arXiv:2302.00704v3 [cs.LG] 9 Jan 2024

深度集成中预测多样性的病态

Taiga Abe *ta2507@columbia.edu*
理论神经科学中心 神经科学系 哥伦比亚大学

E. Kelly Buchanan *ekb2154@columbia.edu*
哥伦比亚大学神经科学系理论神经科学中心

Geoff Pleiss *geoff.pleiss@stat.ubc.ca*
英属哥伦比亚大学统计学系
*Vector*研究所

John Cunningham *jpc2181@columbia.edu*
哥伦比亚大学统计学系

摘要

经典结果表明，在低容量模型集成中，鼓励预测多样性能够提升性能，例如通过装袋法或提升法实现。本文证明，这些直觉并不适用于高容量神经网络集成（深度集成），事实上，情况往往恰恰相反。在一项针对近 600 个神经网络分类集成的大规模研究中，我们考察了多种以组件模型性能换取预测多样性的干预措施。尽管此类干预可以提升小型神经网络集成的性能（符合普遍的直觉），但它们会损害 实践中最常用的大型神经网络集成的性能。令人惊讶的是，我们还发现，在大型网络集成中，抑制预测多样性通常并无不利影响，这完全颠覆了普遍的直觉。即使促进多样性的干预不会牺牲组件模型性能（例如采用异构架构和训练范式），我们仍观察到，追求更高预测多样性会带来机会成本。通过考察 1000 多个集成，我们发现，与采用多样化架构或训练流程所带来的收益相比，仅仅使用更高容量模型带来的收益远胜于前者，尽管这类更高容量模型往往会显著降低预测多样性。总体而言，我们的研究结果表明，最初针对低容量集成提出的有关预测多样性的普遍直觉，并不直接适用于现代高容量深度集成。这项工作阐明了通过提高多样性来改进深度集成所面临的根本挑战，同时提出了一条替代路径：只需使用更强大（且多样性更低）的组件模型来构建集成。

1 引言

成功的集成需要组件模型犯下不同的错误（Dietterich，2000年；Breiman，2001年；Zhou，2012年；Abe 等人，2022a）。这一观点可以通过将集成性能分解为两项来形式化：i）组件模型的平均性能；ii）模型之间的预测多样性（例如参见Breiman，2001；Webb 等人，2020年；Gupta 等人，2022年，以及第3节）。历来，集成方法论的进展

主要集中在后一项上。以决策树集成为例。装袋树（Breiman，1996年）利用自助采样训练数据生成不同的组件树预测；随机森林（Breiman，2001年）还会对组件模型使用的特征进行子采样，从而增加多样性。这些训练流程是在权衡取舍中进行的：它们会降低组件模型的性能，但组件模型之间增加的多样性会带来更好的总体性能（Breiman，1996；2001）。随机森林通常优于装袋树，而装袋树又优于单棵决策树。

相比之下，深度集成（Lakshminarayanan 等人，2017年）并不依赖于主动鼓励组件模型之间产生差异的训练流程。相反，人们通常会集成架构和训练流程完全相同的神经网络，其多样性唯一的来源是参数初始化和随机优化固有的随机性。乍看之下，这种同质性似乎效率极低——如果每个组件模型采用多样化架构、训练流程或数据集/特征子样本，我们肯定能改进深度集成吧？令人惊讶的是，尽管有许多提案试图提高深度集成的多样性——例如，通过改变网络架构、数据集、超参数或训练目标（例如，Wenzel 等人，2020年；Wu 等人，2021年；Gontijo-Lopes 等人，2022年；Ortega 等人，2022年）——但只有少数方法能在标准深度集成基准（Ashukha 等人，2020年；Abdar 等人，2021年；Tran 等人，2022年）上取得超过有限幅度的提升；该基准仍是大规模分类任务的黄金标准集成技术（例如，Szegedy 等人，2015年；Lakshminarayanan 等人，2017年；Tran 等人，2022年）（详细综述见第 2 节和附录 C）。更常见的情况是，以牺牲组件模型性能为代价来换取多样性的标准技术（例如装袋法）会损害集成预测（Lee 等人，2015年；Nixon 等人，2020年）。

决策树与深度神经网络集成之间的这种差异，应促使我们重新审视这样一种观点：促进多样性始终是提升集成性能的有效方法。一方面，标准深度集成相对成功，或许表明我们尚未找到能够进一步促进现代神经网络多样性的合适机制。另一方面，关于集成的直觉可能根本无法从小型决策树沿用到现代神经网络；后者通常运行于训练误差近乎为零的过参数化状态。换言之，对于决策树集成和深度集成而言，组件模型性能与预测多样性之间的权衡可能存在根本差异。

在这项工作中，我们提供了有力证据，支持后一种假设。我们发现，对于现代分类深度集成而言，优先考虑预测多样性而非组件模型性能，会适得其反，甚至可能造成损害。我们的主张基于两项主要结果。第一项结果关注那些以牺牲组件模型性能为代价、明确促进预测多样性的机制。我们通过联合训练机制，在受控环境下调整深度集成的预测多样性（Mishtal 和 Arel，2012年；Zhao 等人，2022年；Ortega 等人，2022年；Abe 等人，2022a）。与许多成熟的集成技术一样，这些机制并非“免费”地促进预测多样性，而是在权衡关系下运作：多样性的提升以牺牲各个组件模型的性能为代价。我们的结果基于约 600 个深度集成，揭示了成熟的集成直觉与现代深度集成行为之间的二分现象。具体而言，在 ResNet 18 这类大型神经网络的集成中促进预测多样性（He 等人，2016）始终会导致更差的集成性能，因为集成成员之间预测多样性的增加，无法抵消组件模型性能所遭受的巨大损失。与此相反，按照标准直觉，小型神经网络集成等低容量集成会因促进多样性而有所改善。令人惊讶的是，我们还发现，抑制多样性不会影响高容量深度集成，有时甚至会带来益处。我们将这些违反直觉的现象归因于：（1）促进多样性对准确且置信度高的预测所产生的、常被忽视的影响；（2）关于过参数化具有降低方差作用的近期证据（Adlam 和 Pennington，2020年；d’Ascoli 等人，2020年）。

我们的第二项结果关注一些技术，例如采用多样化架构或训练流程，这些技术能够在维持组件模型性能的同时提高预测多样性。虽然这并非明确有害，但我们证明，利用这类“免费”的预测多样性来源是有机会成本的。我们从由 304 个独立训练的神经网络组成的模型池中，构建了逾 1000 个同质和异质 深度集成（分别组合架构相同和不同的模型），并测量集成性能如何随预测多样性及组件模型性能变化。在

与先前研究一致（例如 Ashukha 等人，2020年；Abdar 等人，2021年；Gontijo-Lopes 等人，2022年），我们确实发现，异质集成的“免费”预测多样性带来的性能略优于同质集成。同时，我们的结果表明，与仅仅使用更强大的组件模型所带来的好处相比，这些微小提升相形见绌，尽管更强大的模型多样性较低。换言之，最佳深度集成会牺牲预测多样性，以换取组件模型性能，这与通常的直觉恰恰相反。至关重要的是，这种机会成本和权衡取决于组件模型的容量：例如，我们发现，尽管最佳的现代深度集成预测多样性极低，最佳决策树集成却拥有最高的预测多样性。通过对容量范围广泛的神经网络模型进行集成，我们可以观察到这两种状态之间的明显转变。

总而言之，我们的结果强烈表明，现代神经网络的集成并不处于长期以来关于提高预测多样性净收益的传统认知所适用的范围内。此外，结果还凸显了组件模型容量是影响集成动态的一个关键却被忽视的因素。若要获得能够打破我们所观察到的权衡与机会成本的、有意义的预测多样性，可能需要彻底创新的方法论。在此期间，提升深度集成效果的最佳方式很简单：使用更好的、但多样性较低的单个模型。

贡献概要。我们首先在第2节阐明本研究与先前工作的关系，并在第3节介绍实验设置。接着，我们将证明组件模型的规模决定了预测多样性在集成性能中所起的作用。我们将在两个不同的、聚焦于大规模分类的情境中展示这一现象：

- 我们表明，鼓励或抑制集成成员之间预测多样性的优化目标，能够提高小型模型集成的预测准确率；但将这些相同的目标用于大型组件模型集成时则无效。我们在涵盖多种架构、数据集和训练目标的 600 多个深度集成上验证了这一结果，并表明我们的发现调和了传统集成文献与各种旨在改进深度集成的方法之间的分歧（第4节）。

- 此外，我们还表明，研究结果可推广至多种形式的隐式多样性促进，例如通过集成不同的神经网络架构和超参数。我们发现，在这种情况下，集成更多样化的模型或许能带来微小收益，但只需集成能力更强、而多样性较低的组件模型，就能更有效地提升集成性能（第5节）。

2 相关工作

传统集成。从历史上看，集成一直应用于决策树等低容量的基模型（例如，Freund，1995年；Breiman，1996年；Dietterich，1998年；Breiman，2001年；Cutler 和 Zhao，2001年），这与现代深度网络所处的情境截然不同。Breiman （2001年）指出，基学习器的性能以及集成成员之间的相关性，是决定集成性能的关键因素。较新的研究（Gupta 等人，2022年；Wood 等人，2023年）探讨了可直接从评估集成成员预测的方法中推导出的多样性度量。

深度集成。神经网络集成早已受到关注（例如 Hansen and Salamon，1990年；Perrone and Cooper，1992年），但深度集成通常指基于现代深度网络的随机初始化构建的集成。深度集成最常用于分类，有时甚至会将回归重新表述为分类问题（Stewart 等人，2022年）。深度集成的标准做法是训练多个采用相同模型架构的副本（例如 Lee 等人，2015年；Lakshminarayanan 等人，2017年；Fort 等人，2019年；Buchanan 等人，2022年；Tran 等人，2022年）。如果条件允许，对不同架构或训练流程进行集成，也只能适度提升性能（例如 Fort 等人，2019年；Gontijo-Lopes 等人，2022年）。然而，在训练过程中主动鼓励预测多样性（即输出空间中的多样性）（例如 Mishtal 和 Arel，2012年；Lee 等人，2015年；Pang 等人，2019年；Ross 等人，2020年；Thakur 等人，2020年；Webb 等人，2020年；Wen 等人，2020年；Buschjäger 等人，2020年；Jain 等人，2020年；Gong 等人，2022年；Ortega 等人，2022年；Pagliardini 等人，2022年；Teney 等人，2022年；Zhao 等人，2022年），并未被证明能持续带来益处（见表2），

除了不常用于实践的小型网络（Mishtal 和 Arel，2012年；Opitz 等人，2017年；Pang 等人，2019年；Webb 等人，2020年；Ortega 等人，2022年；Pagliardini 等人，2022年）或存在其他训练数据时（例如 Gontijo-Lopes 等人，2022年）。除了指出现有技术未能改善集成性能这一事实之外（例如Lee 等人，2015年；Nixon 等人，2020年），鼓励现代深度集成中的预测多样性可能造成的负面后果也常被忽视。

关于预测多样性的其他研究。其他研究探讨了如何促进深度集成成员进行基于 logit 的平均时的多样性（Buschjäger 等人，2020年；Webb 等人，2020年），但本文提出的机制未必适用于这种情形。尽管近期研究表明，与概率平均相比，logit 平均在理论上依据更充分（Gupta 等人，2022年；Wood 等人，2023年），但概率平均（也是本研究的重点）仍是最常用的深度集成形式（Lakshminarayanan 等人，2017年）。如上所述，我们用于控制深度集成中预测多样性的两种方法，旨在促进集成多样性（Mishtal 和 Arel，2012年；Ortega 等人，2022年）；其他方法则最初用于对大模型训练进行正则化（Zhao 等人，2022年），或分别用于在回归和某些大规模模型中促进或抑制预测多样性（Abe 等人，2022a；Jeffares 等人，2023年）。我们通过在大规模分类任务中系统研究多样性促进／抑制如何随组件模型规模而异，拓展了这些方法的原有用途。值得注意的是，同时开展的研究（Jeffares 等人，2023年）表明，即使在回归任务中或使用低容量组件模型时，我们在此研究的一种机制也可能找到优化的退化解。他们的结论与我们的结论相辅相成，其实验也支持了本文的论点。最后，我们不讨论那些在集成成员的权重空间、特征空间或梯度中促进集成多样性的方法（例如，Dabouei 等人，2020年；D'Angelo 和 Fortuin，2021年；Sinha 等人，2021年），而是聚焦于明确针对集成成员预测空间中的多样性的方法。研究这些其他集成多样性来源虽是值得探索的未来方向，但我们的首要目标是强调组件模型的容量这一关键却研究不足的特征，因为它决定了促进集成预测多样性所产生的影响。

3 权衡　预测多样性与组件模型性能之间的权衡

问题设定。我们主要考虑输入为 $\boldsymbol{x} \in \mathbb{R}^D$、目标为$y \in [1, \dots, C]$的多类别分类问题，这是深度神经网络以及深度集成的典型任务（Hui 和 Belkin，2020年；Stewart 等人，2022年）。标准的深度集成由 M 个模型 $\boldsymbol{f}_1(\cdot) \dots, \boldsymbol{f}_M(\cdot)$ 组成，其中每个 $\boldsymbol{f}_i^-$ 都将输入 $\boldsymbol{x}$ 映射到 C 维概率单纯形。集成预测 $\boldsymbol{f}(\boldsymbol{x})$ 为：

$$\bar{\boldsymbol{f}}(\boldsymbol{x}) \triangleq \frac{1}{M}\sum_{i=1}^{M} \boldsymbol{f}_i(\boldsymbol{x}). \tag{1}$$

通常，组件模型采用相同的神经网络架构。组件模型通常在同一数据集上使用相同的损失函数 $\ell(\boldsymbol{f}_i(\boldsymbol{x}), y)$ 独立训练。此过程等价于最小化经验单模型风险的平均值：

$$\mathcal{L}_{\text{avg}} = \hat{\mathcal{R}}_{\text{avg}} \triangleq \frac{1}{M}\sum_{i=1}^{M} \mathbb{E}_{p_{\text{train}}(\boldsymbol{x},y)} \left[\ell\left(\boldsymbol{f}_i(\boldsymbol{x}), y\right)\right] = \mathbb{E}_{p_{\text{train}}(\boldsymbol{x},y)} \left[\frac{1}{M}\sum_{i=1}^{M} \ell\left(\boldsymbol{f}_i(\boldsymbol{x}), y\right)\right]. \tag{2}$$

集成风险的分解。在测试时，我们使用式（1）中的聚合集成预测。根据式（1）和（2），我们可以将集成风险分解如下：

$$\mathcal{R}_{\text{ens}} \triangleq \mathbb{E}_{p(\boldsymbol{x},y)} \underbrace{\left[\ell\left(\bar{\boldsymbol{f}}(\boldsymbol{x}), y\right)\right]}_{\text{ens. loss}} = \mathcal{R}_{\text{avg}} - \mathbb{E}_{p(\boldsymbol{x},y)} \underbrace{\left[\frac{1}{M}\sum_{i=1}^{M} \ell\left(\boldsymbol{f}_i(\boldsymbol{x}), y\right) - \ell\left(\bar{\boldsymbol{f}}(\boldsymbol{x}), y\right)\right]}_{\text{Jensen gap (pred. diversity)}} \tag{3}$$

（其中 $\mathcal{R}_{\text{avg}}$ 是非经验的平均单模型风险）。我们假设 ℓ 是严格凸损失函数（例如交叉熵、均方误差等），因此最后一项——詹森间隙——非负。本文始终使用詹森间隙作为预测多样性的度量（如 Abe 等人(2022a 所示）：当且仅当所有组件模型作出相同预测时，它才等于 0；随着 $\boldsymbol{f}_i(\boldsymbol{x})$ 之间的差异增大，它也会随之增大。因此，式（3）对集成性能（$\mathcal{R}_{\text{ens}}$）进行了直观分解，分为：（i）

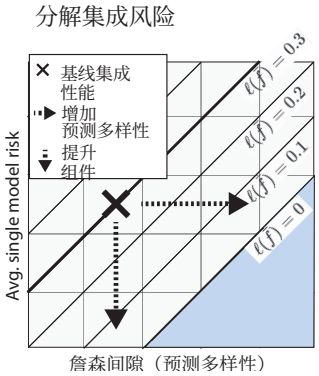


图1：**多样性与组件模型性能权衡**。集成性能以负对数似然（NLL）衡量，可分解为单模型平均负对数似然（纵轴）和詹森间隙预测多样性（横轴）——见式（3）。对角线表示集成 NLL 的等值集（越靠右下越好）。任何集成的性能都可以在图中表示为一个点（×），并对应一个集成性能等值集（粗对角线）。沿着同一等值集，所有集成的 NLL 都相同。要提升给定集成的性能，有两种策略：提高组件模型的预测多样性（向右箭头），或改善组件模型的平均 NLL（向下箭头）。如果由此得到的集成位于粗对角线下方，其性能就会提升。

其组件模型的平均性能（$\mathcal{R}_{\text{avg}}$）以及（ii）模型之间的预测多样性（詹森间隙），正如我们将在后文分析中所使用的。图 1 展示了这种分解：水平轴和垂直轴分别对应预测多样性和组件模型平均风险，而对角线则是集成风险的等值集。在附录 D中，我们验证了詹森间隙与先前研究中使用的其他预测多样性度量指标高度相关。

在多样性与组件模型性能之间进行权衡。 图 1 体现了我们对于多样性与组件模型性能之间权衡的许多直觉。例如，决策树集成表明，可以牺牲组件模型性能，以有利地换取集成多样性。以这种方式得到改进的集成，在图 1 中可表示为位于×右侧、粗对角线下方的区域。如果促进多样性导致更差的 集成性能，那么这必然是由于组件模型性能下降所致（图 1中， ×右侧、粗对角线上方的区域）。在这两种情况下，我们都牺牲了单模型性能来增加集成多样性，但存在不同的情形，在这些情形下，这种权衡可能有助于或损害集成性能。请注意，对于能够提升组件模型性能的权衡，同样存在类似的结论：我们或许可以在不损失过多多样性的情况下改进组件模型（图1中，位于 ×左侧、粗对角线下方的区域），将其作为提升集成性能的另一种方法。

4 操控神经网络集成中的预测多样性

对于由较小模型组成的集成，装袋法和随机森林等方法虽然以牺牲组件模型性能为代价来换取预测多样性，却能提升集成性能。来自这些经过充分研究的范式的直觉，推动了许多研究尝试在现代深度神经网络集成中实现同样有利的权衡。然而，本节将展示，这种权衡在现代深度集成中会产生相反的结果。具体而言，积极促进多样性（以牺牲组件模型性能为代价）会损害深度集成性能，而抑制多样性则无害，甚至有益。我们发现，多样性促进或抑制的有害或无害程度，取决于组件模型的容量。如果大幅缩小组件模型的规模（例如，采用 ResNet 8 这类极小型网络），多样性促进便会再次变得有益，这与普遍的直觉一致。尽管如此，我们强调，实践中使用的大多数现代神经网络，其容量都处于多样性促进会产生危害的范围内。我们将这一发现归因于一个简单却常被忽视的现象：多样性促进对高度自信且准确的预测影响尤为显著，而这类预测在高容量神经网络的预测中占多数。

鼓励/抑制预测多样性的机制。 为了控制这些集成的预测多样性，我们尝试对标准集成训练进行四种修改，以便按不同程度鼓励/抑制预测多样性。我们考虑的前三种鼓励/抑制预测多样性的机制，会对标准集成训练进行显式正则化（公式 2）：

$$\mathcal{L}_{\text{ens.}} = \mathcal{L}_{\text{avg}} - \mathop{\mathbb{E}}_{p_{\text{train}}(\boldsymbol{x},y)} \Big[\gamma \underbrace{\mathbb{D}\left(f_1(\boldsymbol{x}), \ldots, f_M(\boldsymbol{x})\right)}_{\text{diversity regularizer}} \Big] \tag{4}$$

其中， γ 是一个超参数（负值鼓励多样性；正值则抑制多样性）， $\mathbb{D}$是某种预测多样性度量。

1. **方差**正则化（Ortega 等人，2022年）将 $\mathbb{D}$ 设为：

$$\mathbb{D}\left(f_1(\boldsymbol{x}),\ldots,f_M(\boldsymbol{x})\right)=\frac{1}{2(M-1)\max_i f_i^y(\boldsymbol{x})}\sum_{i=1}^{M}\left(f_i^y(\boldsymbol{x})-\bar{f}_i^y(\boldsymbol{x})\right)^2, \tag{5}$$

这本质上对应于集成成员真实类别预测的方差。

2. **1 对全部 Jensen-Shannon 散度**正则化（Mishtal 和 Arel，2012年）将 $\mathbb{D}$ 设为：

$$\mathbb{D}\left(f_1(\boldsymbol{x}),\ldots,f_M(\boldsymbol{x})\right)=\frac{1}{M}\sum_{i=1}^{M}JS\left(\boldsymbol{f}_i(\boldsymbol{x})\,\middle\|\,\frac{1}{M-1}\sum_{j\neq i}\boldsymbol{f}_j(\boldsymbol{x})\right), \tag{6}$$

其中 JS 表示 Jensen-Shannon 散度。

3. **统　一 Jensen-Shannon 散度**正则化（Zhao 等人，2022年）将 $\mathbb{D}$ 设为：

$$\mathbb{D}\left(f_1(\boldsymbol{x}),\ldots,f_M(\boldsymbol{x})\right)=H\left(\bar{\boldsymbol{f}}(\boldsymbol{x})\right)-\frac{1}{M}\sum_{i=1}^{M}H\left(\boldsymbol{f}_i(\boldsymbol{x})\right), \tag{7}$$

其中 H 是香农　熵。请注意，括号内的项等价于多分布
Jensen-Shannon　散度（采用均匀权重）。

4. **詹森间隙**-正则化（由Abe 等人(2022年a)提出）与公式 (3) 密切对应，我们考虑以下目标函数来鼓励/抑制多样性:

$$\mathcal{L}_{\text{ens}}=\mathop{\mathbb{E}}_{p_{\text{train}}(\boldsymbol{x},y)}\Big[\ell\left(\bar{\boldsymbol{f}}(\boldsymbol{x}),y\right)+(\gamma+1)\left(\frac{1}{M}\sum_{i=1}^{M}\ell\left(\boldsymbol{f}_i(\boldsymbol{x}),y\right)-\ell\left(\bar{\boldsymbol{f}}(\boldsymbol{x}),y\right)\right)\Big]. \tag{8}$$

第一项是集成预测损失（而非单模型平均损失），第二项则是预测多样性的詹森间隙概念。

对于所有方法， $\gamma=0$ 对应于标准集成训练， $\gamma<0$ 鼓励多样性，而 $\gamma>0$抑制多样性。因此，我们可以将鼓励/抑制集成多样性视为标准集成训练的正则化版本（公式 2）。与任何正则化机制一样，我们预计操纵预测多样性也会影响组件模型的性能，并将研究这些因素之间的权衡如何影响整体集成性能。第 5 节研究不属于这一范式的多样性来源的影响（例如，使用不同架构、训练范式等来实现多样性）。在附录 F.1 中，我们将进一步讨论这些独立多样性机制之间的差异，并分别评估每种正则化器的性能。

数据集和架构。我们在 CIFAR10、CIFAR100（Krizhevsky 等人，2009年）、TinyImageNet（Le 和 Yang，2015年）和 ImageNet（Deng 等人，2009年）数据集上训练深度集成。除 ImageNet 外，所有数据集上的集成都由$M=4$ 个相同神经网络架构的模型组成；对于 ImageNet， $M=3$。在 CIFAR 数据集上，我们研究 ResNet 18（He 等人，2016）、VGG 11（Simonyan 和 Zisserman，2014）、WideResNet 28-10（Zagoruyko 和 Komodakis，2016）、DenseNet 121（CIFAR100 上使用 DenseNet 161）（Huang 等人，2017）、ResNet 8 和 LeNet（LeCun 等人，1998年）模型。前四种被视为大型架构，后两种则是小型架构。我们选取的大型网络代表现代神经网络，它们能够做出高度准确但过度自信的预测，而小型网络则不会如此（Guo 等人，2017年）。此外，我们还研究用于 TinyI mageNet 的 ResNet 34 架构，以及用于 ImageNet 的 ResNet 18 架构。总计，我们在四个数据集和六种架构上训练了 574 个集成，并使用上述四种机制来鼓励/抑制集成多样性。（训练和实现细节见附录 B。）

4.1 结果

在本节中，我们展示一种具有代表性的小型架构和大型架构（ResNet 8 和 ResNet 18）的结果，大多数结果基于 CIFAR10 数据集。其他架构和数据集的结果请参见附录。F 。

促进多样性有利于小型网络集成，但会损害大型网络集成的性能。图2 （上方， $\gamma<0$）展示了促进多样性对 CIFAR10 ResNet 8 准确率的影响

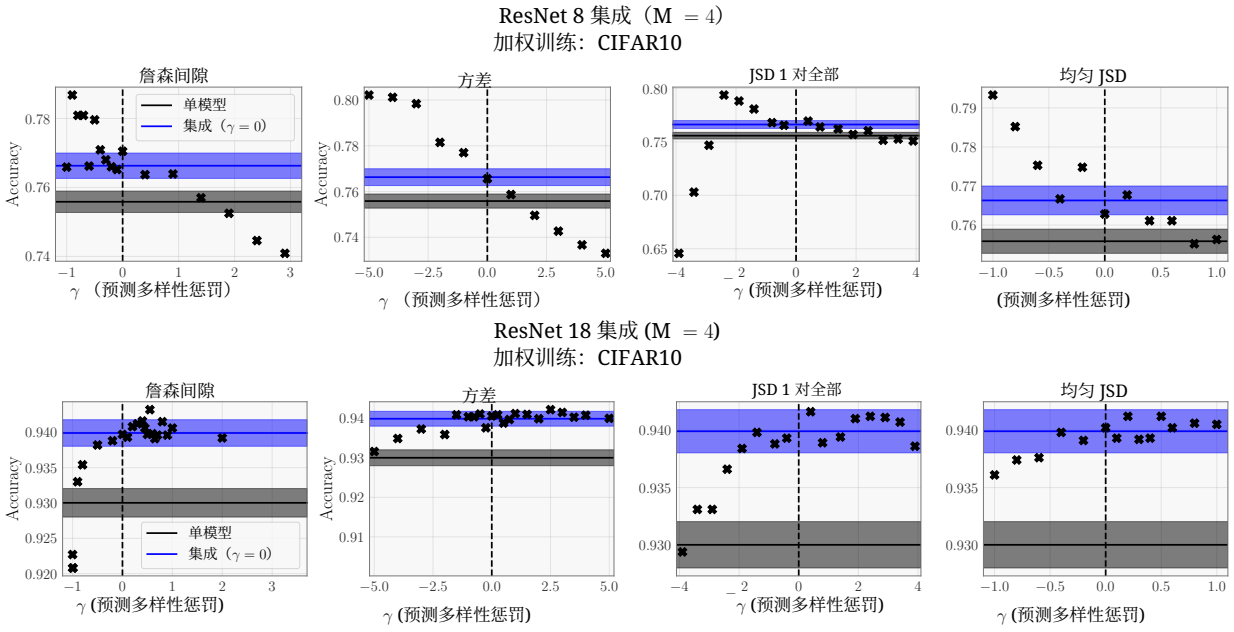


图 2：**促进/抑制多样性对小型与大型神经网络集成的影响各不相同。** 我们在 CIFAR10 上使用多样性机制训练小型 **（上排）**和大型 **（下排）**神经网络的深度集成（分别为 ResNet 8 和 ResNet 18），每列对应一种多样性机制。我们将标准集成训练（虚线）与促进/抑制多样性的训练（分别位于虚线左侧和右侧）进行比较。蓝色线条/色带表示标准集成测试准确率；黑色线条/色带表示组件模型测试准确率。促进多样性（$\gamma < 0$）可提高小型网络集成的测试准确率，但会损害大型网络集成的测试准确率。相反，抑制多样性（$\gamma > 0$）会明显损害小型网络集成的性能，但对大型网络集成似乎没有不良影响。（CIFAR100、TinyImageNet 和其他架构的结果见附录 F 。）

Regularization strength	$\gamma = -0.9$	$\gamma = -0.5$	$\gamma = 0$	$\gamma = 0.1$
Test accuracy (%)	70.16	71.10	71.49	71.37

表 1：**促进多样性在 ImageNet 上无效。** 在 ImageNet 上使用 M = 3 和詹森间隙正则化器对 ResNet 18 模型进行集成，得到的多样性促进结果。与较小的数据集一样，促进预测多样性并无益处，而抑制多样性似乎没有不良影响。

模型。（其他小型架构 LeNet 的结果，以及 CIFAR100 上的结果，见附录F 。）每个面板展示一种不同的多样性促进机制。蓝线及其色带表示标准（未正则化）集成的性能（均值 ± 加减两个标准误），黑线及其色带表示单个模型的错误率。相较于标准集成基线（蓝色色带），每种促进多样性的机制都能在一定范围的 $\gamma < 0$ （基线左侧）内提升性能。在所有多样性机制、数据集和模型架构中，促进多样性都提升了73% 的 86 个小型集成的性能（其他小型架构 LeNet 的结果，以及 CIFAR100 上的结果，见附录F。）

与此形成鲜明对比的是，图2 （底部，$\gamma < 0$）表明，多样性促进对较大型 ResNet 18 模型集成的影响恰恰相反。对于所有机制、数据集和模型架构，多样性促进（$\gamma < 0$）的性能不如标准集成：在 134 个大型网络中，61% 的网络表现更差（有时甚至低于单个网络），另有 29% 未出现统计显著变化。这些结果在大型数据集上也同样成立：在完整的 ImageNet 数据集上训练 ResNet 18 集成（表1），我们发现，多样性促进呈现出与 CIFAR10 上相同的趋势。（其他大型架构——VGG、WideResNet 和 DenseNet——以及 CIFAR100 和 TinyImageNet 上的结果见附录F。）

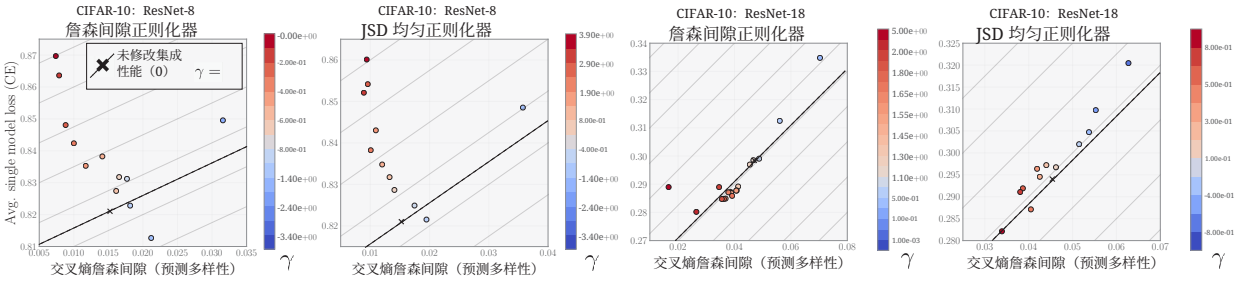


图 3：**在多样性正则化集成中权衡预测多样性与组件模型性能。** 每个标记代表一个在 CIFAR10 上采用多样性干预训练的 ResNet 8（**左侧面板**）或 ResNet 18（**右侧面板**）集成（CIFAR100 见图 12 ）。较暖的颜色对应于正 γ 值（促进多样性），较冷的颜色对应于负 γ 值（抑制多样性）。坐标轴按照图 1中的集成损失分解给出。标准深度集成性能的水平集（$\gamma = 0$）由 $\times$ 和点状对角线表示。在所有集成中，促进或抑制预测多样性会分别导致测试数据上的多样性按比例增加或减少。采用多样性促进的小型网络集成（$\gamma < 0$，蓝色标记）能够获得更高的单模型性能，从而形成更优的集成（位于点状线下方）。然而，对于大型网络集成而言，多样性促进会使平均单模型性能付出高昂代价，整体集成性能更差（位于点状线之上）。对于小型网络集成，抑制预测多样性（$\gamma > 0$；红色标记）同样会导致组件模型性能下降，因此相应的集成性能也比标准集成更差（位于点状线之上）。相比之下，抑制多样性的大型网络集成能够超越标准训练（位于点状线下方）。在大型网络集成中，测试负对数似然最佳的集成是一个多样性被抑制的集成。在图 13中，我们还研究了这些集成的多样性与平均准确率之间的关系。

接下来，我们验证采用多样性促进进行训练是否会提高测试数据上的多样性。图3展示了由 ResNet 8 模型（左侧面板）和 ResNet 18 模型（右侧面板）构成的集成的性能，采用的分解方式如图1所示。我们展示了 CIFAR10 数据集上詹森间隙和均匀 JSD 正则化器的结果。与标准集成训练（以×表示）相比，采用更多多样性促进措施训练的集成（$\gamma < 0$，蓝色圆点）表现出更高比例的预测多样性（位置更靠右）。将组件模型的平均性能纳入考量，可以对我们在图2中观察到的同一权衡关系进行另一种量化。小型模型的集成可能受益于多样性促进（图3中，蓝点位于虚线下方），但对于大型网络集成，促进预测多样性会导致集成损失函数值更高（蓝点位于虚线上方）。附录F.2 中展示了 CIFAR100 上 ResNet 18 多样性正则化集成的其他损失分解。

抑制预测多样性对大型网络集成无害，但会损害小型网络集成。对于小型网络 ResNet 8 集成（图 2，上方，$\gamma > 0$），我们观察到，所有抑制多样性的机制都会增大测试集错误率，这与我们根据传统集成得出的直觉一致。在 82 项小型网络实验中（见此处的图 2和附录 F），抑制多样性在56% 的情况下损害了性能，而在40%的情况下性能保持不变，仅在4%的情况下有帮助。相比之下，图 2 （下方，$\gamma > 0$）显示，对多样性施加惩罚并不会损害大型网络 ResNet 18 集成的性能。在 149 项实验中（见图 2 和附录 F），74% 的大型网络集成在抑制多样性后性能相近或更好（更多结果请参见附录 F 。）

我们再次验证了，抑制预测多样性（$\gamma > 0$，红点）会使集成在测试数据上的多样性降低，如图3所示。与预测准确率方面的结果一致，抑制预测多样性会导致小型模型集成的集成性能严格变差（红点位于虚线之上）。然而，对于大型网络集成，多样性抑制有时会提升集成性能（红点位于虚线之下）。附录F.2 展示了 CIFAR100 上 ResNet 18 多样性正则化集成的更多损失分解。

概率性能指标和校准也呈现出类似趋势。在附录F.5中，我们针对在测试集上测得的负对数似然（NLL）、布里尔分数（BS）和期望校准误差（ECE）指标（Naeini 等人，2015年）重现了图2。我们观察到类似的趋势：鼓励多样性有助于小型网络集成，却有害于大型网络集成；抑制多样性则反之。

分布外准确率也呈现出类似趋势。在附录D 中，我们将图2中的集成应用于 CIFAR10.1（Recht 等人，2018年），这是一个用于 CIFAR10 的分布外（OOD）测试数据集。结果表明，OOD 性能与分布内数据呈现出相同的趋势。

表格数据集上的多层感知机也呈现出类似趋势。最后，我们表明，本文的多样性正则化结果能够推广到其他数据集和网络架构：具体来说，是将多层感知机应用于表格数据集（实现细节见附录B.2.2，结果见附录D.1 ）。有趣的是，在这种设置下，我们还发现了具有中等容量的模型，它们似乎介于本文所考察的大型与小型模型集成的结果之间。

4.2 分析

既然已经验证了上述多样性机制能够成功促进或抑制测试集上的预测多样性，我们必须得出结论：小型与大型神经网络集成在组件模型性能上的权衡方式不同（分别有助于和损害集成性能）。总体而言，我们认为，以牺牲组件模型性能为代价来促进预测多样性，会损害大型网络深度集成，因为其大多数预测都既自信又正确。具体来说，在训练期间促进多样性，不太可能只影响测试时的错误预测；否则，多样性促进技术就相当于预言机错误检测器。不幸的是，对于任何正确且自信的输出，增加预测多样性都只会带来负面影响，因为只有所有模型都产生相同的正确且自信的输出时，集成风险才会最小。小型网络集成不会受到这一问题的影响，因为它们往往会产生更多错误或不自信的预测，而去相关能够改善这些预测。下文将为这些论点提供证据。

训练期间促进多样性会影响全部预测，即使是正确预测也不例外。图4 研究了多样性促进是否更可能影响原本会正确还是错误的预测。我们首先使用采用不同 $\gamma \leq 0$值训练的集成，测量测试集中的每个数据点所促进的多样性。我们绘制了 CIFAR10 测试集中多样性值的分布：左侧为使用詹森间隙多样性促进训练的 ResNet 8 集成，中间为 ResNet 18 集成（顶行），不同线条表示不同 γ值对应的密度。相较于标准（$\gamma = 0$）训练，采用多样性促进（$\gamma < 0$）训练的集成在分布最右侧区域包含更多数据，展现了多样性促进对单个预测的影响。对于这些预测中的每一个，我们随后考察标准（$\gamma = 0$）集成的预测准确率，并将其称为反事实准确率（底行）——即，在没有任何多样性促进的情况下该预测的准确率。理想情况下，我们希望促进多样性的集成预测具有较低的反事实准确率（即，我们使原本会导致错误的预测多样化），但实践中并非如此。对于使用詹森间隙多样性促进训练的 ResNet 8（左侧）和 ResNet 18（中间）集成，我们观察到，对于导致更高预测多样性的干预（颜色较浅的线条），多样性预测中反事实正确的比例（更靠右侧）更高。换言之，大多数集成表现出更高多样性的数据点，在没有多样性干预的情况下原本就会被正确分类。如果多样性促进仅仅使错误去相关，我们预期会看到相反的结果：在预测多样性较低时，反事实集成准确率的值更高；而相对于标准集成表现出更多多样性的数据，其值应更低。值得注意的是，这一趋势同时适用于小型和大型网络深度集成。此外，这一现象还可推广到本节研究的促进多样性的机制之外。在附录I 中，我们展示了随机森林中也存在类似趋势，这表明即使是特征子集选择等传统集成技术，也会影响原本正确的预测。

对于原本正确且有信心的预测，预测多样性有害。 当集成成员已经做出准确度和置信度都很高的预测时（大型分类神经网络通常如此），上述结果具有重要意义（Guo 等人， 2017年）。在这种情况下，组件模型预测

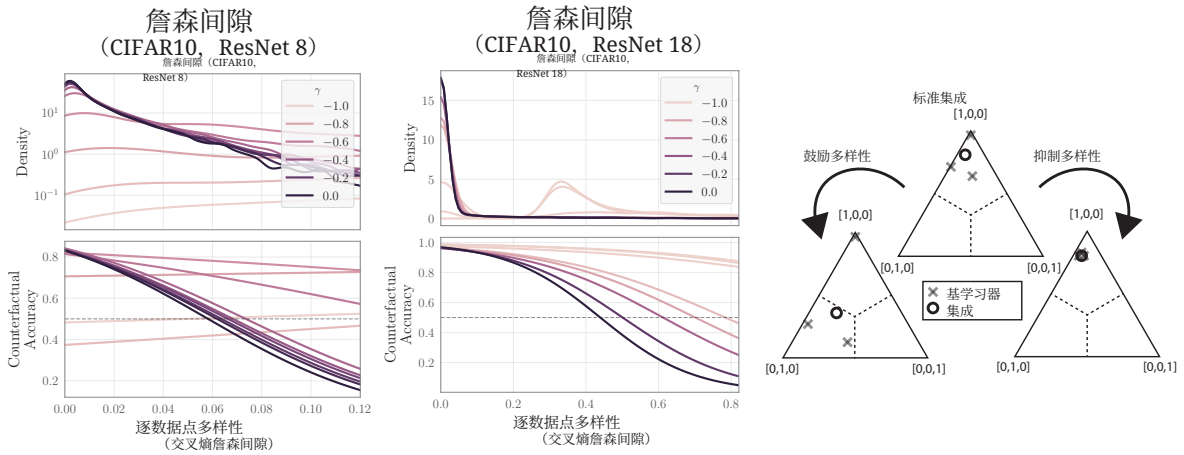


图 4：**鼓励多样性有害于置信度高且准确率高的分类器集成。（左、中：）**我们通过反事实准确率：标准 $\gamma = 0$集成预测给出的逐数据点准确率（详见第4.2 节），研究鼓励多样性对各数据点的影响。评估一个通过鼓励多样性训练的集成（式（8））， 可得到测试集预测的集成多样性分布（**上排**，每行对应一个 γ 值）。在此分布上，我们进一步测量反事实准确率（**下排**），并关注右侧尾部的数据，因为这些数据受多样性促进技术的影响最强。然而，我们发现，受鼓励多样性影响最大的预测具有较高的反事实准确率：也就是说，如果我们用标准集成测试这些数据点，结果原本就会是正确的。这一发现对小型（ResNet 8，**左侧**）和大型网络集成（ResNet 18，**中间**）均成立（其他促进多样性的方法见附录I。）**（右）：**将正确且置信度高的组件模型预测去相关会降低集成性能。正确且置信度高的组件模型预测（×）集中在概率单纯形的一个顶点上（**上图**）。由于单纯形的几何结构，鼓励多样性（**左图**）必然会使集成预测（○）变差，并可能改变类别预测。相比之下，抑制多样性（**右图**）则不一定会损害集成预测。

都会集中在概率单纯形的同一个顶点上（见图 4，“标准集成”），因此不可能在不损害式（3）中的集成风险的情况下提高预测多样性。提高多样性必然要求至少一个组件模型降低置信度，从而降低集成预测的置信度（并可能改变集成预测）（示意图见图 4 右侧）。

这些论断描述了一种机制，可将多样性促进效果的观测差异直接与组件模型容量联系起来。小型神经网络、决策树和其他低容量模型产生的高置信度且正确的预测较少，而预测多样性在这些预测上只会损害组件模型性能。对于剩下的低置信度或错误预测，更多的多样性未必会损害组件模型性能，因此能带来更有利的权衡。作为对这些论断的初步检验，我们预测，任何规模的回归集成都不会受到这种病理现象的影响，因为输出空间没有界限，所以一个组件模型的低估预测可以抵消另一个组件模型的高估预测。在附录B.2.3中，我们确实证实，大型回归深度集成能够从多样性促进中获益，不同于分类集成。类似地，使用抑制高置信度预测的损失函数训练的模型（例如标签平滑损失（Szegedy 等人，2016））较不容易受到这种病理现象的影响，进一步支持了这些论断（附录 D.4）。

为什么抑制预测多样性会有益？尽管前述论点支持了我们关于鼓励多样性的结果，却没有解释为什么在大型网络深度集成中抑制多样性是无害的（有时甚至有益）。在此，我们为观察到的结果提出一种可能的假设。近期的理论证据表明，增大过参数化模型的规模相当于缩减方差（Neal 等人，2018年；Adlam 和 Pennington，2020年；d’Ascoli 等人，2020年）——图6 以及集成蒸馏领域的其他研究（Lan 等人，2018年；Zhao 等人，2022年）的实证结果支持了这一结论。抑制

因此，通过使用正则化器来抑制多样性，也可能成为实现方差缩减的另一条途径，其效果类似于增大模型规模。验证这一假设仍是一个开放问题，我们将留待未来研究。

与现有工作的关系。重要的是，正如第2 节和附录C 节所述，我们在关于集成多样性机制的大量文献中，没有发现任何与我们的主张相矛盾的证据。特别是，在许多旨在提高预测多样性的不同机制中，只有在将其应用于小型网络集成，或使用其他训练数据来源时，才有研究报告集成性能显著提升；而在训练由实践中使用的大型神经网络组成的深度集成时，并未报告此类提升。

5 最佳深度集成的预测多样性较低

迄今为止，我们的结果考察了这样一些方法：提高集成多样性需要付出明确代价，即牺牲组件模型的性能。不过，集成多样性也有无需付出此类代价的来源，例如对异质架构和训练方法进行集成。在这种情况下，可以选用训练方式或架构不同、但性能相近的组件模型，从而“免费”构建多样化的集成成员。即使在这种情况下，我们也证明，选择集成成员以提高这种“免费”的多样性仍会产生机会成本。尽管低性能架构（如 AlexNet、VGG 11 等）之间存在显著的预测多样性，但高性能架构（如 ResNet 152、DenseNet 161 等）在不同数据点上的预测往往非常相似。换言之，随着组件神经网络规模增大，异构架构集成所能提供的多样性会逐渐减少。先前的研究也证实了这一发现：对多样化架构或训练流程进行集成，只能略微提升集成准确率（Gontijo-Lopes 等人，2022年）。另一种做法是，假设我们能够使用一个异质的已训练模型池，那么可以构建只使用最佳组件模型的集成。尽管最佳组件模型的预测几乎完全相同，这类集成的性能仍远远优于那些使预测多样性最大化的集成。这些结果补充了第4节的发现：当试图从候选模型池中构建最佳集成时，我们发现，最佳策略是牺牲预测多样性，以换取更好的组件模型性能，这与通常的直觉恰恰相反。（关键的是，我们还表明，这一结果也使深度集成有别于传统小模型集成；后者在组件模型性能较差但多样性极高时效果最佳。）

实验设置。 我们从包含 31 种 CIFAR10 架构和 92 种 ImageNet 架构的集合中抽取模型，以组合方式构建异质集成，每个集成包含 $M = 4$ 个网络。我们随机选取四个独立训练且属于同一性能分位数的模型。此外，我们还构建由同质集成组成，每个集成包含 $M = 4$ 个采用相同架构/训练流程独立训练的副本。总计，对于 CIFAR10/ImageNet，我们分别构建了 1050/427 个异质集成和 22/3 个同质集成。关键在于，我们所考虑的模型涵盖了现有架构的广泛范围，其中包括Taori 等人 (2020年) 中的标准模型。（架构/训练详情见附录 B 。）我们测量了这些集成在分布内（InD）测试集以及 CINIC10（Darlow 等人，2018年）和 ImageNet-C（Hendrycks 和 Dietterich，2019年）分布外（OOD）数据集上的性能。

5.1 结果

图 5展示了集成性能（测试集交叉熵）随组件模型平均性能的变化。根据公式（3），这些量之间的差值就是衡量预测多样性的詹森间隙。模型越远离恒等线下方，多样性越高；越接近恒等线，预测多样性越低。（更多分布外数据集见附录 G。）

从该图中我们观察到，在对组件模型平均性能进行条件化的情况下（从图示角度看：考虑图 5 中的任意一个垂直切片），异质集成通常比同质集成获得更好的（更低的）性能。这一结果表明，异质集成所带来的架构/训练多样性通常是有益的（正如预期，因为组件模型性能不存在任何代价），从而证实了先前研究的发现（Mustafa 等人，2020年；Gontijo-Lopes 等人，2022年；Agostinelli 等人，2022年）。与此同时，在所有异质集成中，性能最佳的集成并不是多样性最高的集成（即图 5 各面板右下角的集成）。相比之下，最佳集成

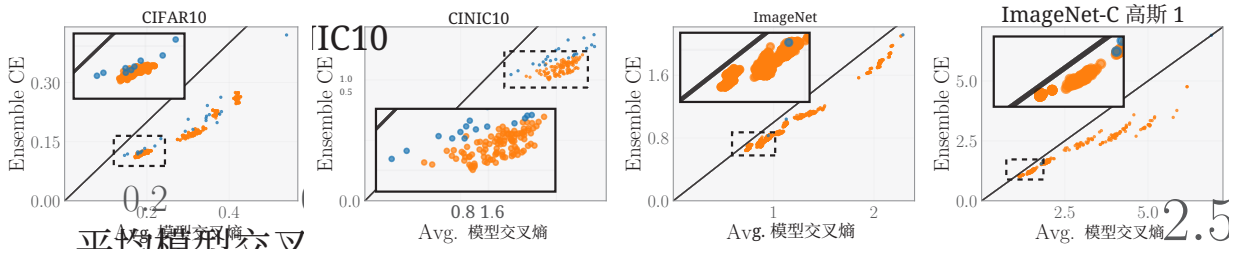


图 5：**预测多样性的机会成本**。图中展示了集成在测试集上的性能与其平均组件模型的性能（以交叉熵衡量）。每个标记代表一个模型集成，其在 CIFAR10/CINIC10 和 ImageNet/ImageNet-C 的分布内/分布外数据集上接受评估。虚线边界标示了插图区域，以便查看细节。集成以颜色区分：同质集成（蓝色）和异质集成（橙色）。在控制组件模型性能后（纵向切片），异质集成比同质集成具有更高的多样性（离恒等线更远）。然而，当集成性能更高的组件模型时（位置更靠左），异质集成所带来的预测多样性会逐渐减少（更接近恒等线）。性能最佳的集成（位置最低）采用了最佳的组件模型（位置最靠左），且预测多样性很低。

通常由性能最高的组件模型构成（位于图 5 的左下角），即使它们的多样性逐渐降低 （趋近于恒等线）。因此，随着我们考察由更强组件模型构成的集成，通过异构架构/训练流程获得的多样性会产生递减收益；尽管这些组件模型的预测多样性低得多，但其性能更佳。在此情况下，预测多样性并不会直接损害集成性能，但这些结果表明，为最大化预测多样性而构建集成存在机会成本。与经典的集成直觉相反，在图 5 中，无论是分布内还是分布外数据集，性能最佳的深度集成（即组件模型平均性能最高的集成）都是通过牺牲预测多样性来选择性能更高的组件模型而构成的。

5.2 分析

接下来，我们研究图 5 中观察到的机会成本是否也取决于组件模型容量。比较深度集成与随机森林和装袋随机特征模型的行为（图 6），我们观察组件模型性能与预测多样性如何随模型规模变化而共同影响集成性能。对于不同深度的随机森林（图 6 左），我们绘制组件模型平均性能（测试集交叉熵）与集成预测多样性（詹森间隙）的关系。圆点越大，表示随机森林的深度越大。与图1一样，集成性能由对角线等值集表示。与深度集成截然相反，性能最佳的随机森林（对角线等值集最低）多样性最高（最靠右）。

接下来，我们研究在 MNIST 上训练的随机特征分类器集成。在这种设置下，我们可以将小规模与大规模网络明确对应到模型容量的概念：根据分类器宽度（即随机特征的数量），网络处于欠参数化区域（无法完美拟合训练数据）或过参数化区域（达到 100% 的训练准确率）。在图6 （中左）中，我们再次绘制组件模型性能与多样性的关系图，对角线等值集表示集成性能。随着组成分类器的宽度增加（图中以点的大小表示），我们仅通过观察平均单模型性能（图6 中左，纵轴）随模型规模（标记大小）变化的情况，来识别过参数化区域和欠参数化区域。随着模型规模增大，平均单模型性能呈现“双重下降”（Belkin 等人，2019年），性能峰值标志着插值阈值。值得注意的是，集成多样性似乎也会在插值阈值处改变趋势：在欠参数化区域，随模型规模增大而增加；在过参数化区域，则随模型规模增大而降低。虽然增大宽度总能提升集成性能，但在欠参数化区域，这种提升源于多样性的增加，即牺牲组件模型性能来换取更高的多样性；这一点与

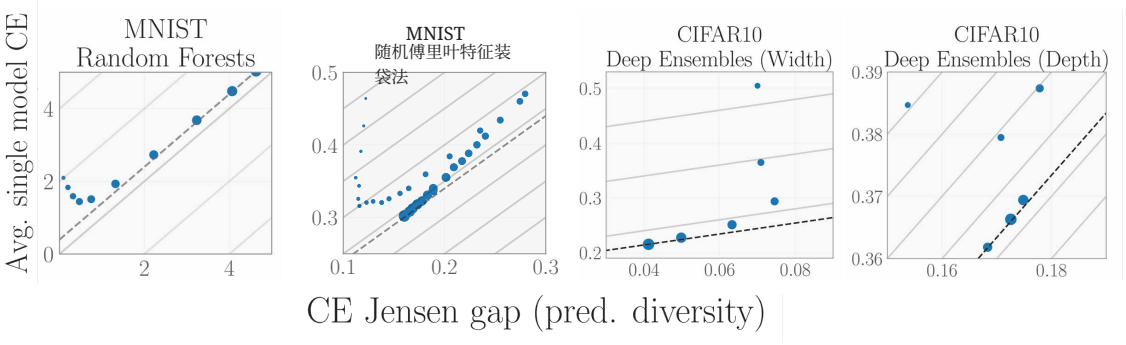


图 6：**小型集成与大型集成通过不同策略提升性能**。决策树集成（**左**）、随机特征分类器（**中左**）和神经网络（**中右/右**）。虚线对角线表示在给定模型类别中所能达到的最佳集成交叉熵（右下方为佳）。点的大小对应组件模型的深度/宽度。**（左）：**随机森林通过提升多样性（使用更深的组件树获得）来改善。**（中左）：**随机特征集成随着分类器宽度的增加而单调改善。当组件模型参数不足（较小的点）时，集成性能的提升对应于预测多样性的增加。当组件模型参数过多（较大的点）时，集成性能的提升对应于预测多样性的降低。**（中右/右）：**深度集成随着宽度/深度的增加而单调改善。达到一定规模后，这种改善完全归因于组件模型性能的提升以及预测多样性的降低。

我们在决策树集成中观察到的情况相同（图6 中左；小点）。与此同时，在过参数化区域，集成性能的提升则源于更优秀的组成分类器；它们牺牲预测多样性，以换取更好的组件模型。

我们通过改变组件模型的宽度（图6 中右）和深度（图6 右），将其作为模型容量的替代指标（详情见附录 B.4 ），从而在 ResNet 集成中重现了这一发现。同样，增加宽度和深度（图中以点的大小表示）会带来性能更好的集成。当网络模型较小时，集成性能的提升与多样性起初呈正相关；但当组件网络足够大后，两者便转为负相关。我们认为，这种转变大致对应于参数不足与参数过多的 ResNet 架构之间的区别，但训练过程中的正则化使我们无法明确确定这一转变点。这些结果表明，鼓励预测多样性所带来的收益递减并非现代深度集成所特有，而是在大型（很可能参数过多的）模型集成中普遍存在的现象；这类集成处于这样一种状态：牺牲预测多样性以换取更好的组件模型性能，反而能带来更好的集成。

6 讨论

我们证明，在构建深度集成（或更广义地说，由高容量分类器组成的集成）时，鼓励或筛选预测多样性是一种次优且可能有害的策略。这一发现源于预测多样性与组件模型性能之间存在明确的权衡或机会成本。我们将这种现象与低容量神经网络集成（以及回归集成）进行对比：在后者中，我们观察到，追求更高的预测多样性能够提升集成性能的整体收益，这表明我们的发现与传统集成文献完全一致。

总体而言，我们的发现表明，传统集成设置中颇具价值的直觉，必须结合组件模型容量这一概念加以限定。其他研究也指出，组件模型的容量是深度集成的一项重要特征。例如，Kondratyuk 等人(2020年) 和 Lobacheva 等人(2020年) 研究了这样一个问题：与参数数量相同的单模型相比，集成在何时更有效；而 Kobayashi 等人(2021年) 发现，随着集成模型的宽度不断增加，深度集成的性能会逐渐下降至单模型的水平。 Theisen 等人(2023年) 则对集成中预测多样性与组件模型平均性能之间的关系提出了另一种刻画。尽管他们没有研究鼓励或追求集成中的预测多样性所带来的影响，

他们观察到，当考察超过插值阈值的模型时，集成的效果会减弱，这与我们在第 5 节中的部分结果一致。同样，理论表明，集成的许多优势或许只是重现了增加组件模型宽度所带来的优势（Adlam 和 Pennington，2020年；Geiger 等人，2020年；Bernstein 等人，2021年）。

后续研究的一个重要目标，是更详细地描述集成成员容量、集成成员性能与我们在此观察到的现象之间的关系。根据第4节的结果，我们预计，多样性促进改善还是降低集成性能，取决于高置信度预测所占的比例及其质量（正确还是错误）。以往研究发现，单个模型的性能和模型容量都与高置信度预测相关（例如，Guo 等人，2017年；Nakkiran 等人，2021年；Bernstein 等人，2021年），因此，如果能知道其中一个（或两个）因素是否可以预测多样性促进会改善还是损害集成性能，将会很有价值。我们或许可以通过对不同容量的未正则化模型施加多样性促进（受图6中 RFF模型的启发），来区分这些因素，从而更好地刻画在不同数据集和正则化器下，多样性促进从有益转为有害的过程。更广泛地说，我们的结果可能与Breiman（1996年）的观察有关；该研究已指出，学习模型预测的“稳定性”（相对于训练数据而言）是决定装袋法成败的关键特征：不稳定预测器可以从装袋法中获益，但对于在不同训练集上都能生成稳定预测的模型，装袋法则不会带来提升。未来一个值得探讨的方向，是在模型稳定性的框架下，结合我们所考察的各种预测多样性来源来审视这些结果。

总而言之，我们的结果以及先前的相关工作表明，在集成中使用更大型的组件模型时，我们理应预期集成内部的多样性会自然降低。此外，我们的结果意味着，目前为深度集成引入额外的预测多样性来源所做的努力（通过联合训练机制、异构架构等）不会改变这一事实，除非这些方法能够克服若干根本性挑战，例如仅在出现反事实错误时才促进预测多样性，并找到“免费”的多样性来源，使其带来的优势超过使用更强大、但多样性较低的模型。尽管集成深度神经网络无疑是提升当前神经网络模型性能的一种实用技术，但随着神经网络架构和训练范式不断改进，我们预计，有意义的预测多样性来源将愈发难以寻觅。

更广泛影响声明

我们的研究得出的一个结论是，与其尝试寻找现有模型中具有多样性的组合，不如使用更强大的模型来构建集成。因此，我们的研究可能带来的一个负面影响是，进一步促使人们投入计算资源来训练新的、更大规模的组件模型，而不是重复利用现有模型。

与此同时，我们的研究发现，采用联合训练策略来创建更多样化的大容量深度集成，益处寥寥，而这类集成本身就计算开销高昂。我们希望这一结论有助于降低创建新集成模型所需的一些计算成本。

致谢

我们感谢 John Miller 分享在 CIFAR10 上训练的模型，也感谢 Taori 等人 (2020年) 将其训练的 ImageNet 模型和代码开源，并使其易于使用。

我们还要感谢 Rich Zemel 提出的宝贵建议，以及 Yixin Wang、Yaniv Yacoby、Zelda Mariet、Dustin Tran 和 Jeremy Bernstein 提供的真知灼见；同时感谢 Zuckerman Mind BrainBehavior Institute 和理论神经科学中心给予的额外支持。

TA 获美国国立卫生研究院训练资助项目 2T32NS064929-11 的支持。EKB 获美国国立卫生研究院 5T32NS064929-13 和美国国立卫生研究院 U19NS107613 的支持。JPC 获西蒙斯基金会、麦克奈特基金会和 Grossman

心智统计中心。TA、EKB 和 JPC 获美国国家科学基金会 1707398 及盖茨比慈善信托 GAT3708 的支持。

参考文献

Moloud Abdar、Farhad Pourpanah、Sadiq Hussain、Dana Rezazadegan、Li Liu、Mohammad Ghavamzadeh、Paul Fieguth、Xiaochun Cao、Abbas Khosravi、U Rajendra Acharya 等。深度学习中不确定性量化的综述：技术、应用与挑战。*Information Fusion*，76:243–297，2021年。

Taiga Abe、E. Kelly Buchanan、Geoff Pleiss 和 John Patrick Cunningham。最佳深度集成会牺牲预测多样性。发表于 *I Can't Believe It's Not Better Workshop*：通过经验性证伪理解深度学习，2022a年。URL https://openreview.net/forum?id=6sBiAIpkUiQ。

Taiga Abe、E Kelly Buchanan、Geoff Pleiss、Richard Zemel 和 John P Cunningham。深度集成有效，但它们是必需的吗？发表于 神经信息处理系统进展，2022b年。

Ben Adlam 和 Jeffrey Pennington。理解双重下降需要细粒度的偏差-方差分解。神经信息处理系统进展，33:11022–11032，2020年。

Andrea Agostinelli、Jasper Uijlings、Thomas Mensink 和 Vittorio Ferrari。用于选择源模型集成的可迁移性度量。载于*IEEE/CVF*计算机视觉与模式识别会议，第 7936–7946 页，2022年。

Arsenii Ashukha、Alexander Lyzhov、Dmitry Molchanov 和Dmitry Vetrov。深度学习中域内不确定性估计与集成的陷阱。*arXiv*预印本 *arXiv:2002.06470*，2020年。

Mikhail Belkin、Daniel Hsu、Siyuan Ma 和 Soumik Mandal。调和现代机器学习实践与经典的偏差–方差权衡。美国国家科学院院刊，116(32): 15849–15854，2019年。

Jeremy Bernstein、Alex Farhang 和 Yisong Yue。作为贝叶斯点机的核插值。*arXiv*，2021年。

Jock Blackard。覆盖 类型。UCI机器学习资料库，1998年。DOI：https://doi.org/10.24432/C50K5 N.

Leo Breiman。装袋预测器。机器学习，24(2):123–140，1996年。

Leo Breiman。随机森林。机器学习，45(1):5–32，2001年。

E Kelly Buchanan、Michael W Dusenberry、Jie Ren、Kevin Patrick Murphy、Balaji Lakshminarayanan 和 Dustin Tran。图像分割的可靠性基准。发表于*NeurIPS 2022* 分布偏移：连接方法与应用研讨会，2022年。

Sebastian Buschjäger、Lukas Pfahler 和 Katharina Morik。用于深度集成的广义负相关学习。*arXiv* 预印本 *arXiv:2011.02952*，2020年。

Adele Cutler 和 Guohua Zhao。Pert-perfect 随机树集成。计算科学与统计，33（4）：90–4，2001年。

Ali Dabouei、Sobhan Soleymani、Fariborz Taherkhani、Jeremy Dawson 和 Nasser M Nasrabadi。利用对抗扰动下的联合鲁棒性。发表于*IEEE/CVF*计算机视觉与模式识别会议，第 1122–1131 页，2020年。

Francesco D'Angelo 和 Vincent Fortuin。排斥性深度集成具有贝叶斯特性。神经信息处理系统进展，34:3451–3465，2021年。

Luke N Darlow、Elliot J Crowley、Antreas Antoniou 和 Amos J Storkey。CINIC-10 既不是 ImageNet，也不是 CIFAR-10。*arXiv*预印本 *arXiv:1810.03505*，2018年。

Ahmet Demirkaya、Jiasi Chen 和 Samet Oymak。探究损失函数在多类别分类中的作用。载于《2020年第54届信息科学与系统年会（CISS）》，第1–5页。IEEE，2020年。Jia Deng、Wei Dong、Richard Socher、Li-Jia Li、Kai Li 和 Li Fei-Fei。ImageNet：一个大规模层次化图像数据库。载于《计算机视觉与模式识别》，第248–255页，2009年。Thomas G Dietterich。构建决策树集成的三种方法的实验比较：装袋法、提升法和随机化。《机器学习》，32：1–22，1998年。Thomas G Dietterich。机器学习中的集成方法。载于《多分类器系统国际研讨会》，第1–15页，2000年。Josip Djolonga、Jessica Yung、Michael Tschannen、Rob Romijnders、Lucas Beyer、Alexander Kolesnikov、Joan Puigcerver、Matthias Minderer、Alexander D'Amour、Dan Moldovan 等。卷积神经网络的鲁棒性与可迁移性。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第16458–16468页，2021年。Stéphane d'Ascoli、Maria Refinetti、Giulio Biroli 和 Florent Krzakala。双重下降中的双重问题：惰性机制中的偏差与方差。在《国际机器学习大会》上发表，2020年。Stanislav Fort、Huiyi Hu 和 Balaji Lakshminarayanan。深度集成：从损失景观的视角看。arXiv预印本 arXiv:1912.02757，2019年。Yoav Freund。通过多数投票提升弱学习算法。《信息与计算》，121(2)：256–285，1995年。Xavier Gastaldi。Shake-shake正则化。arXiv预印本 arXiv:1705.07485，2017年。Mario Geiger、Stefano Spigler、Arthur Jacot 和 Matthieu Wyart。分离深度神经网络中的特征学习与惰性训练。《统计力学杂志：理论与实验》，2020(11)：113301，2020年。Chengyue Gong、Lemeng Wu 和 Qiang Liu。如何填充最优集？具有无害多样性的总体梯度下降。arXiv预印本 arXiv:2202.08376，2022年。Raphael Gontijo-Lopes、Yann Dauphin 和 Ekin Dogus Cubuk。不存在一种表征可以统御一切：训练方法的重叠特征。载于《国际学习表征会议》，2022年。网址 https://openreview.net/forum?id=BK-4qbGgIE3。Yury Gorishniy、Ivan Rubachev 和 Artem Babenko。表格深度学习中数值特征的嵌入。《神经信息处理系统进展》，35：24991–25004，2022年。Chuan Guo、Geoff Pleiss、Yu Sun 和 Kilian Q Weinberger。现代神经网络的校准。载于《国际机器学习大会》，2017年。Neha Gupta、Jamie Smith、Ben Adlam 和 Zelda E Mariet。分类器集成：偏差-方差视角。《机器学习研究汇刊》，2022年。网址 https://openreview.net/forum?id=lIOQFVncY9。Fredrik K Gustafsson、Martin Danelljan 和 Thomas B Schon。评估用于鲁棒计算机视觉的可扩展贝叶斯深度学习方法。载于《计算机视觉与模式识别研讨会》，第318–319页，2020年。Dongyoon Han、Jiwhan Kim 和 Junmo Kim。深层金字塔残差网络。载于《IEEE计算机视觉与模式识别会议论文集》，第5927–5935页，2017年。Lars Kai Hansen 和 Peter Salamon。神经网络集成。《模式分析与机器智能汇刊》，12(10)：993–1001，1990年。

Marton Havasi、Rodolphe Jenatton、Stanislav Fort、Jeremiah Zhe Liu、Jasper Snoek、Balaji Lakshminarayanan、Andrew Mingbo Dai 和 Dustin Tran。训练独立子网络以实现鲁棒预测。载于国际学习表征会议，2021年。网址 https://openreview.net/forum?id=OGg9XnKxFAH。Kaiming He、张祥雨、任少卿和孙剑。用于图像识别的深度残差学习。载于*IEEE*计算机视觉与模式识别会议会刊，第770–778页，2016。Dan Hendrycks 和 Thomas G Dietterich。神经网络在常见损坏和表面变化下鲁棒性的基准测试。载于国际学习表征会议，2019年。Jie Hu、Li Shen 和 Gang Sun。压缩与激励网络。载于*IEEE*计算机视觉与模式识别会议会刊，第7132–7141页，2018年。Gao Huang、Zhuang Liu、Laurens Van Der Maaten 和 Kilian Q Weinberger。密集连接卷积网络。载于*IEEE*计算机视觉与模式识别会议会刊，第4700–4708页，2017年。Like Hui 和 Mikhail Belkin。分类任务中采用平方损失与交叉熵训练的神经网络架构评估。*arXiv* 预印本 *arXiv:2006.07322*，2020年。Siddhartha Jain、Ge Liu、Jonas Mueller 和 David Gifford。最大化整体多样性，以提升深度集成中的不确定性估计。载于*AAAI*人工智能大会会刊，第34卷，第4264–4271页，2020年。Alan Jeffares、Tennison Liu、Jonathan Crabbé 和 Mihaela van der Schaar。深度集成联合训练因学习器串谋而失败。*arXiv* 预印本 *arXiv:2301.11323*，2023年。Sanjay Kariyappa 和 Moinuddin K Qureshi。通过多样性训练提升集成的对抗鲁棒性。*arXiv* 预印本 *arXiv:1901.09981*，2019年。Seijin Kobayashi、Johannes von Oswald 和 Benjamin F Grewe。深度集成中反转的偏差-方差权衡。国际机器学习会议，2021年。Dan Kondratyuk、Mingxing Tan、Matthew Brown 和 Boqing Gong。集成较小模型何时比使用单个大型模型更高效。*arXiv* 预印本 *arXiv:2005.00570*，2020年。Alex Krizhevsky、Geoffrey Hinton 等。通过微型图像学习多层特征，2009年。Ludmila I Kuncheva 和 Christopher J Whitaker。分类器集成中的多样性度量及其与集成准确率的关系。机器学习，51(2):181–207，2003年。Balaji Lakshminarayanan、Alexander Pritzel 和 Charles Blundell。使用深度集成进行简单且可扩展的预测不确定性估计。载于神经信息处理系统进展，2017年。Xu Lan、Xiatian Zhu 和 Shaogang Gong。通过即时原生集成进行知识蒸馏。神经信息处理系统进展，第31卷，2018年。Ya Le 和 Xuan Yang。Tiny ImageNet 视觉识别挑战。*CS 231N*，7(7):3，2015年。Yann LeCun、Léon Bottou、Yoshua Bengio 和 Patrick Haffner。将基于梯度的学习应用于文档识别。*IEEE*会刊，86(11):2278–2324，1998年。Daniel Lee 和 H Sebastian Seung。非负矩阵分解算法。神经信息处理系统进展，第13卷，2000年。Stefan Lee、Senthil Purushwalkam、Michael Cogswell、David Crandall 和 Dhruv Batra。为何 m 个头优于一个：训练多样化的深度网络集成。*arXiv* 预印本 *arXiv:1511.06314*，2015年。Yoonho Lee、Huaxiu Yao 和 Chelsea Finn。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年。

Ekaterina Lobacheva、Nadezhda Chirkova、Maxim Kodryan 和 Dmitry P Vetrov。深度集成中的幂律。载于神经信息处理系统进展，2020年。John P Miller、Rohan Taori、Aditi Raghunathan、Shiori Sagawa、Pang Wei Koh、Vaishaal Shankar、Percy Liang、Yair Carmon 和 Ludwig Schmidt。准确率在线：分布外泛化与分布内泛化之间的强相关性。载于国际机器学习大会，2021年。Aaron Mishtal 和 Itamar Arel。并行训练的神经网络集成中的 Jensen-Shannon 散度。载于*2012*年第*11*届国际机器学习与应用大会，第2卷，第558–562页。IEEE，2012年。Rafael Müller、Simon Kornblith 和 Geoffrey E Hinton。标签平滑何时有帮助？神经信息处理系统进展，第32卷，2019年。

Basil Mustafa、Carlos Riquelme、Joan Puigcerver、André Susano Pinto、Daniel Keysers 和 Neil Houlsby。用于低数据迁移学习的深度集成。*arXiv*预印本 *arXiv:2010.06866*，2020年。

Mahdi Pakdaman Naeini、Gregory Cooper 和 Milos Hauskrecht。使用贝叶斯分箱获得校准良好的概率。发表于*AAAI*人工智能大会，2015年。

Preetum Nakkiran、Gal Kaplun、Yamini Bansal、Tristan Yang、Boaz Barak 和 Ilya Sutskever。深度双重下降：更大的模型和更多的数据何以有害。统计力学杂志：理论与实验，2021年(12):124003，2021年。

Brady Neal、Sarthak Mittal、Aristide Baratin、Vinayak Tantia、Matthew Scicluna、Simon Lacoste-Julien 和 Ioannis Mitliagkas。神经网络中偏差-方差权衡的现代视角。*arXiv*预印本 *arXiv:1810.08591*，2018年。

Jeremy Nixon、Balaji Lakshminarayanan 和 Dustin Tran。为什么自助法深度集成并没有更好？载于“真不敢相信它竟然没有更好！”*NeurIPS 2020* 研讨会。openreview.net，2020年12月。

Michael Opitz、Horst Possegger 和 Horst Bischof。深度神经网络的高效模型平均。载于计算机视觉—*ACCV 2016*：第*13*届亚洲计算机视觉大会，台北，台湾，*2016*年*11*月*20*日至*24*日，修订精选论文集，第*II*部分第*13*卷，第205–220页。施普林格，2017年。

Luis A. Ortega、Rafael Cabañas 和 Andres Masegosa。神经网络集成中的多样性与泛化。载于国际人工智能与统计大会，第 11720–11743 页。PMLR，2022年。Yaniv Ovadia、Emily Fertig、Jie Ren、Zachary Nado、David Sculley、Sebastian Nowozin、Joshua V Dillon、Balaji Lakshminarayanan 和 Jasper Snoek。你能信任模型的不确定性吗？在数据集偏移下评估预测不确定性。载于神经信息处理系统进展，2019年。Matteo Pagliardini、Martin Jaggi、François Fleuret 和 Sai Praneeth Karimireddy。同意彼此意见相左：通过分歧实现多样性，从而提升可迁移性。*arXiv*预印本 *arXiv:2202.04414*，2022年。Tianyu Pang、Kun Xu、Chao Du、Ning Chen 和 Jun Zhu。通过促进集成多样性提升对抗鲁棒性。载于国际机器学习大会，第 4970–4979 页。PMLR，2019年。F. Pedregosa、G. Varoquaux、A. Gramfort、V. Michel、B. Thirion、O. Grisel、M. Blondel、P. Prettenhofer、R. Weiss、V. Dubourg、J. Vanderplas、A. Passos、D. Cournapeau、M. Brucher、M. Perrot 和 E. Duchesnay。Scikit-learn：Python中的机器学习。机器学习研究杂志，12:2825–2830，2011年。Michael P Perrone 和 Leon N Cooper。当网络意见不一致时：用于混合神经网络的集成方法，1992年。Alexandre Rame 和 Matthieu Cord。DICE：通过条件冗余对抗估计实现深度集成的多样性。*arXiv*预印本 *arXiv:2101.05544*，2021年。

Benjamin Recht、Rebecca Roelofs、Ludwig Schmidt 和 Vaishaal Shankar。CIFAR-10分类器能否泛化到CIFAR-10？*arXiv*预印本 *arXiv:1806.00451*，2018年。Benjamin Recht、Rebecca Roelofs、Ludwig Schmidt 和 Vaishaal Shankar。ImageNet分类器能否泛化到ImageNet？载于国际机器学习大会，2019年。Andrew Ross、Weiwei Pan、Leo Celi 和 Finale Doshi-Velez。局部独立预测模型的集成。载于*AAAI*人工智能大会论文集，第 34 卷，第 5527–5536 页，2020年。Karen Simonyan 和 Andrew Zisserman。用于大规模图像识别的超深卷积网络。*arXiv*预印本 *arXiv:1409.1556*，2014。Samarth Sinha、Homanga Bharadhwaj、Anirudh Goyal、Hugo Larochelle、Animesh Garg 和 Florian Shkurti。DIBS：在模型集成中引入多样性的信息瓶颈。载于*AAAI*人工智能大会论文集，第 35 卷，第 9666–9674 页，2021年。Lawrence Stewart、Francis Bach、Quentin Berthet 和 Jean-Philippe Vert。回归即分类：任务表述对神经网络特征的影响。*arXiv*预印本 *arXiv:2211.05641*，2022年。Christian Szegedy、Wei Liu、Yangqing Jia、Pierre Sermanet、Scott Reed、Dragomir Anguelov、Dumitru Erhan、Vincent Vanhoucke 和 Andrew Rabinovich。借助卷积网络更进一步。载于计算机视觉与模式识别，2015年。Christian Szegedy、Vincent Vanhoucke、Sergey Ioffe、Jon Shlens 和 Zbigniew Wojna。重新思考用于计算机视觉的Inception架构。载于*IEEE*计算机视觉与模式识别会议论文集，第 2818–2826 页，2016。Rohan Taori、Achal Dave、Vaishaal Shankar、Nicholas Carlini、Benjamin Recht 和 Ludwig Schmidt。衡量图像分类中对自然分布偏移的鲁棒性。载于神经信息处理系统进展，2020年。Damien Teney、Ehsan Abbasnejad、Simon Lucey 和 Anton Van den Hengel。规避简单性偏差：训练一组多样化模型，以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论文集，第 16761–16772 页，2022年。Sujay Thakur、Cooper Lorsung、Yaniv Yacoby、Finale Doshi-Velez 和 Weiwei Pan。用于贝叶斯回归的具备不确定性感知能力的（UNA）基函数：采用多头辅助网络。*arXiv*预印本 *arXiv:2006.11695*，2020年。Ryan Theisen、Hyunsuk Kim、Yaoqing Yang、Liam Hodgkinson 和 Michael W Mahoney。集成何时才真正有效？*arXiv*预印本 *arXiv:2305.12313*，2023年。Alexandru Tifrea、Eric Stavarache 和 Fanny Yang。使用带正则化分歧的集成进行半监督新颖性检测。载于人工智能中的不确定性，第 1939–1948 页。PMLR，2022年。Dustin Tran、Jeremiah Liu、Du Phan Michael W Dusenberry、Mark Collier、Jie Ren、Kehang Han、Zi Wang、Zelda Mariet、Huiyi Hu、Neil Band、Tim GJ Rudner、Karan Singhal、Zachary Nado、Joost van Amersfoort、Andreas Kirsch、Rodolphe Jenatton、Nithum Thain、Honglin Yuan、Kelly Buchanan、Kevin Murphy、D Sculley、Yarin Gal、Zoubin Ghahramani、Jasper Snoek 和 Balaji Lakshminarayanan。PLEX：借助预训练大型模型扩展提升可靠性。*arXiv*预印本 *arXiv:2207.07411*，2022年。Mehmet Ozgur Turkoglu、Alexander Becker、Hüseyin Anil Gündüz、Mina Rezaei、Bernd Bischl、Rodrigo Caye Daudt、Stefano D'Aronco、Jan Wegner 和 Konrad Schindler。FiLM集成：通过特征级线性调制实现概率深度学习。神经信息处理系统进展，35:22229–22242，2022年。Andrew Webb、Charles Reynolds、Wenlin Chen、Henry Reeve、Dan Iliescu、Mikel Lujan 和 Gavin Brown。要不要采用集成：端到端训练何时会失败？载于欧洲机器学习与数据库知识发现联合会议，第 109–123 页。施普林格，2020年。

Yeming Wen、Dustin Tran 和 Jimmy Ba。BatchEnsemble：一种高效集成与终身学习的替代方法。载于国际学习表征会议，2017年。

Florian Wenzel、JasperSnoek、Dustin Tran 和 Rodolphe Jenatton。用于鲁棒性和不确定性量化的超参数集成。神经信息处理系统进展，33：6514–6527，2020年。

Danny Wood、Tingting Mu、Andrew Webb、Henry Reeve、Mikel Lujan 和 Gavin Brown。集成学习中多样性的统一理论。*arXiv*预印本 *arXiv:2301.03962*，2023年。

Yanzhao Wu、Ling Liu、Zhongwei Xie、Ka-Ho Chow 和 Wenqi Wei。通过重新审视集成多样性指标提升集成准确率。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第16469–16477页，2020年。

Saining Xie、Ross Girshick、Piotr Dollár、Zhuowen Tu 和 Kaiming He。用于深度神经网络的聚合残差变换。载于*IEEE*会刊计算机视觉与模式识别，第1492–1500页，2017年。

Sergey Zagoruyko 和 Nikos Komodakis。宽残差网络。*arXiv* 预印本 *arXiv:1605.07146*，2016。

Hang Zhang、Chongruo Wu、Zhongyue Zhang、Yi Zhu、Haibin Lin、Zhi Zhang、Yue Sun、Tong He、Jonas Mueller、R Manmatha 等。ResNeSt：分裂注意力网络。发表于 *IEEE/CVF*计算机视觉与模式识别会议论文集，第 2736–2746 页，2022年。

Hongyi Zhang、Moustapha Cisse、Yann N Dauphin 和 David Lopez-Paz。mixup：超越经验风险最小化。*arXiv* 预印本 *arXiv:1710.09412*，2017年。

Shuai Zhao、Liguang Zhou、Wenxiao Wang、Deng Cai、Tin Lun Lam 和 Yangsheng Xu。迈向更优的准确率-效率权衡：分而共训练。*IEEE*图像处理汇刊，2022年。

Zhi-Hua Zhou。 集成方法：基础与算法。CRC出版社，2012年。

A 计算资源

我们的大多数实验都在基于云的计算资源（AWS 和 GCP）上进行。图中使用多样性正则项训练的集成见 2、3、6、12、14、15 和17 至22 ，这些模型使用 P3 系列的 EC2 实例进行训练。我们使用 `p3.xlarge` 实例训练模型，除非遇到 GPU 内存问题。在这种情况下，我们会使用 `p3.8xlarge` 实例训练模型。据估计，CIFAR10 集成的大型模型训练平均需要 3 小时计算时间，CIFAR100 集成需要 5 小时，TinyImageNet 集成需要 8 小时。我们估计，所有数据集上的小型模型训练每个集成都需要 30 分钟。因此，我们估计这些实验总共需要 1250 小时的计算时间。

此外，图中采用多样性正则项训练的集成8 和表1 使用 4 台 NVIDIA V100 计算实例进行训练。我们估计这些实验额外需要约 700 个计算小时。
实验结果分析和可视化在 M1 MacBook Ai 上完成 rs.

B 模型训练

B.1 代码可用性。

用于训练深度集成的代码见此处： https://github.com/cellistigs/ensemble_attention/tree/dkl。用于数据分析以及训练非深度神经网络集成的其他代码见此处：
https://github.com/cellistigs/interp_ensembles。除非另有说明，集成训练需要 GPU。对于采用多样性机制的集成，所有模型训练均通过一个 `run.py` 脚本进行管理，该脚本从文件 `ensemble_attention/configs/run_default_gpu.yaml`读取配置参数。其他参数，例如学习率计划、集成成员数量，或是否在分布外数据上进行评估，都可以通过此配置文件进行管理。

分类 器集成与多样性机制可以通过运行以下命令进行训练：

```
python ensemble_attention/scripts/run.py ++classifier={classifier_name}
++gamma={gamma_value} ++module={ensemble_dkl,ensemble_js_avg,ensemble_js_unif,ensemble_p2b}
```

提供的 `module` 参数将分别训练詹森间隙、JSD 1 对全部、均匀 JSD 和方差正则化器，训练完成后将保存集成预测。可用于训练的分类器列在 `ensemble_attention/src/ensemble_attention/module.py`中的字典`all classifiers`内。_

MLP 集成与多样性机制可以通过运行以下命 d:

```
python ensemble_attention/scripts/run.py ++classifier={classifier_name}
++gamma={gamma_value} ++module=tabular_ensemble_jgap
```

带有多样性机制的回归集成可以通过以下方式训练：

```
python ensemble_attention/scripts/run.py ++classifier={classifier_name}
++gamma={gamma_value} ++module=casresgress_ensemble_dkl
```

同质深度集成可以通过以下命令进行训练：

```
python ensemble_attention/scripts/run.py ++classifier={classifier_name}  ++gamma=0
++module=ensemble
```

异质深度集成 包含独立训练的组件模型。可以使用以下命令进行训练：

```
python ensemble_attention/scripts/run.py ++classifier={classifier_name}  ++gamma=0
++module=base
```

宽度/深度逐渐增加的集成残差网络可以使用以下命令进行训练：

```
python interp_ensembles/scripts/run.py ++classifier={classifier_name}            ++gamma=0
++module=ensemble
```

所有分类器的列表可在 `interp_ensembles/src/interpensembles/module.py`的字典 `all_classifiers`中找到。

随机森林可通过以下命令进行训练：

```
python interp_ensembles/scripts/trees/rf_large.py
```

此脚本可在 M1 MacBook 上运行，并会保存所有深度的树的概率输出及对应的测试数据标签。

随机特征模型的集成可通过以下命令进行训练：

```
python interp_ensembles/scripts/random_feature_regression.py
```

此脚本可在 M1 Macbook 上运行，并会将每种宽度下训练的集成模型的标签、logits 和概率输出分别保存在不同目录中。

B.2 多样性正则化实验。

对于所有计算资源允许的多样性正则化实验，我们计算了集成准确率均值的两个标准误，并量化了采用多样性机制训练的集成相对于该范围的性能是更高、更低还是相当。我们在正文和表 3 至 6中报告这些量化结果。

B.2.1 图 2中的模型

对于给定的数据集，图 2、14、15 和 17以及表 1中展示的所有大容量集成和单个网络，为简化起见，均采用相同的训练计划和超参数进行训练。

对于 CIFAR10，我们选择训练 100 轮，批量大小为 256，基础学习率为 $1e-2$，权重衰减为$1e-2$。使用的优化器是带动量的随机梯度下降，并采用 30 轮线性预热和余弦衰减。训练结束后，选择验证数据上性能最佳的检查点用于后续评估。

对于 CIFAR100，我们训练了 160 轮，批量大小为 128，基础学习率为 $1e-1$，权重衰减为$5e-4$。使用的优化器是带动量的随机梯度下降，并在第 60 和第 120 轮将学习率降低至原来的十分之一。训练结束后，选择验证数据上性能最佳的检查点用于后续评估。

对于 TinyImageNet，我们使用了与 CIFAR100 相同的训练超参数，只是训练持续了 200 轮。

对于 ImageNet 模型，我们训练了 90 轮，批量大小为 256，基础学习率为 $1e-1$，权重衰减为 $1e-4$。使用的优化器是带动量的随机梯度下降，学习率调度设定为 $0.1^{(m/30)}$，其中 m 是轮次索引。

在所有实验中，我们创建了大小为（$M=4$）的集成。图 2 中的 SEM 带由（$N=4$）个独立训练的集成或样本生成。对于 CIFAR10，我们独立训练了一组新的集成和单个模型作为基线；图 2 中的集成和单个模型之间没有复用模型；大小为 $M=4$ 的集成则通过从一组由 5 个独立训练模型组成的集合中选取部分模型来创建。对于 CIFAR100 和 TinyImageNet，以及图14 和15，我们通过考察不同正则化器下训练的所有 $\gamma=0$ 个随机种子来创建集成 SEM 带。由于计算限制，我们没有为 ImageNet 创建这些带。

小容量模型采用与图中大容量模型相同的损失函数和训练计划进行训练。2，但在这组实验中，我们选择对所有模型和数据集都保留 CIFAR10 的学习率和超参数。

图 B.2.2 中的模型 7（表格数据集上的多层感知机）

为了为 covertype 数据集找到一组合适的模型超参数，我们首先使用 optuna，按照 Gorishniy 等人 (2022年) 中的参数范围建议，对多层感知机的层数、层大小、学习率和权重衰减进行超参数优化。随后，我们固定了此次超参数搜索得到的学习率和权重衰减（$5e-4$，$2e-5$），并考虑了 4 种不同的模型架构：**更小型**（2 层，每层 32 个神经元）、**小型**（2 层，每层 64 个神经元）、**大型**（8 层，每层 512 个神经元）、**更大型**（8 层，每层 1024 个神经元）。模型使用 AdamW 优化器训练，最多训练 100 轮，并采用提前停止。

图 B.2.3 中的模型 8（使用均方误差损失函数训练的模型的多样性机制）

本节中的模型按照 Hui 和 Belkin （2020年）；Demirkaya 等人 （2020年）的方法，使用均方误差（MSE）损失函数而非 softmax 交叉熵损失函数进行训练。集成采用 WideResNet 28-10 架构（Zagoruyko 和 Komodakis，2016）和 DenseNet 161 架构（Huang 等人，2017年）在 CIFAR10 和 CIFAR100（粗粒度）数据集（Hui 和 Belkin，2020年）上进行训练。WideResNet 28-10 的实现采用了 Zagoruyko 和 Komodakis （2016）中的开源 PyTorch 实现。

在 CIFAR100 上训练的模型使用粗粒度标签进行训练，类别数为 20 而非 100，因为在这种设置下，深度神经网络在类别数较多的数据集上的性能不佳 Hui 和 Belkin （2020年）；Demirkaya 等人 （2020年）。

B.3 图 5 和图 6 中的标准集成模型

MNIST 模型。我们使用 Scikit-learn（Pedregosa 等人，2011年）中提供的标准实现，在 MNIST 上拟合随机森林，并将单棵树抽取的特征最大比例设为0。7；其他所有参数均采用默认值。为获得非零概率估计，我们在所有概率输出中加上一个较小的常数 $\epsilon = 1^{-10}$，并对概率重新归一化。

此外，我们也在 MNIST 上训练随机特征模型。我们构建随机傅里叶特征模型，将任意数据向量 $\vec{x} \in \mathbb{R}^N$ 映射为 $\cos(\mathbf{W}\vec{x} + b)$，其中 $\mathbf{W} \in \mathbb{R}^{D\times N}$，其中 D 是模型的宽度。这里 $\mathbf{W} \sim \mathcal{N}(0, \frac{1}{D})$ 且 $b \sim \text{Unif}[0, 2\pi]$。我们使用 Scikit-learn（Pedregosa 等人，2011年）中提供的标准实现，通过逻辑回归拟合所得特征，并结合 `BaggingClassifer` 生成装袋集成。

CIFAR 模型。用于生成这些图的模型包括三种集成类型的子集。第一个子集由使用与附录B.2.1中相同的超参数训练的模型组成。这些模型的大小均为 $M = 5$，架构如下：

- ResNet 18 He 等人(2016)
- WideResNet 18-2、18-4、28-10Zagoruyko 和 Komodakis(2016)
- GoogleNet、Inception v3Szegedy 等人(2015年)
- 具有 11 层和 19 层的 VGGSimonyan 和 Zisserman(2014)
- DenseNet 121 和 169Huang 等人(2017年)

第二个子集由从随以下论文发布的代码中衍生而来的模型实现组成：Zhao 等人(2022年)，代码见此处：`https://github.com/FreeformRobotics/Divide-and-Co-training`。这些较大模型的集成采用以下架构：

- 金字塔网络 110（M=4）Han 等人(2017年)
- 挤压-激励残差网络 164 (M=5) Hu 等 (2018年)

- ShakeShake 262x96d(M=5）Gastaldi （2017年）
- ResNexst50 4x16d (M=5)Xie 等人 (2017年); Zhang 等人 (2022年); Zhao 等人 (2022年)

For 对于这些模型，我们将学习率改为 $1e-1$，并将权重衰减改为 $1e-4$.

最后，我们使用了来自 Miller 等人 (2021年) 的 36 组独立集成，共包含 137 个单个模型。我们感谢作者慷慨地与我们分享这些结果。总而言之，我们的结果涵盖了 206 个单独的深度网络。

为了生成异质集成，我们首先记录了这些独立集合集合中的集成规模分布。然后，我们将所有单个模型随机打乱，并将它们划分为新的深度集成，同时遵循原始同质集合集合中的集成规模分布。我们对集成成员进行 10 次随机打乱，计算所有感兴趣的度量指标，并在图 6 中以橙色圆点绘出这些随机打乱的结果范围。

ImageNet 模型。这 98 个在 ImageNet 上训练的模型来自Abe 等人(2022b)；Taori 等人(2020年，这些模型均可公开获取。正文中还提供了如何基于这些模型构建异质集成的更多实验细节。

B.4 图6中宽度/深度逐渐增加的 ResNet

在图6右侧的面板中，我们考察了宽度或深度逐渐增加的 ResNet 模型集成，并分析这些集成在预测多样性和单模型平均性能方面的提升。

增加宽度。对于宽度逐渐增加的模型，我们采用来自He 等人(2016) 的“ImageNet 风格”ResNet 18 架构作为基础架构，并将其应用于 CIFAR10（参见 https://github.com/huyvnphan/PyTorch_CIFAR10 ，其中提供了实现）。相较于这一标准实现，我们将每一层的通道数乘以因子$\{1\,16, 1\,8, 1\,4, 1\,2, 1, 2, 4\}$ 以生成宽度各异的网络。模型采用与附录B.2.1中相同的超参数进行训练。

增加深度。对于深度不断增加的模型，我们考虑He 等人(2016描述的“CIFAR10 风格”ResNet 架构，该架构包含三种滤波器尺寸，每种滤波器尺寸对应 $2n$ 层，总计 $2n+2$ 层。在这些实验中，我们考虑 $n=\{1, 2, 3, 4, 5, 6, 7\}$，并且不随深度改变训练超参数：我们训练 160 个训练轮数，批量大小为 128，基础学习率为 $1e-1$，权重衰减为 $5e-4$，并采用学习率调度，在训练到第 60 和第 120 轮时将学习率除以 10。

C 在训练期间鼓励多样性，相较于简单的深度集成只能带来微小收益

在表2中，我们对学界提出的各种集成方法进行了更详细的总结。我们对比依赖训练（集成成员经过明确训练以实现多样化）与独立训练（训练期间每个集成成员都不了解其他集成成员）。正如正文所述，某些独立训练方法可以带来“免费多样性”（例如数据集多样化集成）；而大多数依赖训练机制对准确率的提升有限，不过它们可以改善其他受关注的度量指标，例如期望校准误差（ECE），或对抗攻击下的鲁棒性。我们还对比了两类依赖训练机制：一类是在训练集上明确鼓励多样性，另一类则假设可以使用某个辅助数据集（额外数据多样性）。

D 附加实验

在正文中，我们研究了多样性机制对卷积数据集上小容量与大容量分类集成的一个子集的影响，并以标准测试误差作为衡量指标。本节将介绍其他实验的证据，这些证据同样支持正文中的结论。

方法	依赖训练		独立	优于基线?
	训练集多样性	额外数据多样性	训练	
简单集成Lakshminarayanan 等人 (2017年)			x	基线
专家集成 Mustafa 等人 (2020年)			x	是（但这是相对于微调基线而言）
通过预训练构建的集成 Agostinelli 等人 (2022年)			x	是（通过优化预训练性能）
数据集多样化集成 Gontijo-Lopes 等人 (2022年)			x	是（但这是通过采用不同的训练方法/增加训练数据实现的）。
JSD 集成 Mishtal 和 Arel (2012年)	x			否（训练/推理时间增益）
使用 MCL 损失函数的 TreeNets Lee 等人（2015年）	x			否（训练/推理时间增益）
DivTrainKariyappa 和 Qureshi（2019年）	x			是（考虑特定对抗攻击时）
实证研究* Webb 等人（2020年）	x			No
批量集成 Wen 等人 (2020年)	x			无（训练/推理时间增益）
MIMOHavasi 等人 (2021年)	x			无（推理时间增益）
分而共训练 Zhao 等人 (2022年)	x			是，准确率 $+<1\%$
FiLM 集成Turkoglu 等人（2022年）	x			没有（训练/推理时间增益）
实证研究*Ortega 等人（2022年）	x			No
F_{sum} Gong 等人（2022年）	x			是（期望校准误差改进 1）。
多样性损失Teney 等人（2022年）	x			是，准确率为 $+<2\%$
LITRoss 等人(2020年)		x		是，9/10 个数据集的准确率为 $+<1\%$，1/10 个数据集的准确率为 $+<3\%$
DICERame 和 Cord(2021年)		x		是，准确率为 $+2\%$
ERDTifrea 等人(2022年)		x		是（用于新颖性检测）
D-BATPagliardini 等人(2022年)		x		是，5/6 的数据集准确率为 $+1\%$，1/6 的数据集准确率为 $<10\%$
DivDis Lee 等人(2022年)		x		是（针对最差群体准确率）

表2：**难以提升基线深度集成的性能**。在不同的训练方法中，我们发现，除非假设可以通过某个额外数据源获得更多多样性，否则很难改进标准深度集成方法（Lakshminarayanan 等人、2017年）。

D.1 多层感知机

图7 将正文中的研究从在视觉分类任务上训练的卷积神经网络扩展到在表格数据上训练的多层感知机。使用 covertype（Blackard，1998年）数据集和詹森间隙正则化器，我们观察到，在这种设置下，多样性机制的影响与正文中的结果一致。具体而言，对于大型多层感知机架构（图7 上排，“更大的多层感知机”），在所考察的多样性促进强度范围及性能指标（准确率、负对数似然、布里尔分数）内，多样性促进始终会导致性能变差。然而，当我们考察稍小一些的架构（图7 第二排，“大型多层感知机”）时，会发现多样性促进强度存在一个局部最优点：相较于基线集成训练，性能似乎有所提升，但随着我们进一步促进预测多样性，性能随后会下降至标准集成训练水平以下。如果再考察更小的架构（图7 第三排，“小型多层感知机”），就会发现，随着正则化强度变化，促进预测多样性会使负对数似然和布里尔分数先达到局部最大值，之后随着正则化强度进一步增大而改善，因此呈现出与“大型多层感知机”相反的行为。准确率并未呈现同样的对应关系；在所有多样性促进水平下，“小型多层感知机”的方差似乎都非常高。最后，对于我们考察的最小多层感知机架构（图7 最下排，“更小的多层感知机”），我们观察到了正文中讨论的小型卷积架构的性能：多样性促进能够一致地提升集成性能。神经网络性能中存在这些局部最优点，让人联想到先前的研究（Webb 等人，2020年），该研究在通过平均 logits 构建集成时发现了类似现象。这些局部最优点可能与我们所关注的卷积架构／视觉数据集设置中的离群值集成行为有关（见表3 和4）。进一步理解此类正则化强度最优点如何随模型容量和集成聚合方法而变化，将是未来研究的重要方向。

D.2 回归深度集成

图 8 表明，在回归集成中促进多样性的效果，与小规模分类集成中的效果相似。我们绘制了在 CIFAR10 和 CIFAR100 粗粒度数据集上训练的 WideResNet 28-10 集成（左）和 DenseNet 161 集成（右）的预测性能；这些模型均使用均方误差损失（而非交叉熵）进行训练，并通过詹森间隙正则化器促进或抑制多样性（其他正则化器要么会退化为詹森间隙（附录 F.2），要么是专为分类设计的）。与图 2不同，在通过多样性促进进行训练的集成中，多数（12/16）优于基线性能，且性能提升与正则化强度成正比。在

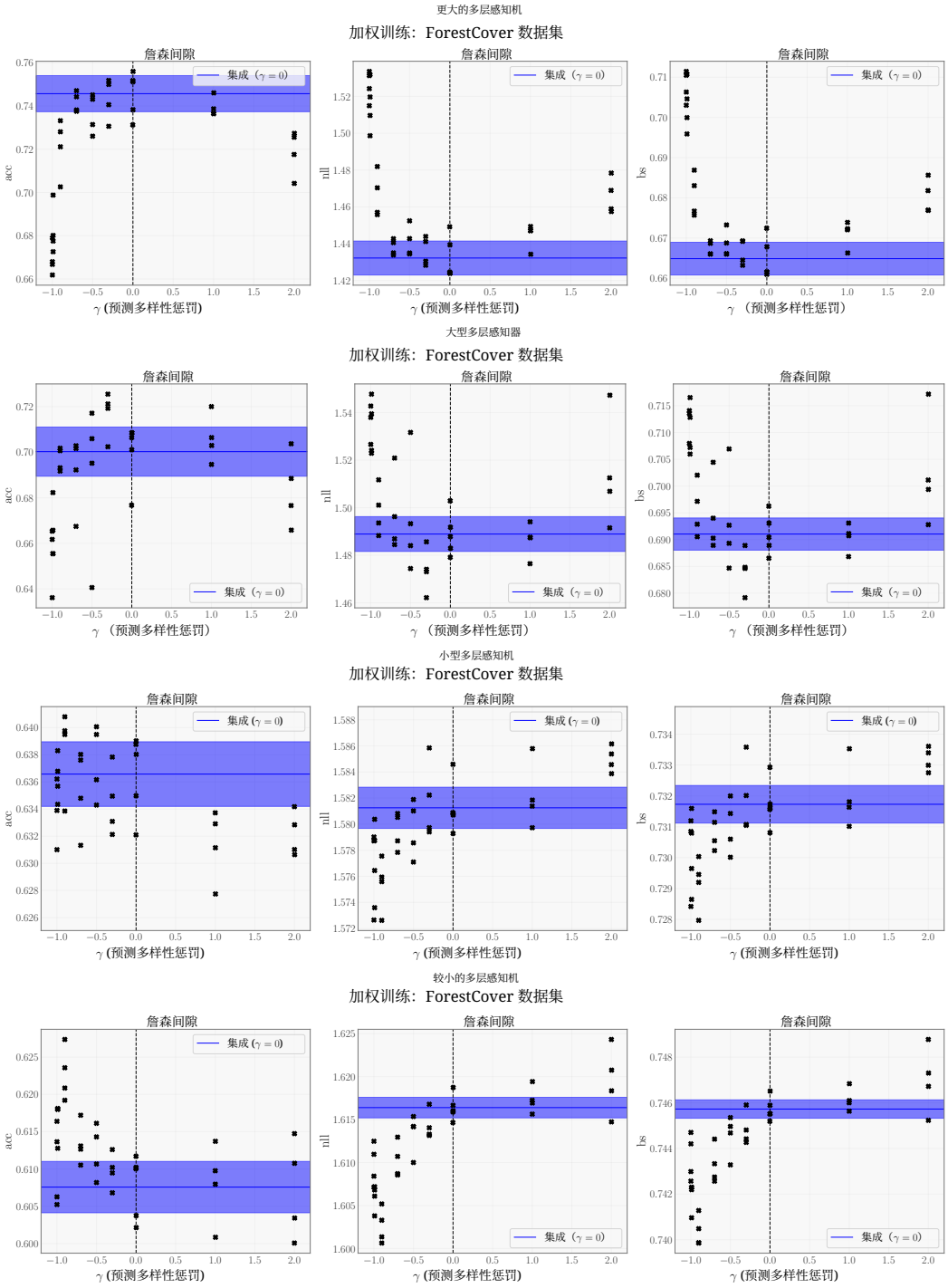


图7：**多层感知机在表格数据集上训练时，多样性机制的效果取决于模型容量**。我们研究了在 covertype（Blackard，1998年）表格数据集上，对不同规模的多层感知机进行聚合的集成，并使用詹森间隙正则化器促进或抑制多样性。对于全部集成，我们报告预测准确率（**左列**）、负对数似然（**中列**）和布里尔分数（**右列**）（坐标轴与图2相同）。从我们考察的最大多层感知机集成（**顶行**）到最小的集成（**底行**），可以看到多样性促进的影响发生了逆转：促进多样性分别有助于较小的模型集成、损害较大的模型集成。详情请参阅附录D.1 。

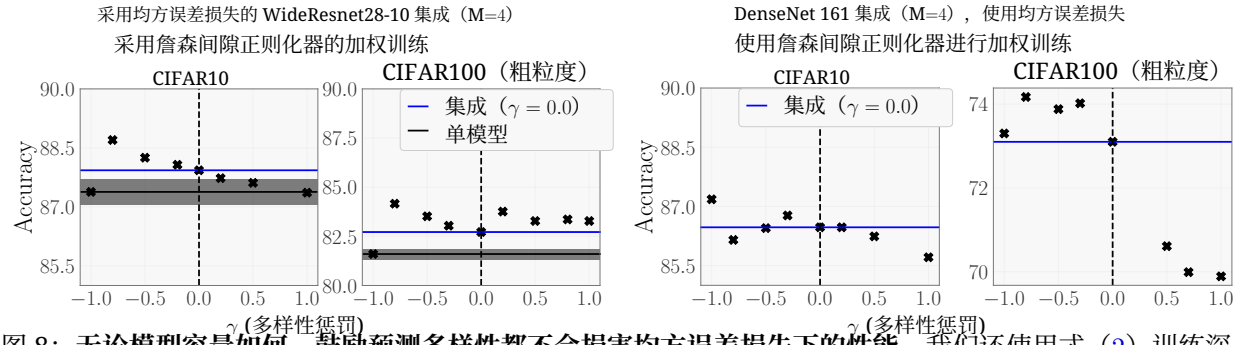


图 8：**无论模型容量如何，鼓励预测多样性都不会损害均方误差损失下的性能**。我们还使用式（2）训练深度回归集成，采用均方误差损失，在 CIFAR10 和 CIFAR100（粗粒度标签）上进行训练。正如我们的假设所预测的那样，在回归任务中鼓励多样性不会损害预测性能。

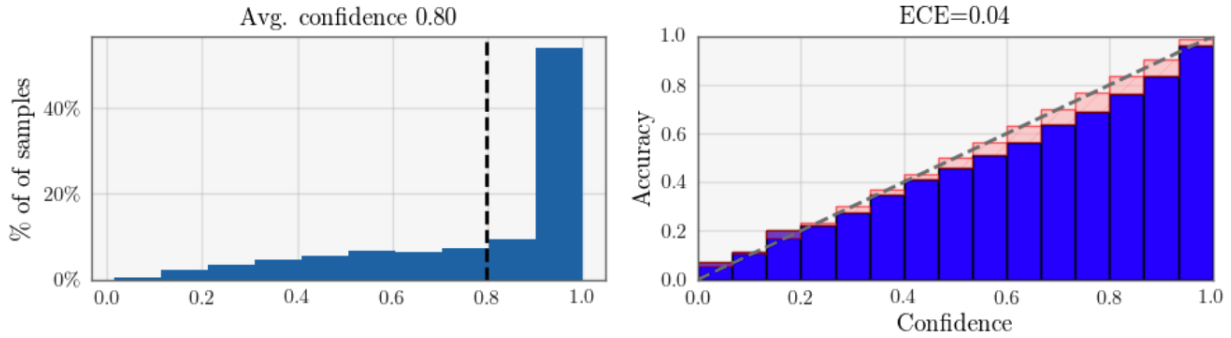


图 9：**标准模型能在 ImageNet 上做出置信度高且准确的预测**。在 ImageNet 上，由 ResNet 50 模型组成的集成，其验证集预测中有多数（~55%）的置信度介于 0.9 和 1 之间，其中 ~95% 是正确预测。这些结果表明，即使训练误差尚未降至 0%，大容量模型也已经容易受到观察 2 的影响。

相比之下，抑制预测多样性的效果似乎因数据集而异：在 CIFAR10 数据上会导致性能普遍变差，但在 CIFAR100 上则能带来持续改进，与大型模型分类集成的情况相同。

D.3 ImageNet 预测

尽管计算资源限制使我们无法在 ImageNet 上开展大规模的多样性正则化实验，但我们提供了初步证据，表明我们预期目前流行的模型在表 1 之外也会呈现相同的趋势。在图 9 中，我们考察了 ResNet 50 模型在 ImageNet 验证集上的预测，并观察到该模型已有很大比例的预测具有高置信度，其中包括许多高度准确的预测。单模型预测的这些特征表明，在 ImageNet 这类大型数据集上训练的标准大型模型，也很可能容易受到我们在表1 中观察到的病态现象的影响。

D.4 标签平滑

在图 2 中，我们展示了在训练期间鼓励模型在所有数据点上保持多样性，会对高容量模型造成不利影响。我们认为，在所有数据点上（而不仅仅是在错误预测上）鼓励多样性，对高容量模型尤其有害，因为这类模型往往会给出高置信度预测，无论预测正确与否。因此，鼓励模型对正确预测保持多样性所造成的损失，超过了鼓励模型对错误预测保持多样性带来的任何收益。由此引出的一个自然问题是，鼓励多样性

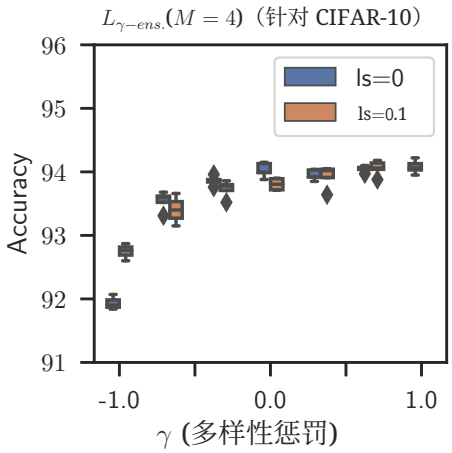


图10：**标签平滑减轻了多样性促进的影响，但保持了相同的趋势**。我们以多样性正则化深度集成为基线（$ls = 0$，蓝色标记），考察标签平滑的影响 $ls = 0.1$。尽管更高程度的标签平滑减轻了性能差距，但预测多样性增加时性能损失的整体趋势依然存在。

如果预测的置信度较低，那么多样性带来的负面影响是否会较小？图 2中的低容量模型就是这种情况。为检验这一假设，我们使用标签平滑（Müller 等人，2019年）训练集成；标签平滑会在神经网络训练期间降低预测的置信度。图 10 显示，在训练集成时通过标签平滑控制置信度，并未显著改变结果中的性能。未来的工作可以探究，更强的标签平滑或其他降低神经网络置信度的技术（例如 mixup（Zhang 等人，2017）或温度缩放（Guo 等人，2017）是否能够带来显著变化。

D.5 CIFAR10 分布外（OOD）数据

除了分布内测试性能之外，深度集成还被视为一种能够保证提升性能的方法，尤其是在分布外（OOD）数据场景下，此时测试数据会显著偏离训练分布（Gustafsson 等人，2020年；Lakshminarayanan 等人，2017年；Ovadia 等人，2019年）。我们在图2中评估了所有使用 CIFAR10 训练的集成和单个模型在 CIFAR10.1 上的表现；CIFAR10.1 是针对使用 CIFAR10 训练的模型的分布偏移测试数据集（Recht 等人，2018年）。

图中的结果16 与图中分布内数据的结果得出了相同结论2：相较于 $0 > \gamma > 1$相对于 $\gamma = 0$ 未经修改的训练，较低的多样性惩罚会导致集成性能大幅下降，而提高惩罚则可能提升性能。这一结果支持了Abe 等人(2022年b的结论：OOD 集成性能与分布内性能密切相关。

E 常见损失函数的詹森间隙

本节将明确推导詹森间隙如何在各种损失函数下衡量预测多样性 ℓ。

E.1 均方误差詹森间隙

均方误差损失的詹森间隙是样本方差的一个缩放版本。下面推导这一结论：

$$
\begin{aligned}
&\tfrac{1}{M}\textstyle\sum_{i=1}^{M}\left(\boldsymbol{f}_i(\boldsymbol{x})-y\right)^2-\left(\tfrac{1}{M}\textstyle\sum_{i=1}^{M}\boldsymbol{f}(\boldsymbol{x})-y\right)^2\\
&=\tfrac{1}{M}\textstyle\sum_{i=1}^{M}\left[(\boldsymbol{f}_i(\boldsymbol{x}))^2-2(\boldsymbol{f}_i(\boldsymbol{x}))y+y^2\right]-\left[(\bar{\boldsymbol{f}}(\boldsymbol{x}))^2-2(\bar{\boldsymbol{f}}(\boldsymbol{x}))y+y^2\right]\\
&=\tfrac{1}{M}\textstyle\sum_{i=1}^{M}\left[(\boldsymbol{f}_i(\boldsymbol{x}))^2-2(\boldsymbol{f}_i(\boldsymbol{x}))y\right]-\left[(\bar{\boldsymbol{f}}(\boldsymbol{x}))^2-2(\bar{\boldsymbol{f}}(\boldsymbol{x}))y\right]\\
&=\tfrac{1}{M}\textstyle\sum_{i=1}^{M}(\boldsymbol{f}_i(\boldsymbol{x}))^2-\tfrac{1}{M}\textstyle\sum_{i=1}^{M}2(\boldsymbol{f}_i(\boldsymbol{x}))y-(\bar{\boldsymbol{f}}(\boldsymbol{x}))^2+2(\bar{\boldsymbol{f}}(\boldsymbol{x}))y\\
&=\frac{1}{M}\sum_{i=1}^{M}\left[(\boldsymbol{f}_i(\boldsymbol{x}))^2-(\bar{\boldsymbol{f}}(\boldsymbol{x}))^2\right]\\
&=\frac{M-1}{M}\left[\frac{1}{M-1}\sum_{i=1}^{M}\left[(\boldsymbol{f}_i(\boldsymbol{x}))^2-(\bar{\boldsymbol{f}}(\boldsymbol{x}))^2\right]\right]\\
&=\frac{M-1}{M}\,\overline{\mathrm{Var}}\left[\boldsymbol{f}_i(\boldsymbol{x})(\boldsymbol{x})\right],
\end{aligned}
$$

其中，$\bar{\boldsymbol{f}}(\boldsymbol{x})=\frac{1}{M}\sum_{i=1}^{M}\boldsymbol{f}_i(\boldsymbol{x})$ 是样本均值，Var 是样本方差。

E.2 交叉熵詹森间隙的分解

在本文中，我们使用交叉熵损失的詹森间隙作为测试数据上多样性的度量，同时将其用作预测多样性的正则化器。这里我们表明，詹森间隙是对多样性进行信息论量化的一种方式。詹森间隙定义如下：

$$
\begin{aligned}
&\frac{1}{M}\sum_{i=1}^{M}\left[\mathrm{CE}(\boldsymbol{f}_i(\boldsymbol{x}),\ y)\right]-\left[\mathrm{CE}(\bar{\boldsymbol{f}}(\boldsymbol{x}),\ y)\right]\\
&=\frac{1}{M}\sum_{i=1}^{M}\left[-\log\boldsymbol{f}_i(\boldsymbol{x})^{(y)}\right]-\left[-\log\bar{\boldsymbol{f}}(\boldsymbol{x})^{(y)}\right]\\
&=\frac{1}{M}\sum_{i=1}^{M}\left[-\log\boldsymbol{f}_i(\boldsymbol{x})^{(y)}\right]+\log\left[\frac{1}{M}\right]+\log\left[\sum_{j=1}^{M}\boldsymbol{f}_j(\boldsymbol{x})^{(y)}\right]\\
&=\sum_{i=1}^{M}\frac{1}{M}\left[\log\frac{1}{M}-\log\left[\frac{\boldsymbol{f}_i(\boldsymbol{x})^{(y)}}{\sum_{j=1}^{M}\boldsymbol{f}_j(\boldsymbol{x})^{(y)}}\right]\right]
\end{aligned}
\tag{9}
$$

公式（9）可解释为两个类别分布之间的 KL散度。第一个分布表示均匀随机抽取一个集成成员的概率（$1/M$）。第二个分布表示按其正确类别预测成比例抽取一个集成成员的概率。当这两个分布相等时，公式（9）将达到最小值；而这只有在所有集成成员对正确类别的预测都相同时才会发生。反之，组件模型预测彼此差异越大，公式（9）的值就越大。

图 11 展示了公式（3）中的詹森间隙多样性与文献中使用的其他多样性度量指标之间的高度相关性，包括 KL 散度（Lee 和 Seung，2000年）和成对相关性

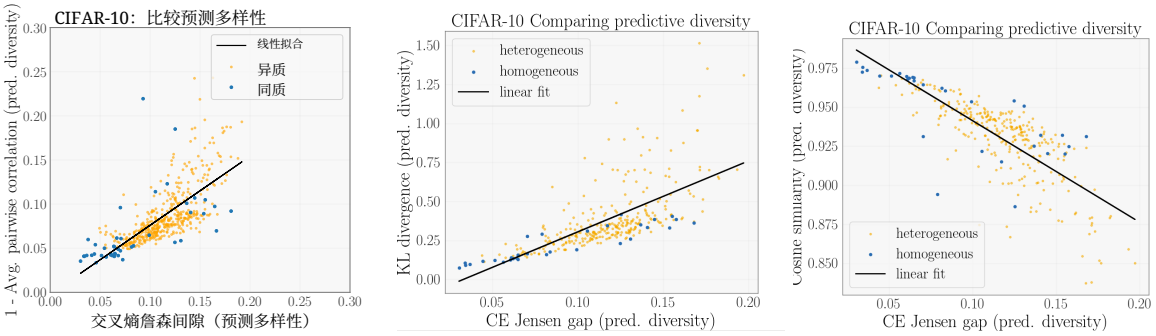


图11：**交叉熵詹森间隙与标准多样性度量指标相关**。我们将詹森间隙这一预测多样性概念（此处使用交叉熵损失）与图中所示的同质和异质 CIFAR10 集成的其他标准预测多样性度量指标进行比较。6。我们展示了交叉熵詹森间隙预测多样性与基于集成成员预测之间平均成对相关性的多样性度量（**左**）、基于 KL散度的多样性（**中**）以及基于余弦相似度的多样性（**右**）高度相关——黑线表示线性拟合。

多样性促进.....	有害	中性	有帮助
小型神经网络分类tion 集成	12/85 (14%)	7/85 (8%)	66/85 (78%)
大幅过度自信神经网络分类集成	82/134 (61%)	38/134 (29%)	14/134 （10%）

表3：**大容量深度集成对多样性促进的响应各不相同**。在大量实验设置中，我们发现，促进预测多样性对大容量深度集成与其他类型集成的影响截然相反。所列集成的分类是相对于标准集成性能的 2 倍标准误计算的。请注意，这些结果不包括多层感知机。

多样性抑制——程度.....	有害	中性	有帮助
小型神经网络分类集成	46/82 （56%）	33/82 (40%)	3/82 (4%)
大型过度自信神经网络分类集成	40/155 (26%)	78/155 (50%)	37/155 (24%)

表 4：**大容量深度集成对多样性抑制的反应有所不同**。在大量实验设置中，我们发现，抑制预测多样性对大容量深度集成与其他类型集成的影响截然相反。所列集成的分类是相对于标准集成性能的 2 个标准误计算的。请注意，这些结果不包括多层感知机。

(Kuncheva 和 Whitaker，2003年），以及余弦相似度（Djolonga 等人，2021年）。图11中的模型包含与图 6 中相同的 CIFAR10 模型。

F 使用多样性机制训练的集成的补充分析 ms

在本节中，我们提供了正文中未展示的模型架构、数据集和正则化器组合的其他复现实验结果，以及测试准确率之外的评估指标。表 3 和4 总结了我们的发现。用于生成图 2所示结果的脚本（以及该图本身）形式如下`ensemble_attention/scripts/vis_scripts/all_weights_{classifier_name}_{dataset_name}.py`。

多样性促进	有害	中性	有帮助
小型神经网络（Jensen Gap）	0/23 (0%)	6/23 (26%)	17/23 (74%)
小型神经网络（方差）	0/20 (0%)	0/20 (0%)	20/20 (100%)
小型神经网络（JSD 1 与 All）	11/23 (48%)	2/23 (9%)	10/23 (43%)
小型神经网络（JSD Uniform）	1/20 (5%)	3/20 (15%)	16/20 (80%)
大型 NN（Jensen Gap）	23/34 (68%)	8/34 (24%)	3/34 (8%)
大型神经网络（方差）	25/35 (72%)	5/35 (14%)	5/35 (14%)
大型神经网络（JSD 1 对比 All）	18/31 (58%)	12/31 (39%)	1/31 (3%)
大型神经网络（JSD 均匀分布）	17/34 (50%)	10/34 (29%)	7/34 (21%)

表 5：**各正则化器的性能（鼓励多样性）**。

F.1 分析多样性机制之间的差异

尽管第 4 节讨论的多样性机制在我们研究的模型中总体上都表现出相似的行为，但它们之间仍存在一些显著差异，或许值得进一步研究。我们将在此讨论其中一些差异。

正则化强度的有效范围。为了进一步研究这些正则化器的行为，考察正则化强度有效范围之间的差异会很有帮助。在Jeffares 等人 (2023年) 中，作者表明，对于 $\gamma < -1$ 的取值，詹森间隙损失会表现出病态行为；这一结论适用于由任何损失函数推导出的情形，而不仅限于我们在此考察的交叉熵损失和平方误差损失。同样，Mishtal 和 Arel(2012年) 从解析角度表明，对于 JSD 一对其余正则化器，只有在平方误差损失下，我们才能确保损失函数的最小值存在。 $\gamma > -4$ 他们提供了实验依据，表明这一界限同样适用于交叉熵损失。据我们所知，此前尚未报告方差正则化器或 JSD 均匀正则化器的界限；在测试这些机制时，我们也未在所测试的正则化强度范围内遇到稳定性问题。若能为所有正则化器（以及多样性抑制）推导出此类界限，将有助于理解如何更细致地比较各个正则化器，同时兼顾其正则化强度有效范围的差异。

对网络预测的影响。这些多样性机制在试图使预测的哪个方面具有多样性上也有所不同。具体而言，詹森间隙和方差正则化器都专门致力于使训练期间分配给正确预测的概率 $f_i^y(\boldsymbol{x})$ 具有多样性，而其他机制则致力于使整个预测向量 $f_i(\boldsymbol{x})$ 具有多样性。虽然在本文所述的实验中，我们没有观察到这些方法之间存在显著差异，但这是一个重要的区别，值得在未来的工作中进一步研究。

各个机制的实证效果。最后，我们将表 3 和 4 中的结果拆分为各个机制的结果，分别列于表 5 和 6中。尽管各个多样性机制对集成性能的影响效果（或无效果）存在一些差异，但我们的主要结论在所有情况下都成立。

F.2 分解 采用多样性机制训练的集成的性能分解（补充结果)

图 1 2 表明，图 3 中所述的趋势同样适用于 CIFAR100 数据集 t.

虽然这并非对集成性能的数学分解，但在图 13 中，我们还分析了预测多样性（以交叉熵詹森间隙衡量）与平均单模型准确率之间的关系，所涉及的模型为

多样性抑制-机制.....	有害	中性	有帮助
小型神经网络（Jensen Gap）	11/19（58%）	7/19（37%）	1/19（5%）
小型神经网络（方差）	19/20（95%）	1/20（5%）	0/20（0%）
小型神经网络（JSD 1 与 All）	14/23（61%）	8/23（35%）	1/23（4%）
小型神经网络（JSD Uniform）	6/20（30%）	12/20（60%）	2/20（10%）
大型 NN（詹森 Gap）	13/50（26%）	23/50（46%）	11/50（22%）
大型神经网络（方差）	13/37（35%）	17/37（46%）	7/37（19%）
大型神经网络（JSD 1 与 All）	5/31（16%）	17/31（55%）	9/31（29%）
大型神经网络（JSD 均匀分布）	9/38（24%）	18/38（47%）	11/38（29%）

表6：**每种正则化器的性能（抑制）**。

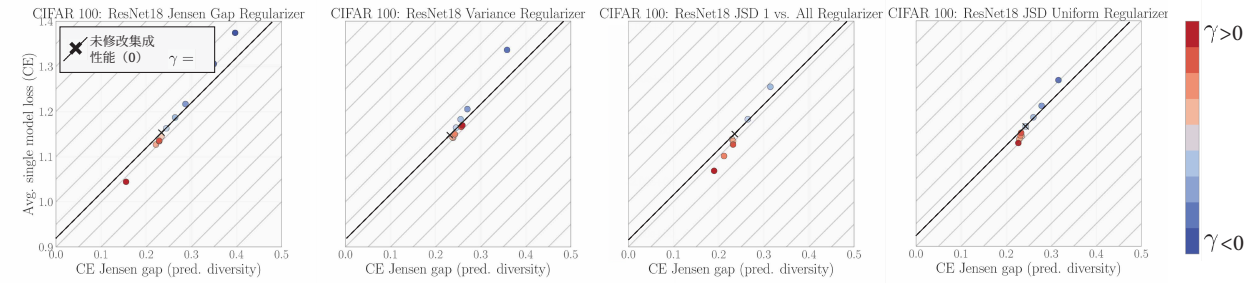


图12：**在 CIFAR100 上测量多样性正则化集成的预测多样性**。我们也将多样性正则化深度集成训练应用于 CIFAR100 上的大容量 ResNet 18 模型。惯例和结论与图3相同。

图 3中所示的相同模型。我们发现，以平均集成成员准确率衡量的集成成员性能，与图 3中给出的分解结果密切吻合。

F.3 图2和3 （针对其他大容量模型的多样性正则化）

我们证明，图2 中的结果同样适用于其他大容量架构（DenseNet 121、VGG 11、WideResNet 18、VGG 13），涵盖 CIFAR10 和 CIFAR100（图14 和15），以及分布外数据（图16）。

F.4 图2（针对其他小容量模型的多样性正则化）

我们证明，图2 中低容量模型的结果同样适用于其他在 CIFAR10 和 CIFAR100 上训练的架构，如图18 和图19所示。

F.5 校准性能指标

我们证明，图2 中的结果可推广至校准性能指标：NLL（图20）、布里尔分数（图21）和期望校准误差（图22）。

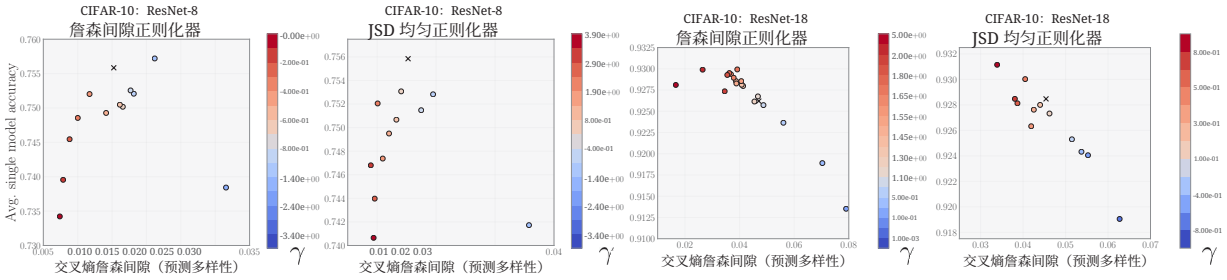


图 13： **衡量多样性正则化集成中的预测多样性和组件模型准确率**。 与图3中一样，每个标记代表一个在 CIFAR10 上通过多样性干预训练的 ResNet 8（**左侧面板**）或 ResNet 18（**右侧面板**）集成，颜色表示 γ 值，而 $\times$表示标准集成性能。我们省略了对角线水平集线，因为无法将集成性能分解为詹森间隙、预测多样性和准确率。尽管如此，我们观察到的模式与图 3相似：对于小型网络集成（**左侧**），多样性促进成功地产生了更多样化的集成，有时也会使集成成员的准确率提高；而多样性抑制则会导致性能下降。与此同时，对于大型网络集成（**右侧**），多样性促进会导致性能下降，但在所测试的各种正则化强度下，多样性抑制似乎都能产生明显更优的集成成员。

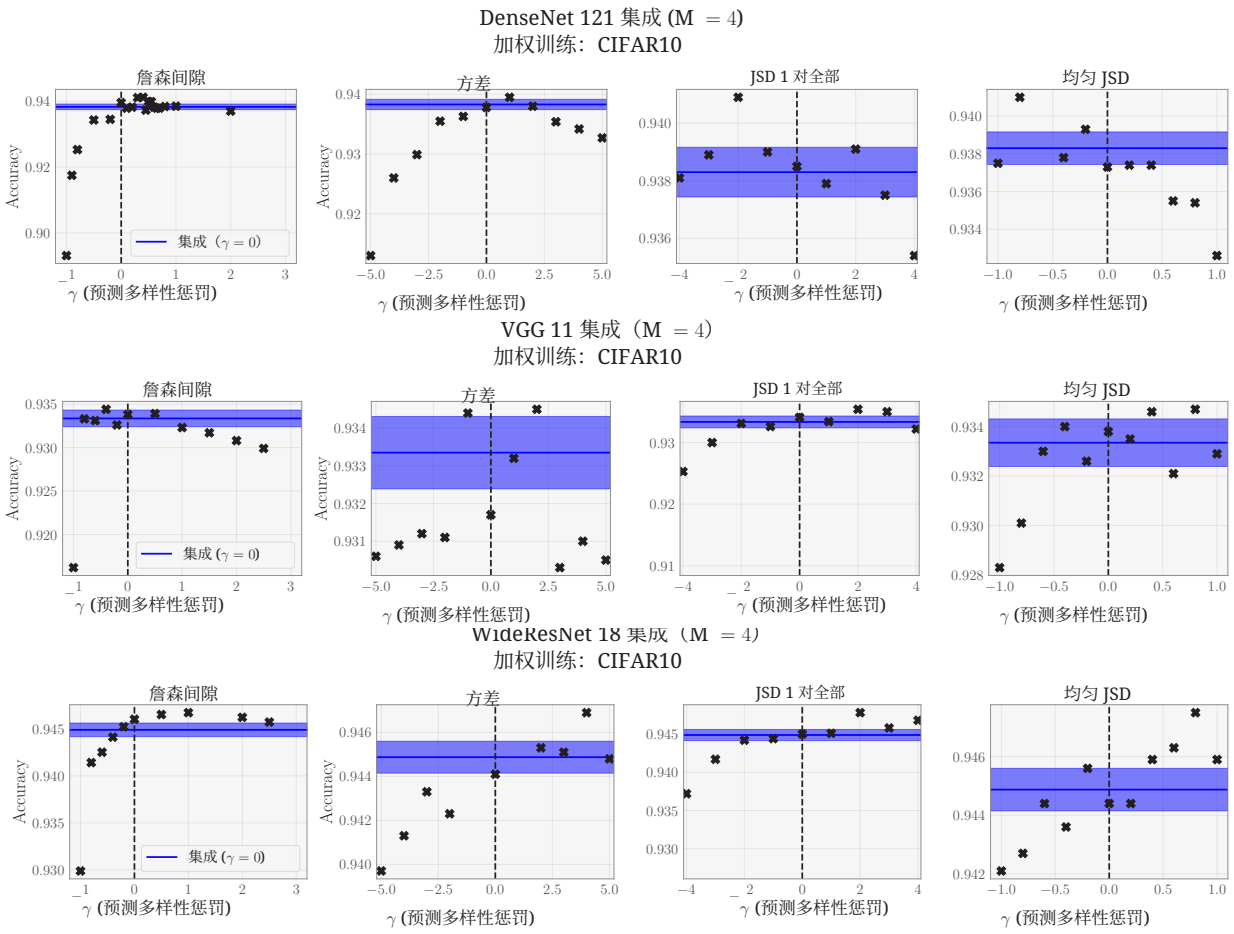


图 14： **多样性促进有害于大容量模型的深度集成准确率：CIFAR10 的其他架构**。我们将多样性正则化深度集成训练应用于 CIFAR10 上的另外 3 种架构。各行展示了 DenseNet 121、VGG 11 和 WideResNet 18 的训练。与图2中的惯例和结论相同。

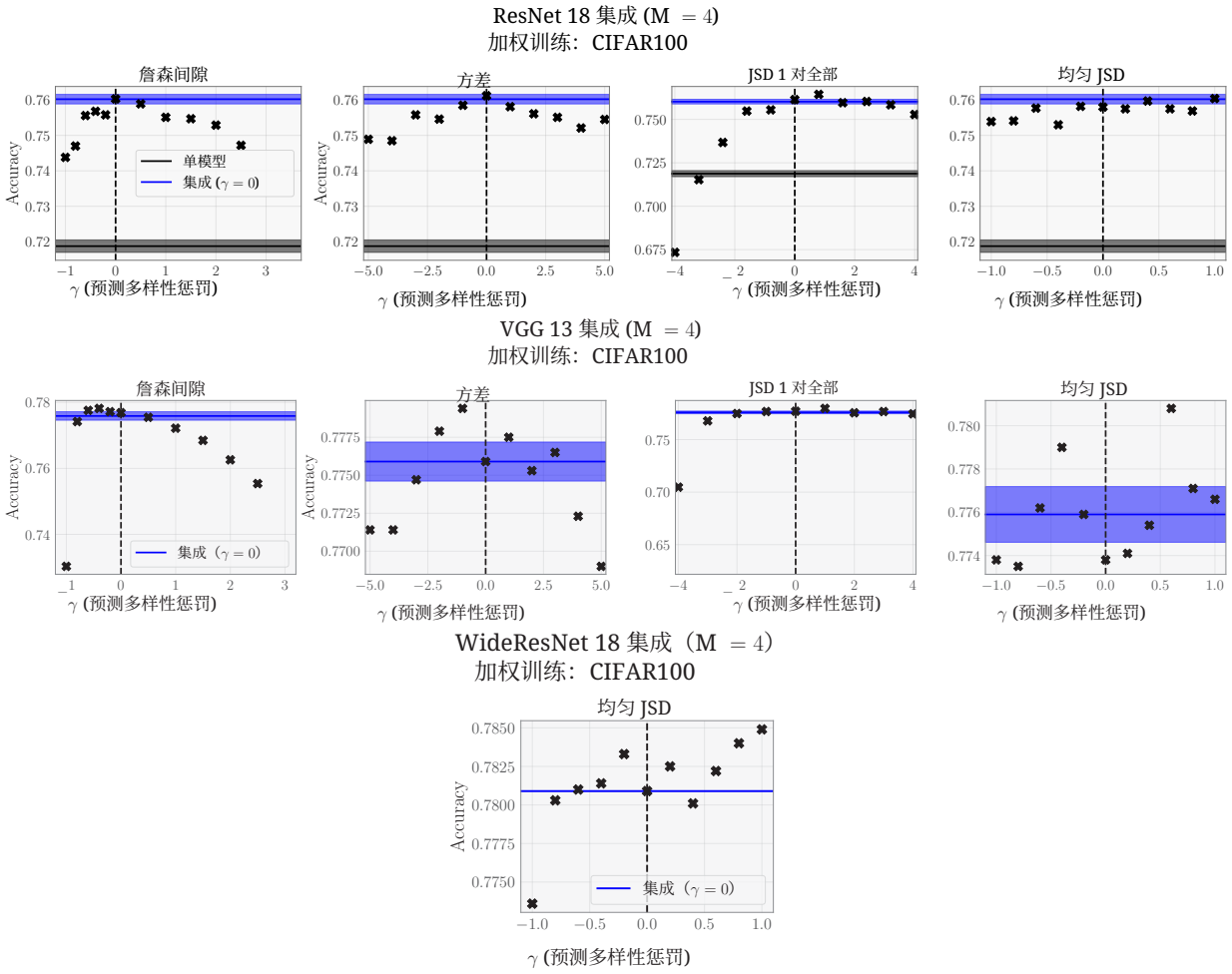


图 15： **多样性促进有害于大容量模型的深度集成准确率：CIFAR100**。我们将多样性正则化深度集成训练应用于 CIFAR100 上的另外两种架构。各行分别展示了使用全部正则化器训练 VGG 11，以及使用 JSD 均匀正则化器训练 WideResNet 18。图中的约定和结论与图 2 相同。

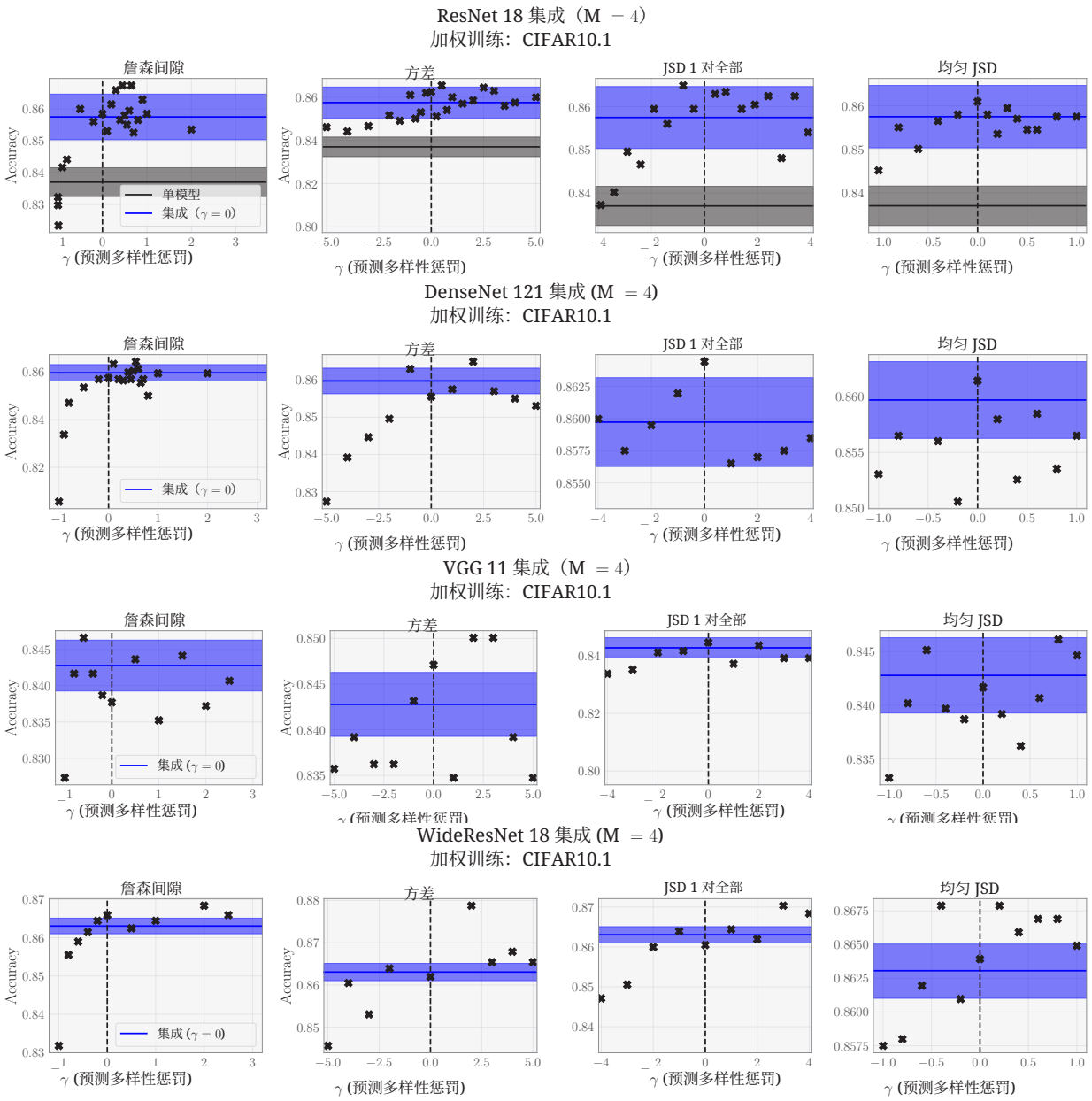


图 16： **多样性正则化在分布外数据上呈现出相同的趋势**。我们在分布外测试数据集 CIFAR10.1 上评估了在 CIFAR10 上训练的大规模多样性正则化模型（与图2相同）。预测性能与标准 CIFAR10 测试集上呈现的趋势一致。

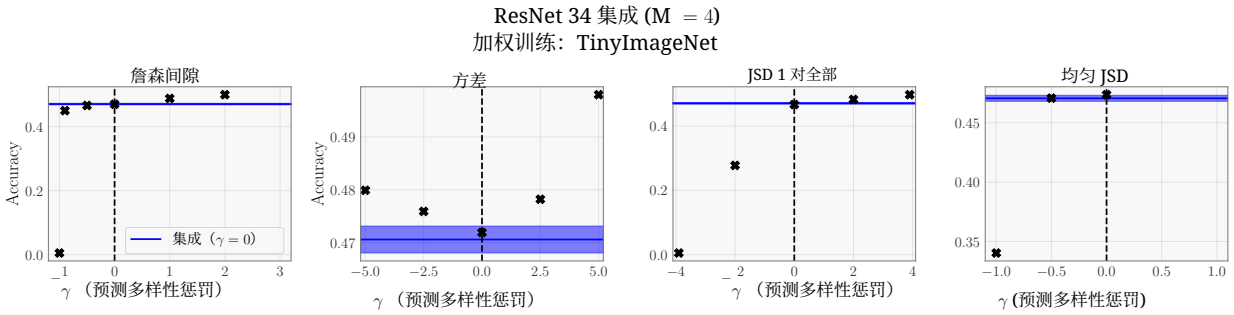


图17：**多样性促进有害于大容量模型的深度集成准确率：TinyImageNet**。我们将多样性正则化深度集成训练应用于 TinyImageNet 上的另一种架构。图中展示了采用全部正则化器对 ResNet 34 进行训练的情况。图例惯例和结论与图2相同。

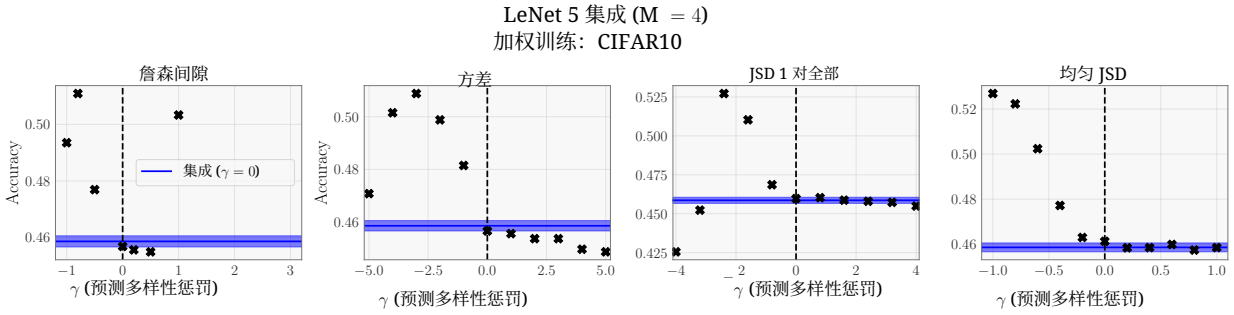


图18：**多样性促进可提升低容量模型的表现：CIFAR10 上的另一种架构**。我们将多样性正则化深度集成训练应用于 CIFAR10 上的 LeNet 5，并测量预测准确率。图例惯例和结论与图2相同。

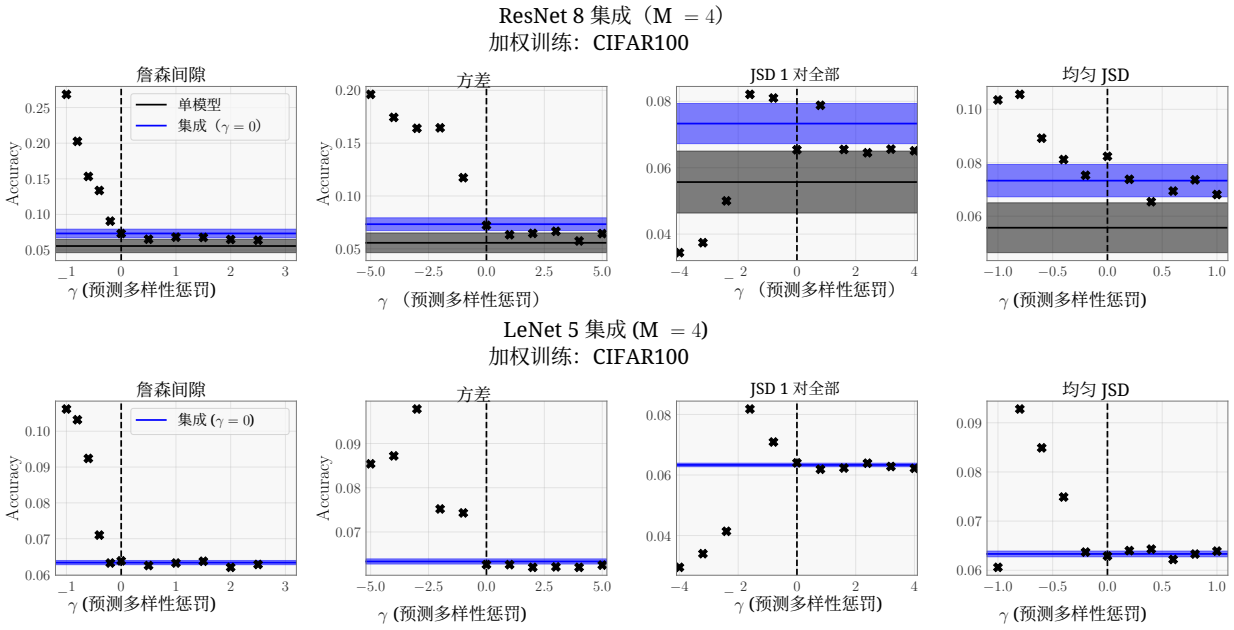


图19：**多样性促进可提升低容量模型的表现：ResNet 8 以及 CIFAR100 上的另一种架构**。图例惯例和结论与图2相同。

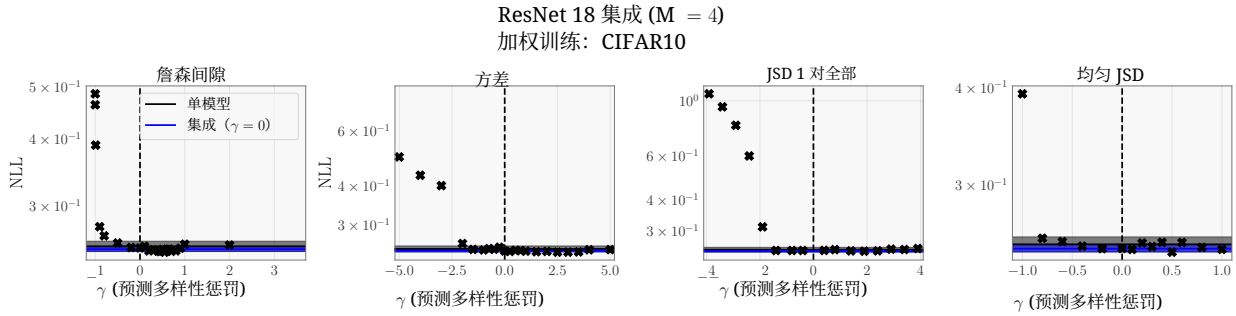


图 20： **多样性促进有害于大容量模型的深度集成负对数似然**。我们在 CIFAR10 上对 ResNet 18 应用多样性正则化深度集成训练，并测量负对数似然。约定和结论与图2相同。

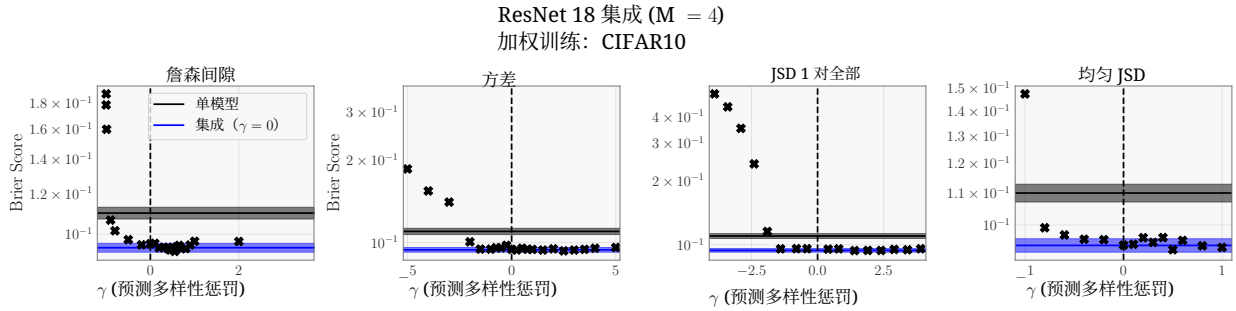


图 21： **多样性促进有害于大容量模型的深度集成布里尔分数（大容量模型）**。我们在 CIFAR10 上对 ResNet 18 应用多样性正则化深度集成训练，并测量布里尔分数。约定和结论与图 2相同。

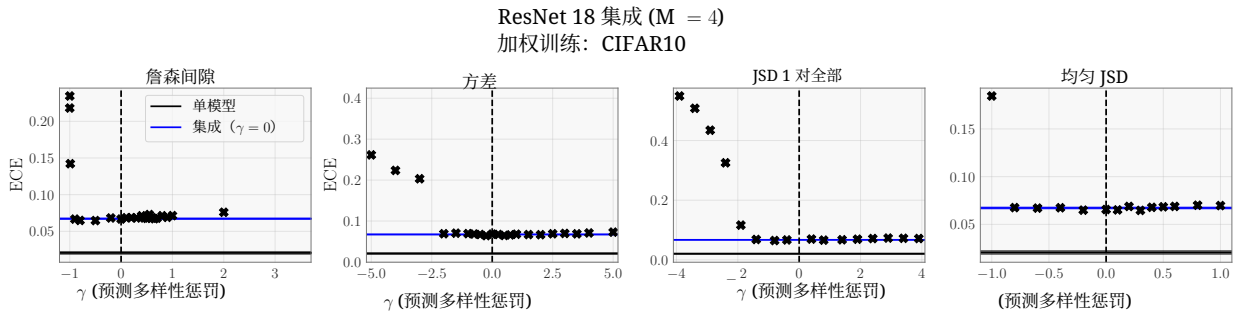


图 22： **多样性促进有害于大容量模型的深度集成期望校准误差**。我们在 CIFAR10 上对 ResNet 18 应用多样性正则化深度集成训练，并测量期望校准误差。约定和结论与图2相同。

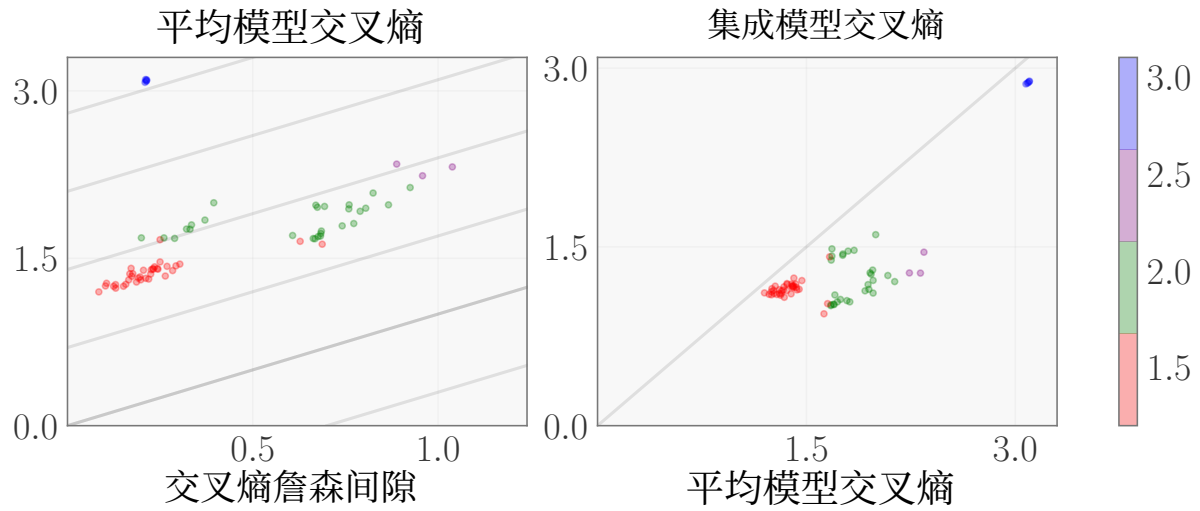


图23： **性能最佳的集成由模型性能最佳的集成组成**。模型与图 5 相同，模型均在 ImageNet V2MF 上进行评估。 **左图：** 多样性与单模型平均性能的分解，如图 1所示。 **右图：** 单模型平均性能与集成性能的分解，如图 5所示。对于每个集成，颜色表示组件模型的平均交叉熵，如色标所示。

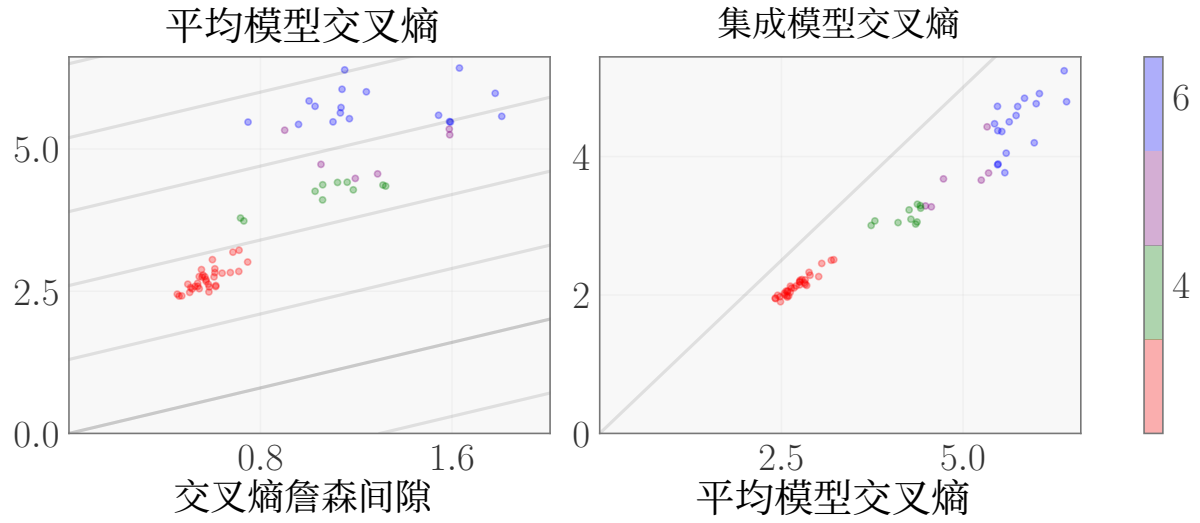


图24： **性能最佳的集成由性能最佳的模型组成**。模型与图 5相同，模型均在 ImageNet C gaussian-noise-3 上进行评估。约定与图 23相同。

G 在各种测试数据集上，最佳集成的预测多样性较低

图 5 展示了多个独立训练的模型中，单模型平均性能与詹森间隙之间的权衡。本节将图 5 中的结果扩展到多种分布外数据集，并比较单模型平均性能与集成损失函数，其中包括 ImageNetV2（Recht 等人，2019年）以及其他 ImageNet-C 数据集（Hendrycks 和 Dietterich，2019年）。总体而言，图 23 至 25表明，由性能更好的组件模型构建性能更好的集成，需要以牺牲预测多样性为代价。

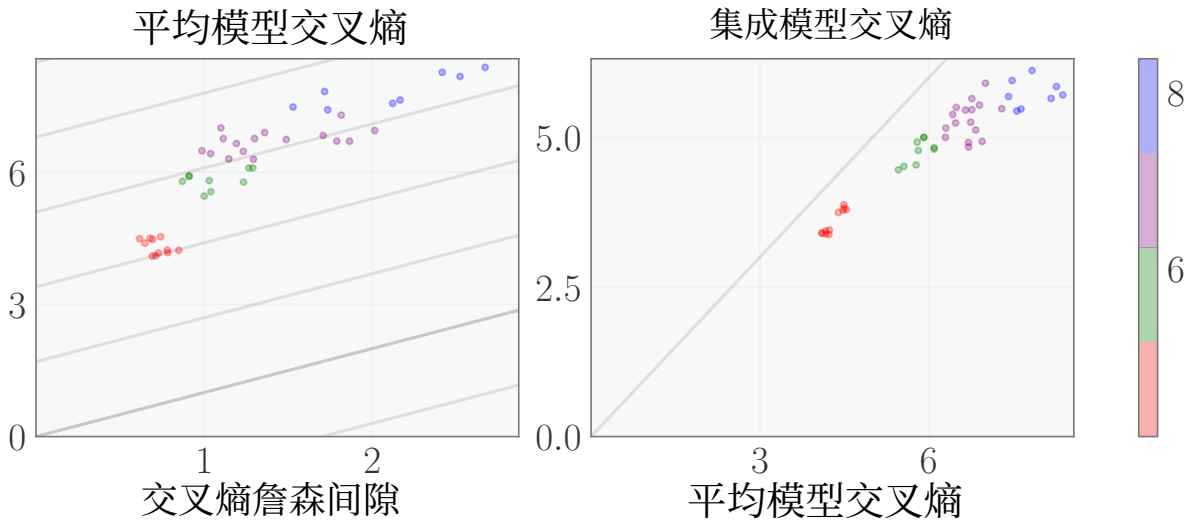


图25： **最佳模型可获得最佳集成性能**。模型与图 5相同，其中模型在 ImageNet-C 高斯噪声-5 上进行评估。约定同图 23

H 玩具实验

对于准确且置信度高的预测，我们应当预期误差去相关会如何影响多样性和单模型性能？图中展示了三种已知的提升集成多样性的方法，以此为这个问题提供更多直观理解。26，我们在其中考察了三种多样性干预各自对单次预测质量的影响。我们按照图中所示，将集成在这次单次预测上的性能进行分解。1。每种多样性干预都由一个尺度参数 s标记，该参数控制所体现的预测多样性程度。所有情况下，我们考察的都是三分类任务中三个集成成员预测 c_i之间的多样性。为简单起见，我们假设正确类别为 0，这意味着完美预测为 e_0，其中 e_i 是标准基中的第 i个成员，位于 $\mathbb{R}^3$。

几何多样性。作为一个极限情形，我们考察了如何从完美预测平滑地插值到具有最大多样性的预测。在这种情况下，我们的三个预测 c_i 为：

$$c_i = (1-s) * (e_0) + s * (e_i) \tag{10}$$

Logit 多样性。接下来，我们考察了在 Logit 空间中增加预测多样性会如何影响概率空间中的预测。在这种情况下，我们的三个预测定义完全相同：

$$c_i = \text{softmax}(10 * e_0 + \epsilon); \epsilon \sim \mathcal{N}(0, s) \tag{11}$$

狄利克雷多样性。最后，我们考虑直接在单纯形中从狄利克雷分布采样预测：

$$c_i \sim \text{Dir}(\alpha); \alpha = \frac{1}{10} * [s, (1-s)/2, (1-s)/2] \tag{12}$$

对于这两种基于采样的模拟，我们在 100 次迭代中计算交叉熵詹森间隙和单模型平均负对数似然，并绘制其平均值。

我们发现，无论选择何种多样性机制，单模型平均损失的恶化速度都快于多样性的增长速度，这为我们的直觉提供了实验验证：只有在单纯形的极端点，多样性才会损害性能。我们在 `interp_ensembles/scripts/visualize_diversity/figure_gen.py`提供了用于复现该图的代码。

I 反事实准确率分析

对于采用多样性干预的树集成和深度集成，我们考察反事实准确率：即不存在干预时集成的准确率，并将其表示为干预所实现的预测多样性的函数。在所有情况下，我们的多样性度量都是针对测试集中的各个数据点计算的交叉熵詹森间隙（附录 E）。具体而言，对于测试集中的每个数据点，我们计算：

- 集成成员在该测试点上表现出的詹森间隙预测多样性。
- 在没有多样性干预的情况下，集成成员的平均准确率（即反事实准确率）。

每幅图的上图展示了集成成员之间预测多样性的直方图，并使用高斯核进行了平滑。为生成下图，我们使用逻辑回归模型，根据预测多样性预测反事实准确率。给定一组测试集 logits 后复现此分析的代码见 `ensemble_attention/notebooks/counterfactual_accuracy.ipynb`。

I.1 树

对于树集成，我们将单棵树视为“基集成”，因为在没有特定多样性干预的情况下，树的训练几乎没有随机性。我们考虑深度为 1 到 12、在 MNIST 数据集上拟合的树。随着组件树深度增加，集成表现出更强的预测多样性。为避免出现二元预测（此时我们的多样性度量未定义），我们在所有预测中填入 $\epsilon = 1^{-10}$，然后重新归一化。

I.2 深度集成

对于深度集成，我们将标准深度集成（基于初始化形成）视为基集成。作为多样性干预，我们考虑第 4 节所述的每一种多样性干预，其取值为$\gamma < 0$。请注意，$\gamma = 0$ 对应于标准深度集成。随着 γ的取值越来越负，集成表现出更强的预测多样性。

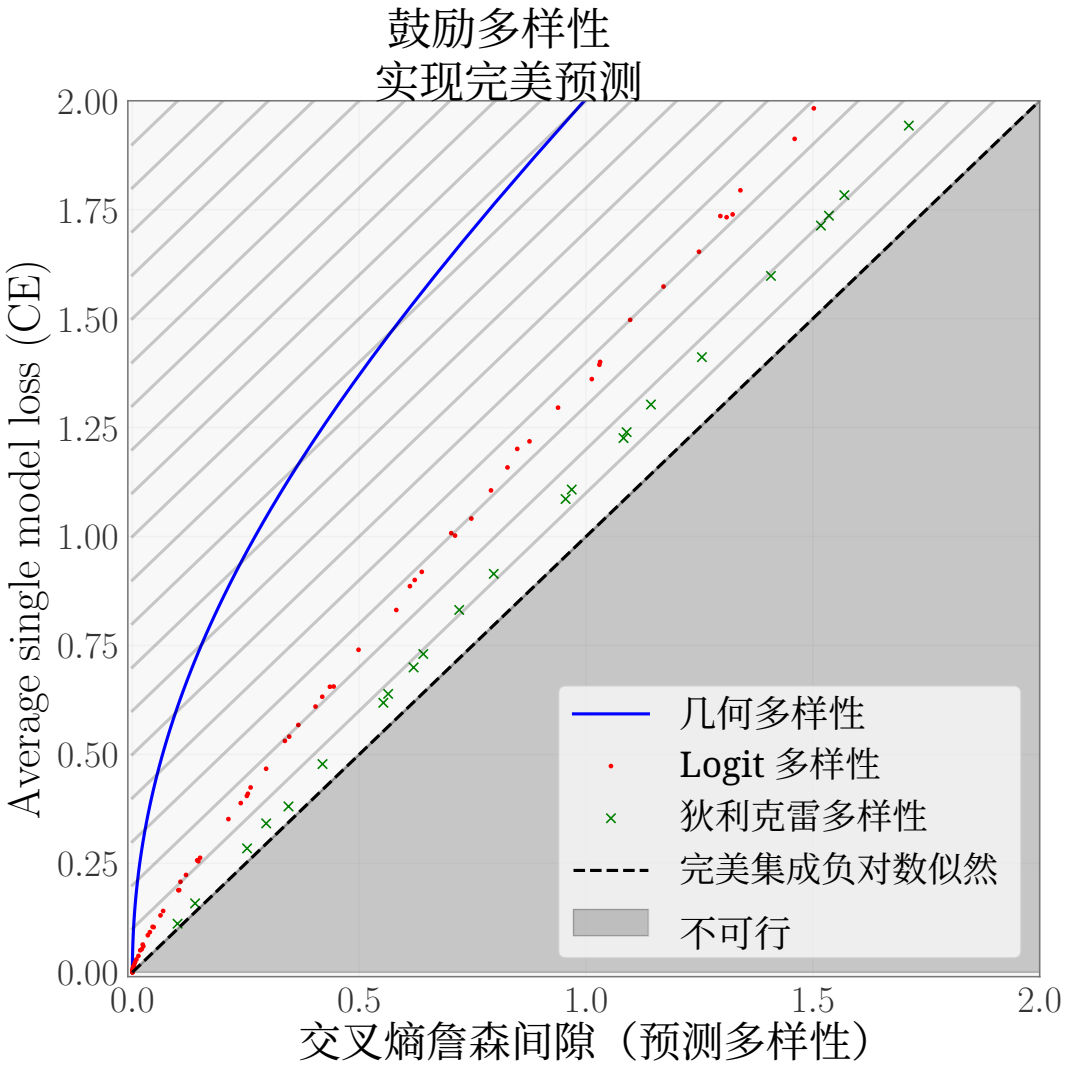


图26：**无论采用何种鼓励策略，在极端点处都不可能实现多样性**。在向完美集成预测添加多样性的 3 种不同方法下（几何缩放、在 Logit 空间中采样，或从狄利克雷分布中采样），任何一种策略都会使单模型平均损失的恶化速度快于预测多样性的提升速度，最终导致整体集成性能变差。

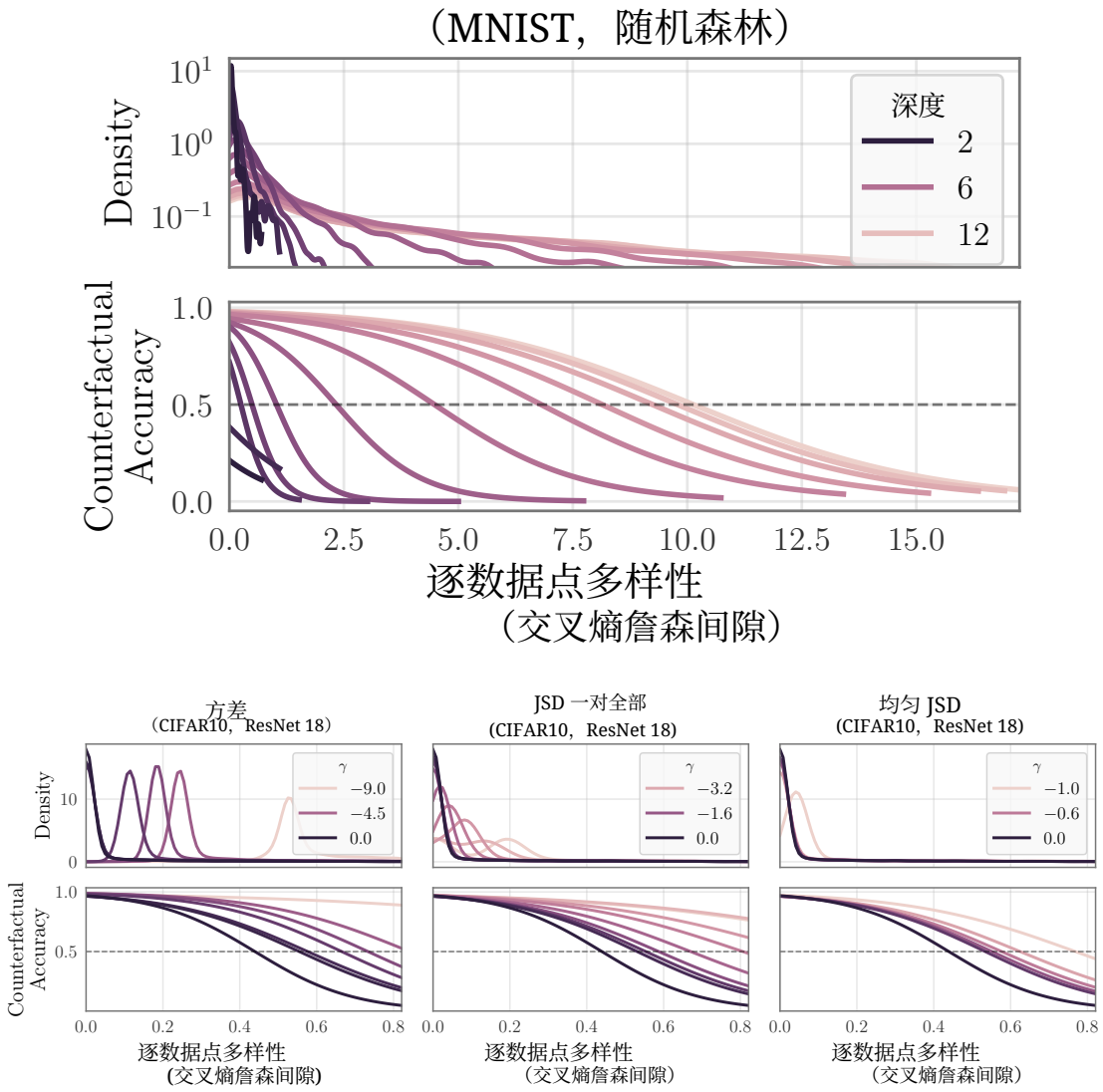


图27： **对于全部多样性干预，多样性促进会使正确的预测去相关**。 与图 4一样，我们发现，促进更高水平预测多样性的方法会使正确的集成预测去相关；此处展示了随机森林（**上**）以及应用于 ResNet 18 的正文中未展示的三种多样性正则项（**下**）的结果。