arXiv:2302.00457v1 [cs.LG] 1 Feb 2023

单隐藏层神经网络中的简单性偏差

Depen Morwani [1] **Jatin Batra** [2] **Prateek Jain** [3 4] **Praneeth Netrapalli** [3 4]

摘要

近期研究（Shah 等人，2020；Chen 等人，2021年）表明，神经网络表现出极强的简单性偏差（SB）。也就是说，即使存在其他更鲁棒但更复杂的特征，它们也只会学习解决当前任务所需的最简单特征。由于缺乏对特征普遍且严格的定义，这些研究在 Color-MNIST、MNIST-CIFAR 等半合成数据集上展示了 SB，因为在这些数据集中，特征相对更容易定义。

在这项工作中，我们对 SB 进行了严格定义，并对其进行了全面验证。具体而言，(i) 我们将 SB 定义为网络本质上是输入的低维投影的函数；(ii) 从理论上证明，当数据线性可分时，即使存在任意多其他更复杂的特征（这些特征本可带来显著更鲁棒的分类器），网络主要仍只依赖于线性可分的（一维）子空间；(iii) 通过实验证明，在 Imagenette 和 Waterbirds-Landbirds 等真实数据集上训练的模型确实依赖于输入的低维投影，从而展示了这些数据集上的 SB；(iv) 最后，我们提出一种自然的集成方法，通过让后续模型在先前模型未使用的特征上进行训练，鼓励模型之间的多样性，并证明该方法能够得到对高斯噪声鲁棒性显著更强的模型。

1. 引言

众所周知，神经网络（NN）容易受到分布偏移以及对抗样本的影响



图 *1*。天鹅与熊的分类。有多种特征，例如背景、动物颜色、动物形状等，每种特征都足以用于分类，但综合使用这些特征将使模型更具鲁棒性。 [2]

（Szegedy 等人，2014；Hendrycks 等人，2021年）。近期的一系列研究（Geirhos 等人，2018年；Shah 等人，2020；Geirhos 等人，2020）提出，简单性偏差（*SB*）——又称捷径学习，即神经网络（NN）倾向于只学习最简单的特征，而不学习其他有用但更复杂的特征——是导致其缺乏鲁棒性的关键原因。大致论据如下：例如，在对天鹅和熊进行分类时，如图1所示，可用于分类的特征有很多，例如背景、动物颜色、动物形状等。然而，只使用其中一个或少数几个特征，可能会导致模型无法应对特定的分布偏移；而使用所有特征，则可能使模型更具鲁棒性。

近期多项研究在多种半真实构造数据集 (Geirhos 等人，2018年；Shah 等人，2020；Chen 等人，2021年）上展示了简洁性偏置（SB），并提出 SB 是神经网络（NN）在分布偏移下表现出脆弱性的关键原因（Shah 等人，2020）。然而，这类观察目前仍仅针对特定的半真实数据集；文献中仍缺少一种能够在给定数据集 和给定模型 上识别简洁性偏置的通用方法。这样的方法不仅有助于评估模型的鲁棒性，也能帮助设计更具鲁棒性的模型。

设计一种能够识别（并可能修正）简洁性偏置的通用方法，面临的一项关键挑战是，特征这一概念本身含糊不清，缺乏严格的定义。现有研究，例如（Geirhos 等人，2018年；Shah 等人，2020；Chen 等人，2021年），通过避免对含糊的特征进行定义来回避这一挑战——

[1]哈佛大学计算机科学系，美国马萨诸塞州剑桥市（部分工作是在印度班加罗尔的 Google Research 完成的） [2]TIFR 技术与计算机科学学院，印度孟买 [3]Google Research，印度班加罗尔。 [4]按字母顺序排列。通信作者：Depen Morwani< dmorwani@g.harvard.edu>。

[2]图片来源：维基百科（swa）、（bea）。

方法是使用精心设计的数据集（例如，将MNIST图像与CIFAR图像拼接），在这些数据集的定义中，某些高级特征（例如MNIST特征和CIFAR特征、形状特征和纹理特征）已经被预先设定，因此直观上很容易论证这些特征的简洁性。

贡献：本工作的主要贡献之一是为单隐藏层神经网络的一种特定简洁性偏置——LD-SB——提供精确定义。具体而言，我们将简洁性偏置刻画为模型对低维输入的依赖。具体来说，

定义 1.1（LD-SB）。一个模型 $f\colon \mathbb{R}^d \to \mathbb{R}^c$，其输入为 $x \in \mathbb{R}^d$，输出为 $f(x) \in \mathbb{R}^c$（例如，c个类别的 logit值），在分布（x，y）$\sim \mathcal{D}$ 上训练，若存在满足以下条件的投影矩阵 $P \in \mathbb{R}^{d\times d}$，则满足 LD-SB：

- 秩（P）$= k \ll d$，
- $f(Px_{(1)} + P_\perp x_{(2)}) \approx f(x_1) \; \forall \; (x_{(1)}\, y_{(1)})$，$(x_{(2)} y_{(2)}) \sim \mathcal{D}$
- 在 $(P_\perp x, y)$ 上训练的独立模型 g（其中 (x, y) 取得了很高的准确率。$\sim \mathcal{D}$

这里 $P_\perp$ 是投影到与 P 正交的子空间上的投影矩阵。

换句话说，LD-SB 表示，在输入空间 $\mathbb{R}^d$ 中存在一个维度较低的子空间（由投影矩阵 P 给出），模型 f在为任何输入点 x赋予标签时只考虑这一子空间。具体而言，如果我们混合两个数据点 x_1 和 x_2，将 x_1投影到 P 上的结果与 x_2 投影到正交子空间 $P_\perp$上的结果结合起来，那么 f对这个混合点 $Px_1 + P_\perp x_2$的输出就会与对 x_1的输出相同。如果子空间 $P_\perp$ 不包含任何有助于分类的特征，这倒也无妨。然而，第三个要点指出，$P_\perp$确实包含有助于分类的特征，因为在 $(P_\perp x, y)$ 上训练的独立模型 g 取得了很高的准确率。

此外，我们从理论上证明，对于一个相当一般的分布类别——1- 隐藏层神经网络具有简洁性偏置，该分布类别称为独立特征模型（IFM），其中，给定标签后，特征（即坐标）相互独立地分布。IFM 历史悠久，且受到广泛研究，尤其是在朴素贝叶斯分类器的研究中（Lewis，1998年）。对于 IFM，我们证明，只要数据中存在一个线性可分的特征，即使另有任意多个特征的数据可分，但其分界是非线性，使用随机梯度下降训练的神经网络仍会学习到几乎完全依赖这个线性可分特征的模型。在三个真实世界数据集上：Imagenette 的二分类和多分类版本（Fas-tAI，2021年），以及 Waterbirds-Landbirds（Sagawa 等人，2020a）数据集。与（Shah 等人，2020）中的结果相比，我们的结果 (i) 从理论上展示了在相当一般的设定下存在简洁性偏置，(ii) 从实证上展示了真实数据集中的简洁性偏置。

最后，基于这些见解，我们提出了一种简单的集成方法——*OrthoP* ——该方法通过投影去除先前神经网络所使用的输入数据主要方向，依次构建神经网络。我们证明，在存在高斯噪声等简单分布偏移的情况下，该方法能够为真实世界数据集带来鲁棒性显著更强的集成。

为何只研究1单隐藏层网络?：有人可能会想，为什么本文的结果仅限于单隐藏层网络，以及这些结果为何值得关注。我们给出两个原因。

1. 从**理论**角度来看，先前的研究已充分刻画了不同初始化方案下无限宽1单隐藏层网络的训练动态（Chizat 等人，2019），并且还确定了此类网络中梯度下降的极限点（Chizat 和 Bach，2020）。我们的结果在很大程度上建立在这些先前研究的基础上。另一方面，我们尚未像对单隐藏层网络那样，清晰地理解更深层网络的动态[3]。

2. 从 **实践** 角度来看，当前机器学习领域的主流范式是：先用大量数据预训练大型模型，再在小型目标数据集上进行微调。人们观察到，由于这些模型接触过大量且多样化的预训练数据，它们确实学到了丰富的特征（Rosenfeld 等人，2022；Nasery 等人，2022）。然而，在目标数据集上进行微调时，可能无法利用预训练模型中的所有特征。因此，能够训练出稳健微调头的方法（例如，在其上叠加一个 1层隐藏层网络）可能非常有效。

从理论和实践角度来看，将我们的结果推广到更深的网络和其他架构，是一个令人振奋的研究方向。

论文结构：本文结构如下。第 2 节介绍相关工作。第 3 节介绍预备知识。关于 LD-SB 的主要结果见第4节。第5 节介绍训练多样化分类器的结果。最后，我们在第 6节作总结。

[3]关于将这些结果推广到深度神经网络的难点，更多讨论请参阅附录 D

2. 相关工作

本节简要介绍与本文密切相关的工作。扩展的相关工作见附录C。

简单性偏差：继 Shah 等，2020 之后，已有多篇论文探讨了各种网络中是否存在简洁性偏置（SB），以及其背后的原因（Scimeca 等，2021）。其中，Huh 等，2021 与我们的工作最为相关。Huh 等，2021 通过实证观察发现，在某些合成数据集上，神经网络在初始化时和训练后得到的嵌入都具有低秩结构。与此不同，我们从理论上证明了 IFM 模型中的 LD-SB，并在真实数据集上进行了实证验证。此外，我们的结果表明，尽管网络权重在丰富 regime 中呈现低秩结构（定义见第 3.2 节），但 LD-SB 在惰性区间中的表现要微妙得多。我们还展示了如何利用 LD-SB 训练第二个多样化模型，并将其组合起来，得到一个鲁棒集成。Galanti 和 Poggio，2022 为各种超参数（如学习率、批量大小等）与所学习权重矩阵的秩之间的关系提供了理论解释，并通过实验证明了这一点。Pezeshki 等，2021 提出，训练开始时的梯度饥饿可能是惰性/NTK 区域中出现简洁性偏置的原因，但相关条件难以解释。相比之下，我们的结果适用于丰富训练 regime 下 IFM 模型中的任意数据集。最后，Lyu 等，2021 考虑了反对称数据集，并证明单隐藏层输入齐次网络（即不含偏置参数的网络）会收敛到线性分类器。然而，我们的结果适用于一般数据集，且不要求输入齐次性。

学习多样化分类器：已有多项研究尝试学习多样化分类器。大多数研究通过确保这些模型的输入梯度彼此不对齐来学习此类模型（Ross 与 Doshi-Velez，2018年；Teney 等人，2022）。（Xu 等人，2022）提出了一种学习多样化／正交分类器的方法，其前提是存在一个使用全部特征的完整分类器，并展示了该方法在风格迁移等多种下游任务中的实用性。（Lee 等人，2022）通过在无标签目标数据上施加多样性约束来学习多样化分类器。

伪相关：大量研究探讨了神经网络中伪相关的成因（Sagawa 等人，2020b），并针对不同场景提出了算法层面的修正方案（Liu 等人，2021年；Chen 等人，2020）。

梯度下降的隐式偏差：关于梯度下降动力学隐式偏差的研究也有大量成果。其中大多数研究针对标准线-性（Ji & Telgarsky，2019）或深度线性网络（Soudry 等人，2018年；Gunasekar 等人，2018年）。对于非线性神经网络，一个广为人知的结果适用于具有齐次激活函数的1层隐藏层神经网络（Chizat 和 Bach，2020），我们在证明中关键性地使用了这一结果。

3. 预备知识

本节将介绍理解本文结果所需的符号记法，以及关于无限宽最大间隔分类器的背景知识。

3.1. 基本概念

单隐藏层神经网络与损失函数。考虑实例 $x \in \mathcal{R}^d$ 和标签 $y \in \{\pm 1\}$，它们的联合分布为 $\mathcal{D}$。用于预测给定实例 x的标签的单隐藏层神经网络模型由参数 $(\bar{w} \in \mathbb{R}^{m\times d}, \bar{b} \in \mathbb{R}^m, \bar{a} \in \mathbb{R}^m)$定义。对于固定的激活函数 ϕ，给定输入实例 x，模型 $f((\bar{w}, b, \bar{a}) x) := \langle \bar{a}, \phi(\bar{w}x + b)\rangle$，其中 $\phi(\cdot)$按元素应用。给定模型 f、输入 x和标签 y时，交叉熵损失 $\mathcal{L}$定义为 $\mathcal{L}(f(x) y)$ def $= \log(1 + \exp(-yf((\bar{w}, b, \bar{a}) x)))$。

间隔。对于数据分布 $\mathcal{D}$，模型$f(x)$ 的间隔定义为 $\min_{(x,y)\sim\mathcal{D}} yf(x)$。

记号。以下是一些我们将反复使用的有用记号。对于矩阵 A，$A(i,.)$ 表示A的第 i 行。对于任意 $k \in \mathbb{N}$，$\mathbb{S}^{k-1}$ 表示 k维单位范数欧几里得球面的表面。

3.2. 初始化

网络的梯度下降动力学在很大程度上取决于初始化尺度。在这项工作中，我们主要考虑丰富机制 初始化。

丰富 regime。在丰富机制初始化中，对于任意 $i \in [m]$，第一层的参数（$\bar{w}(i, .)$ $b(i)$）从 $\mathbb{S}^d$上的均匀分布中采样。每个 $\bar{a}(i)$ 都从均匀分布$\{-1, 1\}$中采样，网络的输出则按 $\frac{1}{m}$ 缩小（Chizat 和 Bach，2020）。这大致相当于 Xavier 初始化 Glorot & Bengio （2010）：当 $m \gg d$ 时，两层中的权重参数均近似初始化为 $\mathcal{N}(0, \frac{2}{m})$。

此外，我们还给出了下文所述惰性机制初始化的一些结果。

惰性区间。在惰性区间中，第一层的权重参数初始化为 $\mathcal{N}(0, \frac{1}{d})$，第二层的权重参数初始化为 $\mathcal{N}(0, \frac{1}{m})$，偏置初始化为 0 (Bietti 和 Mairal，2019；Lee 等人，2019）。这近似等价于 Kaiming 初始化（He

等人，2015年）。

3.3. 无限宽度情形

对于无限宽度极限下采用 ReLU 激活的单隐藏层神经网络，即当 $m \to \infty$时，Jacot 等人(2018年）；Chizat 等人 (2019）；Chizat 和 Bach (2020）对训练后的模型给出了有趣的刻画。如上所述，这些模型的训练过程根据初始化的尺度分为两种区间（Chizat 等人，2019）：

丰富 regime。在无限宽度极限下，神经网络参数可以视为三元组 $(w,\ b,\ a) \in \mathbb{S}^{d+1}$ 上的一个分布，其中 $w \in \mathbb{R}^d,\ b,\ a \in \mathbb{R}$。在丰富机制初始化下，模型计算得到的函数 f 可以表示为

$$f(\nu, x) = \mathbb{E}_{(w,b,a)\sim\nu}[a(\phi(\langle w, x\rangle + b)]\,. \tag{1}$$

（Chizat 和 Bach，2020）证明，采用丰富初始化的训练过程可以视为二阶 Wasserstein 空间上的梯度流，并给出了交叉熵损失下训练后模型的如下刻画 [4] $\mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathcal{L}\ (\nu,\ \ (x,\ \ y)\ )]$。

定理 3.1。（*Chizat 和 Bach，2020*）在无限宽度极限下，采用丰富初始化且使用交叉熵损失时，如果带 *ReLU* 激活的单隐藏层神经网络上的梯度流收敛，则其收敛到一个最大间隔分类器 ν^*，其定义如下

$$\nu^* = \underset{\nu\in\mathcal{P}(\mathbb{S}^{d+1})}{\arg\max} \min_{(x,y)\sim\mathcal{D}} yf(\nu, x)\,, \tag{2}$$

其中 $\mathcal{P}(\mathbb{S}^{d+1})$ 表示$\mathbb{S}^{d+1}$上的分布空间。

这种训练模式被称为“丰富”区间，因为它能够学习数据依赖特征 $\langle w, \cdot\rangle$。

惰性区间。(Jacot 等人，2018年) 表明，在无限宽度极限下，神经网络的行为类似于核机器。这种核通常称为神经正切核（NTK），其定义为 $K(x, x') = \left\langle \frac{\partial f(x)}{\partial W}, \frac{\partial f(\ x'}{\partial W} \right\rangle)$，其中 W 表示所有可训练权重参数的集合。这种初始化模式称为“惰性”区间，因为权重相对于初始化几乎没有变化，且神经正切核几乎保持不变；也就是说，网络不会学习数据依赖特征。我们将采用以下单隐藏层神经网络的神经正切核刻画。

定理 3.2。*(Bietti 和 Mairal，2019）*在无限宽度极限下采用惰性机制初始化时，具有 *ReLU* 激活的单隐藏层神经网络的神经正切核（*NTK*），即：

[4]定理 3.1 是 Chizat 和 Bach 2020年定理 5 的非正式表述。关于精确结果，请参阅附录 E.1 。E。

$\phi(u) = \max(u, 0)$, *is given as*

$$K(x, x') = \|x\|\|x'\|\kappa\left(\frac{\langle x, x'\rangle}{\|x\|\|x'\|}\right)\,,$$

其中

$$\kappa(u) = \frac{1}{\pi}(2u(\pi - cos^{-1}(u)) + \sqrt{1-u^2})\,.$$

二元分类的惰性区间。(Soudry 等人， 2018年）表明，对于线性可分数据集，线性预测器在逻辑损失上的梯度下降会收敛到最大间隔支持向量机(SVM)分类器。这意味着，任何足够宽的神经网络在惰性区间内针对可由有限宽度诱导神经正切核分离的数据集进行有限时间训练时，都会趋向于 $\mathcal{L}_2$ 由以下给出的最大间隔分类器

$$\underset{f\in\mathcal{H}}{\arg\min} \|f\|_{\mathcal{H}} \text{ s.t. } yf(x) \geq 1\ \forall\ (x, y) \sim \mathcal{D}\,, \tag{3}$$

其中 $\mathcal{H}$ 表示与有限宽度核相关联的再生核希尔伯特空间（RKHS）（Chizat，2020）。随着宽度增加，该核趋向于无限宽度神经正切核（它具有通用性（Ji 等人，2020））。因此，在惰性区间中，我们将关注由无限宽度神经正切核诱导的 $\mathcal{L}_2$ 最大间隔分类器。

4. 单隐藏层神经网络中的简洁性偏置

本节首先从理论上刻画梯度下降在独立特征模型（*IFM*）中对线性可分数据集所表现出的简洁性偏置。定理4.1给出了主要结果：对于 $\mathbb{R}^d$中的输入进行二元分类，即使数据仅在一个坐标上是线性可分的，梯度下降动力学也会学习到一个仅依赖该坐标的模型，即便还有任意多个 $d-1$ 坐标使数据可分，但需要借助非线性分类器。换言之，这些神经网络的简洁性偏置体现为对低维输入的依赖，我们将其称为 LD-SB。随后，我们通过实验验证，在一些真实数据集上训练的神经网络确实满足 LD-SB。

4.1. 数据集

我们考虑独立特征模型（IFM）中的数据集，其中 (x, y) 上的联合分布满足$p(x, y) = r(y)\prod_{i=1}^{d} q_i(x_i|y)$，即，给定标签 y后，各特征相互独立地分布。这里，$r(y)$ 是定义在 $\{-1, +1\}$ 上的分布，而 $q_i(x_i|y)$ 表示在给定 y时，i^{th} 坐标 x_i 的条件分布。IFM 在文献中得到了广泛研究，尤其是在朴素贝叶斯分类器的背景下（Lewis，1998年）。我们作出

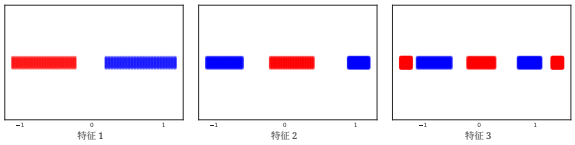


图2。IFM 数据集示意图。给定一个类别 ± 1 分别由 蓝色 和 红色 表示，每个坐标值均独立地从相应的分布中抽取。上方展示了一个示例 IFM 数据集中三个不同坐标上的分布支撑集，分别对应正和 负标签。

以下假设，这些假设表明，至少有两个复杂度不同的特征可用于分类：一个具有线性边界，另有至少一个具有非线性边界。示例见图2。

- 其中一个坐标（例如，第一个坐标，不失一般性）可由具有间隔 γ 的线性决策边界分开（见图2），即 $\exists \gamma > 0$，使得 $\gamma \in Supp(q_1(x_1|y=+1)) \subseteq [\gamma, \infty)$ 且 $-\gamma \in Supp(q_1(x_1|y=-1)) \subseteq (-\infty, -\gamma]$，其中 $Supp(\cdot)$ 表示某个分布的支撑集。
- 其他坐标都不是线性可分的。更准确地说，对于所有其他坐标 $i \in [d] \setminus \{1\}$，$0 \in Supp(q_i(x_i|y=-1))$ 和 $\{-1,+1\} \subseteq Supp(q_i(x_i|y=+1))$。
- 即使不使用线性坐标，也能对数据集进行完美分类。这意味着，$\exists i \neq 1$，使得 $q_i(x_i|y)$ 对于 $y=+1$ 和 $y=-1$ 具有不相交支集。

尽管我们假设特征是轴对齐的，但我们的结果对数据集的任意旋转也成立。虽然我们的结果在一般的 IFM 设置下成立，但相比之下，当前关于简洁性偏置的结果，例如（Shah 等人，2020），是在 IFM 中的非常特殊的数据集上得到的，并不适用于一般的 IFM。

4.2. 主要结果

我们的主要结果指出，对于满足上述条件的任意 IFM 数据集，在采用丰富初始化（第3.2节）时，神经网络会表现出 LD-SB，此外还需满足定理E.1中陈述的一些技术条件。证明见附录A.1。

定理 4.1。对于具有有界密度和有界支撑集、满足上述条件的 *IFM* 模型中的任意数据集 $\gamma \geq 1$，若单隐藏层全连接网络在丰富初始化下、无限宽度极限中的梯度流采用交叉熵损失，且收敛并满足定理*E.1*[5]，中的技术条件，则它收敛到 $\nu^* = 0.5\delta_{\theta_1} + 0.5\delta_{\theta_2}$ 在 $\mathcal{S}^{d+1}$上，其中 $\theta_1 = (\sqrt{\frac{\gamma}{2(1+\gamma^2)}}e1, \sqrt{\frac{1}{2(1+\gamma^2)}}1/\sqrt{2})$ $\theta_2 = (-\sqrt{\frac{\gamma}{2(1+\gamma^2)}}e1, \sqrt{\frac{1}{2(1+\gamma^2)}}-1/\sqrt{2})$，且 e_1 *def*$=[1,0,\cdots,0]$ 表示第一标准基向量。这意味着 $f(\nu^*, Px^{(1)} + P_\perp x^{(2)}) = f(\nu^*, x_1^{(1)}) \forall (x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}) \sim \mathcal{D}$，其中 P表示第一坐标上的（秩为 *1* 的）投影矩阵。

此外，由于至少一个坐标 $\{2,\ldots,d\}$对于 $q_i(x_i|y=+1)$ 和 $q_i(x_i|y=-1)$ 的支撑集不相交，$P_\perp(x)$ 仍能完美地对给定数据集进行分类，从而意味着 LD-SB。

众所周知，丰富 regime 与神经网络的实际性能更为相关，因为它支持特征学习，而惰性区间则不然（Chizat 等人，2019）。尽管如此，在下一节中，我们将通过考虑 IFM 中一个更为特殊的数据集，给出理论证据，表明即使在惰性区间内，LD-SB 仍然成立。

4.3. 惰性区间

在此区间内，我们将使用 IFM 类中的以下数据集：

对于 $y \in \{\pm 1\}$，我们生成 $(x, y) \in D$，如下所示

$$\mathbf{x}_1 = \gamma y$$

$$\forall i \in 2,..,d, \mathbf{x}_i = \begin{cases} \pm 1 & \text{for} \quad y=1 \\ 0 & \text{for} \quad y=-1 \end{cases}$$

尽管上述数据集是一个点质量数据集，但它仍然具有与丰富 regime 数据集共同的重要特征——只有一个坐标是线性可分的，其余坐标均不可分。对于该数据集，我们给出最大间隔 NTK 的刻画（如方程（3）所示）：

定理 4.2. 对于足够小的 $\epsilon > 0$，存在一个绝对常数 N，使得对于所有 $d > N$ 以及 $\gamma \in [7, \epsilon\sqrt{d})$，在数据集 $\mathcal{L}_2$ 上，以 *NTK regime* 联合训练单隐藏层全连接网络两层所得的最大间隔分类器 D，即任何满足方程（$f3$）的，均满足：

$$pred(f(Px^{(1)} + P_\perp x^{(2)})) = pred(f(x^{(1)}))$$
$$\forall\, (x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}) \in D$$

其中，P 表示第一个坐标上的投影矩阵，而 $pred(f(x))$ 表示模型 f 在 x 上的预测标签。

上述定理表明，在混合样本 $Px^{(1)} + P_\perp x^{(2)}$ 上的预测与 $x^{(1)}$上的预测相同，从而确立了 LD-SB。该定理的证明见附录A.2。

[5]注意，定理 E.1 是对 Chizat 和 Bach (2020)，在我们的分析中将其作为黑箱使用

表*1*。 丰富 regime 中 LD-SB 的示例：该表展示了四个数据集上的 $P_\perp$ 和 P随机化准确率（RA）以及 logit 变化（LC）。这些结果证实，将输入 x 投影到由 P张成的子空间，基本决定了模型对 x的预测。↑（分别为 ↓）表示 LD-SB 意味着较大的（分别为较小的）值。

数据集	秩（P）	准确率（f）	$P_\perp$-RA（↑）	P-RA (↓)	$P_\perp$-LC (↓)	P-LC (↑)
b-Imagenette	1	93.05±0.26	89.94±0.22	49.53±0.24	28.57 ± 0.26	92.13 ± 0.24
Imagenette	10	79.52±0.13	75.89±0.25	9.33 ± 0.01	33.64 ± 1.21	106.29 ± 0.53
Waterbirds	3	91.88 ± 0.1	91.47±0.11	62.51±0.07	25.24 ± 1.03	102.35 ± 0.19
MNIST-CIFAR	1	99.69 ± 0.0	94.15±0.21	55.2 ± 0.13	38.97 ± 0.76	101.98 ± 0.31

4.4. 实证验证

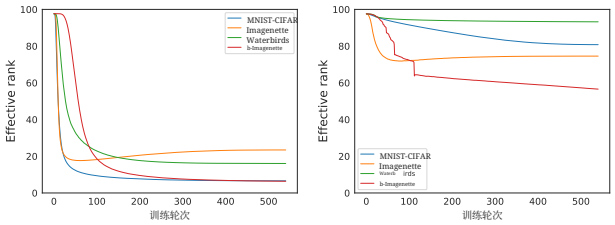
本节将介绍实证结果，展示 LD-SB 在3 个真实数据集上的表现：Imagenette（FastAI，2021年）、Imagenette 的二进制版本（b-Imagenette）和水鸟-陆鸟（Sagawa 等人，2020a），以及一个专门设计的数据集 MNIST-CIFAR（Shah 等人，2020）。有关这些数据集的更多详情，请参见附录B.1。

4.4.1. 实验设置

我们采用在 ImageNet 上预训练的 ResNet-50 模型，提取2048维特征，并在每个数据集上训练一个具有 ReLU 非线性和1个隐藏层的全连接网络（包含100个隐藏单元）进行分类。在微调过程中，我们冻结 ResNet-50 骨干网络，仅训练1层隐藏层的头部（更多细节见附录B.1）。

展示 LD-SB：给定模型 $f(\cdot)$，我们通过找出一个低维子空间来确立其低维 SB，该子空间由其投影矩阵 P确定；若将输入 x_1 和 x_2 混合为 $Px_1 + P_\perp x_2$，则模型对混合输入 $\tilde{x}\text{def}= Px_1 + P_\perp x_2$, $f(\tilde{x})$ 的输出始终接近模型对 x_1 的输出，即 $f(x_1)$。我们采用四种度量衡量接近程度：（1） $P_\perp$随机准确率（$P_\perp$-RA）：在数据集（$Px_1+P_\perp x_2, y_1$）上的准确率，其中（x_1, y_1）和（x_2, y_2）从该数据集中独立同分布采样；（2）P随机准确率（P-RA）：在数据集（$Px_1 + P_\perp x_2, y_2$）上的准确率；（3） $P_\perp$ logit变化（$P_\perp$-LC）：相对于 x_1 的 logit值的变化，即 $\|f(\tilde{x}) - f(x_1)\| / \|f(x_1)\|$；（4）$P$logit变化（$P$-LC）：相对于$x_2$ 的 logit值的变化，即 $\|f(\tilde{x}) - f(x_2)\| / \|f(x_2)\|$。此外，我们还将展示，在（$P_\perp x, y$）上训练的后续模型在这些数据集上取得了显著较高的准确率。

如第4.2 节和4.3节所述，1层隐藏层神经网络的训练轨迹可能因初始化尺度不同而有所差异。因此，对于丰富 regime 和惰性区间，子空间投影矩阵 P 的获得方式也不同。对于丰富 regime，我们将通过实证表明，第一层权重具有定理4.1 所述的低秩结构；而对于惰性区间，我们将表明，尽管第一层权重不呈现低秩结构



（a）丰富 regime　　（b）惰性区间

图*3.* 丰富 regime 和惰性区间中第一层权重矩阵有效秩的演变。

，模型仍如定理4.2所述，对输入的依赖具有低维性。

4.4.2. 丰富 regime

定理 4.1 表明，从渐近角度看，第一层权重矩阵将是低秩的。然而，由于我们只训练有限时间，权重矩阵只能近似为低秩。为了对此进行量化，我们使用有效秩的概念（Roy &Vetterli，2007）来衡量第一层权重矩阵的秩。

定义 4.3. 给定矩阵 M，其有效秩定义为：Eff-rank$(M) = e^{-\sum_i \sigma_i(M)^2 \log \sigma_i(M)^2}$ 其中$\sigma_i(M)$ 表示 M 的 i^{th} 奇异值，且 $\sigma_i(M)^2$ 定义为 $=\sigma_i(M)^2 / \sum_i \sigma_i(M)^2$。

一种理解有效秩的方式是：它等于矩阵$MM^\top$ $\text{Tr}(MM^\top)$ 的冯诺依曼熵（Petz，2001）的指数，其中 Tr (·) 表示矩阵的迹。举例来说，投影到 k维空间上的投影矩阵的有效秩等于 k。

图3a 展示了在四个数据集上训练过程中有效秩的变化。我们观察到，在训练接近结束时，权重矩阵的有效秩急剧下降。为确认这确实会导致 LD-SB，我们将 P 设为由第一层权重矩阵的顶部奇异方向张成的子空间，并计算 P和 $P_\perp$ 随机化准确率以及相对逻辑值变化。表1 中呈现的结果证实了LD-SB在

表 2。 惰性区间中的 LD-SB 演示：此表展示了 $P_\perp$ 和 P 在四个数据集上的随机化准确率以及 logit 变化。这些结果证实，输入 x 在由 P 张成的子空间上的投影，基本决定了模型对 x 的预测。

数据集	秩（P）	准确率（f）	$P_\perp$-随机化准确率（↑）	P-RA (↓)	$P_\perp$-LC (↓)	P-LC (↑)
b-Imagenette	1	92.75±0.06	90.07±0.34	52.09±1.34	36.94±1.01	138.41±1.62
Imagenette	15	79.97±0.44	68.25±1.18	11.92±0.82	55.99±3.86	133.86±5.42
Waterbirds	6	90.46±0.07	89.67±0.42	62.44±4.48	36.89±5.18	105.41±7.06
MNIST-CIFAR	2	99.74 ± 0.0	99.45±0.17	49.83±0.67	24.9 ± 0.61	141.12±1.86

表 3。 独立训练的模型（错误多样性 (f、 f_{ind}) 和类别条件对数几率相关性 (f、 f_{ind})）与在投影去除第一个模型的重要特征后依次训练的模型（错误多样性 (f、 f_{proj}) 和类别条件对数几率相关性 (f、 f_{proj})）的比较。结果表明，与 f 和 f_{ind} 相比， f 和 f_{proj} 具有更高的多样性。

数据集	错误多样性 (f， f_{ind}) (↑)	错误多样性 (f, f_{proj}) (↑)	类别条件对数几率相关性 (f， f_{ind}) (↓)	类别条件对数几率相关性 (f， f_{proj}) (↓)
B-Imagenette	3.87 ± 1.54	21.15 ± 1.57	99.88 ± 0.01	90.86 ± 1.08
Imagenette	6.6 ± 0.46	11.44 ± 0.65	99.31 ± 0.12	91 ± 0.59
Waterbirds	2.9 ± 0.52	14.53 ± 0.48	99.66 ± 0.04	93.81 ± 0.48
MNIST-CIFAR	0.0 ± 0.0	5.56 ± 7.89	99.76 ± 0.17	78.74 ± 2.28

这些数据集上的丰富 regime。此外，在附录B.2的表 4中，我们展示了在 $(P_\perp x, y)$ 上训练的独立模型能够达到显著较高的准确率。

4.4.3. 懒惰机制

对于惰性区间，如图 3b所示，第一层权重矩阵的秩在整个训练过程中始终保持较高。然而，我们能够找到一个低维投影矩阵 P，满足 LD- 简洁性偏置的条件（如定义 1.1 中所述），并将其作为一个优化问题的解。更具体地说，给定一个预训练模型 f 和一个秩为 r 的投影矩阵 P，求解：

$$\min_P \frac{1}{n}\sum_{i=1}^{n}\left(\mathcal{L}\left(f(Px^{(i)}), y^{(i)}\right) + \lambda\mathcal{L}\left(f(P^\perp x^{(i)}), \mathcal{U}[L]\right)\right) \tag{4}$$

其中 $\mathcal{U}[L]$ 表示覆盖所有 L 标签的均匀分布，$(x^{(1)}, y^{(1)}) \cdots$、$(x^{(n)}, y^{(n)})$ 是训练样本， $\mathcal{L}(\cdot,\cdot)$ 是交叉熵损失。我们再次强调，优化仅针对 P 进行，而模型参数 f 保持不变。换言之，上述函数确保神经网络沿着 P 给出正确的预测，并沿着 $P_\perp$ 给出无信息量的预测。表 2 展示了 $P_\perp$ 、 P-RA 以及 LC 的结果。如图所示，即便在这种情况下，我们也能找到低秩投影矩阵，证明存在 LD-SB。与丰富 regime 类似，在附录 B.2 的表 5 中，我们展示了在惰性区间内，基于 $(P_\perp x, y)$ 训练的独立模型也能达到显著高的准确率。

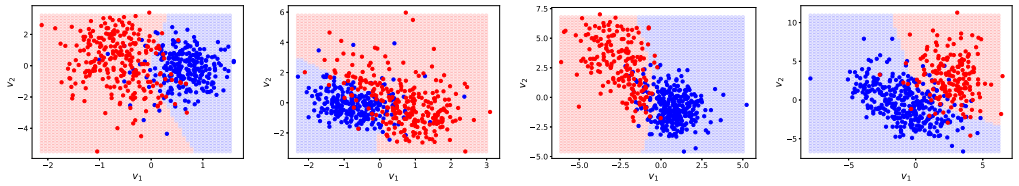
5. 使用*OrthoP*训练多样化分类器

受低维 SB 结果的启发，本节提出一种自然的方法来训练具有多样性的模型，从而使这些模型的集成能够缓解 SB。更具体地说，给定一个使用丰富 regime 初始化训练的初始模型 f，我们首先利用第一层权重的前几个主成分分析分量计算低维投影矩阵 P。

然后，我们通过对输入进行 $P_\perp$ 投影来训练另一个模型 f_{proj} ；也就是说，训练时不使用数据集 $(x^{(i)}, y^{(i)})$，而是使用 $(P_\perp x^{(i)}, y^{(i)})$ 来训练第二个模型（记为 f_{proj}）。我们将这种训练过程称为基于正交投影的 $OrthoP$。首先，我们证明，对于丰富 regime 中 IFM 内的任意数据集，该方法都能保证学到不同的特征集合。然后，我们提出两种自然的多样性度量，并展示 $OrthoP$ 能够在实际数据集上产生多样化模型。最后，我们提出一种自然的方法来集成这些模型，并在真实世界数据集上证明，这类集成的鲁棒性可以显著优于基线模型。

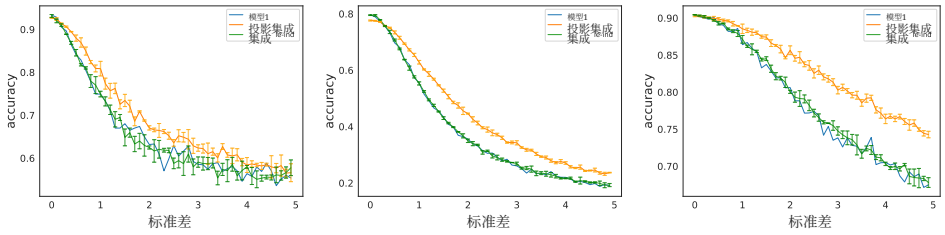
IFM 的理论证明：首先，我们从理论上确立，通过 $OrthoP$ 获得的 f 和 f_{proj} $OrthoP$ 对于 IFM 内的任意数据集都依赖于不同的特征。因此，根据 IFM 的定义， f 和 f_{proj} 在给定类别的条件下具有独立的 logit值。其证明见附录A.1.2。

命题 5.1。 考虑第*4.1*节所述的任意 *IFM* 数据集。令 f 为定理 *3.1*中描述的模型，并令 f_{proj} 为通过 $OrthoP$得到的第二个模型。那么，



(a) b-Imagenette (f) (b) b-Imagenette (f_{proj}) (c) Waterbirds (f) (d) Waterbirds (f_{proj})

图 *4.* 决策边界 对于 f 和 f_{proj}，B-Imagenette 和 Waterbirds 数据集在第一层权重矩阵前两个奇异方向上的可视化结果 e f_{proj} 的决策边界相比 f 的决策边界具有更强的非线性。
第一层权重矩阵。



（a）二元 Imagenette （b）I magenette （c）Waterbirds-Landbirds

图 *5.* 数据集预训练表示中添加的高斯噪声标准差与测试准确率之间的变化。模型1（即 f）保持不变，两个集成的数值均为 3 次运行的平均值。误差棒表示标准差。

输出 f 和 f_{proj} 在 x 上（即 $f(x)$ 和 $f_{proj}(x)$）分别仅取决于 x_1 和 $\{x_2, \cdots, x_d\}$。

~多样性度量：给定任意两个模型 f 和 f，我们使用两种度量对它们的多样性进行实证评估。第一种是错误多样性：错误多样性 $\left(f, \tilde{f}\right)$ 定义为= $1 - |\{i: f(\mathrm{x}^{(i)}) \neq y^{(i)} \& \tilde{f}(\mathrm{x}^{(i)}) \neq y^{(i)}\}| \min(|\{i: f(\mathrm{x}^{(i)}) \neq y^{(i)}\}|, |\{i: \tilde{f}(\mathrm{x}^{(i)}) \neq y^{(i)}\}|)$。这里我们稍微滥用记号，用 $f(x_i)$（分别用 $f(x_i)$）表示模型 f（分别为 $\tilde{f}$）对 x_i 的预测类别。错误多样性 $\left(f, \tilde{f}\right)$越高，说明~$f$ 与 f的错误重叠越少。第二种是类别条件对数几率相关性~，即在给定类别的条件下，f 与 f输出之间的相关性。具体而言，类别条件对数几率相关性 $\left(f, \tilde{f}\right)$ $= \sum_{y \in \mathcal{Y}} \mathrm{Corr}([f(\mathrm{x}_i)], [\tilde{f}(\mathrm{x}_i)]: y_i = y) |\mathcal{Y}|$，其中 $\mathrm{corr}([f(\mathrm{x}_i)], [\tilde{f}(\mathrm{x}_i)] : y_i = y)$ 表示在真实标签为 y 的数据点上，f 与 f 的~logit值之间的经验相关性。表 3 比较了两个独立训练的模型（f 和 f_{ind}）与两个顺序训练的模型（f 和 f_{proj}）的多样性。结果表明，f 和 f_{proj} 相比 f 和 f_{ind}具有更高的多样性。

图 4展示了 f 和 f_{proj} 在权重矩阵的前两个奇异向量所张成的 2维子空间上的决策边界。我们观察到，与第一个模型相比，第二个模型的决策边界非线性更强。

集成：图 5 展示了向预训练模型中添加的高斯噪声强度与测试准确率之间的变化关系

数据集表示。这里，集成是通过对两个模型的logit值取平均得到的。我们可以看到，与 f 和 f_{ind}的集成相比，f 和 f_{proj} 的集成具有强得多的鲁棒性

6. 讨论

在这项工作中，我们从模型对低维输入的依赖性角度，刻画了单隐藏层神经网络所表现出的简单性偏差（SB）。我们从理论上证明，对于一般类别的线性可分数据集，存在低秩简单性偏差。我们还在真实数据集上通过实证进一步验证了这一假设。基于这一刻画，我们还提出了 OrthoP——一种简单的集成技术，用于训练多样化模型，并表明它能使模型的高斯噪声鲁棒性显著提升。

这项工作是朝着严格定义神经网络的简单性偏差或捷径学习迈出的初步一步，后者被认为是导致其脆弱性的关键原因（Geirhos 等人，2020）。为更深层网络提供类似的刻画是一个重要的研究方向，这需要 更深入地 理解损失景观上梯度下降的训练动态和极限点。

致谢

我们感谢 Simons Investigator Fellowship、美国国家科学基金会资助项目 DMS-2134157、美国国防高级研究计划局资助项目

W911NF2010021 以及美国能源部资助项目 DE-SC0022199 的支撑。

参考文献

熊的图片来自维基百科。
https://en.wikipedia.org/wiki/Grizzly_熊#/media/File:GrizzlyBearJeanBeaufort.jpg。访问日期：2022-09-26。

天鹅的图片来自维基百科。
https://en.wikipedia.org/wiki/File:Mute_天鹅_(Cygnus_olor)_和_小天鹅.jpg。访问日期：2022-09-26。

Arora, S., Cohen, N., Hu, W. 和 Luo, Y. 深度矩阵分解中的隐式正则化。载于 Wallach, H., Larochelle, H., Beygelzimer, A., d'Alch´e-Buc, F., Fox, E. 和 Garnett, R.（编），神经信息处理系统进展，第 32 卷。Curran Associates, Inc.，2019。网址：
https://proceedings.neurips.cc/paper/2019/file/c0c783b5fc0d7d808f1d14a6e9c8280d-Paper.pdf。

Bietti, A. 和 Mairal, J. 神经切线核的归纳偏置。载于神经信息处理系统进展，第 32 卷，2019。网址：
https://proceedings.neurips.cc/paper/2019/file/c4ef9c39b300931b69a36fb3dbb8d60e-Paper.pdf。

Chen, T., Luo, C. 和 Li, L. 对比损失的有趣特性。神经信息处理系统进展，34：11834–11845，2021。Chen, Y., Wei, C., Kumar, A. 和 Ma, T. 自训练可避免在领域偏移下使用伪特征。神经信息处理系统进展，33：21061–21071，2020。Chizat, L. 宽双层神经网络的梯度下降——II：泛化与隐式偏置。
https://francisbach.com/gradient-descent-for-wide-two-layer-neu

2020年。Chizat, L. 和 Bach, F. 使用最优传输研究过参数化模型中梯度下降的全局收敛。载于 Bengio, S.、Wallach, H.、Larochelle, H.、Grauman, K.、Cesa-Bianchi, N. 和 Garnett, R.（编），神经信息处理系统进展，第 31 卷。Curran Associates, Inc.，2018年。网址
https://proceedings.neurips.cc/paper/2018/file/a1afc58c6ca9540d057299ec3016d726-Paper.pdf。

Chizat, L. 和 Bach, F.R. 逻辑损失训练的宽双层神经网络中梯度下降的隐式偏置。载于学习理论会议，*2020*，第 125 卷，第 1305–1338 页，2020年。

Chizat, L.、Oyallon, E. 和 Bach, F. 关于可微编程中的惰性训练。在 神经信息处理系统进展，第 32 卷，2019。网址：
https://proceedings.neurips.cc/paper/2019/file/ae614c557843b1df326cb29c57225459-Paper.pdf.

Cook, M.、Zare, A. 和 Gader, P. 通过神经网络的零空间分析进行异常值检测，2020。网址：
https://arxiv.org/abs/2007.01263。

Fang, C.、Lee, J.、Yang, P. 和 Zhang, T. 基于特征建模：过参数化深度神经网络的平均场框架。载于 Belkin, M. 和 Kpotufe, S.（编），第三十四届学习理论会议论文集，第 134 卷，机器学习研究论文集，第 1887–1936 页。PMLR，2021 年 8 月 15–19 日。网址：
https://proceedings.mlr.press/v134/fang21a.html。

FastAI。Imagenette数据集。
https://github.com/fastai/imagenette，2021年。

Galanti, T. 和 Poggio, T. 深度神经网络中的随机梯度下降噪声与隐式低秩偏置，2022。网址
https://arxiv.org/abs/2206.05794。

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wich- mann, F. A. 和 Brendel, W. 在 ImageNet 上训练的卷积神经网络偏向纹理；增强形状偏置可提高准确率和鲁棒性。载于国际学习表征会议，2018年。

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M. 和 Wichmann, F. A. 深度神经网络中的捷径学习。《自然·机器智能》，2(11):665–673，2020。

Glorot, X. 和 Bengio, Y. 理解深度前馈神经网络训练的难点。载于第十三届人工智能与统计国际会议论文集，第249–256页，2010年。

Gunasekar, S., Woodworth, B. E., Bhojanapalli, S., Neyshabur, B.，以及 Srebro, N. 矩阵分解中的隐式正则化。载于 Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. 和 Garnett, R.（编），神经信息处理系统进展，第 30 卷。Curran Associates, Inc.，2017年。网址：
https://proceedings.neurips.cc/paper/2017/file/

58191d2a914c6dae66371c9dcdc91b41-Paper.pdf.

Gunasekar, S., Lee, J. D., Soudry, D. 和 Sre- bro, N. 线性卷积网络中梯度下降的隐式偏差。载于神经信息处理系统进展，第 31 卷，2018年。网址：https://proceedings.neurips.cc/paper/2018/file/0e98aeeb54acf612b9eb4e48a269814c-Paper.pdf.

Hacohen, G. 和 Weinshall, D. 过参数化线性模型中的主成分偏置及其在深度神经网络中的表现。机器学习研究杂志，23(155):1–46，2022年。网址：http://jmlr.org/papers/v23/21-0991.html。

Hacohen, G., Choshen, L. 和 Weinshall, D. 让我们达成共识：神经网络在真实数据集上共享分类顺序。载于 III, H. D. 和 Singh, A.（编），第 *37* 届国际机器学习会议论文集，第 119 卷，机器学习研究论文集，第 3950–3960 页。PMLR，2020 年 7 月 13–18 日。网址：https://proceedings.mlr.press/v119/hacohen20a.html。

He, K., Zhang, X., Ren, S., and Sun, J. 深入研究整流器：在图像分类任务中超越人类水平的性能。载于*IEEE*国际计算机视觉会议论文集，第1026–1034页，2015年。

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., 等。鲁棒性的诸多面向：对分布外泛化的批判性分析。载于国际计算机视觉会议，第8340–8349页，2021年。

Huh, M., Mobahi, H., Zhang, R., Cheung, B., Agrawal, P., and Isola, P. 深度网络中的低秩简单性偏置。*arXiv*预印本*arXiv:2103.10427*，2021年。

Jacot, A., Gabriel, F. 和 Hongler, C. 神经切线核：神经网络中的收敛与泛化。载于神经信息处理系统进展，第 8580–8589 页，2018年。

Ji, Z. 和 Telgarsky, M. 梯度下降在不可分数据上的隐式偏差。载于学习理论会议，第 1772–1798 页，2019。

Ji, Z., Telgarsky, M. 和 Xian, R. 神经切线核、传输映射与通用逼近。载于国际学习表征会议，2020。

Kalimeris, D., Kaplun, G., Nakkiran, P., Edelman, B., Yang, T., Barak, B. 和 Zhang, H. 随机梯度下降在神经网络上学习递增复杂度的函数。载于 Wallach, H., Larochelle, H., Beygelzimer, A., d'Alch´e-Buc, F., Fox, E. 和 Garnett, R.（编），神经信息处理系统进展，第 32 卷。Curran Associates, Inc.，2019。网址：https://proceedings.neurips.cc/paper/2019/file/b432f34c5a997c8e7c806a895ecc5e25-Paper.pdf.

Lee, J., Xiao, L., Schoenholz, S., Bahri, Y., No- vak, R., Sohl-Dickstein, J. 和 Pennington, J. 任意深度的宽神经网络在梯度下降下演化为线性模型。载于神经信息处理系统进展，第 32 卷，2019。网址：https://proceedings.neurips.cc/paper/2019/file/0d1a9651497a38d8b1c3871c84528bd4-Paper.pdf.

Lee, Y., Yao, H., and Finn, C. 多样化并消除歧义：从欠规范数据中学习。*arXiv* 预印本 *arXiv:2202.03418*，2022。

Lewis, D. D. 四十年后的朴素（贝叶斯）方法：信息检索中的独立性假设。载于欧洲机器学习会议，第 4–15 页，1998年。

Li, Z., Luo, Y., and Lyu, K. 探索消除矩阵分解中梯度下降隐式偏差的方法：贪心低秩学习。载于国际学习表征会议，2021。网址：https://openreview.net/forum?id=AHOs7Sm5H7R。

Liu, E. Z., Haghgoo, B., Chen, A. S., Raghunathan, A., Koh, P. W., Sagawa, S., Liang, P., and Finn, C. 只需训练两次：在不使用训练群体信息的情况下提升群体鲁棒性。载于国际机器学习会议，第 6781–6792 页，2021。

Lyu, K.、Li, Z.、Wang, R. 和 Arora, S. 双层神经网络上的梯度下降：间隔最大化与简单性偏差。神经信息处理系统进展，34：12978–12991，2021年。

Nasery, A.、Addepalli, S.、Netrapalli, P. 和 Jain, P. Daft：蒸馏对抗微调模型，以实现更好的分布外泛化。*arXiv* 预印本 *arXiv:2208.09139*，2022。

Ndiour, I.、Ahuja, N. 和 Tickoo, O. 利用子空间技术和特征概率建模进行分布外检测，2020。网址 https://arxiv.org/abs/2012.04250。

Petz, D. 熵、冯·诺依曼与冯·诺依曼熵。载于约翰·冯·诺依曼与量子物理的基础，第83–96页。2001。

Pezeshki, M.、Kaba, O.、Bengio, Y.、Courville, A. C.、Precup, D. 和 Lajoie, G. 梯度饥饿：神经网络中的一种学习倾向。神经信息处理系统进展，34:1256–1272，2021年。

Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F. A., Bengio, Y., and Courville, A. 论神经网络的谱偏置。2018年。doi: 10. 48550/ARXIV.1806.08734。URLhttps://arxiv.org/abs/1806.08734。

Razin, N. and Cohen, N. 深度学习中的隐式正则化可能无法用范数解释。载于 Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (编), 神经信息处理系统进展，第 33 卷，第 21174–21187 页。Curran Associates, Inc., 2020年。
U
R
L
h
t
t
p
s://
p
r
oceedings.neurips.cc/paper/2020/file/f21e255f89e0f258accbe4e984eef486-Paper.pdf。
Ronen, B., Jacobs, D., Kasten, Y., and Kritchman, S. 不同频率学习函数的神经网络收敛速率。载于 Wallach, H., Larochelle, H., Beygelzimer, A., d'Alch´e-Buc, F., Fox, E., and Garnett, R. (编), 神经信息处理系统进展，第 32 卷。Curran Associates, Inc., 2019。
U
R
L
h
t
t
p
s://
p
r
oceedings.neurips.cc/paper/2019/file/5ac8bb8a7d745102a978c5f8ccdb61b8-Paper.pdf。
Rosenfeld, E., Ravikumar, P., and Risteski, A. 领域调整回归：经验风险最小化或许已经能够学习到足以实现分布外泛化的特征。 *arXiv* 预印本 *arXiv:2202.06856*，2022。Ross, A. and Doshi-Velez, F. 通过对输入梯度进行正则化，提升深度神经网络的对抗鲁棒性和可解释性。载于 *AAAI*人工智能会议，第 32 卷，2018年。
Roy, O. and Vetterli, M. 有效秩：有效维度的一种测度。载于 欧洲信号处理会议，第 606–610 页，2007年。
Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. 分布鲁棒神经网络。载于 国际学习表征会议，2020a。Sagawa, S., Raghunathan, A., Koh, P. W., and Liang, P. 探究过参数化为何会加剧伪相关。载于 国际机器学习会议，第 8346–8356 页，2020b。Scimeca, L., Oh, S. J., Chun, S., Poli, M., and Yun, S. 深度神经网络会选择哪些捷径线索？一项来自参数空间视角。载于国际学习表征会议，2021年。Shah, H., Tamuly, K., Raghunathan, A., Jain, P. 和 Netrapalli, P. 神经网络中简单性偏差的陷阱。神经信息处理系统进展，33：9573–9585，2020。Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S. 和 Srebro, N. 梯度下降在可分数据上的隐式偏差。机器学习研究杂志，19(1)：2822–2878，2018年1月。
Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. 和 Fergus, R. 神经网络的奇妙特性。载于国际学习表征会议，2014年。Teney, D., Abbasnejad, E., Lucey, S. 和 van den Hengel, A. 规避简单性偏差：训练多样化模型集，发现具有更优分布外泛化能力的解。载于 *IEEE/CVF*计算机视觉与模式识别会议，第16761–16772页，2022年。Wang, H., Li, Z., Feng, L. 和 Zhang, W. Vim：通过虚拟逻辑值匹配实现分布外检测。载于*IEEE/CVF*计算机视觉与模式识别会议，*2022*年计算机视觉与模式识别会议，新奥尔良，美国路易斯安那州，*2022*年*6*月*18*日至*24*日，第4911–4920页。IEEE，2022年。doi：10.1109/CVPR52688.2022.00487。
U
R
Lhttps://doi.org/10.1109/CVPR52688.2022.00487。
Xu, Y., He, H., Shen, T. 和 Jaakkola, T. 控制与分类器正交的方向。*arXiv*预印本 *arXiv:2201.11259*，2022年。Yang, G. 和 Hu, E. J. 张量程序 IV：无限宽神经网络中的特征学习。载于 Meila, M. 和 Zhang, T.（编），第*38*届国际机器学习会议论文集，第139卷，机器学习研究论文集，第11727–11737页。PMLR，2021年7月18–24日。
U
R
L
h
t
tps://proceedings.mlr.press/v139/yang21c.html。
Zaeemzadeh, A., Bisagno, N., Sambugaro, Z., Conci, N., Rahnavard, N. 和 Shah, M. 使用一维子空间的并进行分布外检测。载于*IEEE/CVF*计算机视觉与模式识别会议论文集（*CVPR*），第9452–9461页，2021年6月。

A. 丰富与惰性区间的证明

A.1. 丰富 regime

我们在下文重述定理4.1并给出证明。

定理 *A.1.* 对于 *IFM* 模型中满足第 *4.1* 节条件的任意数据集，$\gamma \geq 1$ 和 $f(\nu, x)$ 如方程 *(1)* 所示，分布 $\nu^* = 0.5\delta_{\theta_1} + 0.5\delta_{\theta_2}$ 在 $\mathcal{S}^{d+1}$ 上是满足方程 *(2)* 的唯一最大间隔分类器，其中 $\theta_1 = (\frac{\gamma}{\sqrt{2(1+\gamma^2)}}, \frac{1}{\sqrt{2(1+\gamma^2)}})$ $1/\sqrt{2})$ $\theta_2 = (\frac{\gamma}{\sqrt{2(1+\gamma^2)}}, \frac{1}{\sqrt{2(1+\gamma^2)}})$ $-1/\sqrt{2})$，且 $e_1 \overset{def}{=} [1, 0, \cdots, 0]$ 表示第一标准基向量。特别地，这意味着，如果在无限宽度极限下，采用丰富初始化的单隐藏层全连接网络的梯度流在交叉熵损失下收敛，并满足定理 *E.1* 中的技术条件，那么它将收敛到满足 $f(\nu^*, Px_1 + P_\perp x_2) = f(\nu^*, x_1) \forall (x_1, y_1), (x_2, y_2) \in D$ 的 ν^*，其中 P表示第一坐标上的（秩为 *1*）投影矩阵。

定理*A.1*的证明(Chizat 和 Bach，2020) 给出了方程（2）中最大间隔分类器的原始-对偶刻画：

引理 A.2。*(Chizat 和 Bach，2020)* ν^* 满足方程（*2*），当且仅当存在一个数据分布 p^* 使得以下两个互补松弛条件成立：

$$Supp(\nu^*) \subseteq \underset{(w,b,a)\in\mathbb{S}^{d+1}}{\arg\max} \mathbb{E}_{(x,y)\sim p^*} y[a(\phi(\langle w, x\rangle + b))] \quad and \tag{5}$$

$$Supp(p^*) \subseteq \underset{(x,y)\sim\mathcal{D}}{\arg\min}\, y\mathbb{E}_{(w,b,a)\sim\nu^*}[a(\phi(\langle w, x\rangle + b))]\,. \tag{6}$$

我们的计划是构造一个分布 p^*，使其满足上述引理的条件。

唯一性。还需注意，对于固定的 p^*，$\mathbb{E}_{(x,y)\sim p^*} yf(\nu, x)$是任意分类器 ν的间隔上界 $\min_{(x,y)\sim\mathcal{D}} yf(\nu, x)$。因此，要证明唯一性，只需证明 δ_{θ_1}，δ_{θ_2} 是方程（5）右侧目标函数的唯一最大化点，并且 δ_{θ_1}，δ_{θ_2} 在 $\mathcal{D}$ 上的唯一最大间隔凸组合是 ν^*。

我们首先描述 D p^*的支撑集。对于 $y \in \{\pm 1\}$，我们生成 $(x, y) \in D$，使其为

$$\mathbf{x}_1 = \gamma y$$

$$\forall i \in 2, .., d, \mathbf{x}_i = \begin{cases} \pm 1 & \text{for} \quad y = 1 \\ 0 & \text{for} \quad y = -1 \end{cases}$$

现在，对于 $(x, y) \in D$，定义

$$p^*(x, y) = \begin{cases} 0.5 & \text{for} \quad y = 1 \\ 0.5^d & \text{for} \quad y = -1 \end{cases} \tag{7}$$

注意，p^* 由以下对象支撑：2^{d-1} 个正例和一个负样本。我们先证明方程 (6)。

断言 A.3。p^* 如方程 (7) 所示，满足方程 (6)。此外，$\delta_{\theta_1}, \delta_{\theta_2}$ 的唯一最大间隔凸组合为 ν^*。

证明。对于任意 $\nu = \lambda\delta_{\theta_1} + (1-\lambda)\delta_{\theta_2}, 0 \leq \lambda \leq 1$，求 $yf(\nu, x) = y\mathbb{E}_{(w,b,a)\sim\nu^*}[a(\phi(\langle w, x\rangle + b))]$ 的极小化点 $(x, y) \sim \mathcal{D}$。

$yf(\nu, x)$ 对于 (x, y)，其中 $y = -1$（表示 x_1 乘以 $-\alpha_1$，其中 $\alpha_1 \geq \gamma$）为

$$\begin{aligned} yf(\nu, x) = -1\Big[&\lambda * \phi\left(\frac{\gamma}{\sqrt{2(1+\gamma^2)}}\mathbf{e}_1^\top(-\alpha_1\mathbf{e}_1) + \frac{1}{\sqrt{2(1+\gamma^2)}}\right) * \frac{1}{\sqrt{2}} \\ &+ (1-\lambda) * \phi\left(-\frac{\gamma}{\sqrt{2(1+\gamma^2)}}\mathbf{e}_1^\top(-\alpha_1\mathbf{e}_1) + \frac{1}{\sqrt{2(1+\gamma^2)}}\right) * \frac{-1}{\sqrt{2}}\Big]\,, \end{aligned}$$

并且对于 $(x,\ y)$，其中 $y=1$（表示 x_1 乘以 α_2，其中 $\alpha_2 \geq \gamma$）为

$$yf(\nu,x) = 1\Big[\lambda * \phi\left(\frac{\gamma}{\sqrt{2(1+\gamma^2)}}\mathbf{e}_1^\top(\alpha_2\mathbf{e}_1) + \frac{1}{\sqrt{2(1+\gamma^2)}}\right) * \frac{1}{\sqrt{2}} + (1-\lambda) * \phi\left(-\frac{\gamma}{\sqrt{2(1+\gamma^2)}}\mathbf{e}_1^\top(\alpha_2\mathbf{e}_1) + \frac{1}{\sqrt{2(1+\gamma^2)}}\right) * \frac{-1}{\sqrt{2}}\Big].$$

由于 $\gamma \geq 1$，上述表达式分别等于 $\lambda\frac{\sqrt{\gamma\alpha_1+1}}{2}$和 $(1-\lambda)\frac{\sqrt{\gamma\alpha_2+1}}{2}$，因此在 $\alpha_1=\alpha_2=\gamma$ 处达到最小值。因此，ν 的间隔为 $\min(\lambda,\ 1-\lambda)\frac{\sqrt{1+\gamma^2}}{2}$，且在 $\lambda=1/2$ 处取得唯一最大值。此外，对于 $\lambda=1/2,\ D$ 中的所有点都有相同的 $yf(\nu,\ x)$ 值。 □

在证明的其余部分，我们证明方程 (5)。记 $g(w,\ b,\ a) := \mathbb{E}_{(x,y)\sim p^*} y[a(\phi(\langle w,x\rangle + b))]$。我们证明$\delta_{\theta_1},\ \delta_{\theta_2}$是 $g(w,\ b,\ a)$ 在 $\mathbb{S}^{d+1}$上的唯一最大化点。

我们首先找到 $g(\theta_1)$ $g(\theta_2)$。

$$g(\theta_1) = \Pr(y=1)\cdot 1\cdot\frac{1}{\sqrt{2}}\cdot\phi\left(\frac{\gamma}{\sqrt{2(1+\gamma^2)}}\mathbf{e}_1^T(\gamma\mathbf{e}_1) + \frac{1}{\sqrt{2(1+\gamma^2)}}\right) + \Pr(y=-1)\cdot -1\cdot\frac{1}{\sqrt{2}}\cdot\phi\left(\frac{\gamma}{\sqrt{2(1+\gamma^2)}}\mathbf{e}_1^T(-\gamma\mathbf{e}_1) + \frac{1}{\sqrt{2(1+\gamma^2)}}\right) = \frac{\sqrt{\gamma^2+1}}{4},$$

其中，第一项是因为对于 $\theta_1,\ w_2、\ w_3、\ \ldots\ldots、\ w_d$ 均为零。类似地，$g(\theta_2) = \frac{\sqrt{\gamma^2+1}}{4}$。现在我们证明 $g(w、\ a、\ b) < \frac{\sqrt{\gamma^2+1}}{4}$ for $(w,a,b) \notin \{\theta_1,\theta_2\}$.

我们首先证明以下这个简单但有用的结论。

结论 A.4。在 $\mathbb{S}^{d+1}$ 上使 $g(w、\ b、\ a)$ 达到最大值的所有点都满足 $|a| = 1/\sqrt{2}$。

证明。该证明主要基于 ReLU 函数 ϕ 的 $1-$ 齐次性，以及 $g(w、\ b、\ a)$ 的可分离性。注意，$g(w、\ b、\ a) = \sqrt{\|w\|^2+b^2}\,a\cdot g(w'、\ b'、\ 1)$，其中 $\|w'\|^2+b^2=1$。最大化 $g(w、\ b、\ a)$ 等价于分别在 $\mathbb{S}^d$ 上最大化$g(w'、\ b'、\ 1)$，并在 $\mathbb{S}^{d+1}$ 上最大化 $a\sqrt{\|w\|^2+b^2}$。后者在 $|a|=1/\sqrt{2}$ 处取得唯一最大值，证明完毕。 □

现在将 $g(w,\ b,\ a)$ 表示为

$$g(w,b,a) = a\left(\Pr(y=1)\mathbb{E}[\phi(w^Tx+b)|y=1] - \Pr(y=-1)\mathbb{E}[\phi(w^Tx+b)|y=-1]\right) = \frac{a}{2}\left(\mathbb{E}_\sigma\left[\phi(\gamma w_1 + b + \sum_{i=2}^d \sigma_i w_i)\right] - \phi(b-\gamma w_1)\right), \tag{8}$$

其中 σ_i 是相互独立的拉德马赫随机变量。对于 a，有两种情况：

情况 1： $a = 1/\sqrt{2}$。根据方程（8），我们有

$$g(w,b,1/\sqrt{2}) \leq \frac{1}{2\sqrt{2}}\mathbb{E}_\sigma\left[\phi(\gamma w_1 + b + \sum_{i=2}^d \sigma_i w_i)\right].$$

为简化上述内容，定义随机变量 $X=\sum_{i=2}^d \sigma_i w_i$，并将 $\gamma w_1 + b$ 记为 α。注意，$|\alpha| = |\gamma w_1 + b| \leq \sqrt{\frac{\gamma^2+1}{2}}$，这由$\|w\|^2+b^2 = 1/2$推出。上一个表达式中的期望变为

$$\mathbb{E}[\phi(X+\alpha)] = \mathbb{E}[(X+\alpha)\mathbf{1}\{X+\alpha\geq 0\}] = \mathbb{E}[X\mathbf{1}\{X\geq -\alpha\}] + \alpha\Pr(X\geq -\alpha) = \mathbb{E}[X\mathbf{1}\{X\geq\alpha\}] + \alpha(1-\Pr(X\geq\alpha)) \leq \mathbb{E}[X\mathbf{1}\{X\geq\alpha\}] + \alpha,$$

其中，最后一个等式由 X 的对称性推出。注意，$\mathrm{Var}(X) = \sum_{i=2}^{d} w_i^2$，其至多为 $\frac{1}{2} - \frac{\alpha^2}{1+\gamma^2}$（利用 $\gamma w_1 + b = \alpha$ 和 $\|w\|^2 + b^2 = 1/2$）。利用 A.5 对 $\mathbb{E}[X\mathbf{1}\{X \geq \alpha\}]$ 进行上界估计，可得

$$\mathbb{E}[\phi(X+\alpha)] \leq \alpha + \sqrt{\frac{1}{2}\min\left(\frac{1}{2}, \frac{\frac{1}{2} - \frac{\alpha^2}{1+\gamma^2}}{2\alpha^2}\right)\left(\frac{1}{2} - \frac{\alpha^2}{1+\gamma^2}\right)}.$$

我们可以验证，上式右侧在 $\alpha = \sqrt{\frac{1+\gamma^2}{2}}$ 处有唯一的最大化子 $|\alpha| \leq \sqrt{\frac{1+\gamma^2}{2}}$。因此，在这种情况下，$g(w,b,a) \leq \sqrt{\frac{1+\gamma^2}{4}}$。由于任何满足 $\gamma w_1 + b = \sqrt{\frac{1+\gamma^2}{2}}$ 和 $w_1^2 + b^2 \leq 1/2$ 的 (w_1, b) 都有 $b = \frac{1}{\sqrt{2(1+\gamma^2)}}$，至此便完成证明。

情形2： $a = -1/\sqrt{2}$。利用方程（8），我们有 $g(w,b,-1/\sqrt{2}) \leq \phi(b - \gamma w_1)/2\sqrt{2}$，对于 $b^2 + w_1^2 \leq 1/2$，其唯一最大值 $\sqrt{\frac{\gamma^2+1}{4}}$ 在 $b = \frac{1}{\sqrt{2(1+\gamma^2)}}$ 处取得。

最后请注意，训练后的网络的权重（w，b，a）从 ν^*中采样。因此，定理中关于 f（ν^*，$Px_1 + P_\perp x_2$）的最后一项结论成立，因为 w的分布仅支撑在$\mathbf{e}_1$和 $-\mathbf{e}_1$上。

□

A.1.1. 丰富机制的辅助引理

引理A.5。对于任意具有有界方差的对称离散随机变量 X，当 $\alpha > 0$时，

$$\mathbb{E}[X\mathbb{I}(X \geq \alpha)] \leq \sqrt{\frac{1}{2}\min\left(\frac{1}{2}, \frac{Var(X)}{2\alpha^2}\right) Var(X)}.$$

证明。

$$\mathbb{E}[X\mathbb{I}(X \geq \alpha)] = \sum_{x \geq \alpha} xp(x) = \sum_{x \geq \alpha} \sqrt{p(x)}\sqrt{p(x)}x \leq \sqrt{p(X \geq \alpha)\sum_{x \geq \alpha} x^2 p(x)}, \tag{9}$$

其中，最后一个不等式由柯西-施瓦茨不等式得出。由切比雪夫不等式，$p(|X| \geq \alpha) \leq Var(X)/2\alpha^2$。将此与方程（9）结合，并利用的对称性以及 α 的非负性，即得所需引理。 □

A.1.2. 命题 5.1 的证明

我们重述命题 5.1并给出证明

命题 A.6. 考虑任意一个如第 *4.1* 节所述的 *IFM* 数据集。设 f 是定理 *3.1* 中描述的模型，且f_{proj}是通过 $OrthoP$ 得到的第二个模型。那么，输出 f和 f_{proj}在 x上，即 $f(x)$和 $f_{proj}(x)$，分别仅取决于x_1和 $\{x_2, \cdots, x_d\}$。

证明。如定理 3.1 所示，权重的最终分布为 $\nu^* = 0.5\delta_{\theta_1} + 0.5\delta_{\theta_2}$，其中 $\theta_1 = (\frac{\gamma}{\sqrt{2(1+\gamma^2)}}\mathbf{e}_1, \frac{1}{\sqrt{2(1+\gamma^2)}}\ 1/\sqrt{2})$ $\theta_2 = (\frac{\gamma}{\sqrt{2(1+\gamma^2)}}\mathbf{e}_1, \frac{1}{\sqrt{2(1+\gamma^2)}}\ -1/\sqrt{2})$，且 $\mathbf{e}_1$ 表示 $[1, 0, \cdots, 0]$ 表示第一个标准基向量。

由于第一层权重矩阵的支撑集仅沿着$\mathbf{e}_1$ 方向，因此其顶部奇异向量也指向$\mathbf{e}_1$ 方向。因此，$P = \mathbf{e}_1\mathbf{e}_1^\top$且 $P_\perp = I - \mathbf{e}_1\mathbf{e}_1^\top$，其中 I 表示单位矩阵。因此，对于 $y = +1$ 和 $y = -1$，将输入通过 $P_\perp$ 投影后得到的数据集在线性坐标上的取值均为 0。因此，它不能沿线性坐标分离。由此，第二个模型 f_{proj} 依赖其他坐标进行分类。 □

A.2. 惰性机制

定理 4.2 是以下更一般定理的推论。

定理 A.7。考虑一个点 $x \in D$。对于足够小的 $\epsilon > 0$，存在一个绝对常数 N，使得对于所有 $d > N$、$\gamma < \epsilon\sqrt{d}$ 和 $\gamma \geq 7$，在 *NTK regime* 下对单隐藏层全连接网络的两层进行联合训练时，任意形如（ζ，$x_{2:d}$）的点的预测满足以下条件：

1. 对于 $\zeta \geq 0.73$，预测为正。
2. 对于 $\zeta \leq -0.95\gamma$，预测为负。

上述定理表明，对 x_1 施加幅度为 $O(\gamma)$ 的扰动，会改变 $x \in D$ 的 $pred(f(x))$（而存在一个分类器，能在 D 上实现 $\Omega(\sqrt{d})$ 的间隔，因为 $\bar{D}$ 在坐标 $\{2 \cdots d\}$ 上的间隔为 1）。由于 $\gamma = o(d)$，这表明学习得到的模型易受对抗性攻击。

定理的证明 A.7. 证明思路是将标准的核最大间隔支持向量机理论应用于神经正切核 $f(x)$，从而得到其显式表达式，即 3.2。

我们先介绍一些预备知识。我们将样本的第一坐标称为"线性"坐标，其余坐标称为"非线性"坐标。此外，从现在起，我们给所有样本追加一个值为 1 的额外坐标（对应于偏置项）；这是使用无偏置支持向量机时的标准做法，且不失一般性。

的显式表达式为 f。利用最大间隔核支持向量机的表示定理，我们知道 f 可以表示为

$$f(x) = \sum_{(x^{(t)}, y^{(t)}) \in D} \lambda_t y^{(t)} K(x, x^{(t)}) ,$$

对于某些 $\lambda_t \geq 0$（称为拉格朗日乘子）。此外，根据 KKT 条件，具有此类表示（且能正确分类 D）的函数在满足以下条件时具有最大间隔：$y^{(t)} f(x^{(t)}) = 1$，只要 $\lambda_t > 0$（满足 $\lambda_t > 0$ 的训练点 t称为支持向量）。

我们先从一个有用的断言开始。

断言A.8。 对于 D，使用神经正切核的最大间隔核支持向量机将 D 中的所有点都作为支持向 ors.

证明。根据上述讨论，只需证明 $K\alpha = y$ 的（唯一）解 $\alpha \in \mathbb{R}^{|D|}$ 满足：对于所有 i，$\text{sign}(\alpha_i) = y^{(i)}$，其中 K 是 $|D| \times |D|$ 格拉姆矩阵，其（i，j）项为 $K(x^{(i)}, x^{(j)})$，而 $y_i = y^{(i)}$（拉格朗日乘子λ_i 则由 $y_i \alpha_i$给出）。

格拉姆矩阵的结构。 对 D 排序，使正例排在前面。于是，格拉姆矩阵 K 具有如下分块结构：$\begin{pmatrix} B & C \\ C^T & R \end{pmatrix}$ 其中，$B \in \mathbb{R}^{2^{d-1} \times 2^{d-1}}$ 和 $R \in \mathbb{R}$ 分别是正例和负例的格拉姆矩阵，而 $C \in \mathbb{R}^{2^{d-1} \times 1}$ 表示 $K(x^{(i)}, x^{(|D|)})$ 在 $i < |D|$处的取值。

回顾一下，对于神经正切核，$K(x^{(i)}, x^{(j)})$具有形式 $\|x^{(i)}\|\|x^{(j)}\|\kappa(\langle x^{(i)}, x^{(j)} \rangle)$。注意，所有正实例的范数都相同（记为 $\rho_1 = \sqrt{d + \gamma^2}$），且两个正实例之间的内积仅取决于不匹配的非线性坐标的数量（记为 β_i，对应 $0 \leq i \leq d-1$）。因此，B 的各行互为排列，其中条目 $\rho_1^2 \beta_i$ 出现 $\binom{d-1}{i}$ 次。同样，C 中的各个条目都相等，记为 $\rho_1 \rho_2 \beta_d$，其中 β_d 表示任意 $t < |D|$ 和 $\rho_2 = \|x^{|D|}\| = \sqrt{1 + \gamma^2}$的 $\kappa(x^{(t)}, x^{|D|})$。~~R~~ 中唯一的条目是 $\rho_2^2 \kappa(1)$。特别地，

$$\beta_i = \kappa\left(\frac{d - 2i + \gamma^2}{d + \gamma^2}\right) \text{ for } i \in [|D| - 1], \qquad \text{and} \qquad \beta_d = \kappa\left(\frac{1 - \gamma^2}{\sqrt{d + \gamma^2}\sqrt{1 + \gamma^2}}\right) .$$

现在我们可以求解 $K\alpha = y$ 了。根据 K 的结构对称性，α 看起来像 $[a, a, \ldots\ldots, b]$，其中前 $|D| - 1$ 项相同。

展开 $K\alpha = y$，得到以下两个方程：

$$a\rho_1^2 \left(\sum_{i=0}^{d-1} \binom{d-1}{i} \beta_i\right) + b\rho_1 \rho_2 \beta_d = 1 \qquad \text{and} \qquad 2^{d-1} a \rho_1 \rho_2 \beta_d + \rho_2^2 \kappa(1) b = -1 .$$

解得

$$a = \frac{\rho_2 \kappa(1) + \rho_1 \beta_d}{\rho_1^2 \rho_2 \sum_{i=0}^{d-1} \left(\binom{d-1}{i} [\kappa(1)\beta_i - \beta_d^2]\right)} \qquad \text{and} \qquad b = \frac{-1 - 2^{d-1} a \rho_1 \rho_2 \beta_d}{\rho_2^2 \kappa(1)} .$$

现在我们证明 $a > 0$ 和 $b < 0$。注意，当 d充分大时， β_d 可以任意接近 $\kappa(0) = 1/\pi$（因为 κ 在 0 附近是光滑的）。因此， $a > 0$ 意味着 $b < 0$。事实上，我们给出 a的如下估计：

$$a = 2^{1-d} \cdot \frac{\rho_2\kappa(1) + \rho_1\beta_d}{\xi\rho_1^2\rho_2} \qquad \text{where} \qquad \frac{2}{\pi} - \frac{1}{\pi^2} + O\left(\frac{1}{d}\right) \le \xi \le 2 + O\left(\frac{1}{d}\right). \tag{10}$$

对于 ξ的下界，写作

$$\begin{aligned}
&\sum_{i=0}^{d-1} \binom{d-1}{i} [\kappa(1)\beta_i - \beta_d^2] = \kappa(1) \sum_{i=0}^{\lfloor d/2 \rfloor} \binom{d-1}{i} (\beta_i + \beta_{d-1}) - 2^{d-1}\beta_d^2 \\
&\ge \kappa(1) \sum_{i=0}^{\lfloor d/2 \rfloor} \binom{d-1}{i} 2\beta_{d/2} - 2^{d-1}\beta_d^2 \ge 2^{d-1} \left(\kappa(1)\kappa(0) - \kappa^2(0) + O\left(\frac{1}{d}\right)\right),
\end{aligned}$$

其中，第一个不等式使用了 κ 的凸性，第二个不等式使用了 $\beta_{d/2} = \kappa(0) + O(1/d)$ $\beta_d = \kappa(0) + O(1/\sqrt{d})$。对于 ξ 的上界，写作

$$\begin{aligned}
&\sum_{i=0}^{d-1} \binom{d-1}{i} [\kappa(1)\beta_i - \beta_d^2] \le \kappa(1) \sum_{i=0}^{d-1} \binom{d-1}{i} \kappa\left(1 - \frac{2i}{d+\gamma^2}\right) \\
&\le \kappa(1) \sum_{i=0}^{d-1} \binom{d-1}{i} \left(2 - \frac{2i}{d+\gamma^2}\right) = \kappa(1)2^d - \frac{\kappa(1)(d-1)2^{d-1}}{d+\gamma^2},
\end{aligned}$$

其中，第二个不等式使用了 $\kappa(u) \le 1 + u$ （这由凸性以及 $\kappa(-1) = 0$， $\kappa(1) = 2$成立）。 □

现在，我们分析形式为（ζ， $x_{2:d+1}$）的点的预测标签，其中 $x \in D$。根据 x的标签，我们分两种情况讨论。

点$(\zeta,\ x_2^{(t)} : d+1)$ **的预测标签，其中** $x^{(t)} \in D$ **具有正标签**

我们的点（记为 x）具有如下形式：$(\zeta,\ \zeta_1,\ \zeta_2,\ \ldots,\ \zeta_d,\ 1)$，其中 $\zeta_i \in \pm 1$。证明的思路是将 f 显式地写成 ζ 的函数，并在 $\zeta = \gamma$ 附近对其进行一阶泰勒展开，同时做一些额外处理来解决 f 的非光滑性问题。

的显式形式 f。令 τ_i 定义为$= \langle x, x' \rangle / (\|x\|\|x'\|)$，其中正实例 $x' \in D$ 中， x 和 x' 恰有 i 个不匹配的非线性坐标（对于 $0 \le i \le d-1$）。类似地，将 $\langle x, x^{|D|} \rangle / (\|x\|\|x^{|D|}\|)$ 记为 τ_d。特别地，

$$\tau_i = \left(\frac{d - 2i + \gamma\zeta}{\rho_1\|x\|}\right) \qquad \text{and} \qquad \tau_d = \left(\frac{1 - \gamma\zeta}{\rho_2\|x\|}\right).$$

根据上述讨论，我们有

$$f(x) = a\left(\sum_{t=1}^{|D|-1} K(x, x^{(t)})\right) + bK(x, x^{|D|}) = a\rho_1\|x\| \left(\sum_{i=0}^{d-1} \binom{d-1}{i} \kappa(\tau_i)\right) + b\rho_2\|x\|\kappa(\tau_d).$$

代入 b，并将 $f(x)/\|x\|$ 记为 $g(\zeta)$，得到

$$g(\zeta) = a\rho_1 \left[\sum_{i=0}^{d-1} \binom{d-1}{i} \kappa(\tau_i(\zeta)) - \frac{2^{d-1}\beta_d}{\kappa(1)} \kappa(\tau_d(\zeta))\right] - \frac{\kappa(\tau_d(\zeta))}{\rho_2\kappa(1)}. \tag{11}$$

现在尝试在 $\zeta = \gamma$ 附近，利用泰勒级数展开 $g(\zeta)$（注意， $g(\gamma) = 1/\rho_1$）。但要注意， κ' 在 -1 和 1 附近可能无界。为了解决这个问题，令 $g = h + q$，其中 h 的一阶导数和二阶导数有界，而对于我们感兴趣的 ζ， q 的阶低于 h。具体而言，

$$h(\zeta) = a\rho_1 \left[\sum_{i=d/4}^{3d/4} \binom{d-1}{i} \kappa(\tau_i(\zeta)) - \frac{2^{d-1}\beta_d}{\kappa(1)} \kappa(\tau_d(\zeta))\right] - \frac{\kappa(\tau_d(\zeta))}{\rho_2\kappa(1)} \qquad \text{and}$$

$$q(\zeta) = a\rho_1 \left[\sum_{i:|d/2-i|>d/4} \binom{d-1}{i} \kappa(\tau_i(\zeta)) \right] .$$

注意， $q(\zeta) = o(c^d)$ 对于 $c < 1$ ，利用方程 (10) 中针对 a 的估计，以及独立伯努利随机变量之和的集中性。根据泰勒定理,

$$g(\zeta) = h(\gamma) + h'(\gamma)(\zeta - \gamma) + \frac{h''(\theta)(\zeta-\gamma)^2}{2} + q(\zeta) , \tag{12}$$

对于某个 $\theta \in [\gamma, \zeta]$，其中 $h(\gamma) \approx 1/\sqrt{d}$。结果将表明， $|h'(\gamma)| = \Theta(1/\sqrt{d})$， $|h''(\zeta)| = o(1/\sqrt{d})$。这样，我们就可以通过忽略二阶项并 $q(\zeta)$，利用 $g(\zeta)$ 的线性近似来完成证明。现在我们计算 h'， h''，为便于说明，将 $\|x\| = \sqrt{d + \zeta^2}$ 视为常数（证明不依赖于这一近似；或者读者也可以将 γ 视为 $o(\sqrt{d})$）。利用 $\tau_i'(\zeta) \approx \frac{\gamma}{\rho_1 \|x\|}$， $\tau_d'(\zeta) \approx \frac{-\gamma}{\rho_2 \|x\|}$，

$$h'(\zeta) \approx a\rho_1 \left[\sum_{i=0}^{d-1} \binom{d-1}{i} \kappa'(\tau_i(\zeta)) \frac{\gamma}{\rho_1 \|x\|} + \frac{2^{d-1}\beta_d}{\kappa(1)} \kappa'(\tau_d(\zeta)) \frac{\gamma}{\rho_2 \|x\|} \right] + \frac{\kappa'(\tau_d(\zeta))}{\rho_2 \kappa(1)} \frac{\gamma}{\rho_2 \|x\|}$$

$$h''(\zeta) \approx a\rho_1 \left[\sum_{i=0}^{d-1} \binom{d-1}{i} \kappa''(\tau_i(\zeta)) \frac{\gamma^2}{\rho_1^2 \|x\|^2} - \frac{2^{d-1}\beta_d}{\kappa(1)} \kappa''(\tau_d(\zeta)) \frac{\gamma^2}{\rho_2^2 \|x\|^2} \right] - \frac{\kappa''(\tau_d(\zeta))}{\rho_2 \kappa(1)} \frac{\gamma^2}{\rho_2^2 \|x\|^2} .$$

代入 $\|x\| \approx \rho_1 \approx \sqrt{d}$ 并将 a 从方程 (10) 中代入,

$$h'(\zeta) = \frac{(1 + \beta_d^2/\xi)\kappa'(\tau_d(\zeta))\gamma}{\rho_2^2 \kappa(1)\sqrt{d}} + o\left(\frac{1}{\sqrt{d}}\right) \qquad \text{and} \qquad h''(\zeta) = O\left(\frac{1}{d}\right) ,$$

将其代入方程 (12) 并取 $\tau_d(\zeta) \approx 0$， $\beta_d \approx \kappa(0)$ $\kappa'(\tau_d(\zeta)) \approx \kappa'(0)$ 可得

$$g(\zeta) = \frac{1}{\sqrt{d}} \left(1 + \frac{(1 + \kappa^2(0)/\xi)\kappa'(0)\gamma}{\kappa(1)\rho_2^2} (\zeta - \gamma) \right) + o\left(\frac{1}{\sqrt{d}}\right) ,$$

因此，只要上述 $1/\sqrt{d}$ 的系数有正的下界， $g(\zeta) > \bar{0}$ 就成立；对于$g(\zeta) < 0$，也有类似条件。利用方程 (10) 中 ξ 以及 $\kappa'(0) = 1$、 $\kappa(0) = 1/\pi$、 $\kappa(1) = 2$和 $\rho_2^2 = 1 + \gamma^2$ 的估计代入上式，可得 $g(\zeta) > 0$ 对 $\zeta > -0.68\gamma - 1.68/\gamma$ 成立，而 $g(\zeta) < 0$ 对 $\zeta < -0.905\gamma - 1.905/\gamma$成立。

点 $(\zeta, x_2^{(t)}{}_{:d+1})$ 的预测标签，其中 $x^{(t)} \in D$ 具有负标签

按照相同的方案，将我们的点（记为 x）写为 $(\zeta, 0, \ldots, 0, 1)$。

f 的显式形式。首先求出

$$\tau_i = \left(\frac{1 + \gamma\zeta}{\rho_1 \|x\|} \right) \qquad \text{and} \qquad \tau_d = \left(\frac{1 - \gamma\zeta}{\rho_2 \|x\|} \right) .$$

方程 (11) 现在给出

$$g(\zeta) = 2^{d-1} a\rho_1 \left[\kappa(\tau_0(\zeta)) - \frac{\beta_d \kappa(\tau_d(\zeta))}{\kappa(1)} \right] - \frac{\kappa(\tau_d(\zeta))}{\rho_2 \kappa(1)} .$$

在 $\zeta = -1/\gamma$ 附近用泰勒级数展开 $\kappa(\tau_0(\zeta))$,

$$\kappa(\tau_0(\zeta)) = \kappa(0) + \kappa'(\tau_0(\theta))\tau_0'(\theta)(\zeta + \frac{1}{\gamma}) ,$$

对于某个 $\theta \in [-1, 1]$。当 d 很大时， $\tau_0(\theta) \approx 0$ 且 $\tau_0'(\theta) = O(1/\sqrt{d})$。因此我们有

$$\begin{aligned} g(\zeta) &= \frac{\rho_2\kappa(1) + \rho_1\beta_d}{\xi\rho_1\rho_2} \left[\kappa(0) + O\left(\frac{1}{\sqrt{d}}\right) - \frac{\beta_d \kappa(\tau_d(\zeta))}{\kappa(1)} \right] - \frac{\kappa(\tau_d(\zeta))}{\rho_2 \kappa(1)} \\ &= \frac{1}{\rho_2} \left(\frac{\kappa^2(0)}{\xi} - \left(\frac{\kappa^2(0)}{\xi\kappa(1)} + \frac{1}{\kappa(1)} \right) \kappa(\tau_d(\zeta)) \right) + o(1) . \end{aligned}$$

与之前一样， $g(\zeta) > 0$ 只要上文中 $1/\rho_2$ 的系数有一个正的上界，这种情况就会发生；对于 $\zeta \geq 0.73$ （对于 $\gamma \geq 3$）也是如此。类似地， $g(\zeta) < 0$ 对于 $\zeta \leq 0$ 也是如此。 □

B. 实验

在本节中，我们将介绍 实验细节，包括超参数调优设置以及一些额外实验 .

B.1. 实验设置详情

我们将首先介绍本研究中使用的四个数据集。

1. **Imagenette**(FastAI，2021年）：这是 Imagenet 中的一个子集，包含 10 个类别，这些类别相对容易分类。
2. **b-Imagenette**：这是 Imagenette 的二值化版本，只使用其中两个类别（丁鲷和英国史宾格犬）的子集。
3. **Waterbirds-Landbirds** (Sagawa 等人，2020a）：这是一个多数-少数群体数据集，包含水中的水鸟和陆地背景中的水鸟，以及陆地上的陆鸟和水背景中的陆鸟。该数据集可作为基准，用于检验模型在预测鸟类类别时对虚假背景特征的依赖性，因为大多数训练样本中的水鸟都处于水背景中，而陆鸟都处于陆地背景中。
4. **MNIST-CIFAR** (Shah 等人，2020）：这是一个拼贴数据集，由沿某一轴拼接 MNIST 和 CIFAR 图像而成。这是一个合成数据集，用于评估训练后模型的简单性偏差。

设置 本文始终使用上述数据集的预训练表示，这些表示由 ImageNet 预训练的 ResNet-50 提取。我们在这些表示之上微调一个隐藏维度为 100 的单隐藏层全连接网络（骨干网络保持固定），并使用动量为 0.9 的随机梯度下降进行训练。每个模型训练 20000 步，并采用预热和余弦衰减学习率调度器。对于每次实验，我们根据验证准确率调整批量大小、学习率和权重衰减。超参数调优的详细信息如下：

- 批量大小 $\in \{128, 256\}$
- 学习率：– 丰富 regime： $\in \{0.5, 1.0\}$ （因为丰富 regime 中的学习率需要随隐藏维度增大而提高）– 惰性区间： $\in \{0.01, 0.05\}$
- 权重衰减： $\in \{0, 1e^{-4}\}$

最终报告的数值是使用所选超参数进行的 3 次独立实验的平均值。

评估 对于 Imagenette、b-Imagenette 和 MNIST-CIFAR，我们在所有实验中报告标准测试准确率。对于 waterbirds，我们报告训练集调整后的测试准确率，具体如Sagawa 等人(2020a)所述。具体而言，分别计算测试数据中每个群组的准确率，再按相应群组在训练数据集中的比例加权。

B.2. 其他实验结果

本节将介绍一些补充实验结果。

准确率 f_{proj} 在表 4 和5中，我们分别展示了 f_{proj} 在丰富区间和惰性区间下的测试准确率。可以看出，即使投影掉 f所使用的 f_{proj} 主成分，仍能达到相当高的准确率。请注意，在这些实验中，模型1保持固定，而 f_{proj} 的准确率取3次运行的平均值。

奇异值衰减 。在图 6中，我们给出了在丰富区间下训练的第一个模型的权重矩阵的奇异值衰减情况。可以看出，最大的几个奇异值捕获了矩阵弗罗贝尼乌斯范数的大部分。

MNIST-CIFAR 在图 7中，我们展示了以下模型的集成 f 和 f_{proj} 在 MNIST-CIFAR 数据集上的高斯鲁棒性优于以下模型的集成：f 和 f_{ind} 。

表 *4.* 丰富 regime 中 f_{proj} 的训练准确率

数据集	准确率(f)	准确率(f_{proj})
b-Imagenette	93.35	91.35 ± 0.32
Imagenette	79.67	71.93 ± 0.12
Waterbirds	90.29	89.92 ± 0.08
MNIST-CIFAR	99.69	98.95 ± 0.02

表 *5.* f_{proj} 在惰性区间内的训练准确率

数据集	准确率(f)	准确率(f_{proj})
b-Imagenette	93.09	91.77 ± 0.34
Imagenette	80.31	77.34 ± 0.21
Waterbirds	90.4	89.5 ± 0.18
MNIST-CIFAR	99.74	98.54 ± 0.00

决策边界非线性的定量测度 在本节中，我们报告了沿着前两个奇异向量方向的决策边界非线性的定量测度，针对 f 和 f_{proj}。基本上，我们将一个线性分类器拟合到决策边界，并报告其准确率。如表 6所示，线性分类器对 f_{proj} 的测试准确率低于 f。

随深度变化的 LD-SB 在图 8 和 9中，我们展示了深度为 2 和 3 的 ReLU 网络中权重矩阵有效秩的演变。可以看到，秩仍会随着训练而降低，但对于初始层而言，这种效应不那么明显。请注意，这些实验所采用的初始化是 Yang & Hu（2021年）提出的特征学习初始化。Yang& Hu(2021年)。

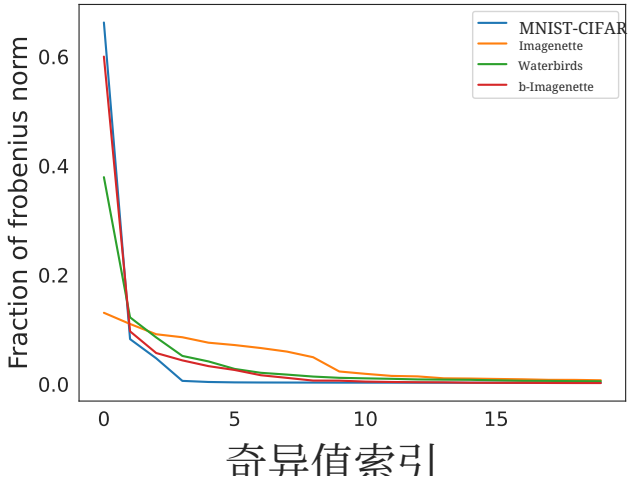


图*6.* 在丰富 regime 下训练的第一层权重矩阵中，前 i^{th} 个奇异值所占弗罗贝尼乌斯范数的比例，即 $\sigma_i^2 / \sum_{j=1}^{d} \sigma_j^2$ 与 i，针对各种数据集的结果。

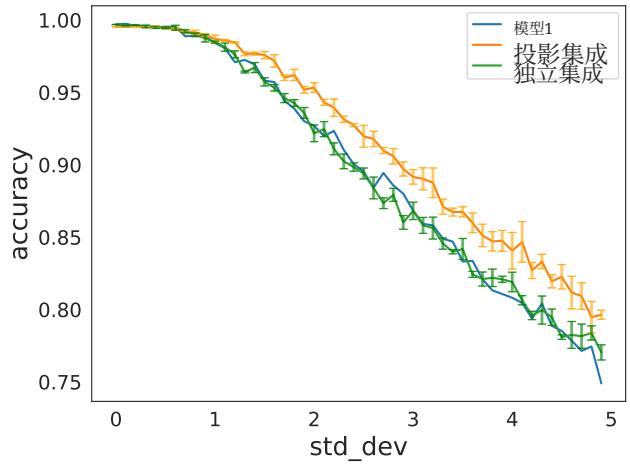
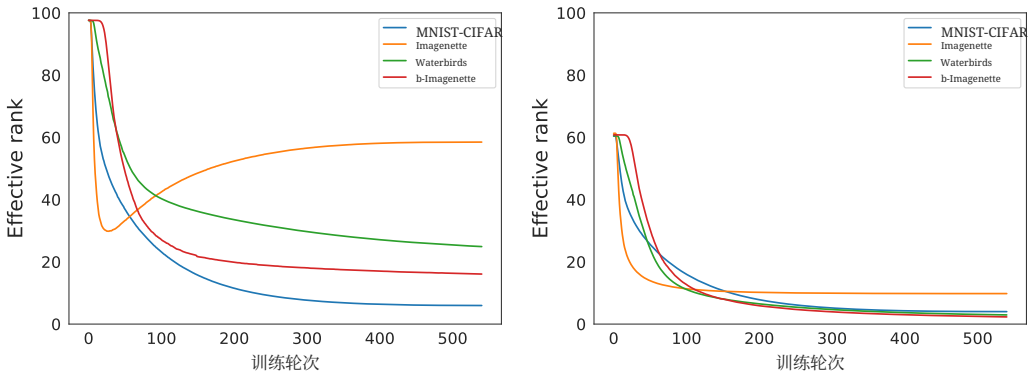


图*7.*在 MNIST-CIFAR 数据集的预训练表示中加入高斯噪声后，测试准确率随其标准差的变化。模型1保持固定，两个集成的数值均为 3 次运行的平均值。

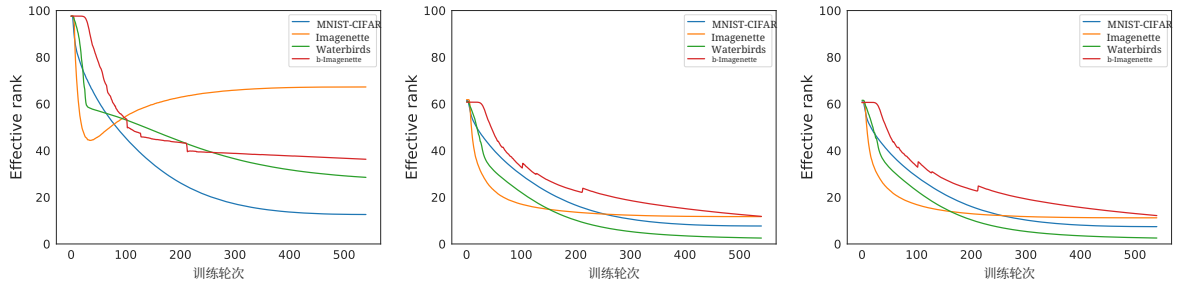
表*6.*决策边界非线性的定量测量——拟合该决策边界的线性分类器的准确率

数据集	线性分类器准确率(f)	线性分类器准确率(f_{proj})
b-Imagenette	96.12	95.28 ± 0.2
Waterbirds	97.28	93.24 ± 0.24



(a) 第1层　　(b) 第2层

图*8*。在数据集的 ResNet-50 预训练表示上，深度为2的ReLU网络权重矩阵有效秩的变化



(a) 第1层　　(b) 第2层　　(c) 第3层

图*9*。数据集的 Resnet-50 预训练表示上深度为3的ReLU网络权重矩阵有效秩的演变

C. 扩展相关工作

在本节中，我们将全面综述本文所依据的各个主题的相关文献。 .

线性网络中的低秩简单性偏差 多项研究表明，对于平方损失和交叉熵损失，线性网络上的梯度下降都具有低秩简单性偏差。对于平方损失，Gunasekar 等人 (2017年) 猜想，针对两层线性网络，该网络倾向于寻找最小核范数解。Arora 等人 (2019) 驳斥了这一猜想，并转而指出该网络倾向于寻找低秩解。Razin 和 Cohen(2020) 通过提供合成示例，为低秩猜想提供了实证支持：在这些示例中，网络使核范数趋于无穷大，却最小化了有效线性映射的秩。 Li 等人(2021年) 证明，对于足够小的初始化，线性网络上的梯度流遵循贪心低秩学习轨迹。对于线性可分数据上的二元分类，Ji & Telgarsky (2019) 表明，随着训练推进，线性网络的权重矩阵最终会变为秩一。

非线性网络中的低秩简单性偏置 关于非线性网络中低秩简单性偏置的研究相对较少。其中最具代表性的两项研究是 Huh 等人 (2021年) 和 Galanti 和 Poggio (2022)。Huh 等人 (2021年)通过实证研究发现，随着训练的进行，采用 ReLU 激活的神经网络所学习到的嵌入的秩会降低。Galanti 和 Poggio (2022)阐释了权重矩阵的秩与批量大小、权重衰减等各种超参数之间关系的直观原理。与这些研究不同，我们从理论和实证两方面证明，对于单层网络，网络依赖于输入的一个极低维投影，并且可以利用这种偏置来开发鲁棒分类器。

与分布外检测的关系 许多近期的分布外检测研究（Cook 等人，2020；Zaeemzadeh 等人，2021年）会显式构造低秩嵌入，从而更容易将其与分布外数据点区分开来。其他研究也会隐式地利用嵌入的低秩特性。Ndiour 等人 (2020)对学习到的特征使用主成分分析，并且只对由少数几个主导方向张成的小子空间内的似然建模。Wang 等人 (2022)利用嵌入的低秩特性，估计给定数据点在该低秩子空间上的正交投影，并将其与 Logit 信息结合，以检测分布外数据点。尽管已有研究隐式利用嵌入的低秩特性，但我们指出，本文（i）展示的是权重的低秩特性，而非嵌入的低秩特性；（ii）表明这是简洁性偏置的结果。

其他关于简洁性偏置的研究 许多研究从经验和理论层面探讨了神经网络中简洁性偏置的本质。 Kalimeris 等人 (2019年）通过实证表明，神经网络上的随机梯度下降会逐渐学会复杂度递增的函数。（Rahaman 等人，2018年）通过实证表明，神经网络倾向于先学习低频函数。（Ronen 等人，2019）从理论上证明，在神经正切核区域，收敛速率取决于核谱的特征值。（Hacohen 等人， 2020）表明，无论架构如何，神经网络学习训练样本和测试样本的顺序几乎总是相同的。（Pezeshki 等人， 2021年）提出，训练初期的梯度饥饿 可能是惰性/NTK 区域中简洁性偏置的原因，但相关条件很难解释。相比之下，我们的结果适用于 IFM 模型在丰富 训练区域中的任意数据集。（Lyu 等人， 2021年）研究了反对称数据集，并表明单隐藏层输入齐次网络（即不含偏置参数）会收敛到线性分类器。然而，与含偏置参数的网络相比，此类网络的表达能力严格较弱。（ Hacohen 与 Weinshall，2022 ）表明，对于深度线性网络，在神经正切核区域中，它们会先学习输入数据的高阶主成分。以往大多数研究都将简洁性偏置视为神经网络泛化能力更强的原因。然而，（Shah 等人， 2020）表明，过强的简洁性偏置也可能导致较差的分布外性能。

学习多样化分类器：已有多项工作尝试学习多样化分类器。大多数工作通过确保这些模型的输入梯度彼此不对齐来学习此类模型（Ross 与 Doshi-Velez，2018年；Teney 等人，2022）。（Xu 等人，2022）提出了一种学习多样化／正交分类器的方法，该方法假设存在一个使用所有特征的完整分类器，并展示了其在风格迁移等各种下游任务中的效用。（Lee 等人，2022）通过在无标签目标数据上施加多样性约束来学习多样化分类器。

伪相关：大量研究探讨了神经网络中伪相关产生的原因（Sagawa 等人，2020b），并针对不同情境提出了算法修正方案（Liu 等人，2021年；Chen 等人，2020）。

梯度的隐式偏差 **梯度下降**：此外，还有大量研究致力于理解梯度下降的隐式偏差 t

动力学。大多数此类研究针对标准线性（Ji& Telgarsky，2019）或深度线性网络（Soudry 等人，2018年；Gunasekar 等人，2018年）。对于非线性神经网络，一个著名的结果针对的是具有齐次激活函数的1层隐藏层神经网络的情况（Chizat 和 Bach，2020），我们在证明中关键地用到了这一结果。

D. 关于将结果推广到深度神经网络的进一步讨论

将我们的理论结果推广到深度神经网络，是一个令人振奋且充满挑战的研究方向。对于浅层和深度网络，即使在训练的平均场机制下，关于收敛到全局极小值的结果也已得到确立（Chizat 和 Bach，2018年；Fang 等人，2021年）。然而，据我们所知，只有对于单隐藏层全连接网络（Chizat 和 Bach，2020），人们才精确刻画了梯度流所收敛到的全局极小值。理解梯度流的这种隐式偏差，对于深度神经网络而言仍是一个未解决的问题；我们认为，这对将我们的结果推广到深度神经网络至关重要。

E. 收敛到 $\mathcal{F}_1$–ReLU 网络的最大间隔分类器

本节将陈述Chizat 和 Bach （2020）关于 ReLU 网络上梯度流渐近收敛点的精确结果。为便于读者阅读，我们将沿用 Chizat 和 Bach （2020）的记号。

神经网络由神经元上的概率测度 μ 参数化，并定义为 y

$$h(\mu, x) = \int \phi(w, x) d\mu(w)$$

其中 $\phi(w,\ x)\ = b(a^\top(x,\ 1))_+$ ($+$ 表示正部分，即 ReLU 激活)，且 $w = (a,\ b)\ \in \mathbb{R}^{d+2}$。由于该网络是二次齐次的，因此可以将测度 μ 投影到单位球面上。测度 μ 在球面上的投影算子 (Π_2) 定义如下：对于球面上的任意连续函数 φ,

$$\int_{\mathbb{S}^{d+1}} \varphi(\theta) d[\Pi_2(\mu)](\theta) = \int_{\mathbb{R}^{d+2}} \|w\|^2 \varphi(w/\|w\|) d\mu(w)$$

现在，令 ρ 表示输入空间 $\mathcal{X}$ 上的输入分布，并令标记函数 $y:\ \mathcal{X} \to \mathcal{Y}$ 为确定性的。接下来，考虑如下总体目标函数:

$$F(\mu) = -\log\left[\int_{\mathcal{X}} \exp(-y(x)h(\mu, x)) d\rho(x)\right]$$

注意，log 不会影响梯度的方向，因此，该损失下梯度流的轨迹与指数损失下的轨迹相同。此外，总体平滑间隔定义如下:

$$S(f) = -\log\left(\int_{\mathcal{X}} \exp(-f(x)) d\rho(x)\right)$$

ˆ 对于这一特定情形，$f(x)\ = y(x)h(\mu,\ x)$。将 $y(x)\cdot h(\mu,\ x)$ 记为 $h(\mu)$。

定理 *E.1.* 假设 ρ 具有有界密度和有界支撑集，且标记函数 y 连续，则存在一个 *Wasserstein* 梯度流 (μ_t)，定义在 F 上，并满足 $\mu_0 = \mathcal{U}(\mathbb{S}^d) \otimes \mathcal{U}\{-1, 1\}$，即输入（分别为输出）权重在球面上（分别在 $\{-1, 1\}$上）均匀分布。若 $\nabla S(\hat{}h(\mu_t))$ 在 $\mathcal{P}(\mathcal{X})$ 中弱收敛，若 $\bar{\nu}_t = \Pi_2(\mu_t)/([\Pi_2(\mu_t)](\mathbb{S}^{d+1}))$ 在 $\mathcal{P}(\mathbb{S}^{d+1})$ 中弱收敛，且 $F'(\mu_t)$ 在 C^1_{loc} 中收敛到满足 *Morse-Sard* 性质的 F'，则 $h(\bar{\nu}_\infty, .)$ 是 $\max_{\|f\|_{\mathcal{F}_1} \le 1} \min_{x \in \mathcal{X}} y(x)f(x)$ 的最大化子。

其中，$\mathcal{P}(\mathcal{X})$ 表示 $\mathcal{X}$ 上概率分布的空间，而 $[\Pi_2(\mu_t)](\mathbb{S}^{d+1})$ 表示测度 $\Pi_2(\mu_t)$ 在 $\mathbb{S}^{d+1}$ 上的总质量。

为理解该定理，请注意:

$$\nabla S(f) = \frac{\exp(-f(x)) d\rho(x)}{\int_{\mathcal{X}} \exp(-f(x')) d\rho(x')}$$

因此，$\nabla S(f)$ 收敛意味着指数化后的归一化间隔收敛。此外，$\bar{\nu}_t$ 与权重的方向收敛类似，不过在此情况下，权重由 $\mathbb{S}^{d+1}$中的方向取代。关于 Morse-Sard 性质和度量 C^1_{loc}的解释，请参阅 Chizat 和 Bach (2020) 的附录 H。