)\right\} + \frac{\gamma F^2}{9\tau^2 n}\right) + 12F\delta\right)^{1/2}.$$

根据算术-几何平均不等式，对于任意 $\zeta > 0$，有

$$\begin{aligned}\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|_{L^2(\nu_X)}^2\right] &\le (1+\zeta)\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|_n^2\right] + 4(2+3\zeta)F\delta \\ &+ \left(1 + \zeta + \frac{1}{4\zeta}\right)\frac{512F^2}{n}\left(\ln(N_\delta/\gamma) + \gamma\max\left\{1, \frac{F^2}{9\tau^2 n}\ln(N_\delta/\gamma)\right\} + \frac{\gamma F^2}{9\tau^2 n}\right) + \zeta\tau^2.\end{aligned}$$

现在令

$$\tau^2 = \frac{512F^2\ln(N_\delta/\gamma)}{n}.$$

于是，

$$\begin{aligned}\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|_{L^2(\nu_X)}^2\right] &\le (1+\zeta)\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|_n^2\right] + 4(2+3\zeta)F\delta \\ &+ \left(1 + \zeta + \frac{1}{4\zeta}\right)\frac{512F^2}{n}\left(\ln(N_\delta/\gamma) + \gamma\left(1 + \frac{9}{512\ln(N_\delta/\gamma)}\right)\right) + \frac{512\zeta F^2\ln(N_\delta/\gamma)}{n}.\end{aligned}$$

化简后可得

$$\begin{aligned}\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|_{L^2(\nu_X)}^2\right] &\le (1+\zeta)\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|_n^2\right] + 4(2+3\zeta)F\delta \\ &+ \left(1 + 2\zeta + \frac{1}{4\zeta}\right)\frac{512F^2\ln(N_\delta/\gamma)}{n} + \left(1 + \zeta + \frac{1}{4\zeta}\right)\frac{512F^2\gamma}{n}\left(1 + \frac{9}{512\ln(N_\delta/\gamma)}\right).\end{aligned}$$

令 $\gamma = \ln(N_\delta)$，则有

$$\begin{aligned}\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|_{L^2(\nu_X)}^2\right] &\le (1+\zeta)\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|_n^2\right] + 4(2+3\zeta)F\delta \\ &+ \left(2 + 3\zeta + \frac{1}{2\zeta}\right)\frac{512F^2\ln(N_\delta)}{n} + \left(1 + \zeta + \frac{1}{4\zeta}\right)\frac{9F^2}{n}.\end{aligned}$$

$f \in \mathcal{F}$ 和 ζ 的任意性给出了所需结果。 □

A.2.2 引理 2 的证明引理 2

证明引理 2。 从回归模型 $Y = f^*(X) + \xi$出发，我们得到

$$\frac{1}{n}\sum_{i=1}^n (f_n(X_i) - Y_i)^2 = \|f_n - f^*\|_n^2 - 2\langle f_n - f^*, \xi\rangle_n + \|\xi\|_n^2.$$

由于 $\|\xi\|_n^2$ 与 f_n相互独立，对于任意与样本独立的 $f \in \mathcal{F}_n$，可得

$$\|f_n - f^*\|_n^2 - 2\langle f_n - f^*, \xi\rangle_n \le \|f - f^*\|_n^2 - 2\langle f - f^*, \xi\rangle_n.$$

因此，

$$\begin{aligned}\mathbb{E}_{\nu^n}\left[\|f_n-f^*\|_n^2\right]=&\mathbb{E}_{\nu^n}\left[\|f_n-f^*\|_n^2-2\langle f_n-f^*,\xi\rangle_n+2\langle f_n-f^*,\xi\rangle_n\right]\\ \le&\|f-f^*\|_{L^2(\nu_X)}^2+2\mathbb{E}_{\nu^n}[\langle f_n-f_K,\xi\rangle_n]+2\mathbb{E}_{\nu^n}[\langle g_K,\xi\rangle_n]\\ \le&\|f-f^*\|_{L^2(\nu_X)}^2+2\delta\mathbb{E}_{\nu}[|\xi|]+2\mathbb{E}_{\nu^n}\left[\|g_K\|_n^2\right]^{1/2}\mathbb{E}_{\nu^n}\left[\left|\left\langle\frac{g_K}{\|g_K\|_n},\xi\right\rangle_n\right|^2\right]^{1/2}.\end{aligned}\tag{13}$$

对于公式 (13)中的第二项， ξ 的次高斯性结合马尔可夫不等式可得

$$\begin{aligned}\mathbb{E}_{\nu}[|\xi|]=&\int_0^\infty(\mathbb{P}_\nu\{\xi>t\}+\mathbb{P}_\nu\{-\xi>t\})dt\\ \le&2\int_0^\infty\inf_{\lambda>0}\exp\left(\frac{\lambda^2\sigma^2}{2}-\lambda t\right)dt\\ =&2\int_0^\infty\exp\left(-\frac{t^2}{2\sigma^2}\right)dt\\ =&\sqrt{2\pi}\sigma.\end{aligned}\tag{14}$$

对于公式 (13)中的最后一项，根据三角不等式，我们有

$$\begin{aligned}\|g_K\|_n^2=&\|f_n-f^*\|_n^2+\|g_K\|_n^2-\|f_n-f^*\|_n^2\\ \le&\|f_n-f_K\|_n^2+\frac{1}{n}\sum_{i=1}^n(f_K(X_i)-f_n(X_i))(f_K(X_i)+f_n(X_i)-2f^*(X_i))\\ \le&\|f_n-f^*\|_n^2+4F\delta,\end{aligned}\tag{15}$$

几乎处处。
在 $X_1,\ldots,X_n$的条件下， $Z_k=\left\langle\frac{g_k}{\|g_k\|_n},\xi\right\rangle_n$ 为零均值且次高斯，方差代理为σ^2/n。因此，

$$\begin{aligned}\mathbb{E}_{\nu^n}\left[\exp\left(\lambda Z_k^2\right)\right]=&1+\int_1^\infty\mathbb{P}_{\nu^n}\left\{e^{\lambda Z_k^2}>t\right\}dt\\ =&1+\int_1^\infty\inf_{\kappa>0}\mathbb{P}_{\nu^n}\left\{e^{\kappa\sqrt{\lambda}|Z_k|}>e^{\kappa\ln^{1/2}(t)}\right\}dt\\ \le&1+2\int_1^\infty\inf_{\kappa>0}\exp\left(\frac{\lambda\sigma^2\kappa^2}{2n}-\kappa\ln^{1/2}(t)\right)dt\\ =&1+2\int_1^\infty\exp\left(-\frac{n\ln(t)}{2\lambda\sigma^2}\right)dt\\ =&1+2\int_1^\infty t^{-\frac{n}{2\lambda\sigma^2}}dt\\ =&1+\frac{2}{\frac{n}{2\lambda\sigma^2}-1},\end{aligned}$$

只要 $2\lambda\sigma^2 < n$。因此，

$$\begin{aligned}\mathbb{E}_{\nu^n}\left[\exp\left(\lambda \max_k Z_k^2\right)\right] \leq& \sum_k \mathbb{E}_{\nu^n}\left[\exp\left(\lambda Z_k^2\right)\right] \\ \leq& N_\delta\left(1+\frac{2}{\frac{n}{2\lambda\sigma^2}-1}\right).\end{aligned}$$

根据詹森不等式，可得

$$\mathbb{E}_{\nu^n}\left[\lambda \max_k Z_k^2\right] \leq \ln\left(\mathbb{E}_{\nu^n}\left[\exp\left(\lambda \max_k Z_k^2\right)\right]\right) \leq \ln(N_\delta) + \ln\left(1+\frac{2}{\frac{n}{2\lambda\sigma^2}-1}\right).$$

令 $\lambda = \frac{n}{4\sigma^2}$，可得

$$\mathbb{E}_{\nu^n}\left[\max_k Z_k^2\right] \leq \frac{4\sigma^2}{n}\ln(3N_\delta). \tag{16}$$

代入公式 (14)、公式 (15)和公式 (16)到公式 (13)中，得到

$$\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|_n^2\right] \leq \|f - f^*\|_{L^2(Q_X)}^2 + \sqrt{2\pi}\sigma\delta + 4\sqrt{\left(\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|_n^2\right] + 4F\delta\right)\frac{\sigma^2}{n}\ln(3N_\delta)}. \tag{17}$$

将公式 (12)应用于公式 (17)可得

$$\begin{aligned}\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|_n^2\right] \leq& \|f - f^*\|_{L^2(\nu_X)}^2 + \sqrt{2\pi}\sigma\delta \\ &+ \frac{8\sigma^2\ln(3N_\delta)}{n} + \left(\frac{16\sigma^2\ln(3N_\delta)}{n}\right)^{1/2}\left(\|f - f^*\|_{L^2(\nu_X)}^2 + \frac{4\sigma^2\ln(3N_\delta)}{n} + \left(\sqrt{2\pi}\sigma + 4F\right)\delta\right)^{1/2}.\end{aligned}$$

根据算术-几何平均不等式，对于任意 $\zeta > 0$，有

$$\mathbb{E}_{\nu^n}\left[\|f_n - f^*\|_n^2\right] \leq (1+\zeta)\|f - f^*\|_{L^2(\nu_X)}^2 + \left((1+\zeta)\sqrt{2\pi}\sigma + 4F\zeta\right)\delta + \left(1 + \frac{\zeta}{2} + \frac{1}{2\zeta}\right)\frac{8\sigma^2\ln(3N_\delta)}{n}.$$

ζ 的任意性给出了所需结果。 □

A.3 命题 A.1的证明

证明命题 *A.1*。设 $f = g_h \circ g_a \circ g_e \in \mathcal{F}$ 与 $f' = g_h \circ g_a' \circ g_e \in \mathcal{F}$。于是，我们有

$$\begin{aligned}\|f - f'\|_{L^\infty(\nu)} =& \|g_h \circ g_a \circ g_e - g_h \circ g_a' \circ g_e\|_{L^\infty(\nu)} \\ =& \sup_{z \in g_e \circ \mathcal{X}} |g_h(g_a(z)) - g_h(g_a'(z))| \\ \leq& C_\alpha \sup_{z \in \mathcal{Z}_i} \|g_a(z) - g_a'(z)\|^\alpha = C_\alpha \|g_a - g_a'\|_{L^\infty}^\alpha,\end{aligned}$$

其中 $g_e \circ \mathcal{X} = \{g_e(x) : x \in \mathcal{X}\} \subseteq \mathcal{Z}_i$。设 $g_1, \ldots\ldots g_N$ 为$\mathcal{G}$关于 $\|\cdot\|_{L^\infty}$ 的一个 $\left(\frac{\delta}{C_\alpha}\right)^{1/\alpha}$-网。则对任意 $f = g_h \circ g_a' \circ g_e \in \mathcal{F}$，我们有

$$\min_{i \in \{1,\ldots,N\}} \|f - g_h \circ g_i \circ g_e\|_{L^\infty(\nu)} \leq C_\alpha \min_{i \in \{1,\ldots,N\}} \|g_a - g_i\|_{L^\infty}^\alpha \leq \delta.$$

因此，$\{g_h \circ g_i \circ g_e : i \in \{1, \ldots, N\}\}$ 是相对于 $\|\cdot\|_{L^\infty(\nu)}$的 δ-网，适用于$\mathcal{F}$。因此， N大于或等于$\mathcal{F}$在 $\|\cdot\|_{L^\infty(\nu)}$-范数下的覆盖数。 □

A.4 命题 A.2 的证明

证明 命题 *A.2*。 设 $f = g_h \circ g_a \circ g_e \in \mathcal{F}$ 和 $f^* = g_h \circ g_a^* \circ g_e$。 那么，我们有

$$\begin{aligned}\|f - f^*\|_{L^2(Q_X)}^2 =& \int ((g_h \circ g_a \circ g_e)(x) - (g_h \circ g_a^* \circ g_e)(x))^2 Q_X(dx) \\ \leq& C_\alpha^2 \int (g_a(g_e(x)) - g_a^*(g_e(x)))^{2\alpha} Q_X(dx) \\ \leq& C_\alpha^2 \int \|g_a - g_a^*\|_{L^\infty}^{2\alpha} Q_X(dx) = C_\alpha^2 \|g_a - g_a^*\|_{L^\infty}^{2\alpha}.\end{aligned}$$

对两边开平方，即得所需结论。 □

A.5 辅助引理

A.5.1 有限平方和的最大不等式

这里给出一个最大不等式，它可用于控制有限个随机变量的上确界；此类问题出现在覆盖数和经验过程的分析中。下面的引理在适当的矩母函数条件下，为有限个独立随机变量的平方值的最大值的期望提供上界。

引理 3. 设 $Z_1, \ldots, Z_N$ 是概率空间 $(\mathcal{Z}, \nu)$ 上的独立随机变量，并且 Z_i 是次伽马随机变量，其方差代理参数为 σ^2、尺度因子为 $F > 0$，即对所有 $\lambda \in (0, 1/F)$，$\mathbb{E}_\nu[e^{\lambda |Z_k|}] \leq \exp\left(\frac{\lambda^2 \sigma^2}{2\ 1-F\lambda()}\right)$。那么，

$$\mathbb{E}_\nu\left[\max_{i \in \{1,\ldots,N\}} Z_i^2\right] \leq 16 \inf_{\gamma > 0} \left(\sigma^2 \ln(N/\gamma) + \gamma \max\left\{\sigma^2, 2F^2 \ln(N/\gamma)\right\} + 2\gamma F^2\right).$$

证明 引理 *3*。期望可以用概率表示为

$$\mathbb{E}_\nu\left[\max_{k \in \{1,\ldots,N\}} Z_k^2\right] = \int_0^\infty \mathbb{P}_\nu\left[\max_{k \in \{1,\ldots,N\}} Z_k^2 > t\right] dt.$$

对于任意 $z > 0$，我们有

$$\begin{aligned}\int_0^\infty \mathbb{P}_\nu\left[\max_{k \in \{1,\ldots,N\}} Z_k^2 > t\right] dt \leq& z^2 + \int_{z^2}^\infty \mathbb{P}_\nu\left[\max_{k \in \{1,\ldots,N\}} |Z_k| > \sqrt{t}\right] dt \\ =& z^2 + \int_{z^2}^\infty \mathbb{P}_\nu\left[\max_{k \in \{1,\ldots,N\}} \exp(\lambda |Z_k|) > \exp\left(\lambda\sqrt{t}\right)\right] dt \\ \leq& z^2 + \sum_{i=1}^N \int_{z^2}^\infty \mathbb{P}_\nu\left[\exp(\lambda |Z_k|) > \exp\left(\lambda\sqrt{t}\right)\right] dt \\ \leq& z^2 + \sum_{i=1}^N \int_{z^2}^\infty \inf_{\lambda \in (0,1/F)} \mathbb{E}_\nu[\exp(\lambda |Z_k|)] e^{-\lambda\sqrt{t}} dt \\ \leq& z^2 + N \int_{z^2}^\infty \inf_{\lambda \in (0,1/F)} \exp\left(\frac{\lambda^2 \sigma^2}{2(1 - F\lambda)} - \lambda\sqrt{t}\right) dt \\ \leq& z^2 + N \int_{z^2}^\infty \exp\left(-\frac{t}{2(\sigma^2 + F\sqrt{t})}\right) dt,\end{aligned}$$

其中，第四个不等式使用了马尔可夫不等式，最后一个不等式则由伯恩斯坦不等式的标准证明得出。对于积分部分，我们有

$$\begin{aligned}\int_{z^2}^{\infty}\exp\left(-\frac{t}{2(\sigma^2+F\sqrt{t})}\right)dt &\leq \max\left\{\int_{z^2}^{\infty}\exp\left(-\frac{t}{4\sigma^2}\right)dt, \int_{z^2}^{\infty}\exp\left(-\frac{\sqrt{t}}{4F}\right)dt\right\}\\ &= \max\left\{4\sigma^2 e^{-z^2/4\sigma^2}, \left(8Fz+32F^2\right)e^{-z/4F}\right\}.\end{aligned}$$

令 $z = 4\max\left\{\sigma\ln^{1/2}(N/\gamma)\ F\ln(N/\gamma)\right\}$，则有

$$\begin{aligned}&\mathbb{E}_{\nu}\left[\max_{i\in\{1,\ldots,N\}} Z_i^2\right]\\ \leq& 16\ln(N/\gamma)\max\left\{\sigma^2, F^2\ln(N/\gamma)\right\} + \gamma\max\left\{4\sigma^2, 32F(\sigma\ln^{1/2}(N/\gamma)+F), 32F(F\ln(N/\gamma)+F)\right\}.\end{aligned}$$

若 $\sigma^2 \leq F^2\ln(N/\gamma)$，则有

$$\begin{aligned}&\mathbb{E}_{\nu}\left[\max_{i\in\{1,\ldots,N\}} Z_i^2\right]\\ \leq& 16\sigma^2\ln(N/\gamma) + \gamma\max\left\{4F^2\ln(N/\gamma), 32F(F\ln(N/\gamma)+F), 32F(F\ln(N/\gamma)+F)\right\}\\ =& 16\sigma^2\ln(N/\gamma) + 32\gamma F(F\ln(N/\gamma)+F)\\ =& 16\left(\sigma^2+2\gamma F^2\right)\ln(N/\gamma) + 32\gamma F^2.\end{aligned}$$

若 $\sigma^2 > F^2\ln(N/\gamma)$，则有

$$\begin{aligned}&\mathbb{E}_{\nu}\left[\max_{i\in\{1,\ldots,N\}} Z_i^2\right]\\ \leq& 16\sigma^2\ln(N/\gamma) + \gamma\max\left\{4\sigma^2, 32(\sigma^2+F^2), 32(\sigma^2+F^2)\right\}\\ =& 16\sigma^2\ln(N/\gamma) + 32\gamma(\sigma^2+F^2).\end{aligned}$$

因此，有

$$\mathbb{E}_{\nu}\left[\max_{i\in\{1,\ldots,N\}} Z_i^2\right] \leq 16\sigma^2\ln(N/\gamma) + 16\gamma\max\left\{\sigma^2, 2F^2\ln(N/\gamma)\right\} + 32\gamma F^2,$$

这等价于所需结果。 □

B 详细示例：复合空间

本节将详细介绍定理 1在第 4.1 节所引入的复合空间中的应用。回顾一下，复合空间定义为

$$\mathcal{F}_H \circ \ldots \circ \mathcal{F}_1 = \{f_H \circ \ldots \circ f_1 : f_i \in \mathcal{F}_i, i = 1, \ldots, H\}.$$

下面，我们给出复合稀疏 Hölder 空间（Schmidt-Hieber，2020；Kohler 和 Langer，2021）以及复合各向异性 Besov 空间（Suzuki 和 Nitanda，2021）的严格定义。

复合稀疏 Hölder 空间。复合稀疏 Hölder 空间对应于每个$\mathcal{F}_i$为稀疏 Hölder 空间的情形。对于多重指标 $\alpha \in \mathbb{N}^d$，我们记 $\partial^\alpha = \partial^{\alpha_1}...\partial^{\alpha_d}$和 $|\alpha| = \|\alpha\|_1$。对于 $\beta > 0$，Hölder 空间定义为

$$H^\beta([0,1]^d) = \left\{f : [0,1]^d \to [0,1] : \|f\|_{H^\beta([0,1]^d)} < \infty\right\}.$$

其中$\|f\|_{H^\beta([0,1]^d)}$ =最大值$_{\alpha\in\mathbb{N}^d:\ |\alpha|\le\beta}$ $\|\partial^\alpha f\|_{L^\infty([0,1]^d)}$ +最大值$_{\alpha\in\mathbb{N}^d:\ |\alpha|=\lfloor\beta\rfloor}$ $|\partial^\alpha f|_{H^{\beta-\lfloor\beta\rfloor}([0,1]^d)}$，

$$|f|_{H^\beta([0,1]^d)} = \sup_{x,y\in(0,1)^d : x\neq y} \frac{|f(x) - f(y)|}{\|x - y\|^\beta}.$$

复合稀疏 Hölder 空间由公式（4）给出，其中

$$\mathcal{F}_i = \mathcal{H}_i = \left\{f = (f_j)_j : [0,1]^{d_j} \to [0,1]^{d_{j+1}} : f_j \in H^{\beta_j}([0,1]^{t_j}), \|f_j\|_{H^{\beta_j}([0,1]^{t_j})} \le 1\right\},$$

对于某些 $\beta_j > 0$ 和 $1 \le t_j \le d_j$。

复合各向异性 Besov 空间。当每个$\mathcal{F}_i$都是各向异性 Besov 空间时，就得到复合各向异性 Besov 空间。对于函数 f：$[0,1]^d \to [0,1]$，其 r 阶有限差分递归定义为

$$\Delta_h^r(f)(x) = \Delta_h^{r-1}(f)(x + h) - \Delta_h^{r-1}(f)(x), \Delta_h^0(x) = f(x),$$

其中 $x \in [0,1]^d$和 $x + rh \in [0,1]^d$；否则 $\Delta_h^r(f)(x) = 0$。对于函数 $f \in L^p([0,1]^d)$ 和 $p \in (0, \infty]$，r 阶光滑模 f 定义为

$$w_{r,p}(f,t) = \sup_{h\in\mathbb{R}^d : |h_i|\le t_i} \|\Delta_h^r(f)(x)\|_{L^p([0,1]^d)},$$

其中 $t = (t_1, ..., t_d)^\top \in \mathbb{R}^d$，$t_i > 0$。对于 $\beta = (\beta_1, ..., \beta_d)^\top \in \mathbb{R}^d$，其中 $\beta_i > 0$，各向异性 Besov 空间定义为

$$B^\beta_{p,q}([0,1]^d) = \{f \in L^p : \|f\|_{B^\beta_{p,q}([0,1]^d)} < \infty\},$$

其中$\|f\|_{B^\beta_{p,\ q}([0,1]^d)} = \|f\|_{L^p} + |f|_{B^\beta_{p,\ q}([0,1]^d)}$，且

$$|f|_{B^\beta_{p,q}([0,1]^d)} = \begin{cases} \left(\sum_{k=0}^\infty \left(2^k w_{r,p}\left(f, \left(2^{-k/\beta_1}, ..., 2^{-k/\beta_d}\right)\right)\right)^q\right)^{1/q} & \text{if } q < \infty, \\ \sup_{k\ge 0} 2^k w_{r,p}\left(f, \left(2^{-k/\beta_1}, ..., 2^{-k/\beta_d}\right)\right) & \text{if } q = \infty. \end{cases}$$

复合各向异性 Besov 空间定义为 式（4），其中

$$\mathcal{F}_i = \mathcal{B}_i = \left\{f : [0,1]^{d_i} \to [0,1]^{d_{i+1}}, f_j \in B^{\beta_i}_{p,q}([0,1]^{d_i}), \|f\|_{B^{\beta_i}_{p,q}([0,1]^{d_i})} \le 1\right\},$$

其中 $\beta_i > 0$。

稀疏ReLU网络。稀疏 ReLU 网络被用于Schmidt-Hieber (2020)；Kohler 和 Langer (2021)；Suzuki 和 Nitanda (2021)以有效逼近复合空间。ReLU网络是一种使用ReLU激活函数的深度神经网络 $\eta(x) = \max(0, x)$。对于向量 $x \in \mathbb{R}^d$， $\eta(x)$ 逐元素应用。具有高度 N_h、宽度 N_w、稀疏约束 S和范数约束 B 的 ReLU 网络类定义为

$$\mathcal{F}_{\mathrm{ReLU}}(N_h, N_w, S, B) = \Big\{ (W^{(N_h)}\eta(\cdot) + b^{(L)}) \circ \ldots \circ (W^{(1)}\eta(\cdot) + b^{(1)}) : \\ W^{(L)} \in \mathbb{R}^{1\times N_w}, b^{(L)} \in \mathbb{R}, W^{(1)} \in \mathbb{R}^{N_w \times d}, W^{(\ell)} \in \mathbb{R}^{N_w \times N_w}, b^{(\ell)} \in \mathbb{R}^{N_w}, \\ \sum_{\ell=1}^{N_h} \Big(\big\| W^{(\ell)} \big\|_\infty + \big\| b^{(\ell)} \big\|_\infty \Big) \le S, \max_{\ell=1,\ldots,N_h} \big\| W^{(\ell)} \big\|_\infty \vee \big\| b^{(\ell)} \big\|_\infty \le B \Big\}.$$

由$\mathcal{F}_H \circ \ldots \circ \mathcal{F}_1$ 在 $H^{\beta_i}([0,1]^{t_i})$ 和 $B_{p',q}^{\beta_i}([0,1]^{d_i})$ 条件下，通过$\mathcal{F}_{\mathrm{ReLU}}(N_h, N_w, S, B)$ 进行逼近的逼近误差，以及$\mathcal{F}_{\mathrm{ReLU}}(N_h, N_w, S, B)$ 的覆盖数，均由以下定理给出。

定理B.1 (Schmidt-Hieber (2020))。 定义 $\alpha_i = \prod_{\ell=i+1}^{H} \min\{1, \beta_\ell\}$。对于每个 $J \in \mathbb{N}$，都存在适当的 N_h, N_w, S, B，使得

$$N(\delta, \mathcal{F}_{\mathrm{ReLU}}(N_h, N_w, S, B), \|\cdot\|_{L^\infty([0,1]^{d_1})}) \le CJ\mathrm{polylog}(J, 1/\delta),$$

并且对于任意 $f^* \in H_H \circ \ldots \circ H_1$,

$$\inf_{f \in \mathcal{F}_{\mathrm{ReLU}}(N_h, N_w, S, B)} \|f - f^*\|_{L^\infty([0,1]^{d_1})} \le C' J^{-\min_{i=1,\ldots,H} \frac{\alpha_i \beta_i}{t_i}},$$

其中某些常数 $C, C' > 0$。

定理B.2 (Suzuki 和 Nitanda (2021))。 定义 $\tilde{\beta}^{(\ell)} = (\sum_{j=1}^{d} 1/\beta_j^{(\ell)})^{-1}$ 以及 $\alpha_i = \prod_{k=\ell+1}^{H} (\min\{1, (\beta^{(\ell)}{}_{\min} - 1/p)\})$，其中 $\beta^{(\ell)}{}_{\min} = \min_{j=1,\ldots,d} \beta_j^{(\ell)}$。假设对于所有 ℓ， $\tilde{\beta}^{(\ell)} > 1/p$。对于每个 $J \in \mathbb{N}$，都存在适当的 N_h, N_w, S, B，使得

$$N(\delta, \mathcal{F}_{\mathrm{ReLU}}(N_h, N_w, S, B), \|\cdot\|_{L^\infty(\lambda)}) \le CJ \ln(J) \big(\ln^2(J) + \ln(1/\delta) \big),$$

且对于任意 $f^* \in B_H \circ \ldots \circ B_1$,

$$\inf_{f \in \mathcal{F}_{\mathrm{ReLU}}(N_h, N_w, S, B)} \|f - f^*\|_{L^\infty(\lambda)} \le C' J^{-\min_{\ell \in \{1,\ldots,H\}} \alpha_i \tilde{\beta}^{(\ell)}},$$

对于某些常数 C、 $C' > 0$。

无预训练模型时的误差。 通过将 定理 B.1 或 定理 B.2 与 定理 1相结合，我们得到如下误差率：

$$\mathbb{E}_{Q^n}[E_Q(f_n)] \le C \max_{i=1,\ldots,H} n^{-\gamma_i} \mathrm{polylog}(n), \tag{18}$$

其中， $\gamma_i = \frac{2\alpha_i\beta_i}{2\alpha_i\beta_i + t_i}$ 表示复合稀疏 Hölder 空间， $\gamma_i = \frac{\alpha_i \tilde{\beta}^{(\ell)}}{2\alpha_i \tilde{\beta}^{(\ell)} + 1}$ 表示复合各向异性 Besov 空间。

使用预训练模型时的误差（填缝）。现在，考虑存在预训练模型 f_{pre} 的情形，且对于某个 $1 < i_e < i_h < H$， f^* 可由 $f_{\text{pre}} = (g_h, m, g_e, m)$进行（$1/n$, $\mathrm{F}_{i_h} \circ \cdots \circ \mathrm{F}_{i_e}$）填缝，其中 $\mathrm{F}_H \circ \ldots \circ \mathrm{F}_1$ 为复合稀疏 Hölder 空间或复合各向异性 Besov 空间。类 $\mathrm{F}_{i_h} \circ \ldots \circ \mathrm{F}_{i_e}$ 的复杂度也由定理 B.1 或定理 B.2 确定。因此，我们有

$$\mathbb{E}_{Q^n}[E_Q(f_n)] \leq C \max_{i=i_e,\ldots,i_h} n^{-\gamma_i} \text{polylog}(n), \tag{19}$$

比较 式（18）和 式（19）后，我们发现，借助预训练模型，索引 i 的取值范围被限制在更小的子集中，这表明使用预训练模型进行经验填缝能够实现更低的误差率。

C 实验详情

C.1 卷积神经网络的微调

我们使用在 ImageNet 上预训练的 ResNet-50（2600 万个参数，He 等人（2016））和 Wide ResNet-50-2（6800 万个参数，Zagoruyko 和 Komodakis（2016）），并将其批归一化层替换为组归一化（Wu 和 He，2018）。这些采用组归一化的卷积神经网络先在 Office-Home 数据集（Venkateswara 等人，2017）中除目标域（剪贴画）以外的其他域上训练，然后在目标域上进行微调。每个域的数据划分为 80% 的训练集和 20% 的验证集。

微调时，在图像特征提取器和最后一个分类器之间插入适配器。适配器是采用 ReLU 激活函数的多层感知机。我们在其他域上进行预训练时采用随机梯度下降（SGD），学习率为10^{-3}、动量为 0.9、权重衰减为 10^{-4}；对适配器进行微调时采用 AdamW，学习率为 10^{-5}、权重衰减为 10^{-2}。每个阶段均对模型训练 50 个训练轮次。每次试验均使用内部集群中的一块 NVIDIA H100 GPU 运行。报告了五个不同随机种子下验证集划分上的最佳结果。

随附脚本用于复现实验。运行该脚本需要 `uv`[1]。

C.2 将视觉能力整合到大语言模型中

本实验使用 `nanovlm v0.2`[2]，除下文所述针对我们问题的设置外，其他设置均遵循其配置。我们采用 Llama-3 风格 Transformer 模型 `SmolLM2-135M`（Allal 等人，2024）作为基础语言模型，并使用 SigLIP2 的视觉编码器Tschannen 等人 (2025)，`siglip2-base-patch16-224` 以及 `siglip2-base-patch16-224`作为视觉特征提取器。提取出的视觉特征经过像素重排和适配器处理后，投影到语言嵌入空间。

使用 AdamW 优化适配器，学习率为 10^{-2}，权重衰减为 10^{-2}；同时使用 AdamW 优化语言模型，学习率为 10^{-5}，权重衰减为 10^{-2}，共训练 64k 次迭代。视觉特征提取器的权重保持固定。每次试验均使用内部集群中的八块 H100 GPU。报告 MMStar 上的最佳测试准确率。

[1]请参阅 https://docs.astral.sh/uv/ 了解安装说明。
[2]https://github.com/huggingface/nanoVLM/tree/v0.2