arXiv:2305.01034v2 [cs.LG] 2 Jun 2023

模型无关的泛化难度度量

阿基兰 Boopathy [1] Kevin Liu [1] Jaedong Hwang [1] Shu Ge [1] Asaad Mohammedsaleh [1] Ila Fiet e [1]

摘要

机器学习算法的度量在于其能够执行的任务的难度，而足够困难的任务是强大机器学习模型的关键驱动力。然而，对机器学习基准的泛化难度进行量化一直颇具挑战。我们提出了据我们所知首个与模型无关的任务固有泛化难度度量。我们的 归纳偏置复杂度度量，量化了在任务上实现良好泛化所需的总信息量减去数据所提供的信息量。具体而言，在给定假设符合训练数据的条件下，通过度量能够在任务上实现泛化的假设所占的分数体积来完成这一度量。该度量随模型必须进行泛化的空间的内在维度呈指数增长，但对于每个维度的分辨率仅呈多项式增长，这表明，需要在许多维度上进行泛化的任务，远比涉及更少维度中更多细节的任务困难。我们的度量可用于计算和比较监督学习、强化学习和元学习之间的泛化难度。我们展示了，将其应用于经验数据后，它能够形式化地量化直观预期的趋势，例如，就所需的归纳偏置而言，MNIST $<$ CIFAR10 $<$ImageNet，以及完全可观测的马尔可夫决策过程（马尔可夫决策过程）$<$部分可观测的马尔可夫决策过程。此外，我们表明，复杂图像的分类 $<$ 简单图像上的少样本元学习。我们的度量提供了一种定量指标，可用于指导构建需要更强归纳偏置的复杂任务，从而促进开发具有更强泛化能力的更先进架构和学习算法。

1. 引言

研究人员提出了许多基准测试，用于训练机器学习模型并测试其泛化能力，涵盖图像识别领域的 ImageNet（Krizhevsky 等人，2012年）到强化学习（RL）领域的 Atari 游戏（Bellemare 等人，2013）。更复杂的基准测试推动了更复杂的学习算法和架构的发展，使其能够实现更好的泛化。

然而，我们缺乏对这些基准测试的泛化难度进行严格且定量的度量。要在一项任务上实现泛化，既需要训练数据，也需要模型的归纳偏置；归纳偏置是任何 能够使模型类实现泛化的约束。归纳偏置可以由模型设计者提供，包括选择架构、学习规则或定义模型类的超参数。尽管先前的研究已对在一项任务上实现泛化所需的训练数据（样本复杂度）进行了量化，但对所需归纳偏置的分析仍然有限。事实上，尽管归纳偏置这一概念被广泛使用，但在一般学习设定中，它尚未得到全面且定量的定义。

本文提出了一种新颖的信息论框架，用于度量任务的归纳偏置复杂度，即归纳偏置的信息内容。正如样本复杂度是模型类固有的属性（无需参照特定训练集），归纳偏置复杂度也是训练集固有的属性（无需参照特定模型类）。据我们所知，我们的度量首次对归纳偏置复杂度进行了量化。正如下文所述，我们的度量量化了整个假设空间中与某项任务的归纳偏置一致的部分所占比例，前提是这些假设能够插值训练数据；图1展示了一个示例，形式化定义见定义3.1 。我们利用这一归纳偏置复杂度量化结果来评估任务的泛化难度；希望我们的度量能够指导开发需要更强归纳偏置的任务。作为一项具体且适用范围广泛的建议，我们发现，向任务的输入添加高斯噪声能够显著提高其所需的归纳偏置。我们的贡献总结如下： [1]:

[1]麻省理工学院。通信请联系：Akhilan Boopathy < akhilan@mit.edu>。

论文集： 40th 国际机器学习会议，美国夏威夷州檀香山。PMLR 202，2023。版权所有 © 2023，作者所有。

[1]我们的代码发布于：
https://github.com/FieteLab/inductive-bias-complexity

• 我们给出了归纳偏置复杂度的形式化信息论定义，以形式化归纳偏置的直观概念。

• 据我们所知，我们提出了一种新颖且首个可量化的归纳偏置复杂度度量，并建议将其用于度量任务的泛化难度。该度量还使我们能够量化应用于某项任务的不同模型架构各自的归纳偏置。

• 我们提出了一种实用算法，用于估计并比较监督学习、强化学习和少样本元学习任务等领域中各项任务的归纳偏置复杂度。通过实证研究，我们发现：1）与完全可观测的强化学习环境相比，部分可观测环境需要大得多的归纳偏置；2）简单的少样本元学习任务所需的归纳偏置，可能远大于复杂且贴近现实的监督学习任务。

2. 相关工作

2.1. 面向模型的泛化难度

机器学习模型的泛化能力传统上通过学习曲线来量化，学习曲线描述任务上的泛化误差与训练数据量（即样本复杂度）之间的关系（Amari，1993；Cortes 等人，1994；Hestness 等人，2017；Murata 等人，1992）。泛化误差可由模型类容量度量加以界定，包括Rademacher复杂度（Koltchinskii 和 Panchenko，2000年）和 VC 维度（Blumer 等人，1989年）；较简单的模型类可实现较低的泛化误差。尽管这些度量很有价值，但它们通常只能为特定任务提供宽松且缺乏针对性的泛化界，因为它们一般不会利用任务的结构。

近期提出的数据依赖的泛化界能够为泛化误差给出更紧的界（Jiang 等人，2021；Kawaguchi 等人，2022年；Lei 等人，2015年；Ne-grea 等人，2019 年；Raginsky 等人，2016）。这些度量利用数据集的性质，表明特定模型类别能够在该数据集上实现良好的泛化。与此相关，近期关于神经网络缩放定律的研究（Bahri 等人，2021；Hutter，2021；Sharma 和 Kaplan，2022年）将学习建模为核回归，并据此得出包括样本复杂度随数据内在维数呈指数级增长在内的结论。我们的方法在思路上与这条研究脉络一致；不过，我们并非利用任务结构来界定泛化误差或样本复杂度，而是界定任务的归纳偏置复杂度；见图1。我们认为，与样本复杂度相比，归纳偏置复杂度更适合作为衡量任务难度的指标。值得注意的是，样本复杂度假设存在一个特定模型类；它并非模型无关。因此，样本复杂度只量化训练数据对于使用特定模型类解决任务所作的贡献；它并不量化选择合适模型类（例如某种模型架构）的难度。

2.2. 泛化难度的信息论度量

近期研究使用信息论分析神经网络的性质；例如，信息瓶颈可用于提取具有良好泛化能力的数据表征（Saxe 等，2019；Shamir 等，2008；Shwartz-Ziv 与 Tishby，2017）。DIME（Zhang 等，2020）使用信息论估计任务难度，根据底层数据分布，对给定数据集上的任意模型所能达到的最佳情况下的错误率进行界定。Achille 等（2021）基于对输入和输出之间关系进行建模所需的 Kolmogorov 复杂度（Kolmogorov，1965）定义任务难度。尽管这些度量可以与泛化相关，但它们并不直接量化实现泛化所需的信息。另一种方式是，将泛化难度表示为：除学习系统所获得的任何训练数据之外，执行一项任务所需的信息量（Chollet，2019）。这与其他评估学习系统中所包含先验知识的文献相一致，无论这些先验知识体现为特定能力（Hern´andez-Orallo，2016）、偏差（Haussler，1988），还是模型架构（Du 等，2018；Li 等，2021）。在本文中，我们采用类似的方法，将泛化难度定义为：在训练数据之上，为了在假设空间内解决一项任务所需的归纳偏置信息量。

3. 衡量归纳偏置信息内容

根据无免费午餐定理（Wolpert，1996），众所周知，没有任何学习算法能够在所有任务上都表现出色；只有具备与特定任务相关的适当归纳偏置的学习算法，才能期望在数据量有限的情况下实现良好的泛化。因此，要在某项任务上实现泛化，既需要训练数据，也需要归纳偏置（即选择一个模型类）（见图1（b））。在第3.1节中，我们正式定义归纳偏置复杂度，即实现泛化所需的归纳偏置量。在第3.2节中，我们根据任务的目标误差率以及训练分布与测试分布之间的距离，对这一数量给出精确界定。在第3.3节中，我们进一步近似该界，从而得出可实际计算的估计值。表1总结了本节中使用的符号。

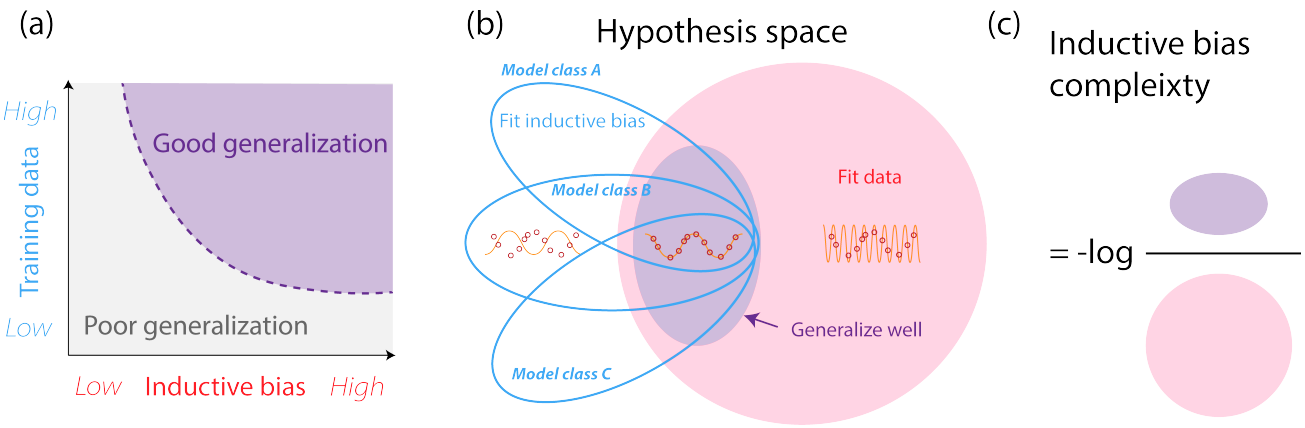


图 1：任务假设空间的划分。(a) 为了实现良好的泛化，所需训练数据量（样本复杂度）与所需归纳偏置量（归纳偏置复杂度）之间的权衡关系。对于固定的泛化性能目标，要降低样本复杂度，就需要提高归纳偏置复杂度，以进一步限制假设空间。尽管许多论文已经对样本复杂度进行了量化，但我们量化的是归纳偏置复杂度。(b) 根据假设是否符合训练数据（红色区域），以及是否包含特定的归纳偏置集合，对假设进行分类；不同的蓝色椭圆代表不同的模型类（即不同的归纳偏置集合）。结合训练数据，特定的模型类选择可以将假设限制在具有良好泛化能力的区域内；紫色表示具有良好泛化能力的假设。训练数据和归纳偏置对于实现良好的泛化都是必不可少的。在实践中，具有良好泛化能力的假设仅占假设空间中符合训练数据的区域的一小部分。(c) 归纳偏置复杂度的计算方法：我们将归纳偏置复杂度量化为在符合训练数据的假设中，能够实现良好泛化的假设所占比例的负对数。

表 1：符号表。

符号	含义
$f(\cdot;\theta)$	假设空间中的函数，由 θ参数化
p	输入分布，表示测试分布
q	输入分布，表示训练分布
$e(\theta,\ p)$	$f(\cdot;\theta)$ 在输入分布 p 上的误差
$\mathcal{L}$	损失函数
ε	期望的测试误差
$\tilde{I}$	归纳偏置复杂度
Θ_q	符合训练数据集的假设集合
$L_{\mathcal{L}}$	$\mathcal{L}$的利普希茨常数
L_f	f上的二阶条件所用的常数
m	数据的内在维度
k	子流形的数量
r	数据点所在球面的半径
d	模型输出 $f(x;\theta)$ 的维数
f_i	f 的单个分量
b	$\|f(x;\theta)\|$ 的上界
n	训练数据集的大小
M	所考虑特征函数的最大频率
δ	f 的空间分辨率；等于 $2\pi r/M$
K	满足 $\sqrt{K(K+m-1)} \leq M$ 的最大整数
E	频率为 $\leq M$的特征函数数量； $O(M^m)$

3.1. 形式化定义归纳偏置复杂度

我们使用假设空间这一概念，正式阐明归纳偏置与泛化之间的关系。假设可能满足训练集，也可能不满足训练集；对于不同的归纳偏置集合，情况也可能不同（见图1 （b））；不同的归纳偏置集合（例如不同的模型架构）对应假设空间中的不同子集。从根本上说，实现泛化既需要归纳偏置，也需要训练数据，以确定一组能够良好泛化的假设。请注意，训练数据量与实现泛化所需的归纳偏置之间存在权衡（见图1（a））。我们的目标是量化在训练数据量固定时，实现泛化所需的最低归纳偏置水平。我们将这一度量称为归纳偏置复杂度（见图1 （c））。样本复杂度衡量在归纳偏置固定（即模型类固定）的情况下，实现泛化所需的训练数据量；与之类似，归纳偏置复杂度衡量在训练数据量固定的情况下，实现泛化所需的归纳偏置量。

我们首先定义一个（非常宽泛且通用的）假设空间，由 $f(x;\theta)$组成，其中参数由一个向量随机变量 θ给出。该假设空间并不等同于用于解决任务的模型类别（例如神经网络）。相反，理想情况下，假设空间应是一个巨大的空间，涵盖所有可能合理用于解决任务的模型类别。我们将

此处构建的假设空间简单地视为一种用于分析任务本身属性的数学工具，而不是与针对某个任务训练的实际模型相关联。我们假设要学习的真实（目标）函数可以在这个大型假设空间中表示，并将其定义为$f^*(x) = f(x;\theta^*)$。特定θ和输入分布p的泛化误差记为$e(\theta, p) = \tilde{e}(f(\cdot;\theta)\ p)$。我们的目标是找到一个假设$\theta$，使其泛化误差低于阈值$\varepsilon$：$e(\theta, p) \leq \varepsilon$。尽管这一表示法类似于监督学习，但正如附录 A所述，我们的框架涵盖了许多不同的学习场景，包括无监督学习、强化学习和元学习。

我们从贝叶斯视角出发，量化指定一个能够良好泛化的假设集合所需的信息 量；这将作为我们对归纳偏置复杂度的度量。具体而言，我们假设假设θ是随机变量，并假设假设空间中的假设服从均匀先验概率分布（通常可以通过重新参数化使这一假设成立）。接着，我们观察到一组训练数据，而假设空间中只有一部分假设与这些数据相一致。因此，训练数据会诱导出一个新的假设后验分布。最后，归纳偏置复杂度将与从该后验分布中抽取的假设属于良好泛化假设集合的概率相关：如果在后验分布下泛化的概率较低，就需要更强的归纳偏置。因此，我们对归纳偏置复杂度的定义可以看作是指定良好泛化假设所需的额外信息，超出训练数据所提供的任何与任务相关的信息。

˜归纳偏置复杂度的形式化定义如下：I

定义 1. 假设在一个假设集合上定义了一个概率分布，并令随机变量θ对应于从该分布中抽取的样本。给定误差阈值ϵ，模型经过优化后在训练集上达到该阈值，则在某项任务上达到错误率ε所需的归纳偏置复杂度为

$$\tilde{I} = -\log \mathbb{P}(e(\theta, p) \leq \varepsilon \mid e(\theta, q) \leq \epsilon) \qquad (1)$$

注意，这恰好是这样一个概率事件的信息内容：一个符合训练集的假设也能很好地泛化。这将直观的概念形式化为：归纳偏置对应于在符合训练数据的模型中找到泛化良好的模型的难度。为便于分析，我们将“符合”训练集定义为对其进行完美插值（设置$\epsilon = 0$）。一个泛化良好的假设虽然错误率较低但不为零，却不一定会对训练集进行插值。尽管如此，完美插值这一情形与训练过参数化深度网络有关；其他情形留待后续工作探讨，并在附录中简要说明如何将方法扩展到这些情形G。

重要的是，我们对归纳偏置复杂度的定义并未指定归纳偏置的具体形式。正如图1(b) 所示，多组归纳偏置（例如，不同的模型架构）可能达到相同的泛化水平。我们的度量量化了任何一组归纳偏置为实现泛化必须提供的信息量。此外，我们对归纳偏置复杂度的定义不同于模型类的表达能力（即模型类所涵盖的假设体积）。归纳偏置复杂度衡量的是模型类为在特定任务上实现泛化而必须受到的约束程度；相比之下，表达能力是模型类特有的度量，不一定取决于任务。因此，表达能力更强（即约束更少）的模型类，针对特定任务提供的归纳偏置可能更多，也可能更少，这取决于表达能力的提升与任务的相关性。例如，我们在第4节的实验中发现，某些视觉变换器架构在 ImageNet 分类任务上的表现优于卷积神经网络（CNN），因而提供了更多归纳偏置；尽管可以认为 Transformer 属于约束更少的模型类，因为其空间不变性约束可能弱于卷积神经网络。

3.2. 归纳偏置复杂度的精确上界

在一些技术性假设下，我们能够为上述归纳偏差复杂度给出一个精确的上界。这些假设未必在所有情形下都成立；然而，我们认为它们适用于许多常见情形（例如，当损失函数为平方损失，且模型f二阶可微，并且输入和参数有界时）。具体而言，我们假设输入空间、参数空间和输出空间都配备了标准距离度量$d(\cdot,\cdot)$和范数$\|\cdot\|$。我们假设损失函数具有平移不变性，即对于所有y、x_1、x_2，$\mathcal{L}(y + f^*(x_2) - f^*(x_1)\ x_2) = \mathcal{L}(y, x_1)$；平方损失函数满足这一假设，因为$||y + f^*(x_2) - f^*(x_1) - f^*(x_2)||_2^2 = ||y - f^*(x_1)||_2^2$。我们还假设损失函数$\mathcal{L}$存在一个利普希茨常数$L_{\mathcal{L}}$，它对应于损失函数对预测输出扰动的敏感度；我们预计，在大多数现实问题中，这种敏感度较小。最后，我们假设模型f满足二阶条件：$\|f(x_1;\theta_1) - f(x_1;\theta_2) - f(x_2;\theta_1) + f(x_2;\theta_2)\| \leq L_f d(x_1, x_2) d(\theta_1, \theta_2)$。$L_f$直观上对应于模型对输入$x$和假设参数化$\theta$的联合敏感度。如果对模型进行轻微扰动不会显著影响输入与输出之间的局部映射，那么L_f就会很小。我们也认为，对于许多问题而言，这一假设是合理的，不过也可能存在不满足该假设的模型（例如存在不连续性的模型），

不满足这一假设的情形。附录C 进一步讨论了这些假设的有效性。

最后，令 p 表示测试分布，q 表示训练分布，并令 $W(\cdot,\cdot)$ 表示 1-Wasserstein 距离。此外，定义 $\Theta_q = \{\theta \mid e(\theta, q) = 0\}$ 为能够完美插值训练数据集的假设集合。

定理 2. 假设损失函数满足$\mathcal{L}(y, x) \geq 0$，且当且仅当 $y = f^*(x)$时等号成立。假设损失函数关于 x 中的平移不变，即 $\mathcal{L}(y + f^*(x_2) - f^*(x_1)\ x_2) = \mathcal{L}(y, x_1)$ 对所有 y、x_1、x_2 均成立。假设 $\mathcal{L}$ 的利普希茨常数为 $L_{\mathcal{L}}$:

$$\mathcal{L}(y + \delta, x) - \mathcal{L}(y, x) \leq L_{\mathcal{L}} \|\delta\| \tag{2}$$

对于所有 y、δ、x。我们还假设对f 满足二阶条件，其中某个常数为 L_f:

$$\|f(x_1;\theta_1) - f(x_1;\theta_2) - f(x_2;\theta_1) + f(x_2;\theta_2)\| \leq L_f d(x_1, x_2) d(\theta_1, \theta_2) \tag{3}$$

对于所有 x_1、x_2、θ_1、θ_2，均成立。由此，我们得到归纳偏差复杂度的如下上界:

$$\tilde{I} \leq -\log \mathbb{P}\left(d(\theta, \theta^*) \leq \frac{\varepsilon}{L_{\mathcal{L}} L_f W(p, q)} \mid \theta \in \Theta_q \right), \tag{4}$$

该界取决于以下关键量：期望错误率 ε、插值假设空间 Θ_q 以及训练分布与测试分布之间的1-瓦瑟斯坦距离 $W(p,\ q)$。证明见附录B。请注意，该证明具有相当的普适性：它不依赖于对假设空间的任何特定假设，也不假定测试分布与训练分布之间存在任何关系。

3.3. 经验归纳偏置复杂度近似

现在，我们对定理2的一般界进行一系列近似，以得到一个可实际计算的估计。我们在监督分类的设定下介绍这种近似，但请注意，正如实验所示，它也可应用于其他设定（经验计算的完整细节见附录D ）。由于我们提出的是首个归纳偏置复杂度度量，因此重点在于捕捉总体趋势（缩放规律），而非给出精确估计。我们希望未来的工作能够进一步改进这些估计，或进一步放宽我们的假设。

监督分类设置 我们假设数据点x 位于半径为 r的 d 互不重叠的 m球面上，其中每个球面对应一个类别。此外，我们假设测试分布 p 在所有 m球面上均匀分布，而训练分布 q 是以下数据点的经验分布 n 从 p中独立同分布抽取的点。这种近似捕捉了实际分类任务中数据分布的关键特征：1）有界域，2）所有点上的密度相近，3）不同类别之间的分布互不重叠。我们假设 $f(x;\ \theta)$ 的维数为 d，即类别数量。令 b 为一个常数，使得对于所有 $x\,\theta$，$\|f(x;\ \theta)\| \leq b$。令 δ 为 f的“空间分辨率”，其定义为：将输入 x 扰动小于 δ 的量时，不会显著改变输出 $f(x;\ \theta)$（即，δ 是 f 能够感知的最小变化）。在我们的监督分类情形中，δ 对应于类别之间的距离。

假设参数化为了计算公式（4），我们必须为 $f(x;\ \theta)$ 选择一种具体的参数化方式。回顾第3.1节，我们希望选择一个广泛且通用的假设空间，涵盖实际适用于某项任务的所有模型类。我们可能不希望考虑对 x 任务相关分辨率以下的变化敏感 θ的函数。因此，我们只考虑频率低于 $M = 2\pi r/\delta$的带限函数。重要的是，这种假设空间参数化的选择并不取决于实际用于训练某项任务的模型类，而仅取决于任务的属性。

更正式地说，为了构造球面上低于某一频率的函数基，我们采用McRae 等人（2020年）的方法，将函数参数化为流形上拉普拉斯-贝尔特拉米算子特征函数的线性组合。拉普拉斯-贝尔特拉米算子是拉普拉斯算子向一般流形的推广；它以定义在流形上的函数为输入，并输出该函数在流形各点处曲率的度量。该算子的特征函数类似于欧几里得空间中的傅里叶基函数，相应的特征值则对应于特征函数的局部曲率（类似于频率）。通过恰当选择频率截断值 M，我们可以构造一个非常庞大且通用的假设空间，它既涵盖深度神经网络及其他常用模型类，又是有限维的，并排除任何包含不合理弯曲分量（即频率过高的分量）的函数。图6 展示了欧几里得情形下的可视化结果。我们还要指出，选择不同的基可能会产生具有不同属性的假设空间：我们选择的基会明确排除所有高频函数，但另一种基可能包含带有高频分量的函数。此外，即使采用我们选择的基，频率截断值 M也会对假设空间的大小及由此产生的归纳偏置复杂度产生关键影响（见附录F.5）。

回想一下，我们的目标是构建一组从 d m-球面映射到 $\mathbb{R}^d$的基函数。我们将为每个 m-球面和每个维度分别使用一组基函数-

的维度 f。对于每个非负整数 k，上的拉普拉斯-贝尔特拉米算子具有 m球面的 $\binom{m+k}{m} - \binom{m+k-2}{m}$个频率为 $\sqrt{k(k+m-1)}$ (Dai & Xu，2013）。令 K 为满足以下条件的最大整数 k： $\sqrt{k(k+m-1)} \leq M$，则有

$$E = \sum_{k=0}^{K} \binom{m+k}{m} - \binom{m+k-2}{m} = \binom{m+K-1}{m} + \binom{m+K}{m} \quad (5)$$

频率至多为 M 的特征函数。注意，对于较大的 M，E 随 $O(M^m)$ 变化，因为 K 随 M 变化。由于每个系数都有实部和虚部， d 维函数 f 在每个 d m维球面 上的每个一维分量都可以用 $2E$ 个基函数参数化。对于单个 m 维球面，参数化完整 d 维函数所需的基函数总数为 $2dE$。对于所有 d m 维球面，基函数总数为 $2d^2E$：这就是 θ 的维数。

每个训练点都会将 θ 可能取值的维数约束 d，因为给定任意一个假设，为了使其拟合一个输出维数为 d 的新训练点，通常只需修改假设的 d 个维度。换言之，只要能够独立控制假设空间中 d 个变化维度，就足以控制假设在新点上的 d 维输出。在线性情形下可以对此作具体理解：假设假设 $f(x;\theta) \in \mathbb{R}^d$ 构造为 $f(x;\theta) = A\theta$，其中 A 是一个维数为 $d \times 2d^2E$ 的满秩矩阵。那么，为了拟合一个点 $(x,\ y)$， θ 必须满足 $y = A\theta$，这将 θ 约束在一个 $2d^2E - d$ 维空间中（具体而言，θ 可以在 A 的零空间中移动）。因此， Θ_q 的维数为 $2d^2E - nd$。

接下来，我们得到关于 $\|\theta\|$的一个界。如果分量 f_k 的参数为 θ_k，帕塞瓦尔恒等式（Stein &Shakarchi，2011）表明 $\|\theta_k\|^2$ 等于 $\|f_k(x;\theta_k)\|^2$ 的平均值，其上界为 b^2。由于共有 d^2个分量，因此有 $\|\theta\| \leq bd$。

接下来，为了近似 Θ_q的体积，我们必须对 Θ_q的形状作出假设。鉴于我们已知 θ的一个上界，我们只需将 Θ_q 近似为维数为 $2d^2E - nd$ 、半径为 bd的圆盘；请注意，该圆盘嵌入在一个 $2d^2E$ 维空间中。这种圆盘近似保留了插值流形的关键性质（低维性和有界性），同时在解析上易于处理；更多讨论请参见附录C.3。

The set $\left\{\theta \in \Theta_q \mid d(\theta,\theta^*) \leq \frac{\varepsilon}{L_{\mathcal{L}} L_f W(p,q)}\right\}$ is a disc with 半径 $\frac{\varepsilon}{L_{\mathcal{L}} L_f W(p,q)}$。因此，

$$\mathbb{P}\left(d(\theta,\theta^*) \leq \frac{\varepsilon}{L_{\mathcal{L}} L_f W(p,q)} \mid \theta \in \Theta_q\right) \approx \left(\frac{\frac{\varepsilon}{L_{\mathcal{L}} L_f W(p,q)}}{bd}\right)^{2d^2E-nd}. \quad (6)$$

为了得到这一近似，我们将假设空间设为流形上拉普拉斯-贝尔特拉米算子的特征函数的线性组合，然后将假设空间中的体积近似为球体。由于我们没有对 x 的流形作出具体假设来构造假设空间，因此我们预期结果的整体缩放关系适用于一般任务。

近似 Wasserstein 距离 我们得到了 Wasserstein 距离 $W(p,q)$ 的一个简单尺度关系；它表示 p 与 q 之间最优传输中的期望传输距离。在最优传输计划中，每个训练点都对应其所在球面上的一个局部区域；该计划会在每个点与其周围的局部区域之间移动概率密度。这些区域必须大小相等、彼此不相交，并且铺满所有球面的整个区域。尽管这些区域的形状取决于具体分布 q，但如果 q 接近均匀分布，我们预计这些区域的形状也会相似。

我们将这些区域近似为超立方体，因为它们是能够铺满 m 维空间的最简单形状。为了填满所有球面的完整区域 $d \cdot \frac{2\pi^{(m+1)/2}}{\Gamma\left(\frac{m+1}{2}\right)} r^m$ ，这些超立方体的边长必须为

$$s = \left(\frac{d}{n} \cdot \frac{2\pi^{(m+1)/2}}{\Gamma\left(\frac{m+1}{2}\right)} r^m\right)^{1/m} = rn^{-1/m} \left(\frac{2\pi^{(m+1)/2} d}{\Gamma\left(\frac{m+1}{2}\right)}\right)^{1/m}. \quad (7)$$

超立方体中随机选取的两点之间的期望距离至多为 $s\sqrt{m}/6$ (Anderssen 等人，1976年）。我们将此距离作为每个点与其对应的局部超立方体之间最优传输距离的近似值。因此，最终的 $W(p,\ q)$ 可以直接近似为每个点对应的最优传输距离：

$$W(p,q) \approx s\sqrt{\frac{m}{6}} = rn^{-1/m}\sqrt{\frac{m}{6}} \left(\frac{2\pi^{(m+1)/2} d}{\Gamma\left(\frac{m+1}{2}\right)}\right)^{1/m} \quad (8)$$

请注意，该表达式按 $O(rn^{-1/m})$ 缩放。这种缩放在直觉上是合理的：训练分布与测试分布之间的距离，会随数据流形 r 的尺度线性变化。此外，给定在一个 m 维流形上随机采样的 n 个点，我们可以预期，流形体积中大约有 $n\rho^m$ 的部分位于距至少一个点 ρ 的半径范围内

；这与 $n^{-1/m}$的缩放规律一致。事实上，Singh & P´oczos (2018年）发现，这种缩放关系普遍适用，与流形结构无关。

在附录C.4中，我们通过实证计算了球面上的 Wasserstein 距离，并发现这些距离符合这一近似。尽管这种特定近似仅在经验训练分布 q 从均匀的底层数据分布中独立同分布抽取时 p才有效，但只需修改我们的 Wasserstein 距离近似以反映训练—测试分布偏移，我们的框架便可应用于非独立同分布情形。

最终近似由于 f 是由系数为 θ的基函数线性组合而成，L_f 给出了基函数梯度范数的上界。具有单个输出且频率为 $\sqrt{k(k+m-1)}$ 的基函数，其梯度幅值至多为 k/r。由于 f的输出维数为 d，我们可以取 $L_f = K\sqrt{d}/r$。将公式（6）的结果以及我们对$W(p,q)$的近似代入公式（ 4 ），可得归纳偏置复杂度的近似值

$$\tilde{I} \approx (2d^2E - nd) \left(\log b + \frac{3}{2}\log d + \log K - \frac{1}{m}\log n - \log\frac{\varepsilon}{L_{\mathcal{L}}} + \log c\right). \tag{9}$$

其中 $c = \sqrt[m]{\frac{6}{}}\left(\frac{2\pi^{(m+1)/2}d}{\Gamma\left(\frac{m+1}{2}\right)}\right)^{1/m}$，当内在维数较大时，该值大致保持不变 m。回顾一下， d是类别数量， n 是训练数据点的数量， ε 是目标误差， $L_{\mathcal{L}}$ 是损失函数的利普希茨常数， b 是模型输出的一个界，而K 和 E 是 m 和最大频率$M = 2\pi r/\delta$的函数。

归纳偏置复杂度的性质 在我们的表达式（公式 9）中，归纳偏置复杂度的主导项为 $2d^2E - nd$。由于 $E \in O(M^m)$，归纳偏置复杂度关于数据的内在维数 m呈指数增长 。直观而言，这是因为假设空间的维数会随内在维数呈指数增长：输入中的每个额外变化维度，都要求假设针对新输入维度中的每个可能取值表达一组新的模型输出值。由于归纳偏置复杂度表示指定一个能够良好泛化的假设的难度，因此它会随假设空间的维数增长。这种指数依赖关系在经验上使我们的归纳偏置复杂度度量具有较大的尺度，而该尺度主要由输入维度决定。有关进一步的直观解释和讨论，请参见附录G 。

由于 $M = 2\pi r/\delta$，对于固定的 m，归纳偏置复杂度是关于 $1/\delta$ 的多项式。直观而言，这是因为改变假设中允许的最大频率，只会沿假设空间的每个维度线性缩放假设的数量，因此总体上产生多项式效应。因此，δ 对归纳偏置复杂度的影响小于 m，但当 m 较大时，影响仍可能显著。归纳偏置复杂度随 d 呈二次增长，这是因为它对假设空间维数产生二次影响，因此其影响小于 m 和 δ。归纳偏置复杂度随训练点数量 n 大致线性下降；直观而言，这是因为每个训练点提供的信息量大致相同。最后，随着期望误差 ε 降低，归纳偏置复杂度呈对数增长；这是因为 ε 对应于假设空间中的一个半径，而归纳偏置是根据假设空间中体积的对数定义的。直观而言，归纳偏置复杂度必须取决于期望的性能水平：例如，在分类任务中，达到随机分类器的性能是轻而易举的。

4. 基准测试的经验泛化难度

在这里，我们将归纳偏置复杂度的量化值用作任务的泛化难度 度量（或任务难度）。归纳偏置复杂度对应于模型设计者必须为在某项任务上实现泛化而付出的努力，而样本复杂度对应于训练数据的贡献；因此，从模型设计者的角度来看，归纳偏置复杂度是衡量任务泛化难度的恰当指标。我们在若干监督分类、元学习和强化学习任务上对该度量进行了实证验证。我们还量化了不同模型架构在不同任务上的相对归纳偏置，并通过实证展示了任务难度随参数化任务变化而呈现的趋势。

˜计算难度 I，即针对指定若干参数的方程（9）。目标泛化误差 ε 根据任务类型设定为固定水平（见附录D表3）。我们通过将训练集中的最近邻距离衰减率与内在维度联系起来，估算数据维度m （Pope 等人，2021）。每个维度所需的细节程度（分辨率） δ 对于分类任务设为类间间隔；对于强化学习任务，则设为状态扰动不会显著影响轨迹的尺度。完整详情见附录D。

4.1. 标准基准测试的任务难度

图像分类我们估算了常用图像分类基准测试 MNIST、SVHN（Netzer 等人，2011）、CIFAR10（Krizhevsky 等人，2009年）和 ImageNet（Deng 等人，2009年）。如图2所示，解决这些任务的模型编码了大量

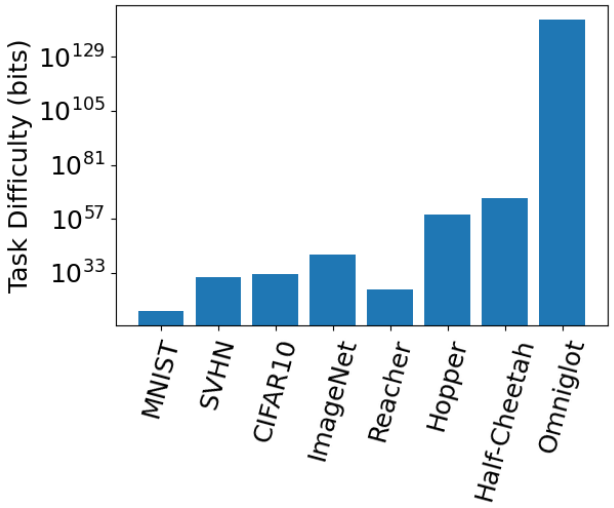


图 2：不同领域中各类基准任务的任务难度。

的归纳偏置，任务越复杂，所需模型的归纳偏置就越多。

少样本学习 我们估算了 Omniglot 上 1-shot、20-way 分类任务的任务难度（Lake 等人，2015年）。模型必须学习一个函数 f，将一个包含某字母表中 20 张图像的数据集 映射为一个分类器 g，而该分类器本身又将一张图像映射为涵盖 20 个类别的概率分布。输入和输出 f 的维数均远大于典型图像分类任务。这导致任务难度极高：Omniglot 的任务难度为 $\approx 10^{145}$ 比特（相比之下，ImageNet 为 10^{41}）；参见图2。

强化学习 我们发现，MuJoCo 中 Reacher、Hopper 和 Half-Cheetah 控制任务的任务难度（Todorov 等人，2012年；Duan 等人，2016）。图 2 表明，正如直观预期的那样，观测空间和动作空间维度更高的复杂环境具有更高的任务难度。此外，任务难度的范围与图像分类相当：不同领域之间的任务难度具有可比性。

4.2. 不同模型对各种任务贡献的归纳偏置

尽管我们的任务难度度量与模型无关，但我们可以用它来比较不同模型为各种任务贡献了多少归纳偏置。为此，我们找出模型类中一个训练好的模型在测试集上的表现（例如分类任务的测试集准确率）。然后，我们用这一表现水平计算 ε 任务难度表达式中的值，见公式（9）（重要的是，我们并未直接使用测试集数据）。这一方法基于这样的观点：模型类通过提供归纳偏置来实现泛化，而其泛化能力可以用测试集表现来量化。因此，我们将模型类的归纳偏置量化为达到该模型类表现所需的最小归纳偏置：表现越好，所需的归纳偏置越大。如表2 所示（完整结果见表7），采用更好的模型架构（例如从残差网络（He 等人，2016）切换到视觉变换器（Dosovitskiy 等人，2021）所带来的信息内容增量，通常在更复杂的任务中更大。换言之，越复杂的任务越能提取模型中蕴含的归纳偏置信息。

表2：不同模型架构为图像分类任务贡献的归纳偏置信息。

CIFAR10	
模型	信息内容（$\times 10^{32}$ 比特）
线性模型 (Nishimoto, 2018年)	2.250
AlexNet（Krizhevsky 等人，2012年）	2.709
ResNet-50（Wightman 等人，2021）	3.195
BiT-L（Kolesnikov 等人，2020年）	3.453
ViT-H/14（Dosovitskiy 等人，2021）	3.513

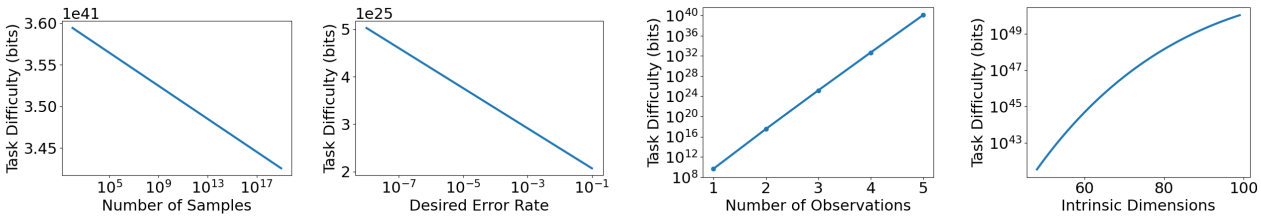
ImageNet	
模型	信息内容（$\times 10^{41}$ 比特）
线性模型（Karpathy，2015年）	2.063
AlexNet（Krizhevsky 等人，2012年）	2.290
ResNet-50（He 等人，2016）	2.362
BiT-L（Kolesnikov 等人，2020年）	2.449
ViT-H/14（Dosovitskiy 等人，2021）	2.461

4.3. 任务难度趋势

在本节中，我们将参数化地改变任务的各个方面，考察任务难度如何随之变化。更多任务变化请参见附录F。

改变训练数据 我们改变 ImageNet 中提供的训练数据量，并在图 3a中发现，随着训练点增多，任务难度会降低。值得注意的是，即使训练数据量增加了许多个数量级，任务难度的量级仍然很大，这表明，相较于归纳偏置，训练数据为解决任务提供的信息相对较少。直观而言，这是因为在缺乏强归纳偏置时，训练数据只能提供每个训练点周围一个局部邻域内有关近似函数的信息；相比之下，归纳偏置则可以为函数提供全局约束。更多讨论请参见附录G 。

改变目标误差率 在图 3b中，我们改变 Reacher 任务的目标误差，发现任务难度会随目标误差率降低，这符合预期。与改变训练数据量的情况一样，即使大幅降低目标误差率，任务难度的量级仍然很大。



(a) 改变训练数据
在 ImageNet 上

(b) 改变目标误差
在 Reacher 上

(c) 改变内在维度
在 Cartpole 上

(d) 变化内在维度
在 ImageNet 上

图 3：基准测试变化中的任务难度 务。

改变内在维度 我们研究 Cartpole 的一个含噪版本（Florian，2007），这是一项倒立摆控制任务，智能体施加二元力来控制摆杆。在含噪的 Cartpole 中，智能体观测到的是带噪声的杆的角位置和速度，因此可能需要进行多次观测才能找到最优动作。智能体必须学习一个输入维度更高的函数，这使任务更加困难。如图3c 所示，任务难度随找到最优动作所需的观测次数呈指数级增长。我们还通过对 ImageNet 分类的图像添加原始图像流形之外的扰动，来增加其内在维度；理论和实证研究均表明，训练出能够抵御此类扰动的分类器十分困难（Madry 等人，2017；Yin 等人，2019 年）。图 3d 显示，随着增加内在维度，任务难度会大幅上升。因此，提高任务的内在维度是增加其难度的一种有效方式。

5. 讨论

对于所有任务，归纳偏置复杂度都远大于实际模型的规模。这是由假设空间的体积造成的：为了尽可能不依赖模型，我们构建了一个庞大的函数空间，其中的函数仅受频率内容带宽约束的限制，见第 3.3 节。泛化良好的函数在该空间中所占的体积比例很小，因此需要在归纳偏置中包含海量信息。进一步讨论见附录 G。未来的一个研究方向是，为特定模型类别（例如深度神经网络）获得任务（归纳偏置）复杂度度量。

我们的度量为设计具有挑战性的基准测试提供了定量指导。它揭示，输入位于更高维流形上的任务，相较于其他变化，需要多得多的归纳偏置。这带来了反直觉的发现：在简单的 Omniglot 手写字符上进行元学习，比对复杂、逼真的 ImageNet 图像进行分类需要多得多的归纳偏置。因此，优先考虑泛化，以应对多种变化，远比在每个变化维度内提升任务的丰富程度更为重要——即使这些基准测试（如 Omniglot）看起来可能很简单。此类基准测试包括少样本学习、对抗鲁棒图像分类，以及观测极为有限的马尔可夫决策过程。作为一个具体但适用范围广的增加变化维度的例子，我们建议像图3d。

我们还要强调，归纳偏置复杂度并非旨在涵盖任务中可能较难的所有方面。例如，在在线强化学习中，我们对归纳偏置复杂度的定义并未捕捉训练数据对策略的依赖性，因此无法量化探索与利用之间的权衡。不过，我们的框架恰好捕捉了任务中与泛化相关的方面：例如，在强化学习中，泛化可以用与监督学习大致相同的方式来理解，区别在于训练和测试期间，不同策略会诱导出不同的分布（Kakade，2003年）。重要的是，在我们的分析中，我们不要求训练和测试必须基于相同的分布（见第 3.2 节）；因此，我们认为，该度量适用于量化在线强化学习等情境下的泛化能力，也适用于其他训练分布与测试分布可能截然不同的情形。

据我们所知，我们首次提出了归纳偏置复杂度的度量，因此该度量并非旨在实现定量上的精确。我们希望未来的研究能够进一步改进我们的估计。尽管如此，我们仍希望这一度量能够推动开发出需要具备良好泛化能力的架构和学习规则的任务，而这些架构和学习规则内置了许多归纳偏置。

致谢

Ila Fiete 获得了美国海军研究办公室、霍华德·休斯医学研究所（HHMI）和美国国立卫生研究院（NIMH-MH129046）的资助。

参考文献

Achille, A., Paolini, G., Mbeng, G. 和 Soatto, S. 学习任务的信息复杂度、任务结构及其距离。*Information andInference: A JournaloftheIMA*，10(1):51–72，2021。

Amari，S. 学习曲线普适定理。神经网络，6(2)：161–166，1993。ISSN 0893-6080。doi: https://doi.org/10.1016/0893-6080(93)90013-M。

An，S.、Lee，M.、Park，S.、Yang，H. 和 So，J. 用于MNIST 数字识别的简单卷积神经网络模型集成。*arXiv* 预印本 *arXiv:2008.10400*，2020年。

Anderssen, R. S., Brent, R. P., Daley, D. J., and Moran, P. A. P. 关于$\int_0^1 \cdots \int_0^1 (x_1^2 + \cdots + x_k^2)^{1/2} dx_1 \cdots, dx_k$以及泰勒级数方法。*SIAM* 应用数学杂志，30(1):22–30，1976年。doi: 10.1137/0130003.

Bahri, Y.、Dyer, E.、Kaplan, J.、Lee, J. 和 Sharma, U. 解释神经网络缩放定律。*arXiv* 预印本 *arXiv:2102.06701*，2021 年。

Bellemare, M. G., Naddaf, Y., Veness, J. 和 Bowling, M.《街机学习环境：通用智能体的评估平台》。*JAIR*，47:253–279，2013 年 6 月。doi: 10.1613/jair.3912。

Blumer, A., Ehrenfeucht, A., Haussler, D. 和 Warmuth, M. K. 可学习性与 Vapnik–Chervonenkis 维数。*J.ACM*，36(4):929–965，1989年10月。doi: 10.1145/76359.76371。

Chollet, F. 关于智能度量。*arXiv* 预印本，2019 年。

Ciregan, D., Meier, U. 和 Schmidhuber, J. 用于图像分类的多列深度神经网络。载于*CVPR*，2012年。

Cortes, C., Jackel, L. D., Solla, S., Vapnik, V. 和 Denker, J. 学习曲线：渐近值与收敛速率。载于*NeurIPS*，1994年。

Dai, F. 和 Xu, Y. 球面与球体上的逼近理论和调和分析，第23卷。Springer，2013年。

Dekkers, A. L. M., Einmahl, J. H. J. 和 Haan, L. D. 极值分布指数的矩估计量。统计年刊，1989年。

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet：大型分层图像数据库。载于*CVPR*，2009年。

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., 等。《一张图像相当于 16x16 个词：用于大规模图像识别的 Transformer》。*ICLR*，2021。Du, S. S., Wang, Y., Zhai, X., Balakrishnan, S., Salakhutdinov, R., and Singh, A. 估计卷积神经网络或循环神经网络需要多少样本？*NeurIPS*，2018年。

Duan, Y., Chen, X., Houthooft, R., Schulman, J. 和 Abbeel, P. 连续控制深度强化学习基准测试。载于*ICML*，2016。

Florian, R. V. 倒立摆系统动力学的正确方程。认知与神经研究中心（*Coneural*），罗马尼亚，2007。

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. 尖锐度感知最小化：高效提升泛化能力。*arXiv* 预印本 *arXiv:2010.01412*，2020年。

Goodfellow, I. J., Bulatov, Y., Ibarz, J., Arnoud, S., and Shet, V. 利用深度卷积神经网络从街景图像中识别多位数字。*arXiv* 预印本 *arXiv:1312.6082*，2013。

Haussler, D. 量化归纳偏置：人工智能学习算法与 Valiant 学习框架。人工智能，36(2)：177–221，1988。

He, K., Zhang, X., Ren, S., and Sun, J. 用于图像识别的深度残差学习。发表于*CVPR*，2016。

Hern´andez-Orallo, J. 人工智能评估：从面向任务到面向能力的度量。人工智能评论，2016年8月。

Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Patwary, M. M.A., Yang, Y., and Zhou, Y. 深度学习缩放在经验上是可预测的。*arXiv* 预印本，2017年。

Huang, G., Liu, Z., Van Der Maaten, L., andWeinberger, K. Q. 密集连接卷积网络。载于*CVPR*，2017年。

Hutter, M. 学习曲线理论。*arXiv* 预印本 *arXiv:2102.04074*，2021年。

Jiang, Y.、Natekar, P.、Sharma, M.、Aithal, S. K.、Kashyap, D.、Subramanyam, N.、Lassance, C.、Roy, D. M.、Dziu- gaite, G. K.、Gunasekar, S. 等。《预测深度学习泛化能力的首次竞赛：方法与分析》。载于*NeurIPS 2020*年竞赛与演示赛道，2021。

Kakade, S. M.《强化学习的样本复杂度》。伦敦大学，伦敦大学学院（英国），2003年。

Karpathy, A. 在 ImageNet 上打破线性模型分类器。*Andrej Karpathy* 博客，2015年。

Kawaguchi, K., Deng, Z., Luh, K., and Huang, J. 鲁棒性通过数据依赖的泛化界意味着泛化。在*ICML*，2022年。

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. 关于深度学习的大批量训练：泛化差距与尖锐极小值。*ICLR*，2017。Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., and Houlsby, N. 大迁移（BiT）：通用视觉表征学习。在欧洲计算机视觉会议，2020。Kolmogorov, A. N. “信息”的定量定义的三种方法。信息传输问题，1965年。Koltchinskii, V. and Panchenko, D. Rademacher 过程与函数学习风险界定。《高维概率 *II*》，第443–457页，2000年。doi: 10.1007/978-1-4612-1358-1 29。Krizhevsky, A., Nair, V., and Hinton, G. CIFAR-10 和 CIFAR-100 数据集。2009年。网址: h t tps://www.cs.toronto.edu/kriz/cifar.html。Krizhevsky, A., Sutskever, I., and Hinton, G. E. 使用深度卷积神经网络进行 ImageNet 分类。在*NeurIPS*，2012年。Lake, B., Salakhutdinov, R., and Tenenbaum, J. 通过概率程序归纳实现人类水平的概念学习。《科学》，350：1332–1338，2015年。Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. 应用于文档识别的基于梯度的学习。*IEEE* 会议录，86(11)：2278–2324，1998年。doi：10.1109/5.726791。Lee, C.-Y., Xie, S., Gallagher, P., Zhang, Z., and Tu, Z. 深度监督网络。在人工智能与统计学国际会议，2015年。Lei, Y., Dogan, U., Binder, A., and Kloft, M. 多分类支持向量机：从更紧致的数据依赖泛化界到新算法。*NeurIPS*，2015年。

Li, H., Xu, Z., Taylor, G., Studer, C., 和 Goldstein, T. 可视化神经网络的损失景观。*NeurIPS*，2018年。

Li, Z.、Zhang, Y. 和 Arora, S. 为什么卷积网络比全连接网络更具样本效率？*ICLR*，2021。

Lin, Z.、Memisevic, R. 和 Konda, K. 不使用卷积，我们能走多远：改进全连接网络。*arXiv* 预印本，2015年。

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., 和 Vladu, A. 构建能够抵御对抗攻击的深度学习模型。*arXiv* 预印本 *arXiv:1706.06083*，2017。Mauch, L. 和 Yang, B. 一种用于简化全连接深度神经网络的新型逐层剪枝方法。发表于 *ICASSP*，2017。

McRae, A., Romberg, J., 和 Davenport, M. 流形回归的样本复杂度与有效维数。*NeurIPS*，2020年。

Mhaskar, H., Liao, Q., 和 Poggio, T. 深度网络何时以及为何优于浅层网络？发表于*AAAI*，2017。mrgrhn。使用 TensorFlow 的 AlexNet，2021。网址 h ttps://medium.com/swlh/alexnet-with-tensorflow-46f366559ce8。

Murata, N., Yoshizawa, S., 和 Amari, S.-i. 神经网络的学习曲线、模型选择与复杂度。发表于*NeurIPS*，1992年。

Negrea, J., Haghifam, M., Dziugaite, G. K., Khisti, A., 和 Roy, D. M. 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。*NeurIPS*，2019 年。Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., 和 Ng, A. 通过无监督特征学习识别自然图像中的数字。*NeurIPS*，2011。Nishimoto, B. E. 线性分类器，2018年。网址: h t t p s:// github.com/brunonishimoto/linear-classifier。Poggio, T., Kawaguchi, K., Liao, Q., Miranda, B., Rosasco, L., Boix, X., Hidary, J., 和 Mhaskar, H. 深度学习理论 III：解释不过拟合之谜。*arXiv* 预印本 *arXiv:1801.00173*，2017。Pope, P., Zhu, C., Abdelkader, A., Goldblum, M., 和 Goldstein, T. 图像的内在维度及其对学习的影响。*ICLR*，2021。

Raginsky, M., Rakhlin, A., Tsao, M., Wu, Y. 和 Xu, A. 学习算法稳定性与偏差的信息论分析。载于*IEEE* 信息论研讨会 *(ITW)*，2016。

S´anchez，J. 和 Perronnin，F. 用于大规模图像分类的高维签名压缩。发表于*CVPR*，2011。

Saxe，A. M.、Bansal，Y.、Dapello，J.、Advani，M.、Kolchinsky，A.、Tracey，B. D. 和 Cox，D. D. 深度学习的信息瓶颈理论。《统计力学杂志：理论与实验》，2019(12):124020，2019。Shamir，O.、Sabato，S. 和 Tishby，N. 信息瓶颈下的学习与泛化。发表于*ALT*，2008。Sharma，U. 和 Kaplan，J. 基于数据流形维度的缩放定律。机器学习研究期刊，23:9–1，2022。Shwartz-Ziv，R. 和 Tishby，N. 通过信息揭开深度神经网络的黑箱。*arXiv*预印本 *arXiv:1703.00810*，2017。Singh，S. 和 P´oczos，B. Wasserstein距离下的极小极大分布估计。*arXiv*预印本，2018。Stein，E. 和 Shakarchi，R. 傅里叶分析：导论。普林斯顿分析学讲义。普林斯顿大学出版社，2011。ISBN 9781400831234。Todorov，E.、Erez，T. 和 Tassa，Y. MuJoCo：用于基于模型的控制的物理引擎。发表于*IROS*，2012。Treves，F. 拓扑向量空间、分布与核：纯粹与应用数学，第*25*卷，第25卷。爱思唯尔，2016。Veeramacheneni，L.、Wolter，M.、Klein，R. 和 Garcke，J. 规范卷积神经网络。*arXiv*预印本 *arXiv:2206.01509*，2022。Wightman，R.、Touvron，H. 和 J´egou，H. ResNet卷土重来：改进的timm训练过程。*arXiv*预印本 *arXiv:2110.00476*，2021。Wolpert，D. H. 学习算法之间缺乏先验区别。神经计算，8(7):1341–1390，1996。Yin，D.、Kannan，R. 和 Bartlett，P. 对抗鲁棒泛化的Rademacher复杂度。发表于*ICML*，2019。Yu，T.、Quillen，D.、He，Z.、Julian，R.、Narayan，A.、Shively，H.、Bellathur，A.、Hausman，K.、Finn，C. 和 Levine，S. Meta-World：多任务与元强化学习的基准测试和评估。*CoRL*，2019。Zagoruyko，S. 和 Komodakis，N. 宽残差网络。*arXiv*预印本 *arXiv:1605.07146*，2016。Zhang，P.、Wang，H.、Naik，N.、Xiong，C. 和 Socher，R. DIME：人工智能数据集的信息论难度度量。*NeurIPSWorkshop DL-IG*，2020。

A. 我们的框架与常见学习设定之间的映射

本节将详细说明如何将各种学习设定映射到我们的框架中。我们首先介绍一套统一的符号来描述所有这些设定，然后逐一讨论每种设定。

我们将任务 定义为由实例 x 和动作 y组成；学习系统旨在学习从实例到动作的映射。我们假设每个实例 x 都由一组特征描述： x 具有 K_c 个连续特征 c_1， c_2， $...c_{K_c}$ 以及K_d 个离散特征 d_1， d_2， $...d_{K_d}$。最后，我们假设这些特征能够唯一描述所有可能的点 x：存在一个可逆函数 A，使得：

$$x = A(c_1, c_2, ...c_{K_c}, d_1, d_2, ...d_{K_d}) \tag{10}$$

对于 x 的所有可能取值以及它们各自流形上的特征。注意，使用单个特征 $c_1 = x$ 或 $d_1 = x$，总能实现对 x 的这种分解；结构更复杂的任务可能包含实例，这些实例还可进一步分解，以反映任务结构。

我们假设，对于每个实例 x，可能动作集合 $\mathcal{Y}$ 中都存在一个唯一的期望动作 $y \in \mathcal{Y}$；最优映射记为 $y = f^*(x)$。一般而言，学习精确的映射 f^* 可能难以实现。相反，在一定误差范围内紧密匹配函数 f^* 可能是可行的。给定损失函数 $\mathcal{L}$，我们将函数 f 在 x 上特定分布 p 下的误差定义为：

$$\hat{e}(f, p) = \mathbb{E}_p[\mathcal{L}(f(x), x)] \tag{11}$$

我们总结 概述该框架与各种学习场景之间的符号映射，见表 6。

监督分类 在监督分类设置中，实例可以视为由两个特征组成：一个离散特征 d_1，对应一个类别；以及一个连续特征 c_1，用于指定该类别中的某个具体样例：

$$x = A(c_1, d_1). \tag{12}$$

最优动作函数用于输出类别 $f^*(x) = d_1$。请注意，连续特征 c_1 与任务无关；此任务的目标是从训练数据中学习如何提取特征 d_1，同时保持对 c_1 的不变性。

举一个具体的例子，考虑 MNIST（Lecun等人，1998年）分类数据集；这里的标签对应于 d_1，而 c_1对应于指定每幅图像所需的所有其他信息（例如线条宽度、旋转角度和整体尺寸）。一般来说， c_1可能无法直接从图像中提取；估计 c_1 表示的一种方法是在 MNIST 图像上训练一个类别条件自编码器，并将每幅图像的潜在表示提取为 c_1。

用于少样本监督分类的元学习 在少样本元学习设置中，实例可视为由两个特征构成：一个连续特征 c_1 对应于一个相关类别集合，另一个连续特征 c_2对应于这些类别中的一组样本。

$$x = A(c_1, c_2) \tag{13}$$

目标是输出一个函数 g，使得 g 将类别集合中的输入映射到相应的标签： $f^*(x) = g$。注意， g 在此仅取决于 c_1；所需的函数 f^* 对 c_2具有不变性。通常，元学习系统会接收来自 c_1 取值子集的实例，并被要求泛化到同一分布内的另一个子集。为了使元学习系统能够将实例与未见过的 c_1 取值上的类别关联起来，会为每个类别提供少量带有未见过的 c_1的输入。

这类元学习问题的一个例子是 Omniglot（Lake 等人，2015年）数据集上的小样本分类。在这里， c_1对应一个字母表，而 c_2 对应该字母表中的一组样本图像。

强化学习 在强化学习设定中，实例可以视为由一个特征构成，该特征指定了实例中的所有信息。用强化学习的术语来说，实例通常对应于完全可观测马尔可夫决策过程（MDP）中环境的状态，或部分可观测 MDP 中的观测序列。描述实例的单个特征通常可以是离散的，也可以是连续的；为便于记述，当特征既可能是连续的也可能是离散的时，我们默认其为连续特征。

$$x = A(c_1) \tag{14}$$

这里， $f^*(x)$ 表示给定状态或观测序列 c_1 时应采取的最优动作。学习系统的目标是对未见过的状态或观测序列进行泛化。不过请注意，智能体无需对底层 MDP 的参数进行泛化；MDP 可以是固定的。

需要注意的是，我们的框架无法捕捉强化学习设置中的某些重要方面；例如，对于在线强化学习，它无法捕捉训练过程中观测到的实例可能取决于智能体先前动作这一事实。分析强化学习任务中的探索—利用权衡时，这种依赖关系非常重要。尽管如此，我们的框架足以评估在强化学习中的泛化难度。训练所依据的状态集或观测序列取决于智能体的行为这一事实，与任务的泛化难度关系有限。

例如，考虑 MuJoCo 套件（Florian，2007年）中实现的 Cartpole（Todorovet al.，2012年）连续控制任务。在此， c_1 对应于环境的观测状态，智能体经过训练后可据此提取最优动作。

无监督自编码 在无监督自编码中，最优动作函数 $f^*(x)$ 只是恒等函数 $f^*(x) = x$。目标是在观察一组训练实例后，在测试集上实现泛化。通常，这并非易事，因为学习系统受到约束，无法学习恒等函数。然而，我们强调，我们的度量与模型无关；在这种情况下，任务难度可以被视为实例集复杂度的度量。此类问题的一个具体例子是对 MNIST 数字进行无监督自编码，其中每个数字都具有单一的连续值特征。

元强化学习 在元强化学习设定中，实例对应于穿过环境的轨迹，这些轨迹由两个特征构成：一个用于参数化环境的连续特征 c_1 ，以及一个用于表示环境中特定轨迹的连续特征 c_2 。一般而言，这两个特征可以是离散的，也可以是连续的；此处为便于记述，我们采用连续特征：

$$x = A(c_1, c_2) \tag{15}$$

目标是输出一个策略 g ，使得 g 能够将环境中的状态映射到动作： $f^*(x) = g$。注意， g仅取决于 c_1；所需的函数 f^* 对于 c_2是不变的。学习系统需要从训练集中的 c_1 进行学习，并泛化到一组未见过的 c_1。

例如，考虑 Meta-World（Yu 等人，2019 年）元强化学习基准测试：在该基准中，控制机械臂的智能体必须从一组训练技能泛化到一组未见过的测试技能。这些技能包括推、转动、放置等简单操作。在我们的表述中， c_1 对应于技能，而 c_2 对应于机械臂执行该技能的特定轨迹。

B. 定理的证明 2

证明。 注意，对于任意 x_1, x_2, θ，我们有

$$\begin{aligned}
\mathcal{L}(f(x_1;\theta), x_1) &= \mathcal{L}\Big(f(x_2;\theta) + (f(x_1;\theta^*) - f(x_2;\theta^*)) \\
&\qquad + (f(x_1;\theta) - f(x_1;\theta^*) - f(x_2;\theta) + f(x_2;\theta^*)), x_1\Big) && (16)\\
&\le \mathcal{L}\Big(f(x_2;\theta) + (f(x_1;\theta^*) - f(x_2;\theta^*)), x_1\Big) \\
&\qquad + L_{\mathcal{L}}||f(x_1;\theta) - f(x_1;\theta^*) - f(x_2;\theta) + f(x_2;\theta^*)|| && (17)\\
&\le \mathcal{L}(f(x_2;\theta)\ x_2) \\
&\quad + L_{\mathcal{L}}\|f(x_1;\theta) - f(x_1;\theta^*) - f(x_2;\theta) + f(x_2;\theta^*)\| && (18)\\
&\le \mathcal{L}(f(x_2;\theta)\ x_2) + L_{\mathcal{L}}L_f d(x_1, x_2) d(\theta, \theta^*), && (19)
\end{aligned}$$

其中，(17) 是由 $\mathcal{L}$的 Lipschitz 条件推导得出的，(18) 是由 $\mathcal{L}$ 的平移不变性推导得出的，而 (19) 则是由我们对 f所施加的条件推导得出的。这使我们能够对差值进行界定。

$$e(\theta, p) - e(\theta, q) = \mathbb{E}_p[\mathcal{L}(f(x;\theta), x)] - \mathbb{E}_q[\mathcal{L}(f(x;\theta), x)] \tag{20}$$

作为 如下：设 γ 是 (x_1, x_2) 上的一个分布，其边缘分布分别为 p 和 q。Th en
fo

$$e(\theta, p) - e(\theta, q) = \mathbb{E}_\gamma[\mathcal{L}(f(x_1;\theta), x_1) - \mathcal{L}(f(x_2;\theta), x_2)] \tag{21}$$
$$\leq \mathbb{E}_\gamma[L_\mathcal{L} L_f d(x_1, x_2) d(\theta, \theta^*)]. \tag{22}$$

根据 Wasserstein距离 的定义，

$$e(\theta, p) - e(\theta, q) \leq L_\mathcal{L} L_f d(\theta, \theta^*) \inf_\gamma \mathbb{E}_\gamma[d(x_1, x_2)] = L_\mathcal{L} L_f d(\theta, \theta^*) W(p, q), \tag{23}$$

其中，下确界取遍所有边缘分布为 p, q 的分布 γ。因此，对于所有满足$d(\theta, \theta^*) \leq \frac{\varepsilon}{L_\mathcal{L} L_f W(p,q)}$的 $\theta \in \Theta_q$，都有 $e(\theta, p) \leq e(\theta, q) + \varepsilon = \varepsilon$（因为 $e(\theta, q) = 0$）。回想一下，我们假设 θ 上服从均匀先验。请注意，通常可以通过适当选择重新参数化来满足这一条件：如果我们在 θ 上的原始密度为$P(\theta)$，那么我们可以选择满足 $|det(\frac{\partial g(\theta)}{\partial \theta})| = P(\theta)$ 的新参数 $\tilde{\theta} = g(\theta)$，从而在 $\tilde{\theta}$ 上得到均匀密度。
利用 θ 上的均匀先验假设，可得

$$\tilde{I} = -\log \mathbb{P}(e(\theta, p) \leq \varepsilon \mid e(\theta, q) = 0) \tag{24}$$
$$\leq -\log \mathbb{P}\left(d(\theta, \theta^*) \leq \frac{\varepsilon}{L_\mathcal{L} L_f W(p, q)} \mid \theta \in \Theta_q\right), \tag{25}$$

正如所需。 □

C. 理论假设的有效性

为了得到公式（9）中任务难度的最终表达式，我们做出了一些假设。在本节中，我们将进一步讨论每个假设的有效性。

C.1. 损失函数的平移不变性

在许多监督分类和回归场景中，平方误差损失函数满足平移不变性假设。为清晰起见，在监督回归中，我们将输入表示为 x，将输出表示为 y。平方误差损失函数构造为 $\mathcal{L}(y, x) = ||y - f^*(x)||_2^2$，其具有平移不变性，因为

$$\mathcal{L}(y + f^*(x_1) - f^*(x_2), x_1) = ||y + f^*(x_1) - f^*(x_2) - f^*(x_1)||_2^2 = ||y - f^*(x_2)||_2^2 = \mathcal{L}(y, x_2). \tag{26}$$

接下来，在少样本元学习的设置中，存在一些相当宽泛的条件，在这些条件下平移不变性成立：实例对应于数据集，而动作对应于将输入映射到内循环任务输出的函数。为清晰起见，我们将内循环任务的输入和输出分别定义为 i 和 o；目标是学习一个从数据集 x 到函数y 的映射，这些函数将 $o = y$映射为i。此外，假设为内循环任务定义了一个损失函数 $L(o, i)$。那么，假设元损失函数定义为 $\mathcal{L}(y, x) = \mathbb{E}_{i \sim x}[L(y(i), i)]$。我们假设函数 y 具有如下意义上的可加性：对于任意 y_1, y_2, x，$(y_1 + y_2)(x) = y_1(x) + y_2(x)$。最后，我们假设内循环损失函数 L满足广义版本的平移不变性：

$$\mathbb{E}_{i \sim x_2}[L(y(i), i))] = \mathbb{E}_{i \sim x_1}[L(y(i) + f^*(x_2)(i) - f^*(x_1)(i), i)] \tag{27}$$

这进而意味着元损失 $\mathcal{L}$具有平移不变性：

$$\mathcal{L}(y, x_2) = \mathcal{L}(y + f^*(x_2) - f^*(x_1), x_1) \tag{28}$$

平移不变性假设意味着动作空间上定义了加法。我们认为，这种可加性适用于大多数学习问题。例如，在分类场景中，动作空间通常对应于所有类别上的逻辑值向量；这些动作空间是欧几里得空间，且动作可以相加。同样，在强化学习中，连续和离散动作空间通常都具有可加性；在离散情况下，输出的是可能动作的概率分布的逻辑值。此外，平移不变性是常用平方误差损失的一项性质，事实上，任何具有 $\mathcal{L}(y, x) = G(y - f^*(x))$ 形式的损失函数都具有此性质，其中 G 是某个函数。

与此同时，我们注意到，许多损失函数可能并不满足平移不变性。例如，错误率（在监督分类中，计算为被错误分类的样本点所占的比例）既不可微，也不具有平移不变性。因此，在训练过程中，通常会使用平方误差或交叉熵损失等可微代理损失函数，以达到类似的效果。换言之，对于不具有平移不变性的损失函数，可能存在具有平移不变性的代理函数。在强化学习中，损失函数可能对应于一个 Q 函数，该函数根据采取特定动作的条件，计算一个回合内的累积负奖励。在这种情况下，损失函数通常无法写成闭式形式，因此，即使这些函数实际上具有平移不变性，也很难证明它们满足平移不变性。

基于这些原因，我们认为平移不变性是一种有用的近似，使我们能够从理论上证明任务难度的性质。我们进一步指出，定理1的结论（我们利用平移不变性证明了该结论）即使对于不具有平移不变性的损失函数也可能成立：

$$\mathcal{L}(f(x_1;\theta),x_1)-\mathcal{L}(f(x_2;\theta),x_2)\leq L_{\mathcal{L}}L_f d(x_1,x_2)d(\theta,\theta^*) \tag{29}$$

直观地说，它表明，当参数接近其最优值时，输入 x 的微小变化只会导致损失发生微小变化，而这似乎是一项非常简单且通用的要求。

C.2. 模型的二阶条件

公式 (3) 中的二阶条件（$\|f(x_1;\theta_1)-f(x_1;\theta_2)-f(x_2;\theta_1)+f(x_2;\theta_2)\|\leq L_f d(x_1,x_2)d(\theta_1,\theta_2)$）假设，当输入和参数发生小幅变化时，模型 f 的变化可以被界定。首先请注意，如果 f 是二阶可微的，那么在假设 x 和 θ 定义在有界域上的情况下，总能选择某个 L_f 使该条件成立。举一个具体的例子，考虑使用 Tanh 激活神经网络进行图像分类的情形，其中每个输入像素的取值范围有界于 $[0,1]$，且网络参数被一个较大的半径所界定（有界参数这一假设是合理的，因为从实际角度来说，训练好的神经网络参数可以由某个半径界定，该半径取决于训练量）。

举例来说，ReLU 激活函数的神经网络可能不满足方程（3）。不过，已有研究发现，ReLU 网络的表现类似于采用平滑多项式激活函数的神经网络（Poggio 等人，2017）；因此，即使对于 ReLU 网络，采用我们的假设也可能是合理的。

C.3. 插值假设空间的形状

关于我们对插值假设空间的近似，首先需要指出，插值假设的维数低于假设空间的完整维数。沿任意方向对一个插值假设进行微小扰动，很可能会破坏插值。相反，我们预计存在某个子空间，沿该子空间进行扰动仍能保持插值（即存在一个由插值假设构成的连续流形）。这类似于神经网络中的情形：局部极小值往往位于平坦区域，在这些区域中，沿相应方向扰动参数不会显著影响损失（Keskar 等人，2017）。

我们将插值假设空间近似为整体假设空间中的一个圆盘（为清晰起见，我们已将术语从“球”改为“圆盘”）；该圆盘的维数低于整体假设空间，但嵌入在一个更高维的空间中。圆盘假设保留了我们可以合理预期插值假设空间具备的关键性质，即 1）低维性和 2）有界性，同时在解析上易于处理。

最后，我们强调，圆盘近似是为了近似插值假设空间的体积。插值假设空间的具体形状本身并不是我们结果的核心，体积的缩放规律才更为重要。我们认为，圆盘假设捕捉到了关键的缩放关系：体积对 b 和 d的依赖关系。

C.4. 实证验证瓦瑟斯坦距离近似

在本节中，我们通过经验验证第D节中对 Wasserstein距离的近似。回顾一下，我们将均匀分布 p（定义在半径为4 的球面上）与由以下数量q 个 n 在球面上随机采样的点构成的经验分布之间的 Wasserstein距离近似为：

$$W(p,q)\in O(rn^{-1/m}) \tag{30}$$

请注意，该表达式随 r 线性增长，随 n次线性衰减； m 越大，衰减越慢，因为可能需要更多点才能使所有区域都位于任意点的固定半径范围内。从理论上讲，我们可以预期，对于较大的 n，这一近似仍然成立，从而实现近乎均匀的 q。如果这种缩放关系成立，那么我们预期 log $W(p,\ q)$ 会随 log n 线性下降（斜率为 $-1/m$）。类似地，我们预期 log $W(p,\ q)$ 会随 $1/m$线性下降（斜率为 –log n）。

我们在单位球面上（$r=1$）进行实验。我们通过在单位球面上采样 10000 个点来近似均匀分布，并计算 10000 个点的经验分布与 $n<10000$ 个点的经验分布之间的精确 Wasserstein距离（我们将 n 从 10 变化到 500）。对于不同的 m 取值计算距离，其中 m 从 3 变化到 19。

如图4所示，我们发现log $W(p,\ q)$ 与log n 之间呈线性关系，且正如预期，较小的 m对应更陡的斜率。我们还观察到log $W(p,\ q)$ 与log m 之间呈非线性关系，对于较大的 m，距离大于线性趋势线所预期的值。尽管如此，线性趋势线仍能很好地拟合经验数据。

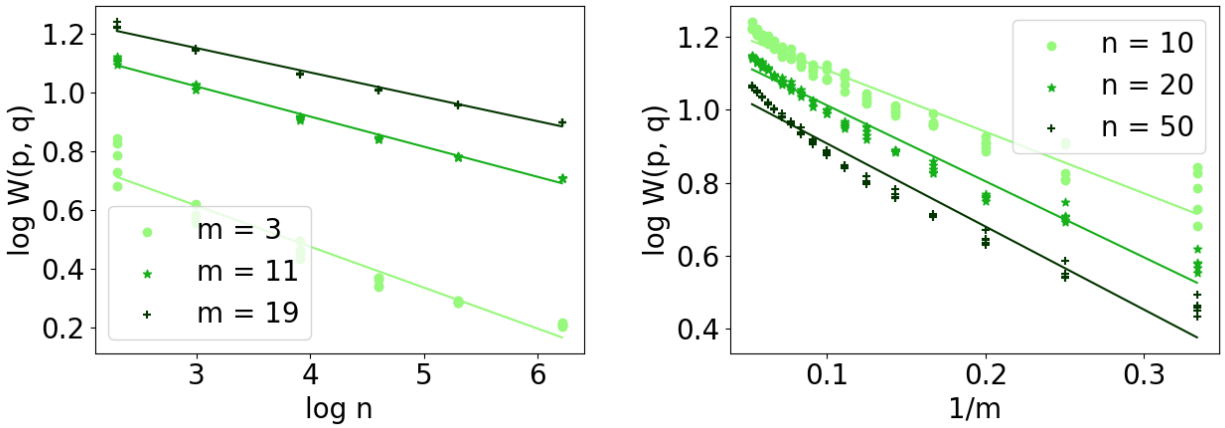


图4：单位球面上的均匀分布与经验采样均匀分布之间的 Wasserstein距离估计值，随球面维数 m 和采样点数量 n 的变化。每种设置展示五次试验的结果，并对结果拟合线性趋势线。

D. 经验近似任务难度

在第 3.3 节中，我们介绍了用于监督分类的归纳偏置复杂度近似方法。这里，我们将进一步说明如何将该近似方法应用于其他情境并进行计算。

首先，我们写出任务难度近似值的一个略微修改版本

$$\tilde{I} \approx (2dzE - nd)\left(\log b + \frac{1}{2}\log z + \log d + \log K - \frac{1}{m}\log n - \log\frac{\varepsilon}{L_{\mathcal{L}}} + \log c\right). \tag{31}$$

其中 $c = \sqrt[m]{\frac{6}{}}\left(\frac{2\pi^{(m+1)/2}z()}{\Gamma\ ^{m+1}_{2}}\right)^{1/m}$，且 K和 E是 m和最大频率 $M=2\pi r/\delta$的函数。在监督分类场景中，$z=d$。

一般而言，控制任务难度的参数（m、 n、 ε、 $L_{\mathcal{L}}$、 r、 b、 d、 z）可以解释如下： m 是实例所在流形的内在维度。 n 是训练点的数量。 ε 是该任务的目标误差率。 $L_{\mathcal{L}}$ 是损失函数的利普希茨常数；对于没有显式可微损失函数的任务（例如强化学习），它可以解释为损失相对于模型输出的整体敏感度。 r和 b 分别是模型输入和模型输出大小的上界。 d 对应于模型输出的维数。 z 对应于实例所在子流形的数量（每种离散特征组合对应一个子流形，

实例对应于一个子流形）。在监督分类场景中，子流形的数量就是类别数量，因此在这些场景中，z 设为等于 d。

为了计算任务的难度 I，我们必须确定这些参数的取值。对于所有任务，我们将 $\varepsilon/L_{\mathcal{L}}$设为目标性能水平，即输出空间中对应于 ε误差的距离。对于每个任务，我们根据任务类别设定一个固定的目标性能水平（即，对于图像分类任务，固定错误率为 1.0%；对于强化学习任务，固定错误率为 0.001）；参见表3。然而，由于文献中在 MNIST/ImageNet 上训练的典型高性能模型所达到的错误率明显优于/劣于 1.0%，我们分别将 MNIST 和 ImageNet 的目标错误率调整为 0.1% 和 10.0%。这样做是为了在这些任务通常采用的错误率范围内，恰当地量化这些任务的难度。不过，我们的任务难度数值对目标性能水平的精确选择相对不敏感：对所有任务使用固定错误率 1.0% 时，我们发现任务难度的尺度并未发生显著变化。具体而言，MNIST 的任务难度从 1.09×10^{16} 比特逐步下降至 9.03×10^{15} 比特，而 ImageNet 的任务难度则从 2.48×10^{43} 比特逐步上升至 2.81×10^{43} 比特。

表3：用于计算归纳偏置复杂度度量的不同任务的目标错误率。对于每项任务，目标错误率设定为文献中提出的典型高性能模型在该任务上训练后所达到的性能。对于分类任务，错误率对应于测试集准确率。对于 Cartpole，错误率对应于任意给定时间步执行错误动作的概率。对于具有连续动作空间的 MuJoCo 连续控制任务，错误率对应于与最优动作之间的期望距离。

数量	MNIST	SVHN	CIFAR-10	ImageNet	Omniglot	Cartpole	Reacher	Hopper	Half-Cheetah
错误率	0.1 %	1.0 %	1.0 %	10.0 %	1.0 %	0.001	0.001	0.001	0.001

对于某些任务（例如 MuJoCo（ m Du 等人，Du 等人，2018年）任务），内在维度Pope 等人（2021）中所述的方法已知内在维度；对于其他任务，则必须对其进行估算；我们采用Pope 等人(2021)中描述的最近邻估算方法。空间分辨率 δ 根据各任务的性质设定。对于分类任务，将空间分辨率设为类别间的间隔是合理的。对于强化学习任务，将其设为一个尺度，在该尺度下，大小为 δ 的扰动所产生的轨迹不太可能存在实质性差异。对于所有任务，我们将 r 近似为训练数据点的最大范数。

D.1. 通过实证计算分类任务难度

对于图像分类任务，模型输出是类别上的概率分布，因此 b 为 1。我们将 δ 设为不同类别点之间的最小距离，因为这是模型必须能够区分的输入之间的最小距离。对于无法实际精确计算 f 的大型数据集，我们使用统计方法对其进行估计。附录E中对此过程进行了说明。

我们估计 m 时采用Pope 等人(2021)所述的技术，该技术基于到最近邻的距离给出最大似然估计。我们在估计中使用 $k = 5$ 个最近邻。对于无法实际计算所有点最近邻的大型数据集，我们采用Pope 等人(2021)提出的锚点近似方法：随机抽取一部分点作为锚点，并在整个数据集中寻找它们的最近邻。为了估计 m 在 ImageNet 上的值，我们随机选取 2000年 个点作为锚点。表4 列出了我们对 m 和 δ 在图像分类基准测试上的估计值。

表 4：各种图像分类基准测试的估计 m 和 δ。这些估计值用于估算解决每项任务所需的归纳偏置程度。

数量	MNIST	SVHN	CIFAR-10	ImageNet
m	14	19	27	48
δ	2.4	1.6	2.8	65

D.2. 通过实证计算元学习任务难度

Omniglot 单样本分类任务是：给定 20 个字母中每个字母的一张示例图像，识别一张图像所属的类别。因此，元学习任务是学习一个函数 f，将包含 20 张图像的集合映射为一个函数 g，该函数可以将单张图像映射为 20 个类别上的概率分布。在接下来的讨论中，对于诸如 z、d 和 E 这样的量，下标 f 将

用于表示我们正在考虑函数 f 的这些值，下标 g 将用于表示我们正在考虑函数 g 的这些值。

在我们的框架中，f 的输入没有离散特征，因此 $z_f = 1$。由于 g 是 f的输出，d_f 等于 g参数化的维数，即 $2z_g d_g E_g$。函数 g 将图像分为20个类别，因此 $z_g = 20$ 和$d_g = 19$。我们可以将 E_g 计算为 $M_g = 2\pi r_g/\delta_g$ 和 m_g 的函数。与图像分类任务一样，我们将 r_g 设为数据集中图像的最大范数，将 δ_g 设为不同类别图像之间的最小距离。我们将 m_g 设为从单个字母表中抽取的图像的维数 m_0。具体做法是估算数据集中每个字母表的维数，并对结果取平均值。

由于 f 的输入是20张图像，$r_f = r_g\sqrt{20}$。要计算 m_f，令 $m_1 > m_0$ 表示从整个数据集中抽取的图像的维数。那么，f输入所依据的字母表的维数为 $m_1 - m_0$，图像本身的维数为 $20m_0$。因此，$m_f = m_1 + 19m_0$。我们希望 f 对20张图像中仅一张图像的变化保持敏感，因此设定 $\delta_f = \delta_g$。

训练数据集由多个字母表组成。我们用 ℓ_a 表示字母表 a中的字母数量。我们将 n 设定为可以从训练数据集中抽取的、由 20 张图像组成的集合数量，这些图像代表同一字母表中的 20 个不同字母。由于训练数据集包含代表每个字母的 20 张图像，

$$n = \sum_a \binom{\max\{\ell_a, 20\}}{20} \cdot 20^{20}, \tag{32}$$

其中，求和遍历所有训练字母表 a。

最后，b_f 是 f输出的最大范数，其中包含 g的参数化。我们发现，这可以由 $b_g\sqrt{z_g d_g}$界定。我们已经计算出 z_g 和 d_g，并且由于 g 输出的是概率分布，因此 $b_g = 1$。然后，我们使用数值 z_f、d_f、m_f、δ_f、n、b_f和最先进的错误率 $\varepsilon/L_{\mathcal{L}}$来计算任务难度。

D.3. 经验计算强化学习任务难度

对于强化学习任务的任务难度评估，我们采用了一些不同于监督分类的近似方法，这些方法更适合我们的特定场景；特别是，我们对实例所处的流形采用了不同的假设，并对 Wasserstein距离 $W(p, q)$采用了不同的近似方法。

首先，我们写出任务难度表达式的一般形式，不采用第 3.3 节中的 Wasserstein 距离近似：

$$\tilde{I} \approx (2dzE - nd)\left(\log b + \frac{1}{2}\log z + \log d + \log K - \log r + \log W(p, q) - \log\frac{\varepsilon}{L_{\mathcal{L}}}\right). \tag{33}$$

在我们评估的强化学习任务中，$z = 1$，因此可以写为：

$$\tilde{I} \approx (2dE - nd)\left(\log b + \log d + \log K - \log r + \log W(p, q) - \log\frac{\varepsilon}{L_{\mathcal{L}}}\right). \tag{34}$$

f的输入是一个 m 维观测，缩放后可视为超立方体 $[-\pi, \pi]^m$ 中的一个点。将 f 参数化为拉普拉斯-贝尔特拉米算子的特征函数之和，

$$f(x; \theta) = \sum_{p\in\mathbb{Z}^m} \theta_p e^{ip^\top x} \tag{35}$$

（Treves，2016）。对应于 p 的分量波长为 $2\pi/\|p\|$，因此我们将参数化限制为满足 $\|p\| \le 2\pi/\delta$ 的分量。于是，我们有 $L_f = 2\pi/\delta$。特征函数 E 的数量等于满足条件 $p \in \mathbb{Z}^m$ 的向量 $\|p\| \le 2\pi/\delta$的数量，我们将其近似为半径为 m的 $2\pi/\delta$ 维球的体积，得到

$$E \approx \left(\frac{2\pi}{\delta}\right)^m V_m, \tag{36}$$

其中，$V_m = \frac{\pi^{m/2}}{\Gamma(\frac{m}{2}+1)}$ 表示半径为 1 的 m 维球体的体积。现在我们求 Wasserstein 距离 $W(p, q)$。在最优传输中，每个训练数据点都与超立方体中一个大小相等的区域相关联。我们对其进行近似。

将这些区域视为边长为 $s = 2\pi n^{-1/m}$ 的超立方体。在这些超立方体中随机选取两点，其期望距离至多为 $s\sqrt{m/6}$ (Anderssen 等人，1976年)，因此

$$W(p,q) \approx \sqrt{\frac{m}{6}} \cdot 2\pi n^{-1/m}. \tag{37}$$

最后，我们注意到，f的输出在离散情形下是 $\{-1,+1\}$ 中的一个力，在连续情形下则是 $[-1,1]^d$ 中的一个元素。因此，在所有情形下，我们都可以设定 $b = \sqrt{d}$。利用这些值，我们得到

$$\tilde{I} \approx d\left(2\left(\frac{2\pi}{\delta}\right)^m V_m - n\right)\left(\log\frac{4\pi^2}{\sqrt{6}} + \log d + \frac{1}{2}\log m - \log\delta - \frac{1}{m}\log n - \log\frac{\varepsilon}{L_{\mathcal{L}}}\right). \tag{38}$$

回顾一下，d 是输出空间的维数，而 m 是观测空间的维数。在带噪声的 Cartpole 任务中，可能需要 $T > 1$ 个观测才能确定最优动作。每个观测都是一个二维状态（包含杆的角位置和速度），因此我们设定 $m = 2T$。对于完全观测任务，请注意 $T = 1$。

剩下要为参数 n、δ和 $\varepsilon/L_{\mathcal{L}}$选择取值。对于所有任务，我们为 δ 选择相对较小的值 0.001，并令 $\varepsilon/L_{\mathcal{L}}$。我们为 δ 选择较小的值，因为初始条件的微小变化可能会随时间推移导致显著变化；我们为 $\varepsilon/L_{\mathcal{L}}$ 选择较小的值，因为智能体应能够执行接近最优的动作来实现任务目标。在带噪声的 Cartpole 任务中，我们设定 $n = 10000$，相当于观测 100 个回合，每个回合包含 100 个时间步。在 MuJoCo 环境中，我们设定 $n = 1000000$，相当于观测 1000 个回合，每个回合包含 1000 个时间步。我们强调，影响任务难度的最重要因素是 m，它由任务规格决定。

E. 分类任务中 δ 的估计

如果我们无法精确计算 δ，就使用极值理论对其进行估计（Dekkers 等人，1989年）。

我们希望找到一个分布的最大值，在本例中，该分布是来自不同类别的随机点对之间距离的倒数的分布。极值分布由极值指数 γ刻画。设 X_1，……，X_n 为从该分布中抽取的值，并设$X_{1,n} \le \cdots \le X_{n,n}$ 为这些值按升序排列后的结果。对于固定的 $k < n$，定义

$$M_n^{(r)} = \frac{1}{k}\sum_{i=0}^{k-1}(\log X_{n-i,n} - \log X_{n-k,n})^r. \tag{39}$$

于是，我们可以将 γ 估计为

$$\hat{\gamma}_n = M_n^{(1)} + 1 - \frac{1}{2}\left(1 - \frac{(M_n^{(1)})^2}{M_n^{(2)}}\right)^{-1}. \tag{40}$$

根据经验，我们通常会发现$\hat{\gamma}_n > 0$。在这种情况下，我们通过估计分布中位于顶部 $1/P$的部分所对应的分位数，来估计 δ；其中 P是来自不同类别的训练数据点对的数量。假设训练数据点在各类别间均匀分布，我们将 P近似为 $n^2(1-1/C)$，其中 C 是类别数量。如果 $a_n = kP/n$，则可得到以下估计：

$$\frac{a_n^{\hat{\gamma}_n} - 1}{\hat{\gamma}_n} \cdot X_{n-k,n} M_n^{(1)} + X_{n-k,n} \tag{41}$$

对于 $1/\delta$。

理论结果（Dekkers 等人，1989年）要求 k 和 n/k 随着 $n \to \infty$趋于无穷。对于 ImageNet，我们选择$n = 40000$和 $k = 200$。为减少估计中的噪声，我们对 $1/\delta$ 计算 10 次估计，并对结果取平均，以获得 $1/\delta$的最终估计。

F. 额外任务变体

在本节中，为了进一步直观理解我们关于归纳偏置复杂度的实证结果，我们计算了任务的其他变化形式的归纳偏置复杂度。

F.1. 任务组合

设置 我们首先考察任务组合的归纳偏置复杂度。假设给定两个任务，分别对应于实例 x_1 到 $y_1 = f_1^*(x_1)$ 以及 x_2 到 $y_2 = f_2^*(x_2)$ 之间的映射。此外，假设训练分布和测试分布 q_1、p_1 以及 p_2、q_2 均已给定。随后，我们将这两个任务组合起来，构造一个从 (x_1, x_2) 到 $f^*(x_1, x_2) = (f_1^*(x_1), f_2^*(x_2))$ 的映射。新任务的实例测试分布构造如下：

$$p(x_1, x_2) = p_1(x_1)p_2(x_2) \tag{42}$$

并且训练分布构造如下：

$$q(x_1, x_2) = q_1(x_1)q_2(x_2) \tag{43}$$

注意，如果 q_1 和 q_2 分别对应大小为 n_1 和 n_2 的训练集，那么 q 对应大小为 $n_1 n_2$ 的训练集。我们假设这两个任务分别具有损失函数 $\mathcal{L}_1$ 和 $\mathcal{L}_2$。在本节的分析中，我们假设这些损失函数满足 $\mathcal{L}_i(y, x) \geq 0$，且当且仅当 $y = f_i^*(x)$（对于 $i = 1, 2$）时等号成立。我们为组合任务构造损失函数 $\mathcal{L}$ 如下：

$$\mathcal{L}((y_1, y_2), (x_1, x_2)) = \alpha\mathcal{L}_1(y_1, x_1) + (1 - \alpha)\mathcal{L}_2(y_2, x_2) \tag{44}$$

对于参数 α in (0,1)。我们将假设表示为 θ，其参数化了一个函数，该函数输入 (x_1, x_2)，并输出每个任务的预测结果：$f(x_1, x_2; \theta) = (f_1(x_1, x_2; \theta), f_2(x_1, x_2; \theta))$。然后，特定假设 θ 在分布 p 下的泛化误差（q 亦同）为：

$$\begin{aligned} e(\theta, p) = \hat{e}(f(\cdot; \theta), p) &= \mathbb{E}_p[\mathcal{L}(f(x_1, x_2; \theta), (x_1, x_2))] \\ &= \alpha\mathbb{E}_p[\mathcal{L}_1(f_1(x_1, x_2), x_1)] + (1 - \alpha)\mathbb{E}_p[\mathcal{L}_2(f_2(x_1, x_2), x_2)] \end{aligned} \tag{45}$$

为便于记号表示，我们还将 $e_i(\theta, p)$ 定义为（q 也进行类似定义）：

$$e_i(\theta, p) = \mathbb{E}_p[\mathcal{L}_i(f_i(x_1, x_2; \theta), x_i)] \tag{46}$$

对于 $i = 1, 2$。因此，我们可以将 $e(\theta, p)$ 表示为：

$$e(\theta, p) = \alpha e_1(\theta, p) + (1 - \alpha)e_2(\theta, p) \tag{47}$$

请注意，$e_1(\theta, p)$ 和 $e_2(\theta, p)$ 并不直接对应于假设 θ 在任务 1 和任务 2 上的误差，因为 θ 参数化了一个同时将 x_1 和 x_2 作为输入的函数。相反，$e_1(\theta, p)$（$e_2(\theta, p)$ 也同理）对应于任务 1 的一个变体的误差：在该变体中，实例 x_1 增加了一个与任务无关的干扰项输入 x_2（生成实例 (x_1, x_2)），但任务输出 y_1 的构造方式与 $y_1 = f_1^*(x_1)$ 相同。为明确起见，我们将此映射定义为 $y_1 = f_1^*(x_1, x_2)$（任务 2 的变体也同理）。需要特别注意的是，任务 1 的这一变体具有更大的内在维度输入。

现在，我们考虑与 f^* 对应的组合任务的归纳偏置复杂度。回顾一下，我们将归纳偏置复杂度定义为：

$$\tilde{I} = -\log \mathbb{P}(e(\theta, p) \leq \varepsilon \mid e(\theta, q) = 0) \tag{48}$$

我们将这种归纳偏置复杂度与分别对应于 f_1^* 和 f_2^* 的任务的归纳偏置复杂度联系起来：

$$\tilde{I}_i = -\log \mathbb{P}(e_i(\theta, p) \leq \varepsilon \mid e_i(\theta, q) = 0) \tag{49}$$

其中，I_i 对应于与 f_i^* 对应的任务的归纳偏置复杂度。在某些条件独立性假设下，我们能够建立 I 与 I_i 之间的联系。具体而言，我们假设，如果 θ 对任务 1 进行插值，那么再对任务 2 进行插值不会改变 θ 在任务 1 上实现泛化的概率（反之亦然）。我们还假设，如果 θ 同时对任务 1 和任务 2 进行插值，那么在任务 1 上实现泛化不会影响 θ 在任务 2 上实现泛化的概率（反之亦然）。如果我们认为任务 1 和任务 2 是独立构建的，这一假设是合理的；这里的独立是指输入 x_2 不会提供关于 $f_1^*(x_1, x_2)$ 的信息，反之亦然；能够在一个任务上进行插值或实现泛化，不会影响另一个任务的泛化难易程度。因此，我们可以作出如下陈述：

定理 3。 假设满足以下条件独立性：

$$e_1(\theta,p)\le\varepsilon \perp\!\!\!\perp e_2(\theta,q)=0|e_1(\theta,q)=0 \tag{50}$$

$$e_2(\theta,p)\le\varepsilon \perp\!\!\!\perp e_1(\theta,q)=0|e_2(\theta,q)=0 \tag{51}$$

$$e_1(\theta,p)\le\varepsilon \perp\!\!\!\perp e_2(\theta,p)\le\varepsilon|e_1(\theta,q)=0,e_2(\theta,q)=0 \tag{52}$$

然后，

$$-\log(e^{-\tilde{I}_1}+e^{-\tilde{I}_2})\le\tilde{I}\le\tilde{I}_1+\tilde{I}_2 \tag{53}$$

证明。 首先，根据条件独立性：

$$\begin{aligned}&\mathbb{P}(e_1(\theta,p)\le\varepsilon\mid e(\theta,q)=0)\mathbb{P}(e_2(\theta,p)\le\varepsilon\mid e(\theta,q)=0)\\&\qquad=\mathbb{P}(e_1(\theta,p)\le\varepsilon\wedge e_2(\theta,p)\le\varepsilon\mid e(\theta,q)=0)\end{aligned} \tag{54}$$

请注意，$e(\theta,p)\le\varepsilon \impliedby e_1(\theta,p)\le\varepsilon\wedge e_2(\theta,p)\le\varepsilon$。这意味着：

$$\begin{aligned}&\mathbb{P}(e(\theta,p)\le\varepsilon\mid e(\theta,q)=0)\ge\mathbb{P}(e_1(\theta,p)\le\varepsilon\wedge e_2(\theta,p)\le\varepsilon\mid e(\theta,q)=0)\\&\qquad=\mathbb{P}(e_1(\theta,p)\le\varepsilon\mid e(\theta,q)=0)\mathbb{P}(e_2(\theta,p)\le\varepsilon\mid e(\theta,q)=0)\end{aligned} \tag{55}$$

此外，请注意，$e(\theta,p)\le\varepsilon \implies e_1(\theta,p)\le\varepsilon\vee e_2(\theta,p)\le\varepsilon$。这意味着：

$$\mathbb{P}(e_1(\theta,p)\le\varepsilon\mid e(\theta,q)=0)+\mathbb{P}(e_2(\theta,p)\le\varepsilon\mid e(\theta,q)=0)\ge\mathbb{P}(e(\theta,p)\le\varepsilon\mid e(\theta,q)=0) \tag{56}$$

接下来，使用我们的前两个条件独立性假设：

$$\mathbb{P}(e_1(\theta,p)\le\varepsilon\mid e_1(\theta,q)=0)\mathbb{P}(e_2(\theta,p)\le\varepsilon\mid e_2(\theta,q)=0)\le\mathbb{P}(e(\theta,p)\le\varepsilon\mid e(\theta,q)=0) \tag{57}$$

$$\mathbb{P}(e_1(\theta,p)\le\varepsilon\mid e_1(\theta,q)=0)+\mathbb{P}(e_2(\theta,p)\le\varepsilon\mid e_2(\theta,q)=0)\ge\mathbb{P}(e(\theta,p)\le\varepsilon\mid e(\theta,q)=0) \tag{58}$$

最后，对两边取负对数：

$$-\log(e^{-\tilde{I}_1}+e^{-\tilde{I}_2})\le\tilde{I}\le\tilde{I}_1+\tilde{I}_2 \tag{59}$$

正如所需。 □

注意，$-\log(e^{-\tilde{I}_1}+e^{-\tilde{I}_2})\approx\min\{\tilde{I}_1,\tilde{I}_2\}$。因此，直观而言，该陈述表示，组合任务所需的归纳偏置至多为分别处理这两个任务所需归纳偏置的总和，且至少需要较容易的~任务所需的归纳偏置。我们再次强调，I_i 并不对应于解决任务 i所需的归纳偏置，而是解决任务 i 的某个变体所需的归纳偏置：该变体在来自另一个任务的实例中加入了一个干扰项。

实验 接下来，我们通过实证计算图像分类任务组合的归纳偏置复杂度。为此，我们计算每个任务的一个变体的任务难度：在每个实例中附加来自另一个任务、与任务无关的干扰项（如上文所述，$\tilde{I}_1$ 或 $\tilde{I}_2$）。然后，我们报告 I的上界 $\tilde{I}_1+\tilde{I}_2$和下界 $-\log(e^{-\tilde{I}_1}+e^{-\tilde{I}_2})$~。

为了计算带有干扰项附加实例的任务的任务难度，我们首先重新审视公式（31）中对归纳偏置复杂度的广义近似：

$$\tilde{I}\approx(2dzE-nd)^{1/m}\left(\log b+\frac{1}{2}\log z+\log d+\log K-\frac{1}{m}\log n-\log\frac{\varepsilon}{L_{\mathcal{L}}}+\log c\right). \tag{60}$$

其中 $c=\sqrt[m]{6}\left(\frac{2\pi^{(m+1)/2}}{\Gamma\left(\frac{m+1}{2}\right)}\right)^{-1}$，当内在维度 m较大时，该值大致保持不变。回顾一下，d是输出维数，z是输入数据所处的流形数量，n 是训练数据点数量，ε 是目标误差，$L_{\mathcal{L}}$ 是损失函数的利普希茨常数，而 K 和 E 是 m 和最大频率$M=2\pi r/\delta$的函数。对于分类问题，通常将 $z=d$ 设为类别数量。

当我们在某个任务的实例中添加与任务无关的干扰项时，五个关键参数会发生变化：训练点数量 n、内在维度 m、流形数量 z、模型输出界限 b 以及模型输入界限 r。重要的是，所有其他参数都保持不变（即 δ、d、ε 和 $L_{\mathcal{C}}$）。训练点数量随不同干扰项实例的数量而变化；如果将任务 2 中的实例（其训练点数量为 n_2）添加到训练点数量为 n_1 的任务 1 中，那么新任务将有 $n_1 n_2$ 个训练点。此外，该任务新的内在维度等于各个任务内在维度之和，因为 (x_1, x_2) 的组合流形是 x_1 和 x_2 各自流形的乘积。因此，如果任务 1 的内在维度为 m_1，任务 2 的内在维度为 m_2，那么增强后任务的内在维度为 $m_1 + m_2$。组合任务输入的流形数量等于对应于 (x_1, x_2) 的流形数量；如果任务 1 和任务 2 的类别数量分别为 z_1 和 z_2，那么该数量就是 $z_1 z_2$。如果任务 1 和任务 2 的模型输出界限分别为 b_1 和 b_2，那么组合模型的界限为 $\sqrt{b_1^2 + b_2^2}$。类似地，r 的新值为 $\sqrt{r_1^2 + r_2^2}$。然后，我们可以使用上述公式计算添加干扰项后的任务难度。

我们计算了 MNIST、SVHN 和 CIFAR-10 两两组合后的任务难度（各自的任务难度以比特计，分别为：1×10^{16}、1×10^{31}、3×10^{32}）。如表5所示，组合任务的难度远高于各个单独任务的难度。一个非常粗略的经验法则是，组合任务的难度约等于各单独任务难度的乘积。这是合理的，因为任务难度随数据的内在维度呈指数增长，而将两个任务组合起来，会将两个任务的内在维度相加（因此其任务难度相乘）。

表5：图像分类任务组合的任务难度界（以比特计）。难度以“下界 / 上界”的形式报告。MNIST、SVHN 和 CIFAR-10 各自的任务难度（以比特计）分别为：$1 \times 10^{16}, 1 \times 10^{31}, 3 \times 10^{32}$。

	MNIST	SVHN	CIFAR-10
MNIST	1×10^{26} / 3×10^{26}	6×10^{39} / 5×10^{45}	3×10^{42} / 8×10^{44}
SVHN	-	4×10^{54} / 8×10^{54}	4×10^{50} / 4×10^{61}
CIFAR-10	-	-	2×10^{55} / 3×10^{55}

为什么将两个简单任务（如 MNIST）组合在一起，会产生一个难度大得多的任务？这是因为，要将组合任务的两个部分区分开来，需要大量的归纳偏置。如果没有这种归纳偏置，求解两个任务的组合就如同求解一个具有更高维度输入的单一任务，而这相应地需要强得多的归纳偏置。这也揭示了神经网络等实用模型类别可能如何为任务提供如此大量的归纳偏置：具体而言，就是将输入分解为维度更低的分量。

F.2. 预测零输出

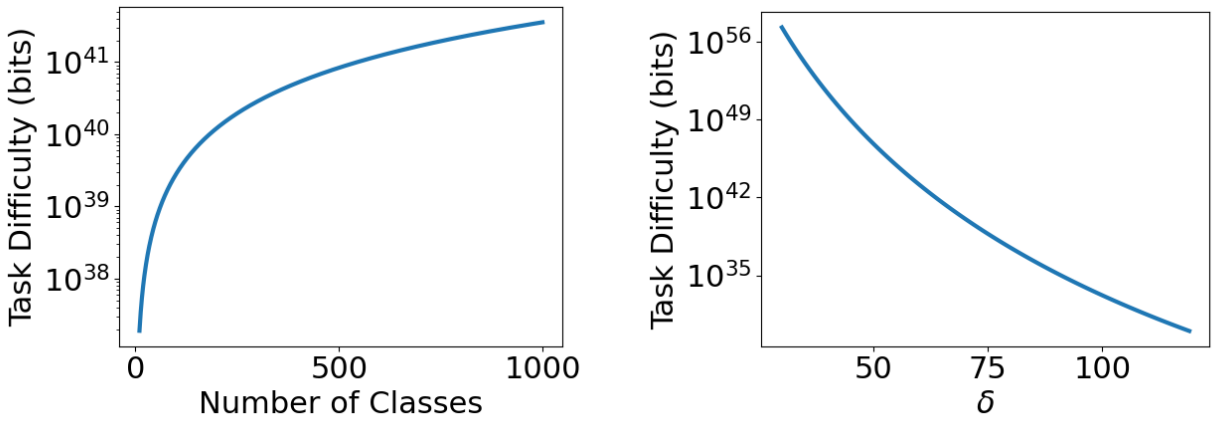
为了直观了解任务难度的性质，我们计算一个任务的任务难度；在该任务中，目标函数的输出始终为 0：$f^*(x)$。为计算任务难度，我们回到定义 3.1 中归纳偏置复杂度的定义：3.1：

$$\tilde{I} = -\log \mathbb{P}(e(\theta, p) \leq \varepsilon \mid e(\theta, q) \leq \epsilon) \tag{61}$$

与本文始终采用的做法一致，固定 $\epsilon = 0$。可以看到，与一般情况一样，归纳偏置复杂度基于能够良好泛化的插值假设所占的比例。对于零输出的情况，良好泛化意味着 $f(\cdot; \theta)$ 在分布 p 上应足够接近 0。请注意，插值假设在分布 p 上的输出未必接近零；因此，解决零输出任务可能需要相当程度的归纳偏置。

确定零输出任务难度的关键在于设定假设空间。回想一下，在第 3.3 节中，我们构造的假设空间由一组基函数的线性组合构成，这些基函数对与任务相关的分辨率（或更大的分辨率）敏感。对于零输出，所有分辨率都与任务无关：真实输出 $f^*(x)$ 对输入的任何变化都不敏感。因此，如果我们采用第 3.3 节的方法来构造假设，就会发现假设空间中只有一个假设，即 f_*。由于 f^* 能够插值训练数据并实现完美泛化，因此在已知插值的条件下实现泛化的概率为1。因此，任务难度为 0。

然而，t 的其他取值　　基础假设空间可能导致极大的任务难度。考虑一个 ImageN　　et



（a）改变类别数量
在 ImageNet 上

（b）改变空间分辨率
在 ImageNet 上

图 5：基准任务参数变化下的任务难度。

版本，其中所有输入都映射到一个 1000 维零向量。如果我们按照构建原始 ImageNet 任务基础假设空间的相同方式来构建，并假设我们希望达到与原始 ImageNet 任务相同的泛化误差目标，那么零输出 ImageNet 的任务难度近似值将与原始 ImageNet 相同： 3.55×10^{41} 比特。

即使目标函数变得简单得多，泛化 ImageNet 所需的归纳偏置为何仍能保持不变（假设假设空间保持不变）？这是因为，归纳偏置复杂度对应于从插值假设中指定一个能够泛化的假设集合的难度。由于基础假设空间规模庞大，可以预期，对于真实的 ImageNet 目标函数和零目标函数，插值假设空间的规模是相同的：我们可以预期，满足零目标训练样本的假设比例，与满足真实 ImageNet 训练样本的假设比例相同。相对于其他假设，基础假设空间并没有过度表示接近零目标函数的假设。这样，在插值假设空间规模相近的情况下，对于这两个任务而言，指定真实假设的难度也相近。

F.3. 预测随机目标

接下来，我们考虑 ImageNet 的一种变化形式：将 ImageNet 目标函数替换为从原始 ImageNet 所使用的基础假设空间中选取的随机函数。在这种情况下，对于每个图像具有随机化输出的训练集，看起来就像是 ImageNet 的一个变体。需要特别注意的是，这不同于为训练集中的每个点独立选择一个随机输出：我们会希望距离足够近的训练点（距离约为 δ 或更小）具有相似的输出。我们考虑采用这种方式随机化 ImageNet 的输出，而不是为每个 ImageNet 输入选择相互独立的随机目标，这是因为为 ImageNet 的所有可能输入构造独立选取的随机目标，可能无法对应一个定义良好的目标函数。

假设我们希望达到与原始 ImageNet 任务相同的泛化错误率。控制任务难度的参数将与公式 9中的相同：具体而言，输入分布特有的参数（m， n， d， r）、模型输出和损失函数相关的参数（b， $L_{\mathcal{L}}$）以及错误率参数（ε）都将保持不变。因此，任务难度将与原始 ImageNet 相同： 3.55×10^{41} 比特。

为什么解决 ImageNet 的随机化版本所需的归纳偏置并不比原始任务更多，尽管其目标函数的结构性较弱？直观地说，这是因为基础假设空间平等地看待所有假设：与其他假设相比，它不会过度表示结构化假设（例如 ImageNet 的目标函数）。因此，对于 ImageNet 的随机化版本和原始版本而言，在假设空间中指定一个良好泛化区域的难度是相同的。

F.4. 类别数量的变化

我们改变 ImageNet 中的类别数量（同时保持其他所有任务参数不变，包括训练点数量 n），并在图 5a 中发现，任务难度会随着类别数量的增加而增大。类别数量从 10 增加到 1000时，任务难度大约增加了 5 个数量级。这主要是因为类别增加后，输入空间的大小也随之增大。结果表明，向分类任务中添加更多类别，可以在一定程度上有效提高任务难度，但效果不如提高任务的内在维度。

F.5. 改变空间分辨率

我们改变用于构建 ImageNet 基础假设空间的空间分辨率（同时保持所有其他任务参数不变），并在图 5b 中发现，任务难度会随着空间分辨率的提高而大幅降低。这是合理的：随着用于构建假设空间的空间分辨率提高，假设空间会缩小，因此，在假设空间中指定区域也会变得困难得多。这些结果强调了基础假设空间的选择对任务难度的关键影响。

G. 补充讨论

G.1. 如何将归纳偏置复杂度度量扩展到非插值情形 ?

我们工作的一个局限是，我们只考虑插值假设（换句话说，我们只考虑训练误差$\epsilon = 0$ 在定义 3.1中）。虽然我们将把形式化地将结果扩展到非插值假设留待未来工作，但我们简要概述一下如何扩展这些结果：

首先，我们可以通过论证非插值假设只要与真实假设的距离更小（相对于插值假设的距离），就可能具有良好的泛化能力，从而为非插值情形给出与定理 1 类似的结果。具体而言，我们可以预期，更一般的半径表达式为 $L_{\mathcal{L}} L_f \overline{W}^{\varepsilon-\epsilon}(p,q)$。

接下来，为了量化假设空间中的概率，并找到归纳偏置复杂度的实用估计，我们必须量化与训练数据的误差不超过 ϵ的假设空间大小。直观而言，我们可能会预期，它是插值假设空间周围的一个区域，其维数与基础假设空间相同。我们还可能会预期，该区域的大小随 ϵ 呈多项式增长。与插值情形一样，接下来我们可以估计近似插值假设空间中足够接近真实假设的部分，从而得出一个通用且可实际计算的任务难度表达式。

G.2. 为什么归纳偏置会随数据维度呈指数级增长?

直观地说，归纳偏置会随数据维度呈指数级增长，是因为归纳偏置与假设空间的维数相关，而假设空间的维数会随数据维度呈指数级增长。

首先，我们来理解为什么归纳偏置会随假设空间的维数增长：回想一下，要实现泛化，就需要在假设空间中指定一组能够良好泛化的特定假设。指定一个 m 维空间中的一个点所需的信息量会随 m 增长，因为它实际上就是该点的坐标数量。因此，假设能够良好泛化的区域集中在某个点周围时，指定该区域所需的信息量也会随假设空间的维数增长。

接下来，我们来理解为什么假设空间的维数会随数据维度增长：考虑一种非常简单的基于“网格”的假设构造方法，其中输入流形被划分为大小相等、铺满整个流形的超立方体。一个假设由超立方体与输出之间的映射构成。请注意，每个假设的维数会随超立方体的数量增长，因为每个假设都可以由其在每个超立方体处的输出来指定。超立方体的数量会随输入流形的维数呈指数级增长；因此，假设空间的维数也会随输入流形的维数呈指数级增长。

我们认为，这种指数级依赖关系是根本性的：无论我们如何对假设进行参数化，对更多变化维度进行泛化都会显著扩大假设空间，从而大幅增加归纳偏置复杂度。

G.3. 为什么训练点提供的归纳偏置如此之少?

在实验中，我们观察到，改变训练数据的数量，即使变化多个数量级，也不会显著影响泛化所需的归纳偏置量。我们可以将其理解为：训练数据对于泛化所提供的信息相对较少。为什么训练点提供的信息如此之少？直观地说，这是因为在缺乏对假设空间的强约束时，训练点只能提供关于我们旨在逼近的函数的信息，而且仅限于每个训练点周围的局部 邻域。相比之下，归纳偏置可以对假设空间提供更多全局 约束，从而大幅缩小假设空间的规模，并实现泛化。没有强归纳偏置时，训练点需要覆盖输入流形的大量区域，才能实现泛化。对于高维流形，这可能需要数量极其庞大的训练点。

通过简单的尺度分析，我们可以估计在缺乏强归纳偏置的情况下，要在 ImageNet 上实现泛化所需的训练样本数量。我们估计 ImageNet 的内在维度为 48，任务相关分辨率 δ 为 65。随后，我们可以将实现泛化所需的训练点数量估计为这样一个数量，使得流形上的任意区域都位于某个训练点的半径 δ 之内。拥有 n 个训练点时，我们可以预期覆盖流形上约 $n65^{48}$ 的体积。给定一个体积约为 $(255 \times 224 \times \sqrt{3})^{48}$ 的流形（对应于 ImageNet 图像的 $[0, 255]$ 像素范围和 $224 \times 224 \times 3$ 外在维数），那么我们将需要约$(255 \times 224 \times \sqrt{3}/65)^{48} \approx 10^{153}$ 个点才能实现泛化。实际上，沿着图 3中的线性趋势，我们可以预期，拥有这么多训练点时，所需的归纳偏置大约会减半。然而，只有 10^{17} 个点时，如图 3所示，要实现泛化就需要强归纳偏置。

G.4. 为什么归纳偏置复杂度的尺度如此之大?

在所有设置中，我们的度量都得出了非常大的归纳信息量，比实际模型的参数维数大许多个数量级。这些大数值可归因于第 3.3 节所构造的假设空间规模极其庞大：它包含数据流形上频率低于某个频率阈值的任意带限函数。典型的函数类可能已经只表示了假设空间中一个非常小的子空间：例如，神经网络具有偏向组合性和光滑性的特性，这可能会显著缩小假设空间（参见 Li 等人(2018年)；Mhaskar 等人(2017)），而标准初始化和优化器还会进一步缩小这个子空间。此外，能够通过我们计算机硬件和软件上规模合理的程序实际实现的函数，本身可能也只占完整假设空间中的一小部分。未来的工作可以采用一种已经包含其中某些约束的基础假设空间，从而降低我们测得的任务难度。

H. 附加表格与图

表 6：在一套通用符号下映射不同的学习设定。

设定	x	特征	$f^*(x)$
监督分类	输入	d_1：类别 c_1：类别内实例	类别
强化学习	状态或观测序列	c_1：状态或观测序列	期望动作
用于……的元学习 **小样本分类**	来自某个集合的样本 相关类别	c_1：相关类别集合 c_2：从 c_1中选择样本	分类器映射输入归属于它们的类别
无监督自编码	输入	c_1：输入	x
元强化学习	穿过……的轨迹 环境	c_1：环境 c_2：环境中的轨迹	策略将状态映射映射为期望动作

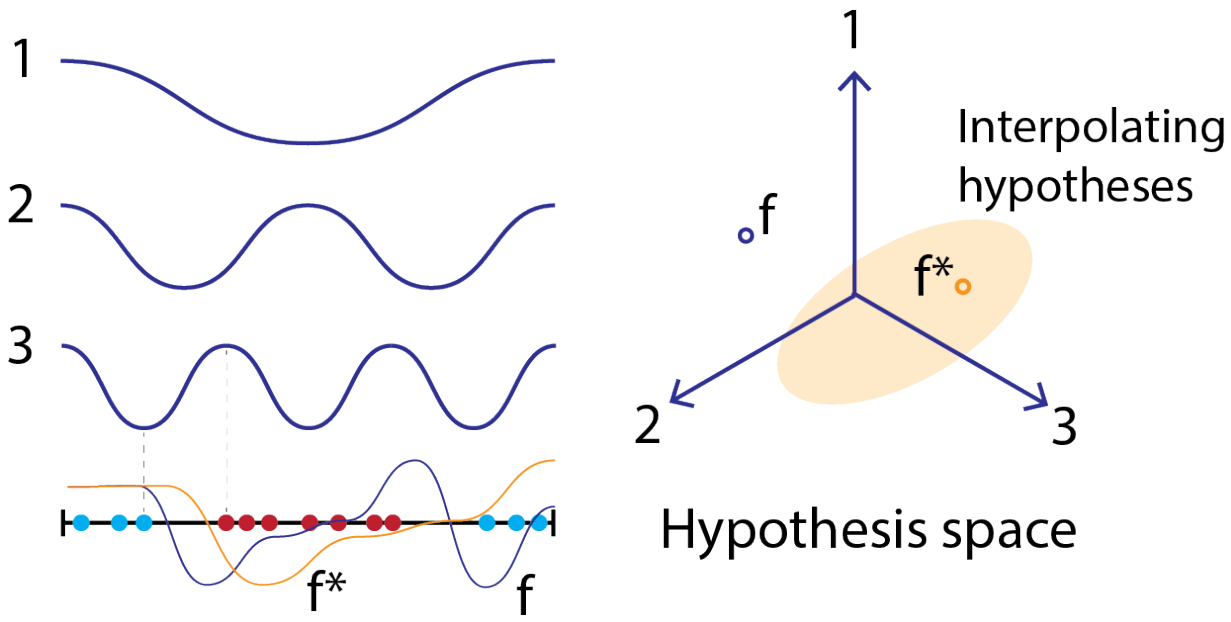


图 6：一个示例，展示如何为输入为一维的二分类任务构建假设空间。黑线表示数据流形，线上的蓝点和红点表示训练点。该假设空间由三种不同频率的基函数构建而成；请注意，所选最高频率的尺度与类别间最小距离相对应。图中展示了假设空间中的两个特定假设：真实假设（橙色），f^* 和另一个假设 f（紫色）。二者都可表示为基函数的线性组合，因此如图所示，它们对应于假设空间中的点。f^* 假设符合训练数据，因此属于浅橙色椭圆所示的插值假设集。

表7：不同模型架构为图像分类任务提供的归纳偏置信息。（FC-N指具有 N 层的全连接网络。）

MNIST

模型	信息 ($\times 10^{16}$ 比特)	内容
线性模型（Lecun等人，1998年）	0.705	
FC-3（Lecun等人，1998年）	0.817	
AlexNet（mrgrhn，2021）	0.888	
LeNet-5（卷积神经网络）（Lecun等人，1998年）	0.907	
DSN（Lee 等，2015年）	0.976	
MCDNN（Ciregan 等人，2012年）	1.020	
集成卷积神经网络（An 等人，2020年）	1.094	

CIFAR10

模型	信息 ($\times 10^{32}$ 比特)	内容
线性模型 (Nishimoto，2018年)	2.250	
FC-4 (Lin 等人，2015年)	2.354	
MCDNN(Ciregan 等人，2012年)	2.704	
AlexNet(Krizhevsky 等人，2012年)	2.709	
DSN(Lee 等人，2015年)	2.786	
DenseNet (Huang 等人，2017)	3.010	
ResNet-50 (Wightman 等人，2021)	3.195	
BiT-L (Kolesnikov 等人，2020年)	3.453	
ViT-H/14(Dosovitskiy 等人，2021)	3.513	

SVHN

模型	信息 ($\times 10^{31}$ 比特)	内容
FC-6（Mauch 和 Yang，2017）	0.870	
AlexNet（Veeramacheneni 等人，2022年）	0.985	
深度卷积神经网络（Goodfellow 等人，2013）	1.049	
DSN（Lee 等人，2015年）	1.060	
DenseNet（Huang 等人，2017）	1.075	
WRN-16-8（Zagoruyko 和 Komodakis，2016）	1.077	
WRN-28-10(Foret 等人，2020年)	1.114	

ImageNet

模型	信息 ($\times 10^{31}$ 比特)	内容
线性模型（Karpathy，2015）	2.063	
SIFT + FVs（S´anchez &Perronnin，2011）	2.261	
AlexNet（Krizhevsky 等人，2012年）	2.290	
DenseNet-121（Huang 等人，2017）	2.348	
ResNet-50（He 等人，2016）	2.362	
DenseNet-201（Huang 等人，2017）	2.363	
WRN-50-2-bottleneck (Zagoruyko & Komodakis, 2016)	2.368	
BiT-L (Kolesnikov 等人，2020年)	2.449	
ViT-H/14（Dosovitskiy 等人，2021）	2.461	