《机器学习研究期刊》24（2023年）1-49 投稿：1/23；修订：12/23；发表：12/23

arXiv:2301.03962v3 [cs.LG] 7 Feb 2024

集成学习中多样性的统一理论

Danny Wood DANNY.WOOD@MANCHESTER.AC.UK†**Tingting Mu** TINGTING.MU@MANCHESTER.AC.UK†**AndrewM. Webb** ANDREW.WEBB@MANCHESTER.AC.UK†
Henry W. J.Reeve HENRY.REEVE@BRISTOL.AC.UK*MikelLuj´an MIKEL.LUJAN@MANCHESTER.AC.UK†**Gavin Brown** GAVIN.BROWN@MANCHESTER.AC.UK†

† 曼彻斯特大学计算机科学系，英国
∗ 布里斯托大学数学学院，英国

编辑： Boaz Nadler

摘要

我们提出了一种集成多样性理论，用以阐释多样性在广泛监督学习场景中的本质。这个难题被称为集成学习的“圣杯”，三十多年来一直是一个悬而未决的研究问题。我们的框架揭示，多样性实际上是集成损失偏差-方差分解中的一个隐藏维度。我们证明，对于回归和分类中的多种损失函数（如平方损失、交叉熵和泊松损失），都存在一类精确的偏差-方差-多样性分解。对于无法进行加性偏差-方差分解的损失函数（如 0/1 损失），我们提出了一种替代方法：量化多样性的影响，结果发现这些影响取决于标签分布。总的来说，我们认为，多样性是衡量模型拟合程度的一项指标，其含义与偏差和方差完全相同，但同时考虑了集成成员之间的统计依赖关系。因此，我们不应像许多研究所追求的那样“最大化多样性”，而应当管理偏差/方差/多样性之间的权衡。

关键词： 集成、多样性、偏差、方差

1 引言

几十年来，集成方法不断带来最先进的成果：从早期工业计算机视觉（Viola 和 Jones，2001年），到深度学习革命（Krizhevsky 等人，2012年），再到跨学科应用（Cao 等人，2020年）。一种公认的观点是，当各个成员的预测具有“多样性”时，集成效果最佳——这种多样性通常由装袋法（Breiman，1996年）等经典方法引入，但文献中也有大量鼓励多样性的启发式方法（Brown 等人，2005年）。鉴于此，我们相信集成能够“平均掉”各个成员的错误。这类方法广受欢迎的原因之一显而易见：集成的理念 本身就具有吸引力，因为它拟人化地让人联想到人类委员会和“群体智慧”。然而，这类类比也有局限。人们也一直在探索更形式化的方法，尤其是量化 多样性的方法。显然，我们不希望所有预测都完全相同；同样显而易见的是，我们也不希望它们仅仅为了不同而不同，从而牺牲整体性能。我们想要的是介于两者之间的某种状态，也就是所谓的

©2023年 Danny Wood、Tingting Mu、Andrew M. Webb、Henry W. J. Reeve、Mikel Luj´an 和 Gavin Brown。许可协议：知识共享署名 4.0 国际许可协议，参见 https://creativecommons.org/licenses/by/4.0/。署名要求见 http://jmlr.org/papers/v24/23-0041.html。

准确率/多样性权衡。然而，我们在这里遇到了一个问题：如何正式定义“多样性”及其与集成性能的关系。总的来说，关于如何量化多样性尚无共识，只有在采用算术平均集成的回归这一有限情形下例外（Krogh 和 Vedelsby，1994年；Ueda 和 Nakano，1996年）。对于分类及其他情形，已有数十种多样性度量被提出（Kuncheva，2014年）。三十多年来，建立一套全面的集成多样性理论始终是一个悬而未决的问题。

动机： 我们的首要动机是填补当前集成理论中的这一“空白”，为理解和研究集成多样性提供坚实基础。不过，推动这项研究还有许多实际原因。与具有相同性能泛化性能的单个大型模型相比，多样化集成的计算效率可能更高（Kondratyuk 等，2020）。多样化集成能够抵御对抗攻击（Biggio 等，2011；Pang 等，2019），还可以应对协变量偏移（Sinha 等，2020）。在重要的应用领域（Cao 等，2020）以及远超监督学习的范围内（Carreira-Perpin´an 和 Raziperchikolaei，2016），也能发现其优势。需要注意的是，这些应用案例并未遵循一种共同的方法：它们要么采用历史文献中选取的某种多样性度量，要么提出自身的新度量指标。因此，有必要发展一种统一理论，从第一性原理出发推导多样性。

这一挑战已被证明并非易事：Dietterich（2000年）、Brown 等人（2005年）、Zhou（2012年）和 Kuncheva（2014年）均对相关进展进行了综述。如今，多样性被视为一种没有精确定义的启发式概念；正如有人所说：

> “毫无疑问，理解多样性是集成学习领域的圣杯”（Zhou，2012年，第5.1节，第100页）。

研究成果概要： 与以往定义新型多样性度量的研究不同，我们对损失函数进行分解，揭示出能够自然体现多样性的项。我们表明，多样性是集成损失偏差-方差分解中的一个隐藏维度。具体而言，我们证明了适用于广泛损失函数的精确偏差-方差-多样性分解，其形式统一如下：

$$\textbf{期望损失} = (\textbf{平均偏差}) + (\textbf{平均方差}) - (\textbf{多样性}),$$

其中，多样性是衡量集成成员分歧的度量，且独立于标签。对于平方损失这一特殊情况，这是一种不同于 Ueda 和 Nakano（1996年）提出的分解方法，但它将多样性的形式化概念扩展到了许多其他损失函数，例如交叉熵和泊松损失。对于不存在加性偏差-方差分解的损失函数（例如 0/1 损失），我们提出了一种精确量化多样性影响的方法，但需注意，这些影响取决于标签分布。

总而言之，我们认为，多样性最好被理解为一种衡量模型拟合程度的指标，其含义与偏差和方差完全相同，但它还考虑了集成成员之间的统计依赖关系。因此，我们不应像许多研究所声称的那样“最大化”多样性；相反，我们需要管理一种权衡。对于单个模型，我们熟知偏差/方差 权衡。对于集成，我们则面临偏差/方差/多样性 权衡，这种权衡会随着单个模型的容量以及模型预测之间的相似性而变化。

符号说明

为便于快速查阅，我们总结了本文中使用的大部分符号约定。文中主体部分会在适当之处给出更详细的定义。粗体字（例如$\mathbf{y}$）表示向量，而 y 表示标量。大写字母（例如 Y）表示随机变量，其中 y 是 Y的实现值。

为简洁起见，在本文大部分内容中，我们不再明言模型 q 是输入 $\mathbf{x}$ 的函数，且依赖于从随机变量 D 中抽取的训练集。

符号	描述
$\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^d$	d 维输入
$\mathbf{y} \in \mathcal{Y} \subseteq \mathbb{R}^k$	k 维标签（或对于标量为 y）
$\{(\mathbf{x}_j, \mathbf{y}_j)\}_{j=1}^n$	n 个带标签样本的训练集
$P(\mathbf{x}, y)$	在 $(\mathbf{x}, y)$ 上的未知数据分布
$\mathbb{E}_{\mathbf{X}Y}[\cdots]$	关于 $P(\mathbf{x}, y)$ 的期望，即对所有可能测试样例 $(\mathbf{x}, y)$ 的期望 $\mathbb{E}_{\mathbf{X}Y}[\cdots] \stackrel{\text{def}}{=} \int P_{\mathbf{X}}(\mathbf{x}) \int P_{Y\mid\mathbf{X}}(y \mid \mathbf{x}) [\cdots] \, \mathrm{d}y \, \mathrm{d}\mathbf{x}$
D	随机变量 $D \sim P(\mathbf{x}, y)^n$，基于大小为 n 的独立同分布训练集
$\mathbb{E}_D[\cdots]$	对从 D 抽取的所有可能训练集 $\{(\mathbf{x}_j, \mathbf{y}_j)\}_{j=1}^n$ 求期望
$q(\mathbf{x})$	预测模型 q，将 $\mathbf{x} \to y$ 映射为.....。如果输出是一个向量，则 $\mathbf{q}(\mathbf{x})$
$\ell(y, q(\mathbf{x}))$	针对标签 y 和预测值 $q(\mathbf{x})$ 的损失函数
$\ell_{0/1}(y, q(\mathbf{x}))$	0-1 损失函数：若 $y \neq q(\mathbf{x})$，则返回 1；否则返回 0。
$R(q)$	模型 q 的风险，定义为 $R(q) := \mathbb{E}_{\mathbf{X}Y}[\ell(y, q(\mathbf{x}))]$
m	集成成员数量
q_i	集成的第 i 个成员，$\{q_i\}_{i=}^m$
$\overline{q}$	质心组合规则，定义为 $q := \arg\min_{z \in \mathcal{Y}} \frac{1}{m} \sum_{i=1}^m [\ell(z, q_i)]$
$\mathring{q}$	关于 D 的模型分布的质心，定义为 $\mathring{q} := \arg\min_{z \in \mathcal{Y}} \mathbb{E}_D[\ell(z, q)]$
$\phi(\mathbf{y})$	布雷格曼散度的严格凸生成函数
$\nabla\phi(\mathbf{y})$	生成函数 ϕ 关于自变量 $\mathbf{y}$ 的梯度向量
$B_\phi(\mathbf{y}, \mathbf{q})$	使用生成函数 ϕ 的点 $\mathbf{y}$，$\mathbf{q} \in \mathbb{R}^k$ 之间的布雷格曼散度

表 1：本文使用的符号说明。

2 问题陈述：什么是集成多样性?

Hansen 和 Salamon（1990年）提出了一种训练多个神经网络的方法，通过为每个网络提供不同的训练数据子集，鼓励模型之间有所差异。此后许多论文沿用了这一“并行”策略：早期工作包括 Perrone 和 Cooper（1992年）的研究；更广为人知的或许是装袋法 （Breiman，1996年）和随机森林（Breiman，2001年）。提升法 算法（Schapire 等，1998年）利用了类似的原理，但以顺序方式构建模型。这些并行和顺序方法（见图1）是学习集成最常见的方案（Kuncheva，2014年），此外还有许多变体，例如从预先构建的候选池中选取成员；候选池中的每个成员都由独立的（并行）团队构建，这些团队处理相同的任务，但使用不同的数据。

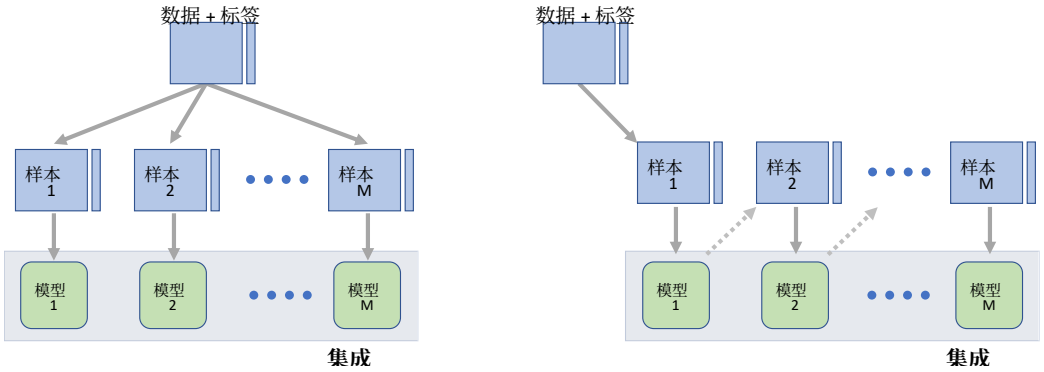


图1：并行与顺序集成构建。两者都可以被视为以某种方式创建“多样化”的模型——要么隐式地 （独立地重新抽样训练数据），要么显式地 （根据先前模型的错误重新抽样）。

那么，这些策略为何有效？从启发式角度看，两者都可以用“多样性”来理解。Opitz 和 Shavlik（1996年）提出的这一概念，指的是一组模型之间泛化行为的差异。在一篇综述中，Dietterich（2000年）解释道：

> “准确的分类器是指在新的 x 值上，其错误率优于随机猜测的分类器。如果两个分类器在新的数据点上产生不同的错误，那么它们就具有多样性。”（Dietterich，2000年）

从这个意义上说，两种方法都能促进多样性——要么隐式地 随机扰动每个模型所用的数据，要么显式地 构建各个数据集，以应对先前模型的错误（Brown 等人，2005年）。隐式方法已广泛应用于深度学习：Goodfellow 等人（2016年）指出，深度网络初始化过程中的随机性往往足以使它们产生部分独立的错误。鉴于集成取得的成功，人们进行了许多尝试，试图从多样性的角度解释它们为何 有效。Goodfellow 等人（2016年）写道：

> “模型平均之所以有效，是因为不同模型通常不会在测试集上犯下完全相同的错误”。（Goodfellow 等人，2016年，第249页）

尽管这确实如此，我们仍希望得到更正式的论述。

什么是 e w正在寻找什么? 多样性理论理想情况下应包含三个关键要素:

1. 多样性的定义，即将多样性视为集成成员之间分歧程度的度量 s;
2. 这种度量应与整体集成误差有明确的关系；并且 ,
3. 该理论应与先前确立的结果有明确联系，并将我们的理解拓展到更广泛的学习场景。

第一点关乎多样性本身的定义。一个值得探讨的问题是，这种度量是否仅由集成成员的输出决定，还是也可以涉及标签分布。Kuncheva（2014年）将依赖真实标签的度量称为“神谕”多样性度量。这两种结果都将是有趣的科学结论：标签无关情形表明，多样性完全是集成本身的函数；标签相关情形则表明，多样性是一种类别条件现象。

第二点确保我们能够理解多样性对最终目标——降低集成误差——的影响。第三点则确保新理论能够增进知识，也就是说，加深或拓展我们的理解。这也关系到目前已知的唯一一种情形：多样性的定义难题或许可以被视为已经“解决”——即采用平方损失和算术平均集成进行回归。现在我们来回顾这一情形。

回归集成的已知结果： Krogh 和 Vedelsby（1994）证明，对于采用平方损失的算术平均组合器，集成损失保证小于或等于个体损失的平均值。

定理 1（歧义分解，Krogh 和 Vedelsby，1994） 给定一个标签$y \in \mathbb{R}$、一个成员预测 $q_i(\mathbf{x})$以及一个集成 $\bar{q}(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} q_i(\mathbf{x})$，有:

$$\big(\bar{q}(\mathbf{x}) - y\big)^2 \;=\; \frac{1}{m}\sum_{i=1}^{m}\big(q_i(\mathbf{x}) - y\big)^2 - \frac{1}{m}\sum_{i=1}^{m}\big(q_i(\mathbf{x}) - \bar{q}(\mathbf{x})\big)^2. \tag{1}$$

左侧是在单个测试点（$\mathbf{x}$ y）上的集成损失。右侧第一项是个体损失的平均值。第二项称为歧义，用于衡量个体之间的分歧，即它们围绕集成预测的离散程度。由于该项非负，因此可以保证集成损失小于或等于平均损失。

这个结果常被错误地引用为所有集成都有效的原因。然而，上述形式仅在使用平方损失和算术平均组合器时适用。如果使用平方损失搭配其他组合器，这一结果便不再成立。Ueda 和 Nakano（1996年）对此有了更深入的理解——不过，他们的结论仍基于相同的损失和组合器假设。他们将 Geman 等人（1992年）的经典偏差-方差理论加以扩展，指出集成的期望平方损失可分解为三项，分别涉及模型的偏差、方差和协方差。在这里，协方差体现了“多样性”的概念。稍后将对此作进一步说明。

如前所述，上述表达式不适用于算术平均组合器下的平方损失以外的情况。研究界投入了大量精力，为分类问题寻找相应的多样性概念。接下来我们将对此进行回顾。

分类器集成的已知结果：对于分类问题，我们可能有类别概率分布的估计值，也可能只有标签——通常通过概率平均或多数投票进行组合。Tumer 和 Ghosh（1996年）表明，平均类别概率两两之间的相关性与整体集成分类误差存在简单的关系，至少在决策边界附近的区域内如此。Brown（2009）以及 Zhou 和 Li（2010年）提出了信息论分析，表明多样性体现为集成成员之间的低阶和高阶交互作用。Buschj¨ager 等人（2020年）对二阶可微损失采用泰勒近似，表明只有在高阶导数为零时（例如平方损失，而非交叉熵）才存在精确分解。类似地，Ortega 等人（2022年）对损失函数的上界进行了分解，也只有在平方损失下才能得到等式。Kuncheva 和 Whitaker（2003年）研究了多样性度量与集成误差之间的相关性。他们考察了多种差异度量，其形式为 $\delta(q_i(\mathbf{x})q_j(\mathbf{x})) \in \mathbb{R}$，这些度量在每个测试点 $\mathbf{x}$ 上对所有集成成员对取平均：

$$\text{diversity}(q_1, .., q_m, \mathbf{x}) = \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i} \delta(q_i(\mathbf{x}), q_j(\mathbf{x})). \tag{2}$$

集成多样性是该量在验证数据集上的平均值。集成多样性根据其与集成性能之间的经验相关性进行评估；如图2所示，相关性越高，通常被认为越成功。

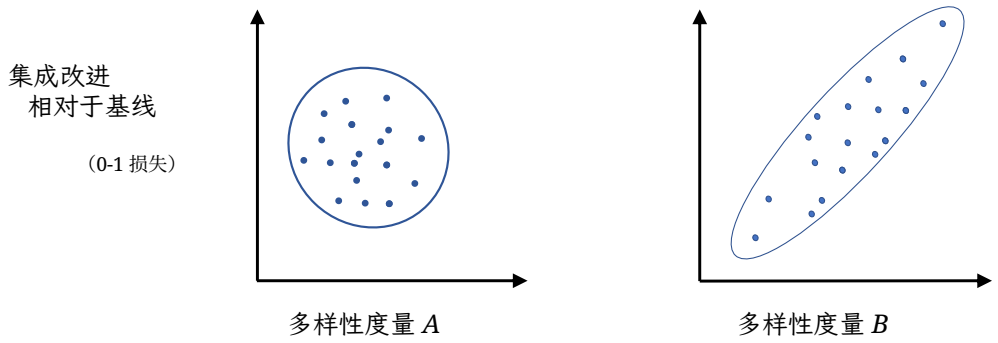


图2：两种（假设的）多样性度量的准确率/多样性。度量B（右侧）更为理想，因为它与性能提升的相关性更强。

研究人员探索了多种度量（也包括非成对度量），但没有任何一种度量被证明比其他度量更成功。近20年过去了，新的多样性启发式方法/度量仍在不断提出，例如 Jan 和 Verma（2019年）；Wu 等（2021年）。

我们处理这一问题的方法：如上所述，多样性通常被置于分类与回归的二分框架中讨论。我们的主要发现之一是，这种高层次的二分法不足以全面描述多样性的本质。相反，我们必须考虑具体的损失函数。对于回归，这可能是平方损失，也可能是泊松损失或其他损失。对于分类，则可能是 0/1 损失或交叉熵。我们以偏差-方差理论的坚实基础为依托（Geman 等人，1992年；James 和 Hastie，1997年；Heskes，1998年；Pfau，2013年），并分解损失函数，自然地揭示多样性的概念。

3 偏差-方差分解简明导论

我们回顾偏差-方差分解相关文献：统计设定是标准的监督学习场景。假设有一个从随机变量 $D \sim P(\mathbf{X}\ Y)^n$ 中抽取的训练集 $\{(\mathbf{x}_j, y_j)\}_{j=1}^n$。根据该训练集，我们学习一个模型 q。为简洁起见（见表 1），在本文余下部分中，我们不再显式说明 q 是 $\mathbf{x}$ 的函数，且始终依赖于从 D 中抽取的数据。我们将风险记作 $R(q) \overset{\text{定义为}}{=} \mathbb{E}_{\mathbf{X}Y}[\ell(Y, q)]$,，其中标签 Y 是一个随机变量，因此可能带有噪声。如果 $\ell(y, q) = (y - q)^2$,，这就是平方风险，而贝叶斯最优预测是条件均值 $Y^* \overset{\text{定义为}}{=} \mathbb{E}_{Y|\mathbf{X}}[Y]$。Geman 等人（1992年）考察了针对从 D 中抽取的各种可能训练集的期望平方风险，并给出了如下三项分解。

$$\underbrace{\mathbb{E}_D\left[\mathbb{E}_{\mathbf{X}Y}[(Y-q)^2]\right]}_{\text{期望平方风险}} = \mathbb{E}_{\mathbf{X}}\Bigg[\underbrace{\mathbb{E}_{Y|\mathbf{X}}\left[(Y-Y^*)^2\right]}_{\text{噪声}} + \underbrace{\left(\mathbb{E}_D[q] - Y^*\right)^2}_{\text{bias}} + \underbrace{\mathbb{E}_D\left[\left(q - \mathbb{E}_D[q]\right)^2\right]}_{\text{方差}}\Bigg]. \tag{3}$$

第一项是问题中不可约的噪声，与任何模型参数无关。第二项是偏差[1],，定义为期望模型相对于贝叶斯最优预测的损失。第三项是方差，用于刻画不同训练集导致的 q 变化。请注意，我们可以用术语偏差（相应地方差）表示某一点（$\mathbf{x}$ y）处的值，也可以表示在某个分布上的平均值——具体含义可从上下文中判断。解释这些概念时，常会用到飞镖靶图，如图 3 所示。

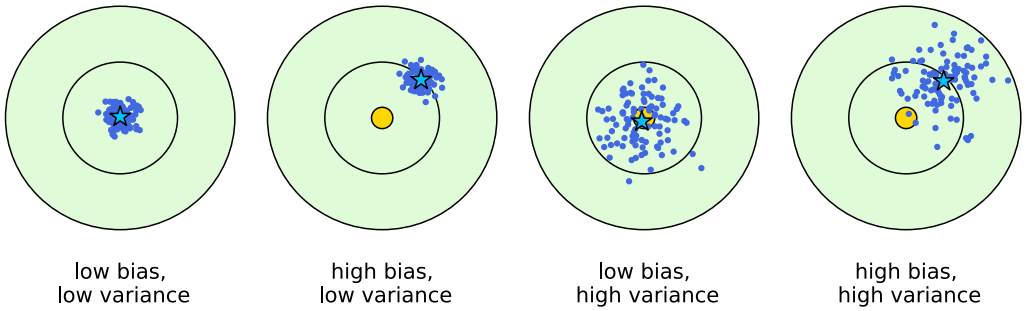


图 3：用于解释偏差和方差的经典飞镖靶类比。

靶心（黄色圆圈）是单个测试点的目标，每个蓝点则是模型在不同训练集上给出的预测。高偏差、低方差的模型平均而言会离目标较远，并且对数据的细微变化不敏感。低偏差、高方差的模型的期望值会接近目标，但对数据变化非常敏感，这意味着任意一个模型都可能过拟合。随机变量 D 不一定要是关于数据的，它可以是学习模型过程中涉及的任何随机量，例如神经网络中的随机权重。

1. *Is it 'squared bias', or just 'bias'?* Geman 等人提出的偏差项通常称为“平方偏差”。然而，平方项实际上是使用平方损失所产生的结果，在其他损失函数的分解中并不存在。因此，“平方偏差”一词 *is a misnomer*，本文全篇仅使用“偏差”一词。附录 B.1 对此作了进一步阐述。

偏差-方差分解不仅适用于平方损失。Geman 等人（1992年）的结果广受认可。鲜为人知的是，类似的分解也适用于其他损失函数，例如类别概率估计的 KL散度（Heskes, 1998年）。假设$\mathbf{y}$ 是一个长度为 k的独热向量，而$\mathbf{q} \in \mathbb{R}^k$ 是我们的模型对 k 个标签的预测分布。期望风险也可以分解[2] 为类似的三部分形式：

$$\underbrace{\mathbb{E}_D\Big[\mathbb{E}_{\mathbf{XY}}\left[K\left(\mathbf{Y} \,||\, \mathbf{q}\right)\right]\Big]}_{\text{expected risk}} = \mathbb{E}_{\mathbf{X}}\Bigg[\underbrace{\mathbb{E}_{\mathbf{Y}|\mathbf{X}}\Big[K\Big(\mathbf{Y} \,||\, \mathbf{Y}^*\Big)\Big]}_{\text{noise}} + \underbrace{K\Big(\mathbf{Y}^* \,||\, \mathring{\mathbf{q}}\Big)}_{\text{bias}} + \underbrace{\mathbb{E}_D\Big[K\Big(\mathring{\mathbf{q}} \,||\, \mathbf{q}\Big)\Big]}_{\text{variance}}\Bigg], \tag{4}$$

其中，$\mathring{\mathbf{q}} \overset{\text{def}}{=} \arg\min_{\mathbf{z}\in\mathcal{Y}} \mathbb{E}_D\Big[K(\mathbf{z} \,||\, \mathbf{q})\Big] = Z^{-1}\exp(\mathbb{E}_D[\ln \mathbf{q}])$ 是每个输入向量 $\mathbf{x}$对应的归一化几何平均。这也可以通过图4中的“飞镖靶”类比来说明。

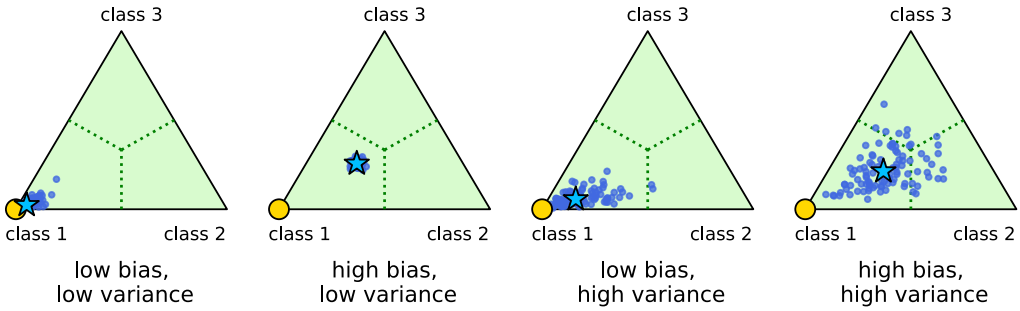


图4：说明 KL散度的偏差/方差的飞镖靶图。

偏差-方差分解并不适用于所有损失函数。上述分解的一个共同点是，方差独立于标签。另一个共同点是，偏差是预测器 $\mathring{q} \overset{\text{def}}{=} \arg\min_{z\in\mathcal{Y}} \mathbb{E}_D\Big[\ell(z, q)\Big]$的损失，而不依赖于任何特定的训练集。我们将这种形式称为加性偏差-方差分解。事实证明，这并不适用于所有损失函数，例如 0/1 损失——在这种情况下，$\mathring{q}$ 是随机变量的众数，但是我们有一个不等式：

$$\mathbb{E}_D\Big[\mathbb{E}_{\mathbf{X}Y}[\ell_{0/1}(Y, q)]\Big] \;\neq\; \mathbb{E}_{\mathbf{X}}\Big[\mathbb{E}_{Y|\mathbf{X}}[\ell_{0/1}(Y, Y^*)] + \ell_{0/1}(Y^*, \mathring{q}) + \mathbb{E}_D[\ell_{0/1}(\mathring{q}, q)]\Big]. \tag{5}$$

许多作者提出了其他的 0/1 分解（Kohavi 等人，1996年；Tibshirani，1996年；James 和 Hastie，1997年；Heskes，1998年；Domingos，2000年）。Geurts（2002年）提供了一篇出色的综述。常见的做法是，要么舍弃简单的加性形式，要么采用一种依赖于标签分布的方差表述。上述形式的加性偏差-方差分解的充要条件，仍是一个开放的研究问题。

综上所述，这些理论已被用于理解模型拟合的本质——即所谓的“偏差-方差”权衡。下一节中，我们将基于这些理论，进一步理解集成多样性这一现象。

2. Heskes（1998年）提出了一种略有不同的形式，不过两者的表达式等价。参见附录B.3。

4 集成多样性的统一理论

集成“多样性”是一个广受关注、但定义各异的概念。偏差和方差的定义明确且精确。在二者之间建立更紧密的联系是有意义的。我们的做法正是如此：揭示多样性是集成偏差-方差分解中的一个隐藏维度。我们认为，多样性可以与偏差/方差完全相同的方式来考量——它只是模型拟合中另一个可量化的方面。

4.1 作为偏差-方差分解隐藏维度的多样性 n

我们的方法之所以称为“统一理论”，是因为相同的 方法适用于分类和回归场景中的多种不同损失函数。我们首先定义一种广义（加法）偏差-方差分解，它是对公式（3）和（4）的推广。

定义 2（广义偏差-方差分解）对于损失函数 ℓ，我们定义 $\mathring{\mathbf{q}} \stackrel{\text{def}}{=} \arg\min_{\mathbf{z}\in\mathcal{Y}} \mathbb{E}_D\Big[\ell(\mathbf{z},\mathbf{q})\Big]$ 为模型分布的“质心”。若满足以下形式，则称之为广义偏差-方差分解。

$$\underbrace{\mathbb{E}_D\Big[\mathbb{E}_{\mathbf{XY}}[\ell(\mathbf{Y},\mathbf{q})]\Big]}_{\text{expected risk}} = \mathbb{E}_{\mathbf{X}}\Big[\underbrace{\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[\ell(\mathbf{Y},\mathbf{Y}^*)]}_{\text{noise}} + \underbrace{\ell(\mathbf{Y}^*,\mathring{\mathbf{q}})}_{\text{bias}} + \underbrace{\mathbb{E}_D\left[\ell(\mathring{\mathbf{q}},\mathbf{q})\right]}_{\text{variance}}\Big]. \tag{6}$$

当 $\ell(y,q) = (y-q)^2$，时，我们有 $\mathring{q} = \mathbb{E}_D[q]$，这就简化为公式（3）。或者，当 $\ell(\mathbf{y},\mathbf{q}) = K(\mathbf{y}\ ||\ \mathbf{q})$ 时，我们有 $\mathring{\mathbf{q}} = Z^{-1}\exp(\mathbb{E}_D[\ln\mathbf{q}])$，这就简化为公式（4）。重要的是，这种相同的加法形式也适用于更广泛的布雷格曼散度族（Bregman，1967年）。下一节将介绍这一族散度；它提供了方便的解析形式，我们将用它推导出若干有趣的性质。

现在我们迈出了探讨集成的第一步。我们指出，歧义分解（公式 (1)）可以用与偏差-方差分解非常相似的方法证明。事实上，只要其中一个存在，另一个也必然存在，详见附录 B.2。对于损失 ℓ，如果公式 (6) 成立，我们就可以给出相应的歧义分解。

命题 3（广义歧义分解）假设损失 ℓ 存在偏差-方差分解，那么对于集成 $\{\mathbf{q}_i\}_{i=1}^m$，其歧义分解为：

$$\underbrace{\ell(\mathbf{y},\overline{\mathbf{q}})}_{\text{ensemble loss}} = \underbrace{\frac{1}{m}\sum_{i=1}^m \ell(\mathbf{y},\mathbf{q}_i)}_{\text{average loss}} - \underbrace{\frac{1}{m}\sum_{i=1}^m \ell(\overline{\mathbf{q}},\mathbf{q}_i)}_{\text{ambiguity}}, \tag{7}$$

其中，$\overline{\mathbf{q}} \stackrel{\text{def}}{=} \arg\min_{\mathbf{z}\in\mathcal{Y}} \frac{1}{m}\sum_{i=1}^m \ell(\mathbf{z},\mathbf{q}_i)$ 是集成组合。

我们强调，集成组合器被定义**为损失的一项属性**。对于平方损失，这会得到 $q = \frac{1}{m}\sum_{i=1}^m q_i$，也就是常用的算术平均组合器。对于 KL散度，这会得到$\mathbf{q} = Z^{-1}\prod_{i=1}^m \mathbf{q}_i^{1/m}$，即归一化几何平均数，也称为“乘积规则”（Kuncheva，2014年）。Audhkhasi 等人（2013年）和 Jiang 等人（2017年）提出了分类损失的歧义分解，不过两者都假设组合器必须是算术平均值。相比之下，这里的组合器是根据损失 ℓ定义的。我们将其称为质心组合规则，其正式定义如下。

定义 4（质心组合规则）对于一个点$(\mathbf{x},\ \mathbf{y})$，给定一组预测$\{\mathbf{q}_i\}_{i=1}^m$，质心组合器$\mathbf{q}$ 是在所有集成成员中，使平均损失 $\ell(\mathbf{z},\ \overline{\mathbf{q}}\ \mathbf{q}_i)$达到最小值的组合器。

$$\overline{\mathbf{q}} \ \stackrel{\text{def}}{=}\ \arg\min_{\mathbf{z}\in\mathcal{Y}} \frac{1}{m}\sum_{i=1}^{m}\ell(\mathbf{z},\mathbf{q}_i). \tag{8}$$

对于每个给定的损失函数，都可以分别推导出质心组合器。然而，在下一节中，我们将介绍布雷格曼散度（Bregman，1967年）这一类散度，它们为质心组合器提供了一种通用的解析形式。现在，我们可以给出主要结果——定理 5，即广义偏差-方差-多样性分解。

定理 5（广义偏差-方差-多样性分解）考虑一组模型 $\{\mathbf{q}_i\}_{i=1}^m$，并使用损失 ℓ 对其进行评估。假设偏差-方差分解满足定义 *2* 的形式，则以下分解也成立。

$$\mathbb{E}_D\Big[\mathbb{E}_{\mathbf{XY}}[\ell(\mathbf{Y},\overline{\mathbf{q}})]\Big] \ =$$

$$\mathbb{E}_{\mathbf{X}}\Bigg[\underbrace{\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[\ell(\mathbf{Y},\mathbf{Y}^*)]}_{\text{noise}}+\underbrace{\frac{1}{m}\sum_{i=1}^{m}\ell(\mathbf{Y}^*,\mathring{\mathbf{q}}_i)}_{\text{average bias}}+\underbrace{\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_D\left[\ell(\mathring{\mathbf{q}}_i,\mathbf{q}_i)\right]}_{\text{average variance}}-\underbrace{\mathbb{E}_D\left[\frac{1}{m}\sum_{i=1}^{m}\ell(\overline{\mathbf{q}},\mathbf{q}_i)\right]}_{\text{diversity}}\Bigg], \tag{9}$$

其中 $\mathring{\mathbf{q}}_{\text{定义为}}=\arg\min_{\mathbf{z}\in\mathcal{Y}}\mathbb{E}_D\Big[\ell(\mathbf{z},\ \mathbf{q})\Big]$，组合器为 $\overline{\mathbf{q}}_{\text{定义为}}=\arg\min_{\mathbf{z}\in\mathcal{Y}}\frac{1}{m}\sum_{i=1}^{m}\ell(\mathbf{z},\mathbf{q}_i)$。

这将期望集成风险分解为：噪声、平均偏差、平均方差和期望歧义。我们将这个期望歧义项视为集成的多样性，并强调它与其他项的符号相反。我们可以得出两个重要观察。

集成组合器 q 是一个质心。在实际情境中，集成的损失取决于具体任务，可以是 0/1 损失、交叉熵、均方误差或其他许多损失。组合规则也几乎可以任意选择，例如投票、算术平均值、加权组合、几何平均数等。上述结果表明，如果我们对这些选择施加约束（即定义 4），就能得到一种分解，并由此自然地显现出多样性项。我们强调，采用这种方法并不意味着对集成的经验表现作出任何断言，而只是涉及其理论性质。

现在存在偏差/方差/多样性权衡。 随着单个模型的容量增加，其平均偏差会降低。若没有正则化，其平均方差则会增加。然而，这些项只能决定集成行为的一部分。最后一部分是多样性。这里的关键点是多样性总是会降低期望风险。这并不意味着多样性越大，期望风险就总会越低——只有在偏差和方差固定时，多样性才会降低期望风险。归根结底，偏差/方差/多样性三者之间的权衡决定了整体集成性能。值得强调的是，多样性的定义与偏差/方差类似，涉及对 D 的期望，而非单次训练过程的属性。我们还强调，该结果适用于依赖训练和独立训练（其中 D 是由 m 随机变量构成的向量，详情见附录C.1），也适用于由 D 定义预先构建模型池的方案。

4.2 如果偏差-方差分解不成立，该怎么办?

定义 2 并不适用于所有损失函数，例如 0/1 损失或绝对损失。这给我们的框架带来了一个明显的挑战。我们依赖这种分解的存在，并据此将多样性推导为一个“隐藏维度”。不过，仍有解决之道。

量化偏差、方差和多样性的影响。James 和 Hastie（1997年）提出了一种适用于任意损失函数且保持简单加法形式的分解。他们通过区分方差的度量（它独立于标签）与其对期望风险的效应（它取决于标签），并对偏差定义了类似的一对量，来实现这一点。度量与效应不一定是同一个量。他们将定义 2 中的度量视为“自然”的形式。随后，他们定义了偏差效应[3], 和方差效应。为简化符号，我们对 $P(\mathbf{x}, \mathbf{y})$ 取平均。令 $R(\mathbf{q}) \overset{\text{定义为}}{=} \mathbb{E}_{\mathbf{X}Y}[\ell(\mathbf{y}, \mathbf{q})]$ 表示 $\mathbf{q}$ 的风险，则相关项为:

$$\text{bias-effect} \quad \overset{\text{def}}{=} R(\mathring{\mathbf{q}}) - R(\mathbf{y}^*), \tag{10}$$

$$\text{variance-effect} \quad \overset{\text{def}}{=} \mathbb{E}_D\Big[R(\mathbf{q}) - R(\mathring{\mathbf{q}})\Big], \tag{11}$$

其中 $\mathring{\mathbf{q}} \overset{\text{定义为}}{=} \arg\min_{\mathbf{z}\in\mathcal{Y}} \mathbb{E}_D\Big[\ell(\mathbf{z}, \mathbf{q})\Big]$，而$\mathbf{y}^* \overset{\text{定义为}}{=} \arg\min_{\mathbf{z}\in\mathcal{Y}} \mathbb{E}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}\Big[\ell(\mathbf{y}, \mathbf{z})\Big]$ 是在每个点$\mathbf{x}$处的贝叶斯最优预测器[4] 。这些量化了使用一个预测器而非另一个预测器对风险的影响。偏差效应是质心 $\mathring{\mathbf{q}}$ 与贝叶斯最优预测器 $\mathbf{y}^*$之间的风险变化。方差效应是模型 $\mathbf{q}$ 相对于质心的风险变化，并对 D 的分布取平均。对于满足定义 2 的损失函数，度量与效应相同，即方差效应等于方差，偏差效应等于偏差。例如，在平方损失和 $y \in \mathbb{R}$ 的情况下，公式（10）等于$\mathbb{E}_{\mathbf{X}Y}[(\mathbb{E}_D[q] - y^*)^2]$。对于一般损失函数（例如 0/1 损失、绝对损失），情况并非如此，此时效应项是不同的数值。根据这些定义，James 和 Hastie 给出了如下分解。

定理 6（偏差-方差效应分解，James 和 Hastie，1997年）对于损失 $\ell: \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, 记质心为 $\mathring{\mathbf{q}} \overset{\text{def}}{=} \arg\min_{\mathbf{z}\in\mathcal{Y}} \mathbb{E}_D[\ell(\mathbf{z},\mathbf{q})]$, , 期望风险可分解如下:

$$\mathbb{E}_D\Big[R(\mathbf{q})\Big] = \underbrace{R(\mathbf{y}^*)}_{\text{噪声}} + \underbrace{R(\mathring{\mathbf{q}}) - R(\mathbf{y}^*)}_{\text{偏差效应}} + \underbrace{\mathbb{E}_D[R(\mathbf{q}) - R(\mathring{\mathbf{q}})]}_{\text{方差效应}} \tag{12}$$

此分解适用于任何损失，因为右侧各项总会相互抵消，从而得到左侧。这里值得注意的是，方差效应取决于标签 y，而方差本身则是一个与标签无关的量。此外，偏差效应总是非负的，而方差效应则可能是一个负 值。由此，利用命题 3 中采用的相同关系，我们可以给出相应的歧义（效应）分解。

3. J&H 将其称为 *systematic* 效应。为保持全文术语一致，我们更倾向于使用 *bias*-效应这一术语。4. 我们承认这里对符号说明略有滥用：根据需要，将贝叶斯预测器 $\mathbf{y}^*$ 表示为每个点 $\mathbf{x}$ 处的函数，或表示为 $\mathbb{R}^k$中的向量。具体含义将根据上下文明确。

命题 7（歧义效应分解）给定损失 ℓ： $\mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$，以及一个集成 $\{\mathbf{q}_i\}_{i=1}^m$，它使用质心组合器 $\mathbf{q} \overset{\text{定义为}}{=} \arg\min_{\mathbf{z}\in\mathcal{Y}} \left[\frac{1}{m}\sum_{i=1}^m \ell(\mathbf{z},\mathbf{q}_i)\right]$，则有如下分解： —

$$\underbrace{R(\overline{\mathbf{q}})}_{\text{ensemble risk}} = \underbrace{\frac{1}{m}\sum_{i=1}^m R(\mathbf{q}_i)}_{\text{average risk}} - \underbrace{\frac{1}{m}\sum_{i=1}^m \left[R(\mathbf{q}_i) - R(\overline{\mathbf{q}})\right]}_{\text{ambiguity-effect}}. \tag{13}$$

我们再次强调，这适用于任何损失函数。这一点也相当显而易见，因为右侧的各项相互抵消后便得到左侧。对于满足定义 2 的损失函数，这一结果推广了前文所述的歧义分解。例如，当 ℓ 为平方损失时，它可化为 Krogh 和 Vedelsby（1994）的结果。更一般地说，它可化为定理 3，即广义歧义分解。

这个表达式的强大之处在于，我们也可以考虑不满足定义 2 的损失函数，例如 0/1 损失或绝对损失。对于 0/1 损失，质心就是随机变量的众数（Domingos，2000年）。对于由有限个集成成员组成的集合，这意味着 q 是集成中各个成员的多数投票结果。或者，对于绝对损失， q是集成成员预测的中位数 。

综上所述，我们现在可以提出一种新颖的分解，用于描述采用质心组合器的集成的期望风险，并分离多样性的影响 。

定理 8（偏差-方差-多样性效应分解）给定由质心组合器 $\mathbf{q} \overset{\text{def}}{=} \arg\min_{\mathbf{z}\in\mathcal{Y}} \left[\frac{1}{m}\sum_{i=1}^m \ell(\mathbf{z},\mathbf{q}_i)\right]$ 组合的模型集成 $\{\mathbf{q}_i\}_{i=1}^m$，使用任意损失 ℓ，该集成的期望风险可分解为，

$$\mathbb{E}_D\left[R(\overline{\mathbf{q}})\right] = \underbrace{R(\mathbf{y}^*)}_{\text{噪声}} + \underbrace{\frac{1}{m}\sum_{i=1}^m \left[R(\mathbf{q}_i) - R(\mathbf{y}^*)\right]}_{\text{平均偏差效应}} + \underbrace{\frac{1}{m}\sum_{i=1}^m \mathbb{E}_D\left[R(\mathbf{q}_i) - R(\mathbf{q}_i)\right]}_{\text{平均方差效应}} - \underbrace{\mathbb{E}_D\left[\frac{1}{m}\sum_{i=1}^m \left[R(\mathbf{q}_i) - R(\overline{\mathbf{q}})\right]\right]}_{\text{多样性效应}}$$

我们得到四项：噪声，以及平均偏差、平均方差和多样性的效应。请注意，这里我们考虑的每一项都包含对$P(\mathbf{x}, y)$的期望。对于满足定义 2 的损失函数，定理 8 可简化为定理 5，即偏差-方差-多样性分解。

与定理 5 中的多样性类似，这里的多样性效应与其他项的符号相反。因此，多样性效应越大，期望风险就越低。然而，多样性效应也可能取负值。这看起来或许有些反直觉，但如果我们将该项视为平均风险与集成风险之差，就更容易理解了。因此，当集成的表现较差于平均水平时，该值为负。

总之，我们得出了与上一小节相同的结论：多样性的作用可以表述为集成风险分解中的一个隐藏维度。所有这些项都可以从数据中估计——接下来我们将对此进行探讨。

4.3 从数据估计多样性

我们展示了一些说明性实验，从数据中估计偏差/方差/多样性项。我们强调，我们并未声称质心组合器优于任何其他组合器，只是用它来进行分解。更多实验见附录 A.1。

平方损失： 定理 5 的一个简单实例是平方损失，$\ell(y,\ q) = (y-q)^2$, 这意味着 $q = \frac{1}{m}\sum_{i=1}^{m} q_i$，且 $\mathring{q} = \mathbb{E}_D[q]$。在每个测试点（$\mathbf{x}\ y$）处的分解为，

$$\mathbb{E}_D\left[(q-y)^2\right] = \underbrace{\frac{1}{m}\sum_{i=1}^{m}(\mathring{q}_i - y)^2}_{\text{平均偏差}} + \underbrace{\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_D\left[(q_i-\mathring{q}_i)^2\right]}_{\text{平均方差}} - \underbrace{\mathbb{E}_D\left[\frac{1}{m}\sum_{i=1}^{m}(q_i-q)^2\right]}_{\text{多样性}}.$$

图5展示了一个使用装袋回归树并改变集成规模的实验。

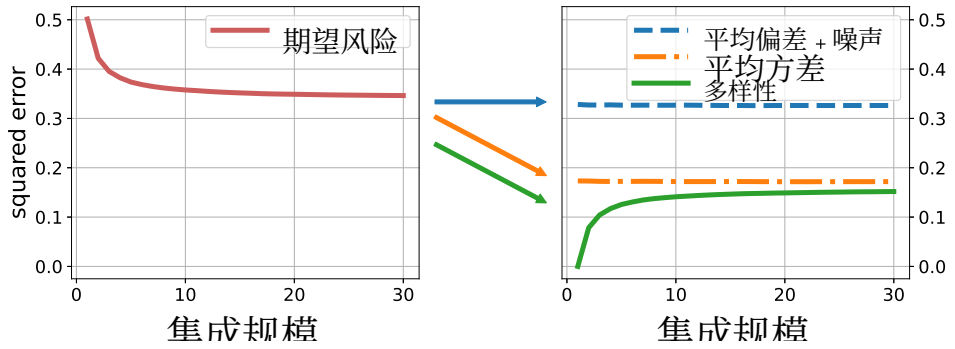


图5：对深度为 8 的决策树使用装袋法，逐渐增大集成规模（加利福尼亚住房数据）。

虽然期望风险有所降低，但平均偏差和平均方差保持不变。这正如我们所预料的，因为个体容量保持不变，唯一改变的是集成规模 m。相比之下，多样性随 增加——从期望风险中扣减——而改进完全由多样性决定。如果改变的不是 m，情况就并非如此。图6固定 $m = 10$，改变树深度——此时三个组成部分都会发生变化，于是形成了偏差-方差-多样性权衡。

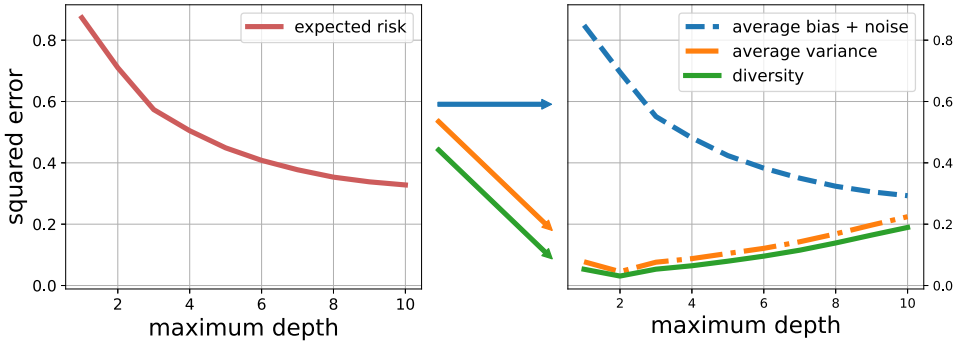


图6：对 $m = 10$ 回归树使用装袋法，逐渐增大最大深度。

交叉熵损失：我们还可以考虑定理 5 中概率向量之间的 KL散度，其中质心组合器$\mathbf{q}$ 是归一化几何平均。对于单个（$\mathbf{x}$，$\mathbf{y}$），其中$\mathbf{y}$ 是独热向量，这给出了交叉熵的分解，因为对于任意$\mathbf{q}$，我们可以写作 $K(\mathbf{y} \,||\, \mathbf{q}) = -\mathbf{y} \cdot \ln \mathbf{q}$，这里采用约定$\mathbf{y} \cdot \ln \mathbf{y} = 0$。

推论 9 取一个测试点 $(\mathbf{x},\mathbf{y})$，其中 $\mathbf{y}$ 是一个独热向量。考虑一个集成$\{\mathbf{q}\}_{i=1}^m$，其中每个模型都预测一个分布，并将它们组合为 $\mathbf{q} := Z^{-1} \prod_i \mathbf{q}_i^{1/m}$。我们有，

$$\underbrace{-\mathbb{E}_D\left[\mathbf{y} \cdot \ln \overline{\mathbf{q}}\right]}_{\text{expected cross-entropy}} = \underbrace{-\frac{1}{m}\sum_{i=1}^m \mathbf{y} \cdot \ln \mathring{\mathbf{q}}_i}_{\text{average bias}} + \underbrace{\frac{1}{m}\sum_{i=1}^m \mathbb{E}_D\left[K(\mathring{\mathbf{q}}_i \,||\, \mathbf{q}_i)\right]}_{\text{average variance}} - \underbrace{\mathbb{E}_D\left[\frac{1}{m}\sum_{i=1}^m K(\overline{\mathbf{q}} \,||\, \mathbf{q}_i)\right]}_{\text{diversity}}.$$

对于神经网络集成，归一化几何平均等价于对网络 logits 求平均，这是一种常见做法（Hinton 等人，2015年），但有利也有弊（Tassi 等人，2022年）。它也称为“乘积规则”（Kuncheva，2014年），并受到广泛研究；研究表明，根据各种因素，其性能可能优于或劣于算术平均值（Tax 等人，1997年）。无论性能如何，要以这种方式分解交叉熵，就必须 使用归一化几何平均。在第 5 节中，我们将探讨使用算术平均值时分解会发生哪些变化。

在图 7 中，我们比较了在 MNIST 上使用装袋法训练的多层感知机：一类是每个网络均只有一层、包含 20 个隐藏节点的小型网络，另一类是每个网络均包含 100 个隐藏节点的较大网络。我们观察到的模式与平方损失下的情况非常相似。

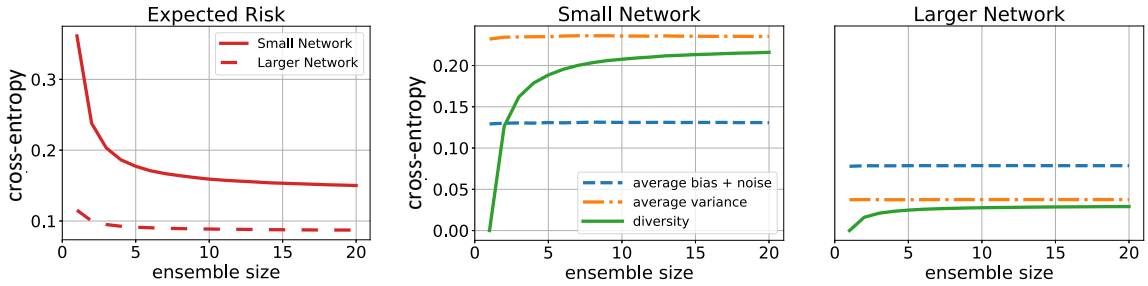


图 7：分解装袋法训练的小型/较大多层感知机的期望集成交叉熵。

总体而言，较大网络集成的表现更好：再次通过考察期望风险分量即可解释这一点。从中间面板可以看到，较小网络集成的多样性显著更高（绿色曲线）。然而，在最右侧面板中，我们可以看到，较大网络的平均偏差/方差显著更低[5]（蓝色/橙色曲线）。这使其在偏差/方差/多样性权衡中占据显著优势。总体而言，小型网络较高的多样性，无法抵消较大网络集成中更强个体带来的优势。

5. 较低的方差与机器学习领域的传统观点相悖；该观点认为，提高模型容量也应增加方差。但这一结果与近期的观察一致（Yang 等人，2020年）。

0/1 损失与多数投票集成：最常见的集成方案之一是多数投票，并使用 0/1 损失进行评估（Kuncheva，2003年；Didaci 等人，2013年）。在这里，定理 5 不成立。因此，我们采用定理 8 的效应分解，而质心组合器 q 即为多数投票。对于从 D 中抽取的可能数据集，使用 $\mathbb{P}(q \neq y) \overset{\text{定义}}{=} \mathbb{E}_D[\ell_{0/1}(y,\ q)]$ 表示集成误差的概率，则在每个测试点（$\mathbf{x}$ y）处有：

$$\mathbb{P}(\bar{q} \neq y) =$$
$$\underbrace{\frac{1}{m}\sum_{i=1}^{m}\Big[\ell_{0/1}(\mathring{q}_i, y)\Big]}_{\text{average bias-effect}} + \underbrace{\frac{1}{m}\sum_{i=1}^{m}\Big[\mathbb{P}(q_i \neq y) - \ell_{0/1}(\mathring{q}_i, y)\Big]}_{\text{average variance-effect}} - \underbrace{\frac{1}{m}\sum_{i=1}^{m}\Big[\mathbb{P}(q_i \neq y) - \mathbb{P}(\bar{q} \neq y)\Big]}_{\text{diversity-effect}}.$$

请注意，平均偏差效应就是质心 $\mathring{q}_i$ 不正确的集成成员所占的比例。多样性效应是个体平均错误概率与集成错误概率之差。人们或许会尝试重新表述多样性效应，寻找一种不涉及标签的改写方式。事实上，对于 0/1 损失，这是不可能的：不仅多数投票如此，任何 组合规则都如此。

定理 10（0/1 损失中与标签无关的多样性效应不存在性）对于 *0/1* 损失，无论采用何种集成组合规则，个体风险的平均值与集成风险之差都必然取决于标签。

因此，多数投票集成中的多样性效应必然取决于标签，这是 0/1 损失不可避免的性质。现在，我们比较 MNIST（10 个类别）上的装袋法和随机森林。在下方最左侧的面板中，可以看到随机森林起初表现较差，但随着集成规模增大，其表现超过了装袋法。

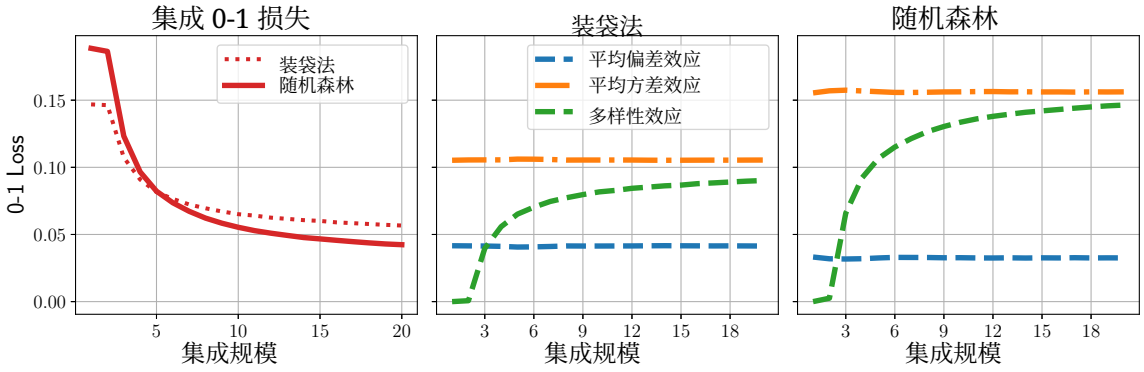


图 8：装袋法与随机森林的偏差/方差/多样性效应。

在右侧面板中，我们分解了期望风险，并观察到一些熟悉的模式。随着 m增加，偏差/方差效应保持不变，而多样性效应则增加。随机森林的方差效应远高于装袋法，但在规模较大的集成中，更高的多样性效应弥补了这一点。我们还可以看到，随机森林的平均偏差效应略低，这一点我们归因于决策树能够避开噪声特征。

定理 8 可以相对容易地扩展到加权投票——如附录C.2所示。在图9中，我们绘制了装袋法和两种不同 Boosting 算法的集成 0/1风险分量，并以决策桩作为基学习器。这里我们采用 Mease and Wyner (2008) 使用的一个合成数据集来分析提升算法。

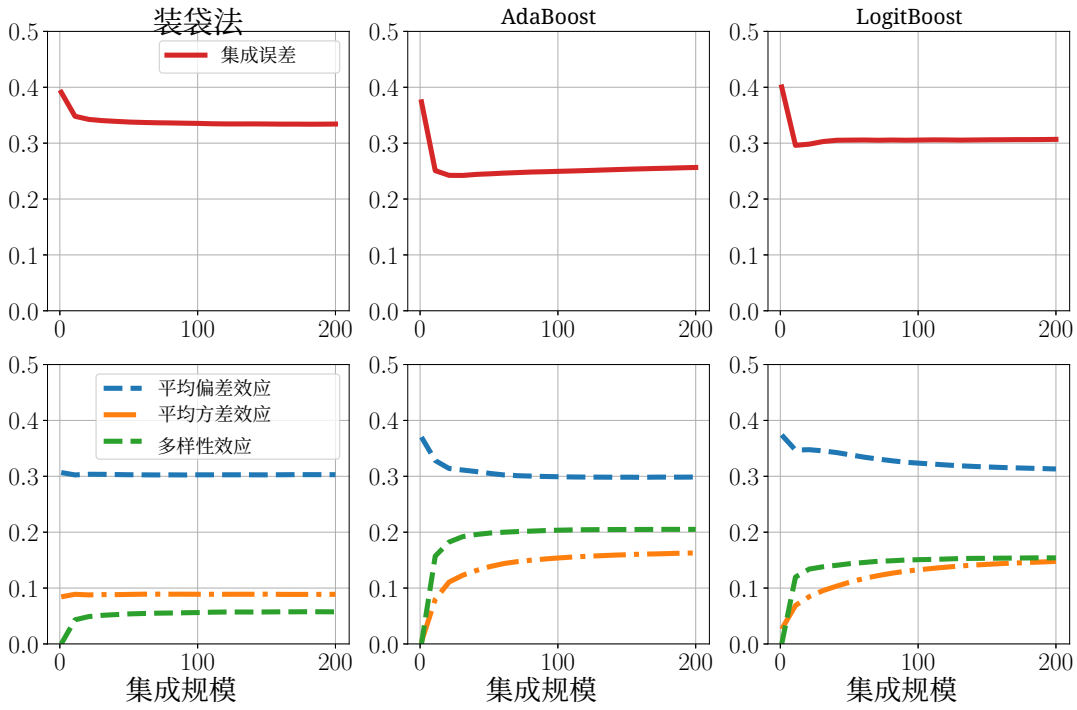


图9：Mease数据，集成决策桩。

我们注意到，并行模型（装袋法）和序列模型（LogitBoost/AdaBoost）之间存在一些有趣的差异。与之前一样，装袋法的偏差/方差保持不变。然而，对于提升法，它们会随 m 变化。这是因为各个个体具有非同质性，专门适应数据的不同部分。对于 0/1 损失，多样性效应可能大于 方差效应；相比之下，多样性本身则受平均方差的上界限制（见第 5 节）。更广泛地说，装袋法的性能来自多样性的提升。而提升法的情况则更为微妙，其性能取决于各个组成部分之间复杂的权衡。

4.4 总结

我们提出了一个用于理解多样性的框架——将其解释为模型拟合程度的一种度量，其含义与偏差/方差完全相同，但符号相反。因此，预期集成风险由偏差/方差/多样性三者之间的权衡决定。当无法进行加性偏差-方差分解时，我们展示了如何转而分离偏差/方差和多样性的影响。在第 5 节中，我们将该框架应用于*Bregman*散度（Bregman，1967年）。这将为我们提供一种便利的解析形式以及额外的洞见：尤其是对质心组合规则的更深入理解。

5 使用布雷格曼散度理解多样性

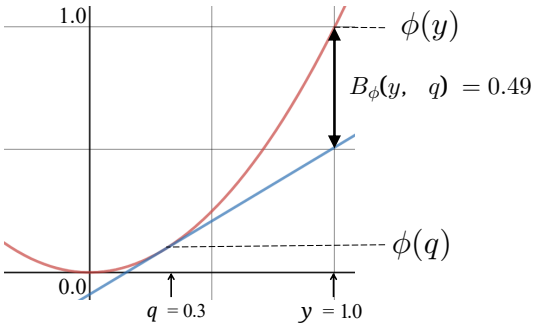
我们将该框架应用于布雷格曼散度（Bregman，1967年）这一族，其中包括许多作为特殊情形的损失函数，并使我们能够推导出若干有趣的性质。

5.1 布雷格曼散度基础

布雷格曼散度根据生成函数 ϕ定义。设 $\phi: \mathcal{Y} \to \mathbb{R}$ 是定义在凸集 $\mathcal{Y} \subseteq \mathbb{R}^k$上的严格凸函数，且 ϕ 在 ri($\mathcal{Y}$)——即 $\mathcal{Y}$的相对内部——上可微。布雷格曼散度 $B_\phi: \mathcal{Y} \times \text{ri}(\mathcal{Y}) \to \mathbb{R}_+$ 定义如下：

$$B_\phi(\mathbf{y}, \mathbf{q}) \stackrel{\text{def}}{=} \phi(\mathbf{y}) - \phi(\mathbf{q}) - \langle \nabla\phi(\mathbf{q}), (\mathbf{y} - \mathbf{q}) \rangle, \tag{14}$$

其中 $\langle \cdot, \cdot \rangle$ 表示内积，$\nabla\phi(\mathbf{q})$ 表示 ϕ 在 $\mathbf{q}$处的梯度向量。选择 ϕ 会得到许多不同的损失函数。当 $\phi(q) = q^2$时，梯度向量 $\nabla\phi(\mathbf{q})$ 是标量导数 $\mathrm{d}\phi(q)/\mathrm{d}q = 2q$，由此可得平方损失，$B_\phi(y, q) = (y - q)^2$。



$$\begin{aligned} B_\phi(y, q) &= \phi(y) - \phi(q) - \langle \nabla\phi(q), (y - q) \rangle \\ &= y^2 - q^2 - \langle 2q, (y - q) \rangle \\ &= y^2 - q^2 - 2yq + 2q^2 \\ &= y^2 + q^2 - 2yq \\ &= (y - q)^2 \end{aligned}$$

图10：生成函数为 $\phi(q) = q^2$时的布雷格曼散度示意图。在此例中，散度为 $B_\phi(y, q) = (y - q)^2 = (1.0 - 0.3)^2 = 0.49$。

或者，对于一个 k 类问题，我们可以取一个向量$\mathbf{q} \in \mathbb{R}^{k-1}$。请注意，这不是一个总和为 1 的概率向量。不过，它是对该分布的最简描述，因为第 k 类的概率为 $1 - \sum_c \mathbf{q}^{(c)}$。采用一个特定的生成函数（见表2最后一行），我们便可得到 $\mathbb{R}^k$ 中分布之间的 KL散度。

损失函数	$B_\phi(\mathbf{y},\mathbf{q})$	**生成函数** $\phi(\mathbf{q})$	**定义域** $\mathcal{Y}$
平方损失	$(y - q)^2$	q^2	$q \in \mathbb{R}$
伊丹–斋藤散度	$\frac{y}{q} - \ln\frac{y}{q} - 1$	$-\ln q$	$q \in \mathbb{R}^+$
泊松损失	$y \ln\frac{y}{q} - (y - q)$	$q \ln q - q$	$q \in \mathbb{R}^+$
KL散度	$\sum \mathbf{y}^{(c)} \ln \frac{\mathbf{y}^{(c)}}{\mathbf{q}^{(c)}} + \left(1 - \sum \mathbf{y}^{(c)}\right) \ln \frac{1 - \sum \mathbf{y}^{(c)}}{1 - \sum \mathbf{q}^{(c)}}$	$\sum \mathbf{q}^{(c)} \ln \mathbf{q}^{(c)} + \left(1 - \sum \mathbf{q}^{(c)}\right) \ln \left(1 - \sum \mathbf{q}^{(c)}\right)$	$\mathbf{q} \in [0, 1]^{k-1}$ $s.t. \sum_c \mathbf{q}^{(c)} \leq 1$

表2：常见损失函数及其布雷格曼生成函数。

5.2 将我们的框架应用于布雷格曼散度族

如第 4 节所述，我们要求存在一种符合定义 2 形式的偏差-方差分解。对于 Bregman散度，Pfau（2013年）证明了以下结论：

$$\underbrace{\mathbb{E}_D\Big[\mathbb{E}_{\mathbf{XY}}[B_\phi\left(\mathbf{Y},\mathbf{q}\right)]\Big]}_{\text{expected ensemble risk}} = \mathbb{E}_{\mathbf{X}}\Big[\underbrace{\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[B_\phi\left(\mathbf{Y},\mathbf{Y}^*\right)]}_{\text{noise}} + \underbrace{B_\phi\left(\mathbf{Y}^*,\mathring{\mathbf{q}}\right)}_{\text{bias}} + \underbrace{\mathbb{E}_D\left[B_\phi\left(\mathring{\mathbf{q}},\mathbf{q}\right)\right]}_{\text{variance}}\Big], \tag{15}$$

其中质心 $\mathring{\mathbf{q}}$ 可以方便地以闭式形式求得，

$$\mathring{\mathbf{q}} \stackrel{\text{def}}{=} \arg\min_{\mathbf{z}\in\mathcal{Y}} \mathbb{E}_D\Big[B_\phi\left(\mathbf{z},\mathbf{q}\right)\Big] = [\nabla\phi]^{-1}\Big(\mathbb{E}_D\left[\nabla\phi\left(\mathbf{q}\right)\right]\Big). \tag{16}$$

这与定义 2 相符。如果 $\phi(q) = q^2$，那么 $B_\phi(y,q) = (y-q)^2$，且 $\mathring{q} = \mathbb{E}_D[q]$，这意味着整个表达式恰好就是 Geman 等人（1992年）所给出的形式。对于其他生成函数/损失函数，该表达式对应于其他偏差-方差分解，例如 Heskes（1998年）的分解。基于这一结果，利用第 4 节中提出的框架，我们可以推导出布雷格曼版本的歧义分解、质心组合器以及偏差-方差-多样性分解。

定理 11（布雷格曼歧义分解）对于标签 $\mathbf{y} \in \mathcal{Y}$ 以及一组预测 $\mathbf{q}_1、\ldots、\mathbf{q}_m \in \mathrm{ri}(\mathcal{Y})$，将其组合为 $\overline{\mathbf{q}} = [\nabla\phi]^{-1}\big(\frac{1}{m}\sum_{i=1}^m \nabla\phi(\mathbf{q}_i)\big)$。则有：

$$B_\phi\left(\mathbf{y},\overline{\mathbf{q}}\right) = \frac{1}{m}\sum_{i=1}^m B_\phi\left(\mathbf{y},\mathbf{q}_i\right) - \frac{1}{m}\sum_{i=1}^m B_\phi\left(\overline{\mathbf{q}},\mathbf{q}_i\right). \tag{17}$$

定义 12（布雷格曼质心组合器）布雷格曼质心组合器是使所有成员的平均散度最小化的组合器。对于集成 $\{\mathbf{q}_i\}_{i=1}^m$，它为

$$\overline{\mathbf{q}} \stackrel{\text{def}}{=} \arg\min_{\mathbf{z}\in\mathcal{Y}} \frac{1}{m}\sum_{i=1}^m B_\phi\left(\mathbf{z},\mathbf{q}_i\right) = [\nabla\phi]^{-1}\Big(\frac{1}{m}\sum_{i=1}^m \nabla\phi(\mathbf{q}_i)\Big). \tag{18}$$

定理13（布雷格曼偏差-方差-多样性分解）
对于模型集成 $\{\mathbf{q}_i\}_{i=1}^m$，令 $\mathring{\mathbf{q}}_i$ 为模型 $\mathbf{q}_i$的左布雷格曼质心，即 $\mathring{\mathbf{q}}_i \overset{\text{定义}}{=} [\nabla\phi]^{-1}\left(\mathbb{E}_D\left[\nabla\phi(\mathbf{q}_i)\right]\right)$，并定义集成输出 $\overline{\mathbf{q}} \overset{\text{定义}}{=} [\nabla\phi]^{-1}\big(\frac{1}{m}\sum_{i=1}^m \nabla\phi(\mathbf{q}_i)\big)$。于是有如下分解

$$\mathbb{E}_D\left[\mathbb{E}_{\mathbf{XY}}\left[B_\phi\left(\mathbf{Y},\overline{\mathbf{q}}\right)\right]\right] =$$
$$\mathbb{E}_{\mathbf{X}}\Bigg[\underbrace{\mathbb{E}_{\mathbf{Y}|\mathbf{X}}B_\phi\left(\mathbf{Y},\mathbf{Y}^*\right)}_{\text{noise}} + \underbrace{\frac{1}{m}\sum_{i=1}^m B_\phi\left(\mathbf{Y}^*,\mathring{\mathbf{q}}_i\right)}_{\text{average bias}} + \underbrace{\frac{1}{m}\sum_{i=1}^m \mathbb{E}_D\left[B_\phi\left(\mathring{\mathbf{q}}_i,\mathbf{q}_i\right)\right]}_{\text{average variance}} - \underbrace{\mathbb{E}_D\Big[\frac{1}{m}\sum_{i=1}^m B_\phi\left(\overline{\mathbf{q}},\mathbf{q}_i\right)\Big]}_{\text{diversity}}\Bigg],$$

其中 $\mathbf{Y}^* = \mathbb{E}_{\mathbf{Y}|\mathbf{X}}[\mathbf{Y}]$。

不同损失（即不同 Bregman 生成元）的样本示例见表 3。这清楚地表明，多样性的数学形式取决于损失函数。

$$\left. - \overline{q}\right)^2 \Big]$$

$$\frac{\overline{q}}{q_i} + (1-\overline{q}) \ln \frac{1-\overline{q}}{1-q_i} \Bigg]$$

$$\overline{\mathbf{q}} \mid\mid \mathbf{q}_i) \Bigg]$$

$$\frac{\overline{q}}{q_i} - \ln \frac{\overline{q}}{q_i} - 1 \Bigg) \Bigg]$$

$$- \prod_{i=1}^{m} q_i^{1/m} \Bigg]$$

$p(\mathbf{x})$ is omitted and

5.3 布雷格曼质心组合器的性质

质心组合器源于一项约束：要求集成组合$\mathbf{q}$ 与质心 °$\mathbf{q}$具有相同的解析形式，正如相关损失的偏差-方差分解所示。对于布雷格曼散度，这种形式（定义 12）在信息几何文献中被称为左布雷格曼质心（Nielsen 和 Nock，2009）。一般而言，这些形式最终都是准算术平均值，其中包括若干广为人知的集成组合规则，表4列出了其中一些。

损失函数	**质心组合器**	**Name**
平方损失	$\frac{1}{m}\sum_{i=1}^{m} q_i$	算术平均值
泊松回归损失	$\prod_{i=1}^{m} q_i^{\frac{1}{m}}$	几何平均数
KL散度	$Z^{-1}\prod_{i=1}^{m}\left(\mathbf{q}_i^{(c)}\right)^{\frac{1}{m}}$	归一化几何平均数
Itakura-Saito损失	$1/\left(\frac{1}{m}\sum_{i=1}^{m}\frac{1}{q_i}\right)$	调和平均数

表4：各种损失函数的质心组合器（即集成的左布雷格曼质心）。

对偶坐标系中的集成平均：质心组合器可以理解为一种在新坐标系中进行的集成平均操作。坐标系之间的映射由 Bregman 生成函数关于其自变量的梯度定义， $\nabla\phi(\mathbf{q})$。图 11 以 KL散度为例对此进行了说明。

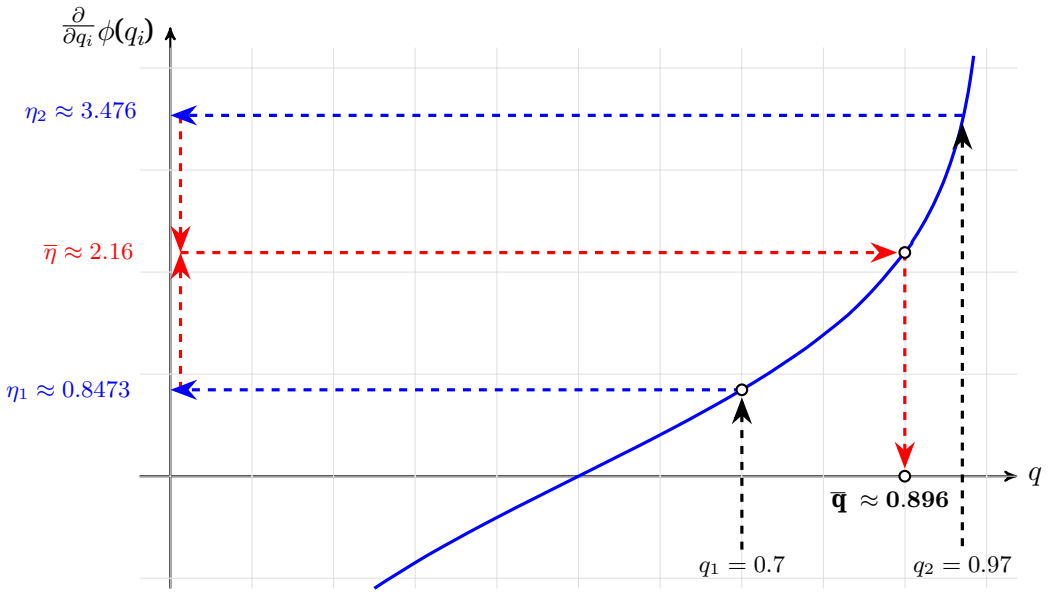


图 11：KL散度定义的几何中的集成平均 .

我们用符号 q 表示原坐标系，用 η 表示对偶坐标系。在这个简单的示例中，我们要预测单个概率 $p \in (0, 1)$。原对偶映射是梯度 $\eta_i = \frac{\partial}{\partial q_i}\phi(q_i) = \ln \frac{q_i}{1-q_i}$，如蓝色曲线所示。原坐标系中的两个点 $\{q_1 = 0.7, q_2 = 0.97\}$被映射到对偶坐标系 $\{\eta_1 \approx 0.8473, \eta_2 \approx 3.476\}$，然后通过算术平均值（$\eta \approx 2.16$）进行组合，最后通过逆运算 $q = \exp(\eta)/(1 + \exp(\eta)) \approx 0.896$映射回来。因此，质心组合器是在对偶坐标系中的算术平均集成，这等价于原坐标系中各模型的左布雷格曼质心。Gupta 等人（2022年）在模型独立同分布的假设下考察了一个等价定义。我们的分析不仅补充并拓展了这一研究，还消除了独立同分布假设，并更全面地刻画了采用这一组合规则的集成的性质。

概率估计的 KL散度示例：对于 KL散度，质心组合器在原坐标系（即概率单纯形）中是归一化的[6] 几何平均数。图 12 展示了 3 类情形。

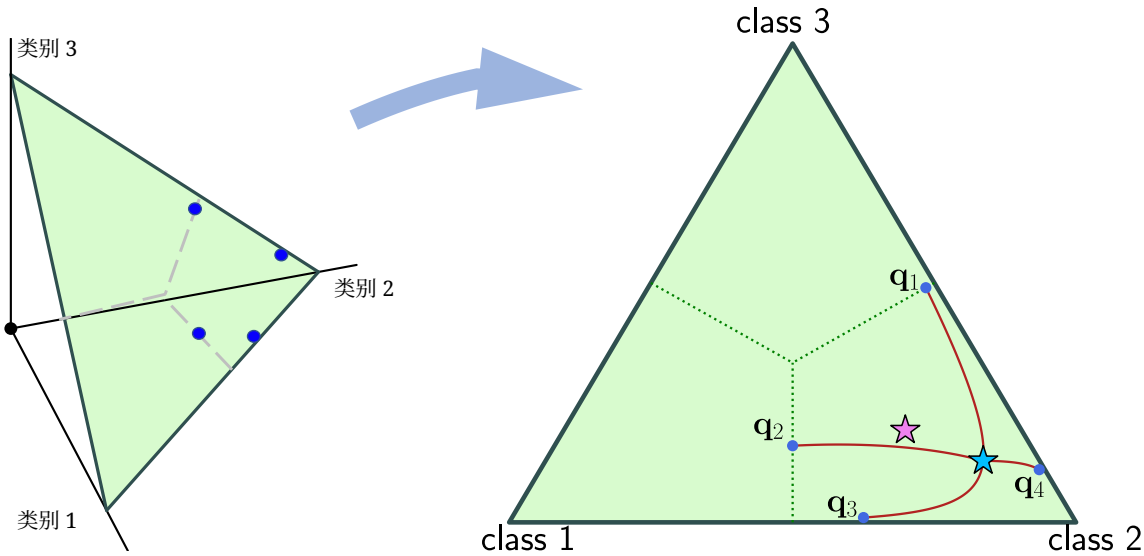


图 12：在概率单纯形中组合 $m = 4$ 预测。

质心组合器以蓝色星形标记表示（也与 4 个单个点相连），而算术平均值以粉色星形标记表示，它是这些点的质心。对于平方损失，所得的布雷格曼几何是欧几里得几何，这意味着质心和质心位置相同。然而，其他散度定义了非欧几里得几何，这意味着质心和质心位置并不相同。请注意，在单纯形中，各点并非通过直线相连，而是通过由 ϕ 的选择所定义的对偶测地线相连（Nielsen 和 Nock，2009）。不同的 ϕ （因此也意味着不同的损失）会产生不同的几何，并意味着不同的集成组合。

6. 请注意，布雷格曼质心仅是一个归一化分布，其 ϕ 如表2所示——参见附录 B.4。

5.4 分解的性质

我们进一步探讨偏差-方差-多样性分解的性质。

5.4.1 Homogeneous vs Heterogeneous Ensembles

集成由一组模型 $\{\mathbf{q}_i\}_{i=1}^m$组成，但也可以将其视为一个独立的"大模型"。因此，我们可以讨论集成本身的偏差/方差。这些偏差/方差项可以通过下述定理与单个模型的偏差/方差联系起来。

定理14 集成偏差和集成方差可以改写为单个模型的函数：

$$\underbrace{B_\phi\left(\mathbf{y}, \mathring{\overline{\mathbf{q}}}\right)}_{\text{ensemble bias}} = \underbrace{\frac{1}{m}\sum_{i=1}^{m} B_\phi\left(\mathbf{y}, \mathring{\mathbf{q}}_i\right)}_{\text{average bias}} - \Delta, \tag{19}$$

$$\underbrace{\mathbb{E}_D\left[B_\phi\left(\mathring{\overline{\mathbf{q}}}, \overline{\mathbf{q}}\right)\right]}_{\text{ensemble variance}} = \Delta + \underbrace{\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_D\left[B_\phi\left(\mathring{\mathbf{q}}_i, \mathbf{q}_i\right)\right]}_{\text{average variance}} - \underbrace{\mathbb{E}_D\left[\frac{1}{m}\sum_{i=1}^{m} B_\phi\left(\overline{\mathbf{q}}, \mathbf{q}_i\right)\right]}_{\text{diversity}}, \tag{20}$$

其中的公共项为 $\Delta = \frac{1}{m}\sum_{i=1}^{m} B_\phi\left(\mathring{\overline{\mathbf{q}}}\ \mathring{\mathbf{q}}_i\right)$。

这里有趣的一项是 Δ，我们称之为模型的"差异度"。它反映了集成中模型族之间的多样性。如果所有单个模型都属于同一模型族（例如，都是采用某种特定架构的神经网络），那么它们预测结果的差异就只由随机变量 D 引起。在这种情况下，它们的质心相同 $\mathring{\mathbf{q}}_i = \mathring{\mathbf{q}}_j = \mathring{\mathbf{q}}$，差异度为零。反之，如果它们具有不同的归纳偏差，质心就不一定相同，差异度也可能不为零。这类集成通常称为同质集成或异质集成。

偏差降低? 在同质集成中，$\Delta = 0$，且集成偏差等于平均值，也就是说，与单个模型相比，偏差并未降低。在包含差异足够大的模型族的异质集成中，$\Delta > 0$，偏差总会降低。

方差降低? 集成通常被称为"方差降低"方法。从公式（20）可以看出，对于任何布雷格曼散度，在同质集成中（$\Delta = 0$），集成方差必然小于平均值，也就是说，只要使用质心组合规则，方差降低这一性质就得以保留。从同一角度来看，多样性有一个由平均方差给出的上界。集成方差降低的幅度恰好等于集成的多样性。对于异质集成，情况则不那么简单，因为其中既有差异项的增加，也有多样性项的减少。

上述性质适用于任意布雷格曼散度。实际上，针对某个学习问题，我们会选择特定的散度——现在来讨论两种常见损失函数的具体性质：交叉熵和平方损失。

5.4.2 Cross Entropy Loss: Averaging estimates of class probabilities

组合类别概率估计时，一种非常流行的策略是取它们的算术平均值，例如 Lakshminarayanan 等人（2017）的方法；但是，如果使用交叉熵，这并非质心组合器。我们可能会想知道这会产生什么影响。下面的命题表明，集成的交叉熵损失仍保证小于其成员损失的平均值，但歧义会依赖于标签。

命题 15 假设真实概率向量为 $\mathbf{y} \in \mathbb{R}^k$，一组模型为 $\{\mathbf{q}_i\}_{i=1}^m$，并通过平均进行组合，即 $\mathbf{q}^\dagger = \frac{1}{m}\sum_{i=1}^m \mathbf{q}_i$，则交叉熵损失为 —

$$\underbrace{-\mathbf{y}\cdot\ln\mathbf{q}^\dagger}_{\text{ensemble cross-entropy}} = \underbrace{-\frac{1}{m}\sum_{i=1}^m \mathbf{y}\cdot\ln\mathbf{q}_i}_{\text{average cross-entropy}} - \underbrace{\sum_{c=1}^k \mathbf{y}^{(c)}\ln\frac{\frac{1}{m}\sum_{j=1}^m \mathbf{q}_j^{(c)}}{\left(\prod_{i=1}^m \mathbf{q}_i^{(c)}\right)^{\frac{1}{m}}}}_{\text{ambiguity (label-dependent)}}, \tag{21}$$

其中第二项非负，因此集成损失保证小于或等于各个模型损失的平均值。

通过求 $-\mathbf{y}\cdot\ln\mathbf{q}^\dagger$ 与 $-\frac{1}{m}\sum_{i=1}^m \mathbf{y}\cdot\ln\mathbf{q}_i$ 两项之差，无需借助我们迄今提出的框架，也能观察到这一性质。事实上，对于两个类别的情况，Ivaşcu 等人（2021年）曾独立观察到这一点。利用上述一般情况，我们可以将其与第 4 节的方法相结合，得到一个类似于偏差-方差-多样性分解的结果。

Pro位置 16（平均概率的多样性取决于标签）
令$\mathbf{q}^\dagger = \frac{1}{m}\sum_{i=1}^m \mathbf{q}_i$，其中$\mathbf{q}_i \in [0,1]^k$。期望交叉熵可分解为：

$$-\mathbb{E}_D\left[\mathbf{y}\cdot\ln\mathbf{q}^\dagger\right] =$$

$$\underbrace{-\frac{1}{m}\sum_{i=1}^m \mathbf{y}\cdot\ln\mathbf{q}_i}_{\text{平均偏差}} + \underbrace{\frac{1}{m}\sum_{i=1}^m \mathbb{E}_D\left[K(\mathbf{q}_i \,||\, \mathbf{q}_i)\right]}_{\text{平均方差}} - \underbrace{\mathbb{E}_D\left[\sum_{c=1}^k \mathbf{y}^{(c)}\ln\frac{\frac{1}{m}\sum_{j=1}^m \mathbf{q}_j^{(c)}}{\left(\prod_{i=1}^m \mathbf{q}_i^{(c)}\right)^{\frac{1}{m}}}\right]}_{\text{依赖项}}.$$

这与定理 9 非常相似，只不过这里的最后一项依赖于标签$\mathbf{y}$。为区分两者，我们避免使用“多样性”这一名称，而将其称为“依赖项”。

Gupta 等人（2022年）还研究了采用算术平均组合器时集成偏差/方差的性质，并表明（在独立同分布模型假设下），集成方差总是会降低。同时，他们也提出了一点担忧：这可能会增加集成偏差（使其高于平均偏差），具体取决于标签分布。我们的命题进一步揭示：总体期望损失总是低于平均值。因此，即使集成偏差有所增加，集成方差的降低也总能充分抵消这种增加，从而使总体期望损失降低。

5.4.3 Relation to the Bias-Variance-Covariance Decomposition

对于平方损失，我们的分解可以与 Ueda 和 Nakano（1996年）的偏差-方差-协方差分解进行对比。该分解指出，对于点（$\mathbf{x}$ y）：

$$\mathbb{E}_D\left[\ (\overline{q}-y)^2\ \right]= \tag{22}$$
$$\underbrace{(\mathbb{E}_D\left[\overline{q}\right]-y)^2}_{\text{bias}(\overline{q})}+\underbrace{\frac{1}{m}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_D\Big[(q_i-\mathbb{E}_D\left[q_i\right])^2\Big]}_{1/m\ \times\ \overline{\text{variance}}}+\underbrace{\frac{1}{m^2}\sum_{i,j}\mathbb{E}_D\Big[(q_i-\mathbb{E}_D\left[q_i\right])(q_j-\mathbb{E}_D\left[q_j\right])\Big]}_{(1-1/m)\ \times\ \overline{\text{covariance}}}.$$

与我们的定理13相比，可以看出偏差组成部分存在差异。我们使用的是个体平均偏差，而 Ueda 和 Nakano 使用的是集成偏差:

$$\overline{\text{bias}} \ = \ \frac{1}{m}\sum_{i=1}^{m}(\mathbb{E}_D\left[q_i\right]-y)^2, \tag{23}$$
$$\text{bias}(\overline{q}) \ = \ (\mathbb{E}_D\left[\overline{q}\right]-y)^2. \tag{24}$$

Ueda 和 Nakano 注意到，他们的项可以改写为：$(\mathbb{E}_D\left[q\right]=y)^2=(\ \frac{1}{m}\sum_{i=1}^{m}[\mathbb{E}_D\left[q_i\right]-y])^2$,并将其描述为“平均偏差的平方”。然而，我们提醒读者，平方项是使用平方损失函数所产生的人为结果。在偏差-方差分解的广义形式中并不存在这个平方项——参见附录 B.1。因此，$(\mathbb{E}_D\left[q\right]-y)^2$应被简单地称为“集成的偏差”。偏差与偏差($q$)之间的差异本身就是一个有趣的量——如果针对平方损失将式（19）具体化，我们会看到偏差(q) $=$ 偏差 $-\ \Delta$。这表明，他们的项实际上由两个部分组成：个体偏差的平均值，以及上文引入的差异项。当然，如果模型是同质的，即属于同一模型族，那么差异为零，并且偏差(q) $=$偏差。

Ueda 和 Nakano 的分解依赖于随机变量线性组合的一个性质： $Var(aX_1+bX_2)$ $=a^2Var(X_1)+b^2Var(X_2)+2abCov(X_1,\ X_2)$，其中 X_1，X_2表示两个模型输出，aX_1+bX_2 表示集成组合。当集成组合规则是非线性的 （即不是简单的算术平均值）时，这一性质（以及由此作为理解多样性的途径）便不再适用。

值得注意的是，协方差既可以为正，也可以为负。然而，我们的多样性项始终非负，并且会随着集成决策周围分歧的增加而增大。协方差本质上是一种成对 计算——这种形式很可能启发了许多已发表的成对 多样性度量（Kuncheva，2014年）。布雷格曼情形下的多样性以非成对 的方式表示——即集成预测周围的期望平均偏差。我们推测，这一项无法仅通过成对运算来表示，这意味着成对度量可能存在根本性局限。

5.5 总结

我们将第 4 节中提出的框架应用于广泛的布雷格曼散度。这类散度具有特别便于分析的形式，使我们能够推导出若干有趣的性质。

6 相关工作讨论

We 讨论相关文献，并概述一些可能的未来研究方向 .

6.1 多样性的多种形式

关于集成多样性，已有许多 度量方法。Kuncheva（2014，第 8.5 节）将这些度量方法分为三类：（1）将多样性视为分类器的特征，不考虑组合器或真实标签；（2）将多样性视为分类器和组合器的特征，但不 使用真实标签；（3）将多样性视为分类器、组合器以及真实标签的特征。对于满足定义 2 的损失函数（例如布雷格曼散度），我们的框架显然支持第二种观点；对于不满足定义 2 的损失函数（例如 0/1 损失、绝对损失），则支持第三种观点。

6.2 多数投票集成中的良性与恶性多样性

对于 0/1 损失，定理 5 不适用。因此，我们必须使用定理 8 中的效应 分解。提醒一下，我们仍以与其他损失函数相同的形式度量 多样性：$\mathbb{E}_D[\frac{1}{m}\sum_{i=1}^{m}\ell_{0/1}(q,\ q_i)]$。不过，这一量对集成风险产生的效应 由多样性效应 项给出。这与“良性”和“恶性”多样性的概念有关（ Brown 和 Kuncheva，2010年）。

“有益”与“有害”多样性。 Brown 和 Kuncheva（2010年）指出，在将标签限制为 $y \in \{-1,+1\}$，并将 q 作为多数投票的情况下，以下结论成立：

$$\underbrace{\ell_{0/1}(y,\overline{q})}_{\text{ensemble loss}} \quad = \quad \underbrace{\frac{1}{m}\sum_{i=1}^{m}\ell_{0/1}(y,q_i)}_{\text{average loss}} - \underbrace{y\overline{q}\frac{1}{m}\sum_{i=1}^{m}\ell_{0/1}(\overline{q},q_i)}_{\text{good/bad diversity}}. \tag{25}$$

与前文的歧义分解相似之处显而易见。区别在于第二项之前的 yq。 **的符号决定了多样性的效应**。如果 $yq = +1$ （等价于说 $q = y$），该项为非负，因此会降低平均误差。出现这种情况时，称为“有益”多样性。当 $yq = -1$ 时，多样性会增加误差，称为“有害”多样性。将其与歧义效应分解（定理7）进行比较，可得二者的关系如下。

推论17 对于二分类问题， $y \in \{-1,+1\}$，歧义效应就是 *Brown* 和 *Kuncheva* 的有益/有害多样性项在数据分布上的平均值。

$$\underbrace{\mathbb{E}_{\mathbf{X}Y}\Big[Y\overline{q}\frac{1}{m}\sum_{i=1}^{m}\ell_{0/1}(\overline{q},q_i)\Big]}_{\text{good/bad diversity}} = \underbrace{\frac{1}{m}\sum_{i=1}^{m}\Big[R_{0/1}(q_i) - R_{0/1}(\overline{q})\Big]}_{\text{ambiguity-effect}}. \tag{26}$$

Didaci 等人（2013年）将有益/有害多样性的概念推广到了多分类情形，并且同样的关系仍然成立。Bian 和 Chen（2019年）运用这一思路定义了加权集成中的多样性，得出的形式与 Didaci 等人（2013年）相同。

当多样性效应为负时。 多样性效应的定义很简单：个体平均风险与集成风险之差的期望。如果集成的表现比个体更差，这一差值就为负，因此多样性的作用是增加0/1风险。假设一个由分类器组成的多数投票集成在一个 k 类问题上进行预测，各分类器的错误彼此独立，且每个模型预测正确类别的概率为 p。如果 $p > 0.5$，那么可以保证集成的 0/1风险低于个体的 0/1风险（Lam and Suen，1997年）。在我们的语境中，这意味着多样性效应非负，因此多样性的作用是降低集成风险。然而，这种理想化情形显然并非总能实现。Kuncheva（2014年，第 8.3 节）详细介绍了经过精心构造的情形：多样性越大，集成风险反而越高，也就是说，有害多样性的影响超过了有益多样性。

6.3 多样性与分类准确率的相关性

图2是一个示意性的“准确率/多样性”散点图，采用了 Kuncheva and Whitaker（2003年）推广的风格；它展示了多样性度量与准确率（0/1 损失）之间的相关性。横轴表示某种多样性度量，纵轴表示 $\frac{1}{m}\sum_{i=1}^{m} R_{0/1}(q_i) - R_{0/1}(q)$，即个体平均误差与集成误差之差。该值越高，集成的表现就越优于个体模型的平均表现。与多样性的相关性越高，说明多样性度量越有效，因为它能够解释性能提升。图13展示了真实数据中交叉熵多样性的两个此类图表。我们在验证数据上估计多样性，并在最终测试集上计算0/1 损失。这里需要注意的是，我们假设可以获取类别概率估计，以计算交叉熵多样性项。但这并非总是可行的。

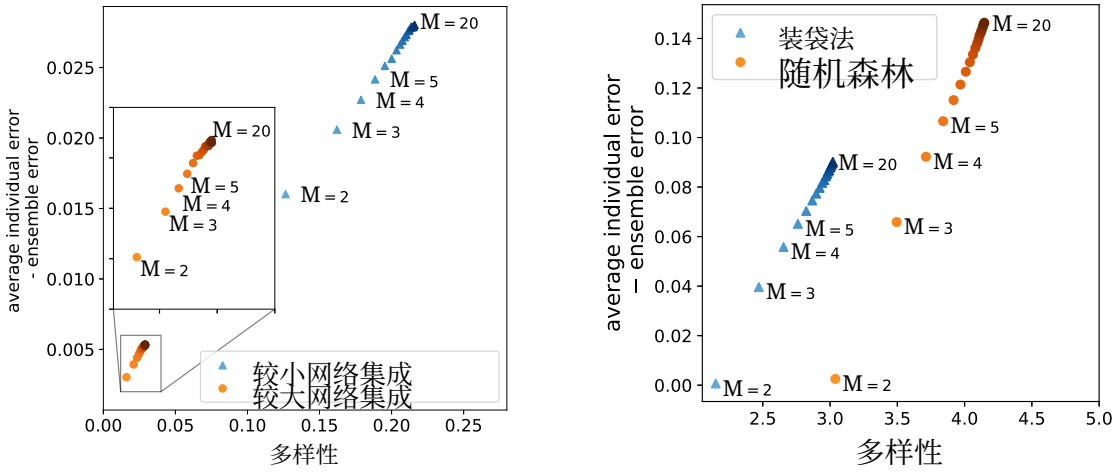


图13：准确率/多样性图。左：装袋多层感知机（对应图 7 中的实验）。右：决策树集成，装袋法与随机森林对比。

我们看到两种集成都呈现出很强的相关性。较小网络（蓝色三角形，皮尔逊相关系数为 $r^2 = 0.998$）比大网络（橙色圆形， r^2 =0。992）具有更高的多样性。解读该图时必须结合整体情况：较大网络显著

优于较小网络——这仅仅表明，其性能提升来自更强大的基础模型，而非它们的多样性。

在右侧图中，我们比较了 MNIST 上的装袋决策树（深度不限）和随机森林。两种情况下，都是先获得各决策树的概率，再通过归一化几何平均数进行合并。我们再次看到，多样性与性能提升之间存在很强的相关性。装袋法的相关性为 $r^2 \approx 0.996$，而随机森林的相关性为 $r^2 \approx 0.999$。对于固定的 m，我们可以比较相应的点；两者唯一的区别是，随机森林额外采用了分裂点随机化。在 $m = 20$ 时，随机森林将泛化误差降低了 $\approx 14.5\%$，而装袋法仅降低了 $\approx 9\%$。值得注意的是，这完全是由随机特征划分所带来的多样性增加造成的。

有人可能会想知道为什么这两种情况下都存在如此强的相关性。如果我们回想同一实验的另一种呈现方式——图 7，就会发现偏差/方差保持不变，变化的只有多样性。整体集成交叉熵出现任何变化时，我们都知道这是由多样性的变化造成的。因此，如果我们可以假设集成交叉熵与 0/1 损失之间存在很强的相关性，那么多样性与 0/1 损失之间也会存在类似的强相关性。

现在，借助这一多样性度量，我们可能会想知道，多样性与 *0/1* 损失的降低之间是否总会呈现很强的相关性？答案是否定的，原因非常充分，也凸显了我们理解多样性时的一个关键点。在图14中，我们固定使用 $m = 10$ 棵装袋树，并改变其深度。期望损失有所降低——然而，这一次并非完全由多样性导致。此时，偏差和方差也在快速变化，而 0/1 损失与多样性之间的相关性要低得多。

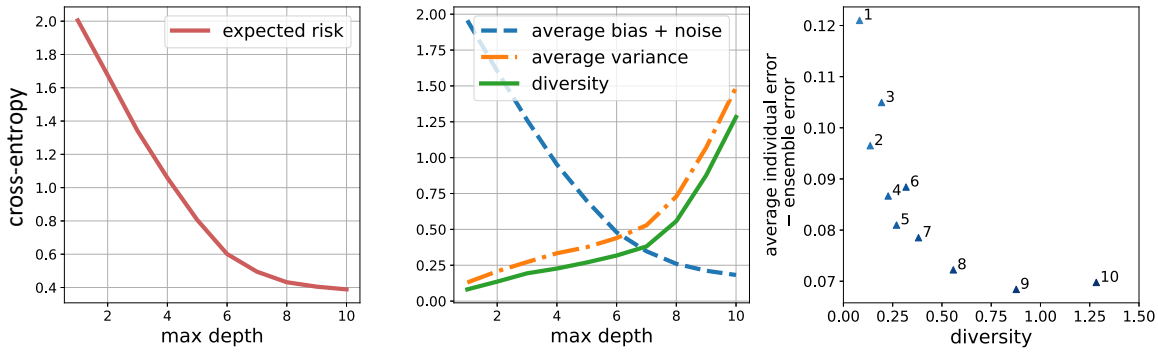


Figure 14: 装袋法 $m = 10$ 决策树，改变深度，此时相关性为 $r^2 \approx 0.59$.

改变除 m 以外的任何参数时，不应期望性能提升与多样性之间存在很强的相关性。这是因为，如果改变了会影响个体容量的参数，那么平均偏差/方差也会随之改变，而多样性并非唯一发挥作用的因素。总体性能由三方权衡决定，正如单个模型中存在偏差/方差的两方权衡一样。值得探究在非常深的神经网络集成中这一现象是否成立，因为其偏差/方差权衡似乎并不像经典理论所预测的那样运作（Belkin等人，2018年）。

6.4 未来工作

重新审视涉及启发式多样性度量的既有研究。在引言中，我们概述了若干采用启发式多样性度量的领域。例如，多样性有助于抵御对抗攻击、抵消协变量偏移，或构建计算效率更高的模型。我们可以基于我们的框架，使用从损失函数导出的多样性，重新开展其中任意一项研究。

强制多样性。 一个自然的研究方向是以某种方式强制实现多样性，即在构建集成时将我们的多样性度量用作正则化项。负相关（NC）学习（Liu and Yao，1999）鼓励回归集成具有多样性。Brown 等人（2006）表明，这实际上利用了公式 (1) 中的平方损失歧义分解。公式 (17) 中的布雷格曼歧义分解表明，NC 算法是更广泛的一类鼓励多样性的损失函数中的特例；Webb 等人（2021）探讨了其中交叉熵的情形。

间隔损失的多样性。 间隔损失（例如逻辑损失或指数损失）是一类重要的损失函数，尤其以其在梯度提升算法中的应用而闻名。Wood 等人（2022年）分析了这类损失的偏差-方差分解。利用这些结果，可以推导出间隔损失的偏差-方差-多样性分解，从而有望深入理解提升算法。然而，Mease and Wyner（2008）指出，有充分证据表明，AdaBoost/LogitBoost 中的加性模型形式导致替代间隔损失与我们的真实目标——0/1 损失——之间出现脱节。具体而言，替代损失可能会上升 （有时以指数速度上升），而留出样本上的 0/1 损失却在下降。这意味着，对替代损失进行的任何分析（包括损失分解）未必能为 0/1 损失提供有意义的见解。此外，在提升法中，相较于单个模型各自表现良好，更自然的理解是它们通过学习纠正先前集成成员的错误来发挥作用，这使得平均偏差项难以解释。将我们的框架扩展到这些情形，或许能为这类流行算法带来有趣的新解释。

专家混合模型中的多样性。目前，我们的框架并不适用于专家混合模型；在这种模型中，各模型由一个额外的输入相关模型进行门控，而该模型本身也是从数据中学习得到的。未来的研究或许可以考虑这一情形下的多样性——例如，测量空间中较小区域内的预测多样性，此时各成员的门控权重大致保持不变。

理解或编码概率假设。多样性是衡量预测之间依赖关系的一种方式。明确假设预测条件独立（有时称为*Dawid-Skene* 模型）看似意味着存在一定程度的多样性。不过，目前尚不清楚我们所提出的多样性损失函数视角与这一概率视角有何关联。鉴于损失与对数似然之间的对偶性，以及布雷格曼散度可视为概率密度之间的 KL 散度，这种关联应该是可以建立的，并且可能带来有趣的见解。

7 结论

我们提出了集成多样性的统一理论。一个关键见解是，重要的并非任务（例如分类或回归），而是损失函数。我们证明，多样性可视为集成损失偏差-方差分解中的一个隐藏维度。从这一视角来看，多样性自然涌现——其具体函数形式取决于所用的损失，但这些分解具有一种适用于多种损失的共同结构：

$$\textbf{期望损失} \;=\; (\textbf{平均偏差}) \;+\; (\textbf{平均方差}) \;-\; (\textbf{多样性})。$$

这清楚地揭示了集成性能与其多样性之间的关系。此前，只有平方损失搭配算术平均组合器的情形（Ueda 和 Nakano，1996年）存在这种关系。在这种情况下，我们的框架提供了另一种方案，但它将多样性的概念推广到了范围广得多的损失函数，包括交叉熵和泊松回归损失，更广义地说，适用于任何布雷格曼散度。该分解要求使用针对具体损失函数的特定组合规则，我们称之为质心组合器。 事实证明，这种组合器对应于文献中已有的几种知名组合器——例如，对于交叉熵，它就是归一化几何平均。这将集成“平均”的概念推广到了许多其他情形，可将其解释为在由布雷格曼散度定义的对偶坐标系中进行平均。

对于不存在加性偏差-方差分解的损失函数，我们采用 James 和 Hastie（1997年）的方法，转而衡量偏差、方差和多样性的影响。这些影响取决于标签分布。0/1 损失的情形尤其有趣——我们证明，多样性的影响不仅对于多数投票如此，对于任何组合器而言，都必然是依赖于标签的量。

因此，我们为广泛的监督学习场景提出了一种兼具广泛适用性和精确性的多样性定义。这一挑战被称为集成学习的“圣杯”（Zhou，2012年，第100页），三十多年来一直是一个悬而未决的问题。我们给出的答案将多样性表述为一种模型拟合度量，其含义与偏差/方差完全相同。因此，我们不应像许多研究所追求的那样，以“最大化多样性”为目标。正如偏差和方差会随模型特征而变化，多样性也同样如此；我们必须管理偏差/方差/多样性三者之间的权衡。

C所有实验的代码见：https://github.com/EchoStatements/Decompose

致谢与经费披露

直接支持本研究的经费：英国工程与自然科学研究委员会项目 EP/N035127/1（LAMBDA 项目）和 EP/T026995/1（EnnCore 项目）。Mikel Luj´an 获得了 Arm/英国皇家工程院研究主席奖以及英国皇家学会沃尔夫森奖学金。

GB 谨此感谢 LK 和 FR 在其职业生涯中给予的启发与支持。 .

附录 A. 补充实验细节

A.1 补充结果

我们在多个数据集上展示了其他结果。需要强调的是，我们并未声称任何一种方法在经验优势方面优于其他方法。我们仅将这些玩具数据集作为说明性样本，用于展示如何估计风险分量。

平方损失： 我们使用加利福尼亚住房数据——表 5 展示了三个集成（每个包含 $M = 30$ 棵回归树）的结果，并与单棵树进行比较。我们使用装袋法搭配受约束树（最大深度为 8），并将其与无限深度树和随机森林进行比较。

单棵树（深度 8）	装袋法（深度 8）	装袋法（不受限制）	随机森林
0.47	0.35	0.30	0.28

Tab图5：加利福尼亚住房数据：单棵树与由30棵树组成的集成的均方误差 s.

我们发现，随机森林是这里的最佳选择，其次是紧随其后的无约束装袋法。图15通过将风险分解为偏差、方差和多样性，解释了它们的性能，同时还展示了随着集成规模增大，各组成部分如何变化。

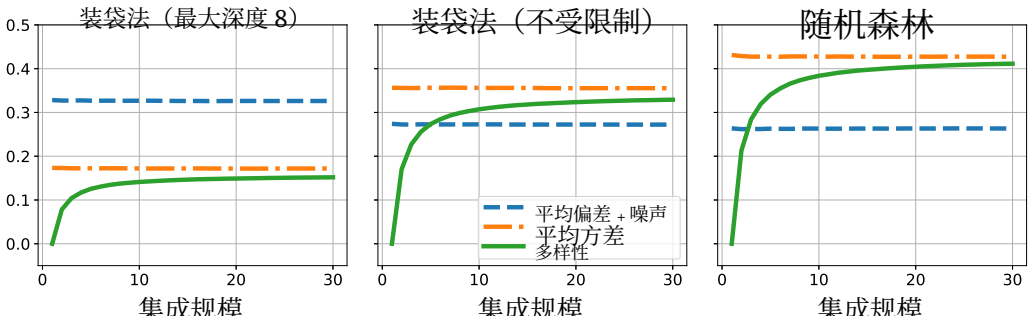


图15：将三种决策树集成的期望风险分解。

我们观察到的行为与图5中的相同。多样性随 m增加，并且其上限为平均方差。平均方差越高，多样性能够达到的“上限”也越高。从深度受限的决策树转为不限制深度的决策树时，平均方差会增大；在随机森林中加入随机分割点后（此处使用特征数量的平方根），平均方差还会进一步增大。更高的平均方差由多样性所补偿，因此随机森林是最佳选择。值得注意的是，对于大型集成，集成的期望风险几乎完全取决于平均偏差的大小（对于不受约束的决策树，$\approx$ 0.28），而多样性实际上抵消了单个模型的平均方差。这种行为并非该数据集的特例；事实上，只要所有单个模型都来自同一个模型族，也就是说，集成是同质的，这种情况就成立——一般情形将在第5.4节中讨论。

交叉熵：图16展示了分类问题中神经网络集成的更多准确率/多样性图（n 个样本， d 个特征， k 个输出类别）。数据集：音素数据集（$n = 5404$、 $d = 4$、 $k = 2$）、Landsat（$n = 6435$、 $d = 36$、 $k = 6$）、垃圾邮件数据集（$n = 4601$、 $d = 57$、 $k = 2$）、南德信用数据集（2019年版本： $n = 1000$、 $d = 20$、 $k = 2$）。

每种情况下，图例中均显示皮尔逊相关系数的平方。所有多层感知机实验均采用以下配置：

- 学习率：0.1（随机梯度下降）
- 训练轮数：50（MNIST），200（其他数据集）
- 隐藏层大小（20 个较小/100 个较大）
- 试验次数：100

其中每次试验使用完整训练数据的 90% 作为子样本，如图17所示。

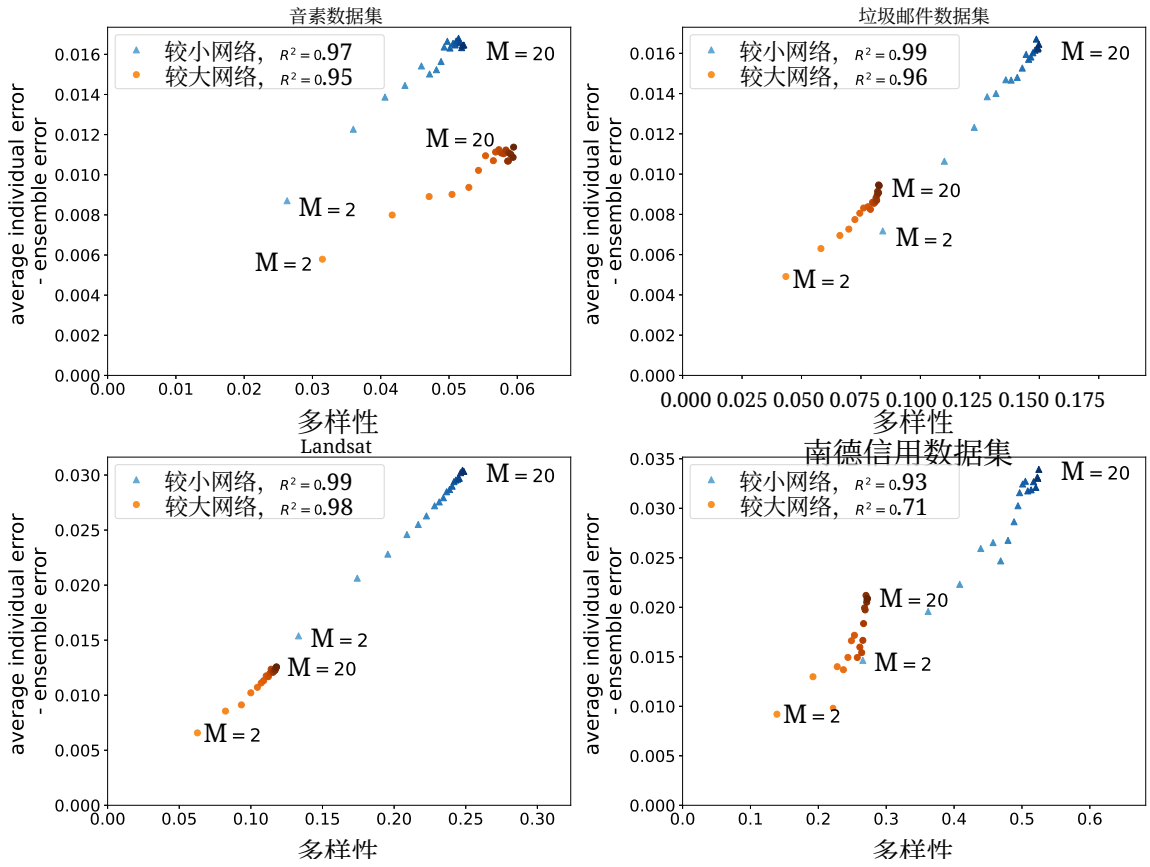


图16：四个数据集上的误差/多样性图，对比小型多层感知机集成（20 个隐藏节点，蓝点）与大型多层感知机集成（100 个隐藏节点，橙点）。

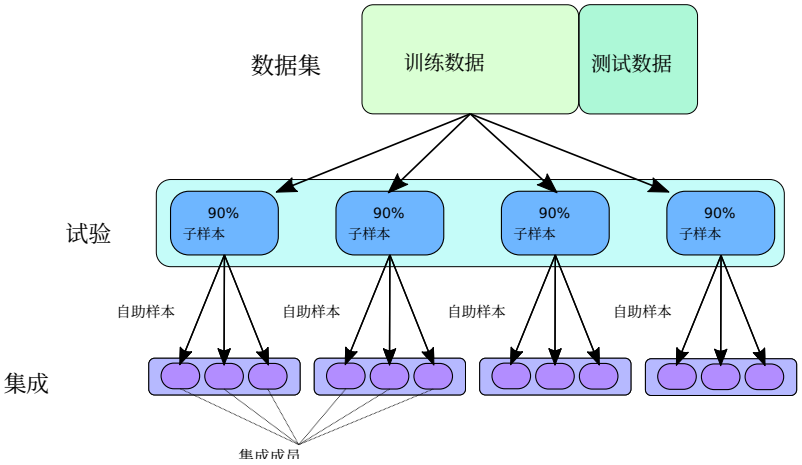
A.2 偏差、方差和多样性的估计方法

我们介绍了从数据中估计偏差、方差和多样性的方法。算法1展示了在不同规模的集成中估计多样性的实验流程，例如装袋法和随机森林的实验。值得注意的是，规模为 $m+1$ 的集成是利用规模为 m 的集成成员创建的，而不是从头创建一个规模为 $m+1$ 的新集成。我们还在图17中展示了装袋法所用子采样方案的可视化结果。

m ： 集成成员数量n ： 测试点数量t ： 试验次数D_k ： 试验 k 中使用的训练集

参数： $model$, $train\ data$, $test\ data$**输出**： $test\ preds$：大小为 $t \times m \times n$ 的模型预测数组。**对于** $k \in \{1, \ldots, t\}$ **依次对于** $j \in \{1, \ldots, n\}$ **执行**$D_k \leftarrow 90$ %，从 $train\ data$ 中无放回抽样；**对于** $i \in \{1, \ldots, m\}$ **执行** $member\ data \leftarrow$ 从 D_k进行自助抽样；i第 $\leftarrow$ 个集成成员是 $model$ 的副本，在 $member\ data$ 上训练；$test\ preds$[k, i, j] $\leftarrow$ 第 k 次试验中，第 j 个测试数据点由第 i 个集成成员给出的预测 $test\ data$ ；**结束 结束 结束**

算法1： 用于收集数据的算法，随后用于估计装袋法集成的多样性，同时改变集成规模 m。



Fi图 17：装袋法集成所使用的子采样方案示意图。

算法1的结果是一个大小为（t, m, n）的数组。对于布雷格曼散度，平均偏差和平均方差通过首先估计每个集成成员的左布雷格曼质心来近似，将随机变量 D 的期望替换为在 t 次试验上的平均值。我们使用符号 D_k 表示第 k 次试验中的完整训练数据，并可据此应用 Bagging 算法。因此，模型 i 在单个测试点 $\mathbf{x}$ 上的质心估计为：

$$\mathring{q}_i^{est}(\mathbf{x}) = [\nabla\phi]^{-1}\left(\frac{1}{t}\sum_{k=1}^{t}\nabla\phi(q_i(\mathbf{x}; D_k))\right) \approx \mathring{q}_i(\mathbf{x})$$

利用对 $\mathring{q}_i(\mathbf{x})$ 的这一估计，可计算平均偏差和平均方差：

$$\text{average bias} \approx \frac{1}{m}\frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n} B_\phi\left(y_j,\, \mathring{q}_i^{est}(\mathbf{x}_j)\right), \tag{27}$$

$$\text{average variance} \approx \frac{1}{t}\frac{1}{m}\frac{1}{n}\sum_{k=1}^{t}\sum_{i=1}^{m}\sum_{j=1}^{n} B_\phi\left(\mathring{q}_i^{est}(\mathbf{x}_j),\, q_i(\mathbf{x}_j; D_k)\right). \tag{28}$$

多样性的估计方式类似，其中 q 定义为应用于给定试验中m 集成成员的质心组合规则：

$$\text{diversity} \approx \frac{1}{t}\frac{1}{m}\frac{1}{n}\sum_{k=1}^{t}\sum_{i=1}^{m}\sum_{j=1}^{n} B_\phi\left(\overline{q}(\mathbf{x}_j, D_k),\, q_i(\mathbf{x}_j; D_k)\right). \tag{29}$$

附录 B. 进一步说明

我们将进一步阐述论文中的某些要点。本附录中的材料并非新颖内容，而是对部分方面进行扩展，以支持我们的论证。

B.1 是“平方偏差”，还是仅仅是“偏差”?

Geman 等人（1992年）提出了平方损失的偏差-方差分解，即公式（3）。其中一项是 $(\mathbb{E}_D[q] - y^*)^2$。他们以及此后许多论文都将其称为“平方偏差”或“偏差平方”。这种叫法并不恰当，下面我们将对此进行说明。

问题的根源在于，Geman 等人借用了经典统计学中的术语。$\hat{}$ 对于某个总体值 θ 和一个估计量 θ，经典统计学将量$\mathbb{E}[\hat{\theta}] - \theta$ 称为估计量 $\hat{\theta}$ 的偏差。因此，类比之下，他们将 $(\mathbb{E}_D[q] - y^*)^2$ 称为估计量 q 的平方偏差，其中 y^* 被假定为总体值。

之所以说这种叫法并不恰当，是因为其他损失函数也有类似的分解，而其中并不存在平方项，例如 Heskes（1998年）、Pfau（2013年）以及 Wood 等人（2022年）的研究。与本文相关的是，Ueda 和 Nakano（1996年）提出的集成偏差-方差-协方差分解也不幸沿用了“平方”这一不恰当的说法，正如我们在第 5.4.3 小节中所解释的。这些结果是在 Geman 等人（1992年）发表很久之后才得到证明的，因此我们不能责怪这些作者，尤其是“偏差”一词在科学领域有多种含义（例如归纳偏差、抽样偏差、选择偏差和确认偏差）。简而言之，“平方偏差”这一说法并不恰当，因此本文始终只称其为“偏差”。

B.2 歧义分解等价于偏差-方差分解

在偏差-方差分解中，通常假设随机变量 D 取遍所有大小固定为 n的可能训练集。这是 Geman 等人（1992年）提出的一种惯例。然而，情况不一定如此， D 也可以取遍构建模型时涉及的任何随机量（例如神经网络的随机初始权重），甚至可以取遍一个由有限个预先构建的模型组成的集合。在后一种情况下，我们可以看到偏差-方差分解与歧义分解之间存在等价关系——也就是说，如果其中一种分解存在，那么另一种也必然存在。考虑不含噪声的偏差-方差分解的一个更简单形式，最容易看出这一点：

$$\mathbb{E}_D[\ell(y, q)] = \ell(y, \mathring{q}) + \mathbb{E}_D[\ell(\mathring{q}, q)]. \tag{30}$$

加入噪声后，这一关系仍然成立，但在上述假设下更容易解释。现在，将 D 定义为一个离散随机变量，其事件空间由 m 个不同的预先构建模型 $\{q_i\}_{i=1}^m$组成，并为每个事件赋予均匀概率，即对于每个事件 q_i， $1/m$。这意味着期望 $\mathbb{E}_D$ 变成了平均值 $\frac{1}{m}\sum_{i=1}^m$。质心的定义仍然相同，但我们使用符号 q 表示它涉及的是有限个预测：$\overline{q} \overset{\text{定义}}{=} \operatorname{arg\,min}_{z \in \mathcal{Y}} \frac{1}{m}\sum_{i=1}^m \ell(z, q_i)$。这意味着上面的表达式变为：

$$\frac{1}{m}\sum_{i=1}^m \ell(y, q_i) = \ell(y, \overline{q}) + \frac{1}{m}\sum_{i=1}^m \ell(\overline{q}, q). \tag{31}$$

对各项进行简单重排即可得到歧义分解。

B.3 Heskes 分解在交叉熵中的性质

Heskes（1998年）提出了概率密度之间 KL散度 的偏差-方差分解，并将其扩展为显式包含噪声项，从而得到交叉熵的噪声-偏差-方差分解。对于多项伯努利分布，我们认为，这与我们在公式（4）中提出的分解等价；这一点十分重要，因为它使该分解与我们统一的偏差-方差形式相一致。为清晰起见，我们在此重述 KL散度 分解：

$$\underbrace{\mathbb{E}_D\Big[\mathbb{E}_{\mathbf{XY}}\left[K\left(\mathbf{Y}\ ||\ \mathbf{q}\right)\right]\Big]}_{\text{期望风险}} = \mathbb{E}_{\mathbf{X}}\Bigg[\underbrace{\mathbb{E}_{\mathbf{Y}|\mathbf{X}}\left[K\Big(\mathbf{Y}\ ||\ \mathbf{Y}^*\Big)\right]}_{\text{噪声}} + \underbrace{K\Big(\mathbf{Y}^*\ ||\ {}^{\circ}\mathbf{q}\Big)}_{\text{bias}} + \underbrace{\mathbb{E}_D\left[K\Big({}^{\circ}\mathbf{q}\ ||\ \mathbf{q}\Big)\right]}_{\text{方差}}\Bigg], \tag{32}$$

其中，$\mathbf{Y}$ 是随机变量，取值范围为输入点$\mathbf{x}$对应真实类别的独热编码指示变量。若采用 Heskes 的表达式（Heskes，1998年，式 6），并假设分布为多项伯努利分布，则可得到单个点$\mathbf{x}$处真实类别分布$\mathbf{y}^*$ 与模型$\mathbf{q}$之间期望交叉熵的分解。用我们的符号表示如下：

$$-\mathbb{E}_D\Big[\mathbf{y}^* \cdot \ln \mathbf{q}\Big] = \underbrace{-\mathbf{y}^* \cdot \ln \mathbf{y}^*}_{\text{noise}} + \underbrace{K(\mathbf{y}^*\ ||\ \mathring{\mathbf{q}})}_{\text{bias}} + \underbrace{\mathbb{E}_D\Big[K(\mathring{\mathbf{q}}\ ||\ \mathbf{q})\Big]}_{\text{variance}}, \tag{33}$$

其中，$^{\circ}\mathbf{q}\ \propto \exp(\mathbb{E}_D[\ln \mathbf{q}])$ 是模型分布的归一化几何平均。

为证明二者等价，我们首先注意到，(32) 对 $P(\mathbf{X})$ 取期望。我们可以将其改写为单个$\mathbf{x}$处的形式：

$$\underbrace{\mathbb{E}_D\Big[\mathbb{E}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}\left[K\left(\mathbf{Y}\ ||\ \mathbf{q}\right)\right]\Big]}_{\text{expected loss}} = \underbrace{\mathbb{E}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}\left[K\Big(\mathbf{Y}\ ||\ \mathbf{y}^*\Big)\right]}_{\text{noise}} + \underbrace{K\Big(\mathbf{y}^*\ ||\ \mathring{\mathbf{q}}\Big)}_{\text{bias}} + \underbrace{\mathbb{E}_D\left[K\Big(\mathring{\mathbf{q}}\ ||\ \mathbf{q}\Big)\right]}_{\text{variance}}, \tag{34}$$

其中，$\mathbf{y}^* = \mathbb{E}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}[\mathbf{Y}]$ 是$\mathbf{x}$处的真实类别分布。右侧的第二项和第三项与 Heskes 分解中的相同。对于噪声项，我们注意到，

$$\begin{aligned}
\mathbb{E}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}\left[K\Big(\mathbf{Y}\ ||\ \mathbf{y}^*\Big)\right] &= \mathbb{E}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}\left[\mathbf{Y} \cdot \ln \frac{\mathbf{Y}}{\mathbf{y}^*}\right] && (35)\\
&= \mathbb{E}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}\left[\mathbf{Y} \cdot \ln \mathbf{Y}\right] - \mathbb{E}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}\left[\mathbf{Y} \cdot \ln \mathbf{y}^*\right] && (36)\\
&= -\mathbb{E}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}\left[\mathbf{Y}\right] \cdot \ln \mathbf{y}^* && (37)\\
&= -\mathbf{y}^* \cdot \ln \mathbf{y}^*, && (38)
\end{aligned}$$

其中我们采用约定 $0 \ln 0 = 0$。由于两边的右侧相等，因此左侧也等价。

B.4 KL散度中参数编码的重要性

考虑两个长度为 k 的离散概率向量之间的 KL散度，

$$K(\mathbf{p} \,||\, \mathbf{q}) = \sum_{c=1}^{k} \mathbf{p}^{(c)} \ln \frac{\mathbf{p}^{(c)}}{\mathbf{q}^{(c)}} \tag{39}$$

将其表示为布雷格曼散度有两种方式：一种是使用完整长度的概率向量，$\mathbf{p} \in \mathbb{R}^k$，并使用生成函数

$$\phi_{\text{full}}(\mathbf{p}) = \sum_{c=1}^{k} \mathbf{p}^{(c)} \ln \mathbf{p}^{(c)} \tag{40}$$

另一种是使用最小参数化向量，$\widetilde{\mathbf{p}} \in \mathbb{R}^{k-1}$，其中省略最后一个元素：

$$\phi_{\min}(\widetilde{\mathbf{p}}) = \sum_{c=1}^{k-1} \widetilde{\mathbf{p}}^{(c)} \ln \widetilde{\mathbf{p}}^{(c)} + (1 - \sum_{c'=1}^{k-1} \widetilde{\mathbf{p}}^{(c')}) \ln(1 - \sum_{c'=1}^{k-1} \widetilde{\mathbf{p}}^{(c')}). \tag{41}$$

给定两个形式适当的概率向量，无论采用哪种表示方式，所得的布雷格曼散度都等价于类别概率估计向量之间的 KL散度（NielsenandNock，2009），即，

$$B_{\phi_{\text{full}}}(\mathbf{p}, \mathbf{q}) = B_{\phi_{\min}}(\widetilde{\mathbf{p}}, \widetilde{\mathbf{q}}) = K(\mathbf{p} \,||\, \mathbf{q}).$$

我们采用最小参数化，因为它具有一个理想的性质。具体而言，必须采用第二种方式，才能确保布雷格曼质心始终是概率单纯形上的有效分布。在这种情况下，质心组合器就是归一化几何平均数，下面将对此进行说明。

最小参数化向量：为说明这一点，我们考虑 m 最小参数化向量，$\mathbf{q}_i \in \mathbb{R}^{k-1}$（注意，为简便起见，我们省略了上方$\mathbf{q}$ 上的波浪号）。我们的主张是，质心组合器，即归一化几何平均数$\overline{\mathbf{q}} = [\nabla\phi]^{-1}\left(\frac{1}{m}\sum_{i=1}^{m} \nabla\phi(\mathbf{q}_i)\right)$，具有如下形式

$$\overline{\mathbf{q}}^{(c)} = [\nabla\phi_{\min}]^{-1}\left(\frac{1}{m}\sum_{i=1}^{m} \nabla\phi_{\min}(\mathbf{q}_i)\right) = \frac{\prod_{i=1}^{m} \mathbf{q}_i^{(c)\,\frac{1}{m}}}{\sum_{c'=1}^{k}\prod_{i=1}^{m} \widehat{\mathbf{q}}^{(c')\frac{1}{m}}}, \tag{42}$$

其中，$\widehat{\mathbf{q}}$ 表示将 $k-1$ 长度向量扩展为完整的 k 长度概率向量。代入表6中的梯度后，我们从以下内容开始

$$\overline{\mathbf{q}}^{(c)} = \frac{\exp\left(\frac{1}{m}\sum_{i=1}^{m} \ln \frac{\mathbf{q}_i^{(c)}}{1-\sum_{c'=1}^{k-1}\mathbf{q}_i^{(c')}}\right)}{1 + \sum_{c'=1}^{k-1}\exp\left(\frac{1}{m}\sum_{i=1}^{m} \ln \frac{\mathbf{q}_i^{(c')}}{1-\sum_{c''=1}^{k-1}\mathbf{q}_i^{(c'')}}\right)}.$$

注意，这里的分子可以重新排列：

$$\exp\left(\frac{1}{m}\sum_{i=1}^{m} \ln \frac{\mathbf{q}_i^{(c)}}{1-\sum_{c'=1}^{k-1}\mathbf{q}_i^{(c')}}\right) = \prod_{i=1}^{m}\left(1 - \sum_{c'=1}^{k-1}\mathbf{q}_i^{(c')}\right)^{-\frac{1}{m}} \prod_{i=1}^{m}\left(\mathbf{q}_i^{(c)}\right)^{\frac{1}{m}},$$

并且分母可以写成

$$
\begin{aligned}
1+\sum_{c'=1}^{k-1}\exp\left(\frac{1}{m}\sum_{i=1}^{m}\ln\frac{\mathbf{q}_m^{(c')}}{1-\sum_{c''=1}^{k-1}\mathbf{q}_m^{(c'')}}\right) &= 1+\sum_{c'=1}^{k-1}\prod_{i=1}^{m}\mathbf{q}_i^{(c')\frac{1}{m}}\left(1-\sum_{c''=1}^{k-1}\mathbf{q}_i^{(c'')}\right)^{-\frac{1}{m}} \\
&= 1+\prod_{i'=1}^{m}\left(1-\sum_{c''=1}^{k-1}\mathbf{q}_{i'}^{(c'')}\right)^{-\frac{1}{m}}\sum_{c'=1}^{k-1}\prod_{i=1}^{m}\mathbf{q}_i^{(c')\frac{1}{m}} \\
&= \prod_{i'=1}^{m}\left(1-\sum_{c''=1}^{k-1}\mathbf{q}_{i'}^{(c'')}\right)^{-\frac{1}{m}}\left(\prod_{i=1}^{m}\left(1-\sum_{c''=1}^{k-1}\mathbf{q}_i^{(c'')}\right)^{\frac{1}{m}}+\sum_{c'=1}^{k-1}\prod_{i=1}^{m}\mathbf{q}_i^{(c')\frac{1}{m}}\right).
\end{aligned}
$$

将分子和分母代回方程（42）的第二个表达式，并利用 $\hat{}\mathbf{q}$ 的定义，我们发现两个乘积中的首项相互抵消，剩下的就是所需结果。

完整长度 k 概率向量：如果我们不使用最小参数化向量，就会得到 Bregman 生成函数 $\phi(\mathbf{p}) = \sum_{c=1}^{k}\mathbf{p}^{(c)}\ln\mathbf{p}^{(c)}$。这会得到几何平均数，而不是归一化版本。要理解这一点，我们首先注意到

$$
\begin{aligned}
\left(\nabla\phi(\mathbf{p})\right)^{(c)} &= 1+\ln\mathbf{p}^{(c)} = \boldsymbol{\eta}^{(c)} \\
\left(\left[\nabla\phi\right]^{-1}(\boldsymbol{\eta})\right)^{(c)} &= \exp\left(\boldsymbol{\eta}^{(c)}-1\right),
\end{aligned}
$$

因此质心组合器为

$$
\begin{aligned}
\left(\left[\nabla\phi\right]^{-1}\left(\frac{1}{m}\sum_{m=1}^{m}\nabla\phi(\mathbf{q}_m)\right)\right)^{(c)} &= \exp\left(\frac{1}{m}\sum_{i=1}^{m}1+\ln\mathbf{q}_i^{(c)}-1\right) \\
&= \exp\left(\frac{1}{m}\sum_{i=1}^{m}\ln\mathbf{q}_i^{(c)}\right) = \prod_{i=1}^{m}\mathbf{q}_i^{(c)\frac{1}{m}}.
\end{aligned}
$$

请注意，这意味着$\mathbf{q}$ 不一定是有效的概率向量。事实上，仅当q $\mathbf{q}_1 = \ldots = \mathbf{q}_m$时，它才是有效的概率向量。

B.5 布雷格曼质心列表

为便于快速查阅（有关更多详细信息，请参见 Nielsen 和 Nock（2009）），我们列出了机器学习界所关注的不同损失函数对应的质心。左布雷格曼质心定义如下：

$$\mathring{\mathbf{q}} \stackrel{\text{def}}{=} \arg\min_{\mathbf{z}\in\mathcal{Y}} \mathbb{E}_D\Big[B_\phi\left(\mathbf{z},\mathbf{q}\right)\Big] \;=\; \left[\nabla\phi\right]^{-1}\Big(\mathbb{E}_D\left[\nabla\phi\left(\mathbf{q}\right)\right]\Big). \tag{43}$$

质心 $\mathring{\mathbf{q}}$ 的形式因所用生成函数而异。

Loss	**梯度** $\eta = \nabla\phi(\mathbf{q})$	**逆梯度** $\mathbf{q} = [\nabla\phi]^{-1}(\eta)$	**左布雷格曼质心** $\mathring{\mathbf{q}} := [\nabla\phi]^{-1}\Big(\mathbb{E}_D[\nabla\phi(\mathbf{q})]\Big)$
平方损失	$2q$	$\frac{1}{2}\eta$	$\mathbb{E}_D[q]$
伊丹–斋藤散度	$-\frac{1}{q}$	$-\frac{1}{\eta}$	$1/\Big(\mathbb{E}_D\left[1/q\right]\Big)$
泊松损失	$\ln q$	$\exp(\eta)$	$\exp\big(\mathbb{E}_D\left[\ln q\right]\big)$
KL散度	$\ln\frac{\mathbf{q}^{(c)}}{1-\sum_{c'=1}^{k-1}\mathbf{q}^{(c')}}$	$\frac{\exp(\boldsymbol{\eta}^{(c)})}{1+\sum_{c'=1}^{k-1}\exp(\boldsymbol{\eta}^{(c')})}$	$\frac{1}{Z}$指数$\Big(\mathbb{E}_D[$ 自然对数 $\mathbf{q}]\Big)$

表6：常见损失函数（见表2）及其布雷格曼质心。对于 KL 散度，Z 是一个归一化因子，用于确保分布有效。

附录 C. 第 4 节的证明

C.1 偏差-方差-多样性与效应分解的证明

定理 5（广义偏差-方差-多样性分解）考虑一组模型 $\{\mathbf{q}_i\}_{i=1}^m$，并使用损失 ℓ 对其进行评估。假设偏差-方差分解以定义 *2* 的形式成立，则以下分解也成立。

$$\mathbb{E}_D\Big[\mathbb{E}_{\mathbf{XY}}[\ell(\mathbf{Y},\overline{\mathbf{q}})]\Big] \;=\; \mathbb{E}_{\mathbf{X}}\Bigg[\underbrace{\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[\ell(\mathbf{Y},\mathbf{Y}^*)]}_{\text{noise}} + \underbrace{\frac{1}{m}\sum_{i=1}^m \ell(\mathbf{Y}^*,\mathring{\mathbf{q}}_i)}_{\text{average bias}} + \underbrace{\frac{1}{m}\sum_{i=1}^m \mathbb{E}_D\left[\ell(\mathring{\mathbf{q}}_i,\mathbf{q}_i)\right]}_{\text{average variance}} - \underbrace{\mathbb{E}_D\left[\frac{1}{m}\sum_{i=1}^m \ell(\overline{\mathbf{q}},\mathbf{q}_i)\right]}_{\text{diversity}}\Bigg], \quad (9)$$

其中 $\mathring{\mathbf{q}} \stackrel{\text{def}}{=} \arg\min_{\mathbf{z}\in\mathcal{Y}} \mathbb{E}_D\Big[\ell(\mathbf{z},\ \mathbf{q})\Big]$，组合器为 $\mathbf{q} \stackrel{\text{def}}{=} \underline{\arg\min}_{\mathbf{z}\in\mathcal{Y}} \frac{1}{m}\underline{\sum_{i=1}^m \ell(\mathbf{z},\ \mathbf{q}_i)}$。

证明 取$\mathbf{q}$的期望风险，并应用命题 *3*，即广义歧义分解：

$$\mathbb{E}_D\left[\mathbb{E}_{\mathbf{XY}}\left[\ell(\mathbf{Y},\overline{\mathbf{q}})\right]\right] = \mathbb{E}_D\left[\mathbb{E}_{\mathbf{XY}}\left[\frac{1}{m}\sum_{i=1}^m \ell(\mathbf{Y},\mathbf{q}_i)\right]\right] - \mathbb{E}_D\left[\mathbb{E}_{\mathbf{XY}}\left[\frac{1}{m}\sum_{i=1}^m \ell(\overline{\mathbf{q}},\mathbf{q}_i)\right]\right]. \quad (44)$$

现在对右侧第一项应用定义 2，即广义偏差-方差分解：

$$\mathbb{E}_D\Big[\mathbb{E}_{\mathbf{XY}}\Big[\frac{1}{m}\sum_{i=1}^m \ell(\mathbf{Y},\mathbf{q}_i)\Big]\Big] = \mathbb{E}_{\mathbf{X}}\Bigg[\mathbb{E}_{\mathbf{Y}|\mathbf{X}}\left[\ell(\mathbf{Y},\mathbf{Y}^*)\right] \;+\; \frac{1}{m}\sum_{i=1}^m \ell(\mathbf{Y}^*,\mathring{\mathbf{q}}_i) \;+\; \frac{1}{m}\sum_{i=1}^m \mathbb{E}_D\left[\ell(\mathring{\mathbf{q}}_i,\mathbf{q}_i)\right]\Bigg]. \quad (45)$$

将公式 (45) 代入 (44)，证明即告完成。 ■

关于依赖/独立训练方案的进一步说明：对于单个模型 q 的偏差-方差分解，我们定义了一个随机变量 $D \sim P(\mathbf{x}\ y)^n$，即针对大小为 n的独立同分布训练集。当我们拥有一个模型集成时，通常每个模型都有各自的训练集，例如 Bagging 算法。在这种情况下，D 被重新定义为一个向量 $[D_0, D_1, D_2, \ldots, D_m]$，其中 $D_0 \sim P(\mathbf{x}\ y)^n$ 定义了提供给集成的整体训练集。对于 $1..m$，每个 D_i都是一个随机变量，表示从 D_0定义的同一个固定训练集中进行的独立同分布有放回抽样。因此，每个 D_i 都依赖于 D_0，但不依赖于任何其他 D_j。不过，D_t 也可能依赖于 D_{t-1}，例如 Boosting 算法就是如此。无论哪种情况，如果我们考虑（不失一般性）第一个模型的任意函数 q_1，那么根据全期望定律，$\mathbb{E}_Y\Big[\mathbb{E}_{X|Y}[X]\Big] = \mathbb{E}_X[X]$，我们有：

$$\mathbb{E}_{D_{0\ldots m}}\Big[q_1(D_1)\Big] = \mathbb{E}_{D_0,D_{2\ldots m}}\Big[\mathbb{E}_{D_1|D_0,D_{2\ldots m}}[q_1(D_1)]\Big] \;=\; \mathbb{E}_{D_1}\Big[q_1(D_1)\Big]. \quad (46)$$

因此，模型的函数仅取决于该模型自身的训练集。

定理 8（偏差-方差-多样性效应分解）给定一个由质心组合器 $\mathbf{q}_i\}_{i=1}^m$ 组合的模型集成 $\{\mathbf{q}_i\}_{i=1}^m$ $\mathbf{q} \stackrel{\text{def}}{=} \arg\min_{\mathbf{z}\in\mathcal{Y}} \left[\frac{1}{m}\sum_{i=1}^m \ell(\mathbf{z},\mathbf{q}_i)\right]$，采用任意损失 ℓ，该集成的期望风险可分解为，

$$\mathbb{E}_D\Big[R(\overline{\mathbf{q}})\Big] =$$

$$\underbrace{R(\mathbf{y}^*)}_{\text{noise}} + \underbrace{\frac{1}{m}\sum_{i=1}^m \Big[R(\mathring{\mathbf{q}}_i) - R(\mathbf{y}^*)\Big]}_{\text{average bias-effect}} + \underbrace{\frac{1}{m}\sum_{i=1}^m \mathbb{E}_D\Big[R(\mathbf{q}_i) - R(\mathring{\mathbf{q}}_i)\Big]}_{\text{average variance-effect}} - \underbrace{\mathbb{E}_D\left[\frac{1}{m}\sum_{i=1}^m \Big[R(\mathbf{q}_i) - R(\overline{\mathbf{q}})\Big]\right]}_{\text{diversity-effect}}$$

P证明 注意，右侧的若干项相互抵消，化简后得到左侧 . ■

定理 10（0/1 损失下标签无关的多样性效应不存在性）对于 *0/1* 损失，使用任意集成组合规则时，个体平均风险与集成风险之差必然依赖于标签。

证明 我们证明，$\frac{1}{m}\sum_{i=1}^m \ell_{0/1}(y,q) - \ell_{0/1}(y,q)$ 在点 $(\mathbf{x}\ y)$ 处必然依赖于y的取值。对 $P(\mathbf{x}\ y)$ 求期望即可证明结论。我们证明这样一种情形：不存在任何集成组合规则能使该表达式与 y 无关。

二分类情形：定义 $\mathcal{Y} = \{1,2\}$。对于固定的 $\mathbf{x}$，不失一般性，令$p = 0.6$ 为预测类别 $\mathbf{1}$ 的 m 模型所占比例。组合器 q 给出一个标签 $\in \mathcal{Y}$。对于两种可能的 q，量 $\frac{1}{m}\sum_{i=1}^{\overline{m}} \ell_{0/1}(y,q) - \ell_{0/1}(y,q)$ 都依赖于 y。首先，若假设 $q = 1$，则有：

$$\frac{1}{m}\sum_{i=1}^m \ell_{0/1}(y,q_i) - \ell_{0/1}(y,\overline{q}) = \begin{cases} 0.4 - 0 = 0.4 & \text{if } y = 1 \\ 0.6 - 1 = -0.4 & \text{if } y = 2 \end{cases}$$

或者，当 $q = 2$ 时，

$$\frac{1}{m}\sum_{i=1}^m \ell_{0/1}(y,q_i) - \ell_{0/1}(y,\overline{q}) = \begin{cases} 0.4 - 1 = -0.6 & \text{if } y = 1 \\ 0.6 - 0 = 0.6 & \text{if } y = 2 \end{cases}$$

对于 bot $\mathbf{h}$ 有 q，$\frac{1}{m}\sum_{i=1}^m \ell_{0/1}(y,\ q_i) - \ell_{0/1}(y,\ q)$ 的值取决于真实标签 y .

多分类情形：定义 $\mathcal{Y} = \{1,2,\ldots,k\}$，以及 m 模型预测每个类别的比例为 p_1、.....、p_k。我们令 $p_1 = 0.6$、$p_2 = 0.4$，并将所有其他类别设为零。从二分类情形可知，当 $q \in \{1,2\}$ 时，存在对 y 的依赖。对于 $q \in \{3,\ldots,k\}$，这种情况仍然成立，此时我们有

$$\frac{1}{m}\sum_{i=1}^m \ell_{0/1}(y,q_i) - \ell_{0/1}(y,\overline{q}) = \begin{cases} 0.4 - 1 = -0.6 & \text{if } y = 1 \\ 0.6 - 1 = -0.4 & \text{if } y = 2 \\ 1.0 - 0 = 1 & \text{if } y = \overline{q} \\ 1.0 - 1 = 0 & \text{for all other c类别。} \end{cases}$$

对于所有 $q \in \mathcal{Y}$，表达式 $\frac{1}{m}\sum_{i=1}^m \ell_{0/1}(y,\ q_i) - \ell_{0/1}(y,\ q)$ 依赖于标签 y .

■

C.2 偏差/方差/多样性效应分解：加权投票 ng

现在，我们证明加权多数投票组合的偏差-方差-多样性效应分解。给定一个由分类模型 q_1、...、q_m 组成的集成，其中每个模型都输出来自 $\{1,\ldots,k\}$ 的标签预测，并为这些模型赋予权重 $\alpha_1(D)$.....、$\alpha_m(D)$，我们考虑加权相对多数投票的歧义效应分解。加权多数投票为

$$\overline{q}(\mathbf{x};D) = \underset{c\in\{1,\ldots,k\}}{\arg\min} \sum_{i=1}^{m} \frac{\alpha_i(D)}{\sum_{j=1}^{m}\alpha_j(D)} \ell_{0/1}(c,q_i),$$

类似地，集成成员的质心可表示为

$$\mathring{q}_i = \underset{c\in\{1,\ldots,k\}}{\arg\min}\, \mathbb{E}_D\left[\frac{\alpha_i(D)}{\mathbb{E}_D\left[\alpha_i(D)\right]}\ell_{0/1}(c,q_i)\right],$$

平局时随机打破平局（可以将平局处理程序视为随机变量 D 的一部分，因为 D 隐含地包含了与模型相关的所有随机性来源）。

定理18（加权多数投票的歧义效应分解）据此，我们可以定义加权效应分解如下

$$\ell_{0/1}(y,\overline{q}) = \underbrace{\sum_{i=1}^{m} a_i(D)\ell_{0/1}(y,q_i)}_{\text{weighted average loss}} - \left[\underbrace{\sum_{i=1}^{m} a_i(D)\ell_{0/1}(y,q_i) - \ell_{0/1}(y,\overline{q})}_{\text{ambiguity-effect}}\right],$$

其中 $a_i = \frac{\alpha_i}{\sum_{j=1}^{m}\alpha_j}$。

只需消去等式右侧的各项，即可验证该定理成立。该定理表明，集成的损失可分解为一个非负项（集成成员的加权平均损失）和一个歧义效应项；后者量化了集成的性能相较于成员平均损失有多好（或多差）。运用相同的原理，我们还可以构造一个考虑权重 $\alpha(D)$ 的偏差 / 方差效应分解。

定理19（加权多数投票的偏差/方差效应分解）

$$\mathbb{E}_D\left[\alpha\ell_{0/1}(y,q)\right] = \mathbb{E}_D\left[\alpha\right]\ell_{0/1}(y,\mathring{q}) + \left[\mathbb{E}_D\left[\alpha\ell_{0/1}(y,q)\right] - \mathbb{E}_D\left[\alpha\right]\ell_{0/1}(y,\mathring{q})\right]$$

同样，只需考察右侧哪些项会相互抵消，即可立即证明该结果。不过，值得探讨一下这种分解是如何起作用的，以及各项的含义。考虑将 α 替换为归一化版本 $\frac{\alpha}{\mathbb{E}_D[\alpha]}$ 时的分解，我们得到：

$$\mathbb{E}_D\left[\frac{\alpha}{\mathbb{E}_D\left[\alpha\right]}\ell_{0/1}(y,q)\right] = \ell_{0/1}(y,\mathring{q}) + \left[\mathbb{E}_D\left[\frac{\alpha}{\mathbb{E}_D\left[\alpha\right]}\ell_{0/1}(y,q)\right] - \ell_{0/1}(y,\mathring{q})\right].$$

这正是我们之前见过的偏差-方差-效应分解，只不过重新加权了不同数据集的贡献。事实上，两者是等价的，这些权重定义了一个新的概率密度函数。将 $P_D(\mathcal{D})$ 视为数据集上的概率密度函数，上述分解恰好就是采用新概率密度函数 $Q_D(\mathcal{D}) = P_D(\mathcal{D})^{\alpha(\mathcal{D})}\,\mathbb{E}_D[\alpha(D)]$ 的偏差-方差-效应分解。我们还可以轻松地重新引入标签噪声并显式写出噪声项，同时将偏差转化为偏差效应：

$$\mathbb{E}_Y\left[\mathbb{E}_D\left[\frac{\alpha}{\mathbb{E}_D[\alpha]}\ell_{0/1}(Y,q)\right]\right] = \underbrace{\mathbb{E}_Y\left[\ell_{0/1}(Y,Y^*)\right]}_{\text{noise}} + \underbrace{\mathbb{E}_Y\left[\ell_{0/1}(Y,\mathring{q}) - \ell_{0/1}(Y,Y^*)\right]}_{\text{weighted bias-effect}} \underbrace{\left[\mathbb{E}_D\left[\mathbb{E}_Y\left[\frac{\alpha}{\mathbb{E}_D[\alpha]}\ell_{0/1}(Y,f) - \ell_{0/1}(Y,\mathring{q})\right]\right]\right]}_{\text{weighted variance-effect}}.$$

现在，我们可以应用类似的证明方法，得到加权多数投票的以下偏差-方差-多样性效应分解。

命题20（加权投票的偏差-方差-多样性效应）给定m 个分类器 $q_1,\ldots,q_m$，其中集成为加权多数投票，即$q = \arg\min_{z\in\mathcal{Y}}\sum_{i=1}^m \alpha_i\ell_{0/1}(z,q_i)$，其权重为 $\alpha_1,\ldots,\alpha_m \in \mathbb{R}_+$，则集成损失可分解如下，其中归一化权重为 $a_i = \alpha_i/\sum_{j=1}^m \alpha_j$。

$$\mathbb{E}_Y\left[\mathbb{E}_D\left[\ell_{0/1}(Y,\overline{q})\right]\right] = \underbrace{\mathbb{E}_Y\left[\ell_{0/1}(Y,Y^*)\right]}_{\text{noise}} + \underbrace{\sum_{i=1}^m \mathbb{E}_D\left[a_i\ \mathbb{E}_Y\left[\ell_{0/1}(Y,\mathring{q}_i) - \ell_{0/1}(Y,Y^*)\right]\right]}_{\text{weighted average bias-effect}}$$
$$+ \underbrace{\sum_{i=1}^m \mathbb{E}_D\left[a_i\ \mathbb{E}_Y\left[\ell_{0/1}(Y,q_i) - \ell_{0/1}(Y,\mathring{q}_i)\right]\right]}_{\text{weighted average variance-effect}} - \underbrace{\mathbb{E}_D\left[\mathbb{E}_Y\left[\sum_{i=1}^m a_i\ \ell_{0/1}(Y,q_i) - \ell_{0/1}(Y,\overline{q})\right]\right]}_{\text{diversity-effect}},$$

其中 $\mathring{q}_i = \arg\min_{z\in\mathcal{Y}}\mathbb{E}_D\left[\alpha_i\ell_{0/1}(z,q_i)\right]$，注意到 α_i 和 q_i 都依赖于 D。

与前文一样，通过消去等式右侧的项即可看出这一结果的正确性。AdaBoost 会生成一组二分类器 $h_i \in \{-1,+1\}$ 及其对应的权重 $\alpha_i(D) \in \mathbb{R}$，因此令 $q_i = h_i$ 即可直接应用该分解。LogitBoost 不会生成分类器/权重对，而是生成一组回归模型，每个模型 $g_i \in \mathbb{R}$。我们可以将这些模型拆分为符号/幅度分量来应用该分解，从而得到一个分类 $q_i = \text{sign}(g_i)$ 和一个权重：$\alpha_i(D) = |g_i|$。

附录 D. 第 5 节的证明

D.1 证明 布雷格曼歧义 / 偏差-方差-多样性分解的证明。

严格来说，本小节的前两个定理是以下两者的推论：布雷格曼散度存在偏差-方差分解（Pfau，2013年），以及定理 5。不过，为了完整性和教学需要，我们在此分别给出证明。第一个证明将用到布雷格曼三点性质（Banerjee 等人，2005年）：

定义 21（布雷格曼三点性质）给定一个凸集 $\mathcal{S} \subseteq \mathbb{R}^k$、一个严格凸函数 ϕ： $\mathcal{Y} \to \mathbb{R}$，以及任意的 $\mathbf{x}_1 \in \mathcal{Y}$、$\mathbf{x}_2$,$\mathbf{x}_3 \in \mathrm{ri}(\mathcal{S})$，以下恒等式成立：

$$B_\phi\left(\mathbf{x}_1, \mathbf{x}_3\right) = B_\phi\left(\mathbf{x}_1, \mathbf{x}_2\right) + B_\phi\left(\mathbf{x}_2, \mathbf{x}_3\right) + \langle \mathbf{x}_1 - \mathbf{x}_2 \ , \ \nabla\phi(\mathbf{x}_2) - \nabla\phi(\mathbf{x}_3)\rangle. \tag{47}$$

定理 11（布雷格曼歧义分解）对于标签 $\mathbf{y} \in \mathcal{Y}$ 和一组预测 $\mathbf{q}_1$、.....、$\mathbf{q}_m \in \mathrm{ri}(\mathcal{Y})$，组合为 $\overline{\mathbf{q}} = [\nabla\phi]^{-1}\big(\frac{1}{m}\sum_{i=1}^m \nabla\phi(\mathbf{q}_i)\big)$。则有：

$$B_\phi\left(\mathbf{y}, \overline{\mathbf{q}}\right) = \frac{1}{m}\sum_{i=1}^m B_\phi\left(\mathbf{y}, \mathbf{q}_i\right) - \frac{1}{m}\sum_{i=1}^m B_\phi\left(\overline{\mathbf{q}}, \mathbf{q}_i\right). \tag{17}$$

P证明 我们将三点性质应用于 $\mathbf{x}_1 = \mathbf{y}$、$\mathbf{x}_2 = \mathbf{q}$ 和 $\mathbf{x}_3 = \overline{\mathbf{q}}_i$.

$$B_\phi\left(\mathbf{y}, \mathbf{q}_i\right) = B_\phi\left(\mathbf{y}, \overline{\mathbf{q}}\right) + B_\phi\left(\overline{\mathbf{q}}, \mathbf{q}_i\right) + \langle \mathbf{y} - \overline{\mathbf{q}} \ , \ \nabla\phi(\overline{\mathbf{q}}) - \nabla\phi(\mathbf{q}_i)\rangle \tag{48}$$

现在，对该表达式两边的 m 模型取平均：

$$\begin{aligned}
\frac{1}{m}\sum_{i=1}^m B_\phi\left(\mathbf{y}, \mathbf{q}_i\right) &= B_\phi\left(\mathbf{y}, \overline{\mathbf{q}}\right) + \frac{1}{m}\sum_{i=1}^m B_\phi\left(\overline{\mathbf{q}}, \mathbf{q}_i\right) + \frac{1}{m}\sum_{i=1}^m \Big[\langle \mathbf{y} - \overline{\mathbf{q}} \ , \ \nabla\phi(\overline{\mathbf{q}}) - \nabla\phi(\mathbf{q}_i)\rangle\Big] \\
&= B_\phi\left(\mathbf{y}, \overline{\mathbf{q}}\right) + \frac{1}{m}\sum_{i=1}^m B_\phi\left(\overline{\mathbf{q}}, \mathbf{q}_i\right) + \langle \mathbf{y} - \overline{\mathbf{q}} \ , \ \nabla\phi(\overline{\mathbf{q}}) - \frac{1}{m}\sum_{i=1}^m \nabla\phi(\mathbf{q}_i)\rangle \\
&= B_\phi\left(\mathbf{y}, \overline{\mathbf{q}}\right) + \frac{1}{m}\sum_{i=1}^m B_\phi\left(\overline{\mathbf{q}}, \mathbf{q}_i\right), \qquad (49)
\end{aligned}$$

其中，在最后一步中，我们重新整理了$\overline{\mathbf{q}}$ 的定义，以使用$\nabla\phi(\overline{\mathbf{q}}) = \frac{1}{m}\sum_{i=1}^m \nabla\phi(\mathbf{q}_i)$。最后，重新整理公式 (49) 中的各项，即可得到所需结果公式 (17)。

■

定理13（布雷格曼偏差-方差-多样性分解）
对于模型集成 $\{\mathbf{q}_i\}_{i=1}^m$，令 $\mathring{\mathbf{q}}_i$ 为模型$\mathbf{q}_i$的左布雷格曼质心，即 $\mathring{\mathbf{q}}_i \overset{\text{定义为}}{=} [\nabla\phi]^{-1}(\mathbb{E}_D[\nabla\phi(\mathbf{q}_i)])$，并定义集成输出$\mathbf{q} \overset{\text{定义为}}{=} [\nabla\phi]^{-1}\left(\frac{1}{m}\sum_{i=1}^m \nabla\phi(\mathbf{q}_i)\right)$。于是我们得到如下分解

$$\mathbb{E}_D\left[\mathbb{E}_{\mathbf{XY}}\left[B_\phi\left(\mathbf{Y},\overline{\mathbf{q}}\right)\right]\right] =$$

$$\mathbb{E}_{\mathbf{X}}\Bigg[\underbrace{\mathbb{E}_{\mathbf{Y}|\mathbf{X}}B_\phi\left(\mathbf{Y},\mathbf{Y}^*\right)}_{\text{noise}} + \underbrace{\frac{1}{m}\sum_{i=1}^m B_\phi\left(\mathbf{Y}^*,\mathring{\mathbf{q}}_i\right)}_{\text{average bias}} + \underbrace{\frac{1}{m}\sum_{i=1}^m \mathbb{E}_D\left[B_\phi\left(\mathring{\mathbf{q}}_i,\mathbf{q}_i\right)\right]}_{\text{average variance}} - \underbrace{\mathbb{E}_D\Big[\frac{1}{m}\sum_{i=1}^m B_\phi\left(\overline{\mathbf{q}},\mathbf{q}_i\right)\Big]}_{\text{diversity}}\Bigg],$$

其中$\mathbf{Y}^* = \mathbb{E}_{\mathbf{Y}|\mathbf{X}}[\mathbf{Y}]$。

证明 取$\mathbf{q}$的期望风险，并应用布雷格曼歧义分解：

$$\mathbb{E}_D\left[\mathbb{E}_{\mathbf{XY}}\left[B_\phi\left(\mathbf{Y},\overline{\mathbf{q}}\right)\right]\right] = \mathbb{E}_D\left[\mathbb{E}_{\mathbf{XY}}\left[\frac{1}{m}\sum_{i=1}^m B_\phi\left(\mathbf{Y},\mathbf{q}_i\right)\right]\right] - \mathbb{E}_D\left[\mathbb{E}_{\mathbf{XY}}\left[\frac{1}{m}\sum_{i=1}^m B_\phi\left(\overline{\mathbf{q}},\mathbf{q}_i\right)\right]\right]. \tag{50}$$

现在对右端第一项应用 Pfau 分解（公式（15）），得到

$$\mathbb{E}_D\Big[\mathbb{E}_{\mathbf{XY}}\Big[\frac{1}{m}\sum_{i=1}^m B_\phi(\mathbf{Y},\mathbf{q}_i)\Big]\Big] = \mathbb{E}_{\mathbf{X}}\Bigg[\mathbb{E}_{\mathbf{Y}|\mathbf{X}}\left[B_\phi\left(\mathbf{Y},\mathbf{Y}^*\right)\right] + \frac{1}{m}\sum_{i=1}^m B_\phi\left(\mathbf{Y}^*,\mathring{\mathbf{q}}_i\right) + \frac{1}{m}\sum_{i=1}^m \mathbb{E}_D\left[B_\phi\left(\mathring{\mathbf{q}}_i,\mathbf{q}_i\right)\right]\Bigg]. \tag{51}$$

将公式（51）代入（50），即可完成证明。 ■

定理14 集成偏差和集成方差可以改写为单个模型的函数：

$$\underbrace{B_\phi\left(\mathbf{y},\mathring{\overline{\mathbf{q}}}\right)}_{\text{ensemble bias}} = \underbrace{\frac{1}{m}\sum_{i=1}^m B_\phi\left(\mathbf{y},\mathring{\mathbf{q}}_i\right)}_{\text{average bias}} - \Delta, \tag{19}$$

$$\underbrace{\mathbb{E}_D\left[B_\phi\left(\mathring{\overline{\mathbf{q}}},\overline{\mathbf{q}}\right)\right]}_{\text{ensemble variance}} = \Delta + \underbrace{\frac{1}{m}\sum_{i=1}^m \mathbb{E}_D\left[B_\phi\left(\mathring{\mathbf{q}}_i,\mathbf{q}_i\right)\right]}_{\text{average variance}} - \underbrace{\mathbb{E}_D\left[\frac{1}{m}\sum_{i=1}^m B_\phi\left(\overline{\mathbf{q}},\mathbf{q}_i\right)\right]}_{\text{diversity}}, \tag{20}$$

其中的公共项为 $\Delta = \frac{1}{m}\sum_{i=1}^m B_\phi\left(\mathring{\overline{\mathbf{q}}},\mathring{\mathbf{q}}_i\right)$。

证明
公式（19）可通过将定理 11 应用于一组质心模型 $\{\mathring{\mathbf{q}}_i\}_{i=1}^m$ 来证明。公式（20）可通过应用定理13，并将 $\mathbf{y} = \mathring{\overline{\mathbf{q}}}$ 代入来证明。 ■

D.2 第5.4节的证明

命题 15设真实概率向量为 $\mathbf{y} \in \mathbb{R}^k$，一组模型为 $\{\mathbf{q}_i\}_{i=1}^m$，并通过平均进行组合，即 $\mathbf{q}^\dagger = \frac{1}{m}\sum_{i=1}^m \mathbf{q}_i$，则交叉熵损失为

$$\underbrace{-\mathbf{y}\cdot\ln\mathbf{q}^\dagger}_{\text{集成交叉熵}} = \underbrace{-\frac{1}{m}\sum_{i=1}^m \mathbf{y}\cdot\ln\mathbf{q}_i}_{\text{平均交叉熵}} - \underbrace{\sum_{c=1}^k \mathbf{y}^{(c)}\ln\frac{\frac{1}{m}\sum_{j=1}^m \mathbf{q}_j^{(c)}}{\left(\prod_{i=1}^m \mathbf{q}_i^{(c)}\right)^{\frac{1}{m}}}}_{\text{歧义（依赖于标签）}}, \tag{52}$$

其中第二项非负，因此可以保证集成损失小于或等于各个模型损失的平均值。

证明 取平均交叉熵，并减去集成交叉熵:

$$\begin{aligned}
-\frac{1}{m}\sum_{i=1}^m \mathbf{y}\cdot\ln\mathbf{q}_i - \left[-\mathbf{y}\cdot\ln\mathbf{q}^\dagger\right] &= \sum_{c=1}^k \mathbf{y}^{(c)}\ln\mathbf{q}^{\dagger(c)} - \frac{1}{m}\sum_{i=1}^m\sum_{c=1}^k \mathbf{y}^{(c)}\ln\mathbf{q}_i^{(c)} \\
&= \sum_{c=1}^k \mathbf{y}^{(c)}\ln\mathbf{q}^{\dagger(c)} - \sum_{c=1}^k \mathbf{y}^{(c)}\ln\left(\prod_i \mathbf{q}_i^{(c)}\right)^{1/m} \\
&= \sum_{c=1}^k \mathbf{y}^{(c)}\ln\left(\frac{\mathbf{q}^{\dagger(c)}}{\prod_{i=1}^m\left(\mathbf{q}_i^{(c)}\right)^{\frac{1}{m}}}\right)
\end{aligned}$$

利用$\mathbf{q}^\dagger$ 的定义并重新整理，即可完成推导。根据算术—几何平均不等式，$\mathbf{q}^{\dagger(c)} \geq \prod_{i=1}^m\left(\mathbf{q}_i^{(c)}\right)^{1/m}$，这意味着对数内部的项大于或等于 1，因此整体项非负。

■

命题 16 设$\mathbf{q}^\dagger = \frac{1}{m}\sum_{i=1}^m \mathbf{q}_i$，其中$\mathbf{q}_i \in [0,1]^k$。期望交叉熵可分解为:

$$\begin{aligned}
&-\mathbb{E}_D\left[\mathbf{y}\cdot\ln\mathbf{q}^\dagger\right] = \\
&\underbrace{-\frac{1}{m}\sum_{i=1}^m \mathbf{y}\cdot\ln\mathring{\mathbf{q}}_i}_{\text{average bias}} + \underbrace{\frac{1}{m}\sum_{i=1}^m \mathbb{E}_D\left[K(\mathring{\mathbf{q}}_i \,||\, \mathbf{q}_i)\right]}_{\text{average variance}} \underbrace{-\mathbb{E}_D\left[\sum_{c=1}^k \mathbf{y}^{(c)}\ln\frac{\frac{1}{m}\sum_{j=1}^m \mathbf{q}_j^{(c)}}{\left(\prod_{i=1}^m \mathbf{q}_i^{(c)}\right)^{\frac{1}{m}}}\right]}_{\text{dependency}}.
\end{aligned}$$

证明 从公式（52）开始，并对 D 取期望，我们有

$$-\mathbb{E}_D\left[\mathbf{y}\cdot\ln\mathbf{q}^\dagger\right] = \mathbb{E}_D\left[-\frac{1}{m}\sum_{i=1}^m \mathbf{y}\cdot\ln\mathbf{q}_i\right] - \mathbb{E}_D\left[\sum_{c=1}^k \mathbf{y}^{(c)}\ln\frac{\frac{1}{m}\sum_{j=1}^m \mathbf{q}_j^{(c)}}{\left(\prod_{i=1}^m \mathbf{q}_i^{(c)}\right)^{\frac{1}{m}}}\right].$$

现在，将 KL偏差-方差分解应用于第一项，即可得证。 ■

参考文献

Kartik Audhkhasi、Abhinav Sethy、Bhuvana Ramabhadran 和 Shrikanth Narayanan。《用于理解集成多样性的广义歧义分解》。*arXiv:1312.7463*，2013年。Arindam Banerjee、Srujana Merugu、Inderjit S Dhillon 和 Joydeep Ghosh。《使用布雷格曼散度进行聚类》。《机器学习研究期刊》，6(10月)：1705–1749，2005年。Mikhail Belkin、Daniel Hsu、Siyuan Ma 和 Soumik Mandal。《调和现代机器学习与偏差-方差权衡》。*arXiv*预印本 *arXiv:1812.11118*，2018年。Yijun Bian 和 Huanhuan Chen。《分类集成中的多样性何时有助于泛化？》*arXiv*预印本 *arXiv:1910.13631*，2019年。

Battista Biggio、Igino Corona、Giorgio Fumera、Giorgio. Giacinto 和 Fabio Roli。用于抵御对抗分类任务中投毒攻击的装袋分类器。载于国际多分类器系统研讨会，第 350–359 页。Springer，2011年。

Lev M Bregman。求取凸集公共点的松弛法及其在解决凸规划问题中的应用。苏联计算数学与数学物理，7(3)：200–217，1967年。

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。

Gavin Brown。多分类器系统的信息论视角。载于国际多分类器系统研讨会，第 344–353 页。Springer，2009。

Gavin Brown 和 Ludmila Kuncheva。多数投票集成中的“好”与“坏”多样性。载于多分类器系统研讨会，第 124–133 页。Springer，2010年。

Gavin Brown、Jeremy Wyatt、Rachel Harris 和 Xin Yao。多样性创建方法：综述与分类。信息融合，6(1):5 – 20，2005年。

Gavin Brown、Jeremy L Wyatt 和 Peter Tiˇno。管理回归集成中的多样性。《机器学习研究期刊》，6(Sep):1621–1650，2006。

Sebastian Buschj¨ager、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。Yue Cao、Thomas Andrew Geddes、Jean Yee Hwa Yang 和 Pengyi Yang。生物信息学中的集成深度学习。*Nature Machine Intelligence*，第1–9页，2020年。Miguel A Carreira-Perpin´an 和 Ramin Raziperchikolaei。用于有监督二进制哈希的集成多样性方法。载于神经信息处理系统，2016年。Luca Didaci、Giorgio Fumera 和 Fabio Roli。分类器集成中的多样性：富有成效的概念还是死胡同？载于多分类器系统研讨会。Springer，2013年。

Thomas G Dietterich。机器学习中的集成方法。载于国际多分类器系统研讨会，第1–15页。Springer，2000年。Pedro Domingos。统一的偏差-方差分解。载于国际机器学习会议，第231–238页，2000年。Stuart Geman、Elie Bienenstock 和 Ren´e Doursat。神经网络与偏差/方差困境。神经计算，4(1):1–58，1992年。Pierre Geurts。决策树归纳的研究：偏差/方差权衡与时间序列分类。博士论文，比利时列日大学，2002年。Ian Goodfellow、Yoshua Bengio 和 Aaron Courville。深度学习。麻省理工学院出版社，2016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。URL https://openreview.net/forum?id=lIOQFVncY9。Lars K. Hansen 和 Peter Salamon。神经网络集成。模式分析与机器智能，*IEEE*汇刊，12:993–1001，11 1990年。doi：10.1109/34.58871。Tom Heskes。基于似然的估计量的偏差/方差分解。神经计算，10(6):1425–1433，1998年。Geoffrey Hinton、Oriol Vinyals 和 Jeff Dean。神经网络中的知识蒸馏。*arXiv*预印本 *arXiv:1503.02531*，2015年。Carina Iva˛scu、Richard Everson 和 Jonathan Fieldsend。优化分类树集成分类器的多样性。载于进化计算应用国际会议。，2021年。Gareth James 和 Trevor Hastie。预测误差偏差/方差分解的推广。技术报告，斯坦福大学统计学系，斯坦福，加利福尼亚州，1997年。Muhammad Zohaib Jan 和 Brijesh Verma。一种用于生成集成分类器的新型多样性度量与分类器选择方法。*IEEE Access*，7，2019年。Zhengshen Jiang、Hongzhi Liu、Bin Fu 和 Zhonghai Wu。用于分类的广义歧义分解及其在主动学习和无监督集成剪枝中的应用。载于第三十一届*AAAI*人工智能大会，2017年。Ron Kohavi、David H Wolpert 等。零一损失函数的偏差加方差分解。载于*ICML*，第96卷，第275–83页，1996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、Ilya Sutskever 和 Geoffrey Hinton。使用深度卷积神经网络进行ImageNet分类。载于神经信息处理系统，2012年。

Anders Krogh 和 Jesper Vedelsby。神经网络集成、交叉验证与主动学习。载于神经信息处理系统，第 231–238 页，1994年。Ludmila Kuncheva。分类器集成中那种难以捉摸的多样性。载于伊比利亚模式识别与图像分析会议，第 1126–1138 页。Springer，2003年。Ludmila Kuncheva。模式分类器的组合：方法与算法。第 2 版，约翰·威立父子出版公司，2014年。Ludmila Kuncheva 和 Christopher Whitaker。分类器集成的多样性度量及其与集成准确率的关系。机器学习，51(2)，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，2017年。Louisa Lam 和 SY Suen。多数投票在模式识别中的应用：对其行为和性能的分析。*IEEE*《系统、人类与控制论汇刊》*A*辑：系统与人类，27(5):553–568，1997年。Yong Liu 和 Xin Yao。通过负相关实现集成学习。神经网络，12 (10):1399–1404，1999年。David Mease 和 Abraham Wyner。与提升算法统计观点相悖的证据。《机器学习研究期刊》，9(Feb):131–156，2008年。Frank Nielsen 和 Richard Nock。有向与对称化布雷格曼质心。*IEEE*《信息论汇刊》，55(6):2882–2904，2009。David W Opitz 和 Jude W Shavlik。生成准确且多样化的神经网络集成成员。载于神经信息处理系统，第 535–541 页，1996年。Luis A Ortega、Rafael Caba˜nas 和 Andr´es R Masegosa。神经网络集成中的多样性与泛化。*AISTATS 2022，arXiv* 预印本 *arXiv:2110.13786*，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。机器学习国际会议，2019年。Michael P Perrone 和 Leon N Cooper。当网络意见不一致时：混合神经网络的集成方法。技术报告，布朗大学脑与神经系统研究所，1992年。David Pfau。布雷格曼散度的广义偏差-方差分解。技术报告，哥伦比亚大学，2013年。Robert Schapire、Yoav Freund、Peter Bartlett 和 Wee Sun Lee。提升间隔：对投票方法有效性的新解释。《统计年鉴》，26(5): 1651–1686，1998年。

Samarth Sinha、Homanga Bharadhwaj、Anirudh Goyal、Hugo Larochelle、Animesh Garg 和 Florian Shkurti。Dibs：模型集成中的多样性诱导信息瓶颈。*arXiv:2003.04514*，2020年。Cedrique Tassi、Jakob Gawlikowski、Auliya Fitri 和 Rudolph Triebel。对神经网络集成而言，对 logits 求平均与对概率求平均的影响。载于*2022*年人工智能安全研讨会，*AISafety 2022*，2022年。David MJ Tax、Robert PW Duin 和 Martijn Van Breukelen。乘积分类器组合规则与平均分类器组合规则的比较。载于统计模式识别研讨会论文集，布拉格，捷克，第 39 页。Citeseer，1997年。Robert Tibshirani。分类规则的偏差、方差与预测误差。技术报告，多伦多大学，1996年。Kagan Tumer 和 Joydeep Ghosh。集成分类器中的误差相关性与误差降低。*Connection Science*，8(3-4):385–404，1996年。Naonori Ueda 和 Ryohei Nakano。集成估计器的泛化误差。载于国际神经网络会议 *(ICNN'96)*，第 1 卷。IEEE，1996年。Paul Viola 和 Michael Jones。使用简单特征的级联提升实现快速目标检测。载于*IEEE*计算机视觉与模式识别会议，2001年。Andrew Webb、Charles Reynolds、Wenlin Chen、Henry Reeve、Dan Iliescu、Mikel Lujan 和 Gavin Brown。是否采用集成：端到端训练何时会失败？载于欧洲机器学习会议，第 109–123 页，2021年。Danny Wood、Tingting Mu 和 Gavin Brown。间隔损失的偏差-方差分解。载于国际人工智能与统计会议，2022年。Yanzhao Wu、Ling Liu、Zhongwei Xie、Ka-Ho Chow 和 Wenqi Wei。重新审视集成多样性指标以提升集成准确率。载于计算机视觉与模式识别会议，2021年。Zitong Yang、Yaodong Yu、Chong You、Jacob Steinhardt 和 Yi Ma。重新思考神经网络泛化中的偏差-方差权衡。载于机器学习国际会议，2020年。Zhi-Hua Zhou。集成方法：基础与算法。Chapman & Hall/CRC，2012年。Zhi-Hua Zhou 和 Nan Li。多信息集成多样性。载于国际多分类器系统研讨会，第 134–144 页。Springer，2010年。