超越集成聚合中的混合与乘积：广义均值的似然视角

Raphaël Razafindrala **mbo**[1] **Rémy Sun**[1] **Frédéric Precioso**[1] **Damien Garreau**[2] **Pierre-Alexandre Matt** **ei**[1]

[1]蔚蓝海岸大学、Inria、法国国家科学研究中心、I3S/LJAD、Maasai，尼斯，法国[2]维尔茨堡尤利乌斯-马克西米利安大学，计算机科学研究所 / CAIDAS，维尔茨堡，德国

arXiv:2603.04204v1 [stat.ML] 4 Mar 2026

摘要

密度聚合是机器学习中的核心问题，例如在结合深度集成的预测时便会遇到这一问题。聚合方式的选择仍未有定论，目前常见的两种方法是线性池化（概率平均）和几何池化（logit 平均）。在这项工作中，我们从对数似然这一机器学习中的标准评估准则出发，研究阶数为 $r \in \mathbb{R} \cup \{-\infty, +\infty\}$ 的归一化广义平均，以探讨这一问题。由此，我们提出一种统一的聚合形式，并展示了在不同情形下的不同最优配置。我们证明，$r \in [0, 1]$ 这一范围是唯一能够确保相较于单个分布实现稳定改进的范围，从而为线性（$r = 1$）和几何（$r = 0$）池化的可靠性及其广泛的实际应用提供了有理有据的解释。相反，我们通过明确的反例表明，$r \notin [0, 1]$ 的聚合规则可能无法带来稳定的收益。最后，我们在图像和文本分类基准上使用深度集成进行实证评估，验证了理论发现。

1 引言

现代机器学习（ML）越来越依赖于概率模型的集合，而非单一的整体式预测器（Lakshminarayanan 等人，2017年；Liu 等人，2024）。这自然引出了一个问题：如何将多个概率分布整合为一个连贯的分布（French，1983；Genest 和 Zidek，1986 年）。分布整合出现在机器学习的许多场景中，包括分类器输出的集成（Kuncheva，2014，第5节）、Mixup 训练技术（El-Laham 等人，2025年）、贝叶斯推断（Carvalho 等人，2023）、高斯过程专家的聚合（Liu 等人，2018；Cohen 等人，2020）、大语言模型中的混合专家架构（Liu 等人，2024；Cai 等人，2025年）以及生成建模（Liu 等人，2022；Razafindralambo 等人，2026）。

两种典型方法尤为普遍。第一种是构造分布混合（Jacobs 等人，1991年；Jordan 和 Jacobs，1994年），对应于密度的算术平均。第二种依赖于乘性聚合，即通过归一化乘积来组合分布（Hinton，2002年）。这些方法具有截然不同的性质（Winkler，1968年）：混合起到逻辑“或”的作用，通过汇集支撑集来刻画异质性；而乘性模型则起到逻辑“与”的作用，在各分量重叠处使密度更为尖锐，从而偏向共识区域。这一观点在组合式生成建模中反复出现（Du 等人，2023；Liu 等人，2022年）。更一般地，在机器学习中，二者之间的选择一直是一个根本性且具有实践意义的问题；例如，Buchanan 等人（2023）在多类分类中对 logit 平均与概率平均进行了实证比较。

混合方法的一个关键理论动机在于其方差降低效应，使集体评估能够优于个体的平均表现（Hastie 等人，2009，第15章；Mattei 和 Garreau，2025年），这一现象与更广泛的“群体智慧”原则相关（Surowiecki，2004年）。当预测器具有多样性时，个体误差往往相互补偿，从而提升集体表现。

在经典的算术平均数和几何平均数之外，已有若干工作试图进一步超越并推广对概率分布求平均的概念（Genest 和 Zidek，1986年；Amari，2007年），以揭示不同汇集规则背后新的聚合性质与根本权衡（Elkin 和 Pettigrew，2025年）。这些努力催生了汇集算子的连续族，它们扩展了标准平均，并涵盖从调和平均数和算术平均数到更为

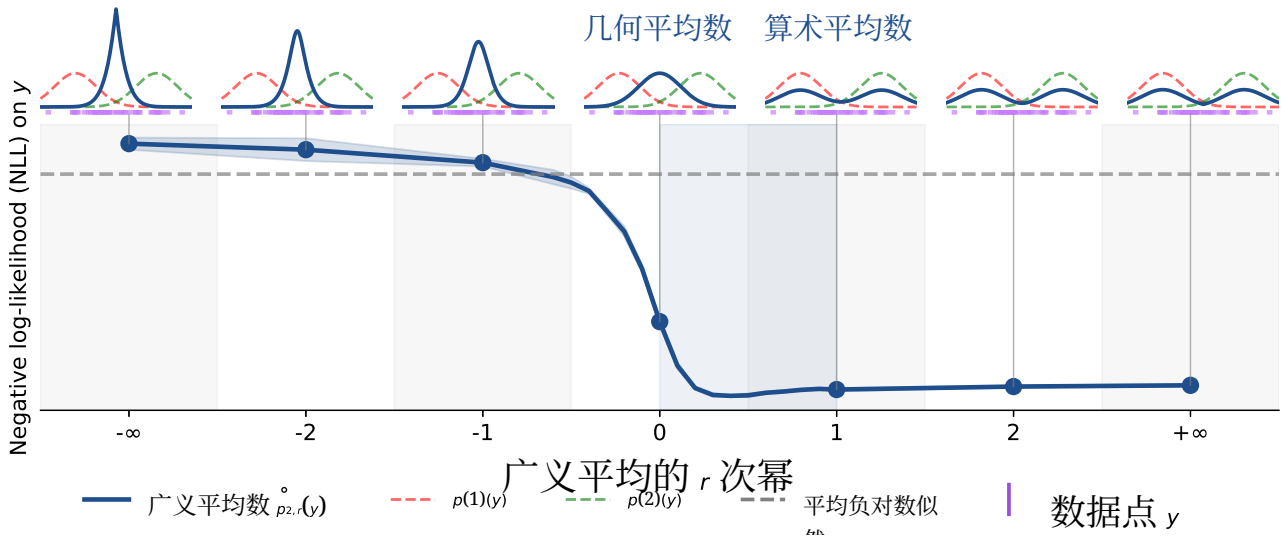


图1：**只有中间阶数 r，尤其是 $r \in [0,1]$，才能带来一致的负对数似然改进，而极端值可能导致性能下降。上：由两个高斯专家在不同下得到的**聚合密度 $\mathring{p}_2$。较小的 r 会将质量集中在单个区域，而较大的正 r 则保留专家的双峰结构。**下：** 在样本 $y \sim \mathcal{N}(0, 4^2)$ 上评估的负对数似然（紫色刻度表示测试样本）。在此设置下，负 r（最小值型行为）的表现劣于单个模型平均负对数似然（虚线），而正 r 取得最低值。我们的理论（定理 3.1）确定了 $r \in [0,1]$ 这一可证明可靠的区间。不同的行为见附录B。

极端的算子，例如基于最大值或基于最小值的池化。特别地，Amari（2007年）引入了由自由参数 α参数化的 α族算子，并针对 α散度为该族的每个成员确立了最优性性质。

我们考察通过阶数为 $r \in \mathbb{R}$ 的广义均值：$= \mathbb{R} \cup \{-\infty, +\infty\}$（Hardy 等人，1952年）进行密度聚合，是否能够为模型集成提供改进的聚合规则。在机器学习中，评估聚合分布的一项自然标准是数据对数似然，因为它直接衡量聚合模型对观测数据的解释能力。这一目标是现代学习过程的核心，因为它既支撑监督分类中的交叉熵损失最小化，也支撑更广泛场景中的基于KL散度的目标，例如生成建模和变分推断（Bishop，2006年，第10章；Theis 等人，2015年；Goodfellow 等人，2016年，第5章）。因此，这一视角促使我们从理论和实证两方面考察：方差降低是否能够扩展到广义池化算子，并改善相较于单个模型的对数似然。

在本研究中，我们证明，所定义的阶数为 r 的归一化广义平均，在对数似然方面是一种可靠的聚合规则；理论保证和经验证据均表明，对于 $r \in [0,1]$，其表现始终优于单个分布。图1概述了这一思想，展示了归一化广义平均密度的形状如何随 r 变化，以及负的r 值所导致的负面对数似然的不可靠表现。

我们的主要贡献在于从对数似然的角度分析阶数为 $r \in \mathbb{R}$的广义平均，该平均定义了有效密度（第3.1节），其性质由 r 决定（第2节和第3节）。具体而言，这一分析得出以下结果：

- 我们在第 3.2 节中确定了一个理论区间 $r \in [0,1]$，在该区间内，对于任意一组密度，在所有数据点上都能一致地获得对数似然收益。该区间以算术平均和几何平均数为边界情形，解释了它们在实践中的成功。
- 我们在第 3.3 节中，根据密度值的一致与分歧，刻画了该区间之外的失效点，并突出了 $r < 0$ 和 $r > 1$ 的不同表现。这解释了与区间内的保证相比，$[0,1]$ 之外缺乏可靠性；图1中也观察到了这一点。

- 我们通过在视觉和情感分析分类任务上开展深度集成实验（第4节），从经验上验证了理论发现。我们还观察到，最优的 r 并非显而易见，但接近 $[0,1]$ 范围。

2 广义均值的背景与记号

我们考虑一个配备参考测度 dx 的数据空间 $\mathcal{X}$。我们关注那些存在相对于

dx 的密度的概率分布。我们用 $\mathcal{P}$ 表示满足以下条件的概率密度集合 $p:\ \mathcal{X}\to\mathbb{R}_{\geq 0}\ \int_{\mathcal{X}} p(\mathbf{x})\mathrm{d}\mathbf{x}\ =1$。给定 $k\geq 1$ 个密度 $p^{(1)},\ldots,p^{(k)}\in\mathcal{P}$，本文研究如何通过广义平均算子组合这些分布来构造聚合密度 $\mathring{p}$，并且在大多数情况下，随后进行归一化步骤，以确保 $\mathring{p}\in\mathcal{P}$。

我们首先介绍混合聚合和乘法聚合的形式化记号，因为它们是本文结果最常用的参照聚合方式。随后，我们转向 $\mathbb{R}$ 中的 r 阶广义平均，它是本文的核心研究对象，并涵盖了这些算子。

2.1 经典示例：线性池化和对数池化

下面介绍概率密度函数 $p^{(1)}$ 和 $p^{(2)}$ 在 $\mathcal{P}$ 中的两种标准组合，并回顾它们的一些基本性质。

混合聚合，也称为 线性池化 / 意见池 （Stone，1961；Genest 等，1986年；Clemen 和 Winkler，1999年；Elkin 和 Pettigrew，2025 年），对于两个概率 ..... 定义如下：

$$\mathring{p}_{\text{lin}}(\mathbf{x}) \coloneqq \frac{1}{2}\Big(p^{(1)}(\mathbf{x})+p^{(2)}(\mathbf{x})\Big). \tag{1}$$

这一规则高度民主，因为每个密度在聚合中都保留了影响力，往往会产生多峰分布。该混合具有一种颇具吸引力的变分刻画（Amari，2007）：它最小化平均（正向）Kullback–Leibler散度，

$$\mathring{p}_{\text{lin}} = \arg\min_{q\in\mathcal{P}} \frac{1}{2}\sum_{i=1}^{2}\mathrm{KL}\Big(p^{(i)}\,\|\,q\Big). \tag{2}$$

乘法式聚合定义为归一化乘积

$$\mathring{p}_{\text{mult}}(\mathbf{x}) \coloneqq \sqrt{p^{(1)}(\mathbf{x})\,p^{(2)}(\mathbf{x})}/Z, \tag{3}$$

其中 $Z=\int_{\mathcal{X}}\sqrt{p^{(1)}(\mathbf{x})p^{(2)}(\mathbf{x})}\mathrm{d}\mathbf{x}$ 为归一化常数。这种运算也通常称为专家乘积（PoE）（Hinton，2002；Cohen 等人，2020；Razafindralambo 等人，2026），但在文献中也称为对数池化 / 意见池（Carvalho 等人，2023；Genest 等人，1986；Clemen 和 Winkler，1999；Elkin 和 Pettigrew，2025）或指数混合（Amari，2007）。PoE 是定义良好的，因为 $\mathcal{P}$ 中密度的几何平均数仍然可以归一化（Genest 等人，1986；Carvalho 等人，2023）。

PoE 具有若干与混合模型形成鲜明对比的理论性质。PoE 最小化相对于单个的平均反向 KL 散度模型（Amari，2007年），

$$\mathring{p}_{\text{mult}} = \arg\min_{q\in\mathcal{P}} \frac{1}{2}\sum_{i=1}^{2}\mathrm{KL}\Big(q\,\|\,p^{(i)}\Big). \tag{4}$$

此外，它通常表现得更加悲观 （甚至具有独裁性）：任何单个模型赋予低密度的区域都会受到严厉惩罚；如果对于某个 k，$p^{(k)}(x)=0$，那么 $\mathring{p}_{\text{mult}}(x)=0$ 也成立。因此，与混合分布相比，PoE 聚合往往会产生更加集中的分布，且通常更接近单峰分布（Winkler，1968；Genest 和 Zidek，1986）。图1对此作了说明（见 $r=0$）。

PoE 的其他显著性质包括独立性保持 （Laddaga，1977；Cooke，1991，第11章）以及外部贝叶斯性质（Genest 等人，1986；Cooke，1991，第11章）。Elkin 和 Pettigrew（2025）（第4节）还对线性池化和对数池化的性质作了更全面的概述。

这两种聚合方案也有信息几何解释：它们分别对应连接 $p^{(1)}$ 和 $p^{(2)}$ 的两条自然测地线的中点（参见 Amari，2016年，第 2.4 节）。

2.2 广义幂平均

现在，我们讨论一种广义聚合框架，它扩展了上述定义的算子。在本研究中，我们采用 Hardy 等人（1952年）定义的广义平均，称为普通平均，并将其扩展到概率密度的情形。给定 k 正面值$a_1,\ldots,a_k$，我们将阶广义（或幂）平均 r 定义为

$$M_r(a_1,\ldots,a_k) \coloneqq \left(\frac{1}{k}\sum_{i=1}^{k}a_i^r\right)^{1/r},\quad r\in\mathbb{R}\setminus\{0\}. \tag{5}$$

在 $r=0$ 处的连续延拓对应于几何平均数，它自然地出现在 r 趋近于零时 M_r的极限中。我们定义

$$M_0(a_1,\ldots,a_k) \coloneqq \left(\prod_{i=1}^{k}a_i\right)^{1/k}. \tag{6}$$

极端情况下可还原最小值和最大值，

$$M_{-\infty}(a_1,\ldots,a_k) \coloneqq \lim_{r\to-\infty} M_r(a_1,\ldots,a_k) = \min_i a_i,$$

$$M_{+\infty}(a_1,\ldots,a_k) \coloneqq \lim_{r\to+\infty} M_r(a_1,\ldots,a_k) = \max_i a_i.$$

经典选取可还原为众所周知的平均数：$r=1$ 对应算术平均数（AM），$r=0$ 对应几何

平均数（GM），$r=-1$ 对应调和平均数。实际上，参数 r 控制广义平均的行为，可解释为乐观参数。r 的值越大，越强调较大的输入值，并形成乐观聚合；而较小（负面）的值则越强调较小的输入值，导致更悲观的行为。更确切地说，r 的值越大，平均数就越接近最大值的平滑近似；值越小，则越接近最小值的平滑近似。

广义平均的一个关键性质是其关于阶数 r 的单调性，这一性质将在我们的分析中发挥核心作用。

引理 2.1（幂平均不等式，Hardy 等人，1952年；第 2 章）。设 $a_1,\dots,a_k$ 为正值。若 $r<s$，

$$M_r(a_1,\dots,a_k)\le M_s(a_1,\dots,a_k). \tag{7}$$

等号成立当且仅当 $a_1=a_2=\cdots=a_k$。该结果推广了不等式链 $\frac{2}{\frac{1}{a_1}+\frac{1}{a_2}}\le\sqrt{a_1a_2}\le\frac{a_1+a_2}{2}$，其中左侧的不等式称为算术-几何平均不等式。

Hardy 等人（1952年）列举了幂平均的诸多基本性质，包括基本不等式。幂平均还可进一步推广为 f-均值，其聚合通过变换函数 f 定义；De Carvalho（2016年）将广义均值刻画为 f-均值，并论述了其统计性质。

概率密度的广义均值。现在，我们将这一概念扩展到概率密度的情形。给定 k 个概率密度 $p^{(1)},\dots,p^{(k)}\in\mathcal{P}$，定义

$$M_{k,r}(\mathbf{x}) \coloneqq M_r\big(p^{(1)}(\mathbf{x}),\dots,p^{(k)}(\mathbf{x})\big). \tag{8}$$

这会在 $\mathcal{X}$ 上得到一个非负函数，但它不一定积分为 1。例如，两个高斯密度的几何平均通常不会产生归一化密度，因此需要进行归一化步骤。

定义 2.1（密度的广义幂平均数）。与 r 相关联的阶数为 $p^{(1)}$、……、$p^{(k)}\in\mathcal{P}$ 的平均数定义为

$$\overset{\circ}{p}_{k,r}(\mathbf{x}) \coloneqq \frac{1}{Z_{k,r}}\,M_{k,r}(\mathbf{x}), \tag{9}$$

其中，归一化常数为

$$Z_{k,r} \coloneqq \int_{\mathcal{X}} M_{k,r}(\mathbf{x})\,\mathrm{d}\mathbf{x}. \tag{10}$$

根据构造，当 $Z_{k,r}$ 有限时，$\overset{\circ}{p}_{k,r}\in\mathcal{P}$，这一条件对所有 r 都成立（见命题 3.1）。

2.3 相关工作

与我们的方法密切相关的是，Cooke（1991年）研究了一类形式为 $\overset{\circ}{p}_{k,r}\propto M_{k,r}$ 的密度聚合算子，这与我们的 r 聚合族相一致。他们推导出了这一族的若干直接性质，包括与第 2.1 节中所述典型情形的性质类似的性质，以及公式（14）。

不同的是，Amari（2007）引入了一类 α-族积分，其中 $\alpha\in\mathbb{Z}\cup\{-\infty,+\infty\}$，它基于信息几何中密度的 α- 表示（Amari 和 Nagaoka，2000），也将经典平均数作为特殊情形纳入其中。他们将其平均数与 α-散度$D_\alpha(p\,\|\,q)$ 联系起来，对任意 $p,\ q\in\mathcal{P}$ 均成立（Chernoff，1952年；Amari 和 Nagaoka，2000），其中包括 α KL 散度、反向 KL 散度或海灵格距离平方等特殊情形。Amari（2007，定理 2）表明，在 α-散度准则下，α-积分是最优的，即它最小化了到各个分布的加权平均 α-散度。该结果推广了公式（2）和（4），并与关于广义预测密度的早期贝叶斯表述相关联（Corcuera 和 Giummolè，1999年）。

最后，一系列密切相关的研究以意见汇集（Elkin 和 Pettigrew，2025年）之名，探讨分布聚合的一般理论。在这一框架中，聚合被抽象地定义为一个无参数算子$F\colon(P_1,\dots,P_n)\mapsto P^\star$，它将一组概率分布映射为一个集体分布。相关文献刻画了这类算子所需的公理性质（例如，一致性保留、独立性、与贝叶斯更新的兼容性，……）。

3 广义均值聚合的似然保证

现在，我们通过分析广义幂平均聚合，并确定其在何种情况下能够保证对数似然系统性提升，来阐述本文的核心理论贡献。在第 3.1 节证明广义平均定义了一个可归一化密度后，我们在第 3.2 节证明，当阶数为 $r\in[0,1]$、恰好处于几何平均数与算术平均数之间时，这种提升能够实现，从而解释了它们为何具有稳健性。最后，我们在第 3.3 节通过反例表明，这一保证在该区间之外不再成立，且对于 $r<0$和 $r>1$，失效模式在性质上有所不同。

3.1 每个阶数 r 都定义一个密度吗?

一个需要考察的重要问题是，对于所有 r 的取值，归一化常数是否有限；下述命题表明确实如此。

命题 3.1 (阶数为 r 的广义均值是良定义的)。 假设 $p^{(1)}$，…，$p^{(k)}$ 是定义在 $\mathcal{X}$ 上的 k 正面概率密度函数。那么对于所有 $r \in \mathbb{R}$

$$\int_{\mathbb{R}^d} M_{k,r}(\mathbf{x})\mathrm{d}\mathbf{x} < \infty. \tag{11}$$

证明。 我们先应用引理 2.1，再根据第 2.2 节中$M_{k,+\infty}$的定义。对于所有 $r \in \mathbb{R}$,

$$\int_{\mathbb{R}^d} M_{k,r}(\mathbf{x})\,\mathrm{d}\mathbf{x} \le \int_{\mathbb{R}^d} M_{k,+\infty}(\mathbf{x})\,\mathrm{d}\mathbf{x} \tag{12}$$

$$\le \int_{\mathbb{R}^d} \sum_{i=1}^{k} p^{(i)}(\mathbf{x})\,\mathrm{d}\mathbf{x} = k < +\infty. \tag{13}$$

此外，值得注意的是，当 $r \le 1$ （例如几何平均数）时，

$$\int_{\mathbb{R}^d} M_{k,r}(\mathbf{x})\mathrm{d}\mathbf{x} \le \int_{\mathbb{R}^d} M_{k,1}(\mathbf{x})\mathrm{d}\mathbf{x} = 1. \tag{14}$$

□

据我们所知，尚无先前研究针对 r的所有实数取值给出归一化常数的界。几何池化（$r = 0$）的证明见（Genest 等人，1986年，第 489 页；Carvalho 等人，2023，定理 2.1）。Cooke（1991年，第 11 章）也提及的情形 $r \le 1$ 尤为值得关注：公式（14）中的归一化常数界意味着 $-\log Z_{k,r} \ge 0$。 $-\log Z_{k,r}$ 的符号对于分析 $\mathring{p}_{k,r}$ 至关重要，因为对数似然可以写成 $\log \mathring{p}_{k,r} = \log M_{k,r} - \log Z_{k,r}$。因此，它直接决定归一化会使总体似然增加还是减少，这对我们的论证至关重要（见定理 3.1）。

3.2 广义平均何时能可靠地带来正面效果?

我们探讨一个重要问题：与个体对数似然的平均值相比，聚合何时能够保证有所提升？这对应于“群体智慧”效应。我们证明，这种情况恰好在$r \in [0,1]$时发生。

定理 3.1 (对数似然上的群体智慧)。设 $\mathbf{x} \in \mathcal{X}$ 且 $k \in \mathbb{N}^*$。假设所有密度函数$p^{(1)}$，……，$p^{(k)}$ 均为正面。若 $0 \le r \le 1$,

$$\log \mathring{p}_{k,r}(\mathbf{x}) \ge \frac{1}{k}\sum_{i=1}^{k} \log p^{(i)}(\mathbf{x}). \tag{15}$$

证明。 令 $\mathbf{x} \in \mathcal{X}$，且 $0 < r \le 1$。根据詹森不等式（log 函数的凹性），并且由于 $r > 0$,

$$\log M_{k,r}(\mathbf{x})/Z_{k,r} = \frac{1}{r}\log\left(\frac{1}{k}\sum_{i=1}^{k}[p^{(i)}(\mathbf{x})]^r\right) - \log Z_{k,r} \tag{16}$$

$$\ge \frac{1}{rk}\sum_{i=1}^{k}\log([p^{(i)}(\mathbf{x})]^r) - \log Z_{k,r}. \tag{17}$$

因此，

$$\log \mathring{p}_{k,r}(\mathbf{x}) \ge \frac{1}{k}\sum_{i=1}^{k}\log p^{(i)}(\mathbf{x}) - \log Z_{k,r}. \tag{18}$$

注意，对于 $r = 0$，利用 $\log\prod = \sum\log$，公式（18）仍然成立。

由于 $r \le 1$，由公式（14）可直接得出

$$\log M_{k,r}(\mathbf{x})/Z_{k,r} \ge \frac{1}{k}\sum_{i=1}^{k}\log p^{(i)}(\mathbf{x}) \tag{19}$$

由于 $\log Z_{k,\,r} \le 0$。

□

对数似然衡量模型 p 对数据的解释能力，值越大表示拟合效果越好。因此，定理 3.1 通过保证对于任意数据点$\mathbf{x}$，聚合模型的对数似然至少与各个模型对数似然的平均值一样大，确立了 $r \in [0,1]$的聚合可靠性。因此， $[0,1]$ 可视为一个“安全”区间。

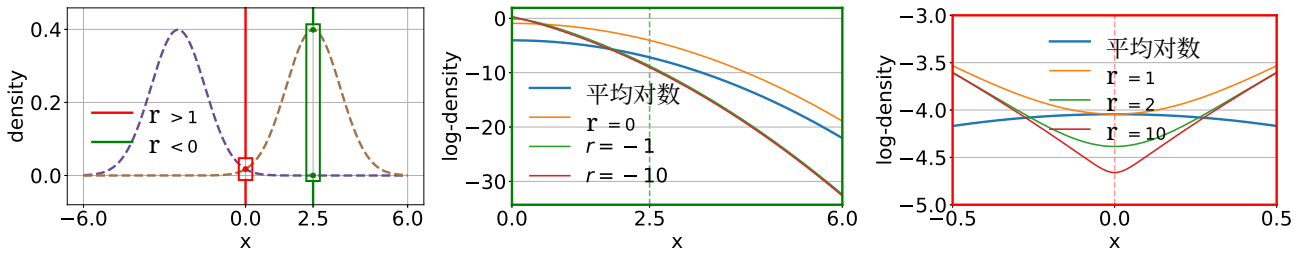
这一结果尤为揭示了一个重要特点：几何平均数（$r = 0$）是仍能可靠带来益处的聚合方式中最悲观、最不民主的一种，而算术平均数（$r = 1$）则是实现这一点的最不悲观、最民主的聚合方式。两者共同涵盖了安全聚合的整个范围，这也解释了为何文献中特别推崇这两种经典规则。

3.3 可靠性范围之外的失败案例

当 $r < 0$或 $r > 1$时，“群体智慧”效应不再成立，因为公式 (15) 并非对所有$\mathbf{x}$都成立。以下定理将对此作出形式化表述。

定理 3.2(群体智慧的失效)。 令 $r \notin [0,1]$。则存在两个正密度 $p^{(1)}$、$p^{(2)} \in \mathcal{P}$，满足 $p^{(1)} \ne p^{(2)}$，以及一个点$\mathbf{x} \in \mathcal{X}$，使得

$$\log \mathring{p}_{2,r}(\mathbf{x}) < \frac{1}{2}\sum_{i=1}^{2}\log p^{(i)}(\mathbf{x}). \tag{20}$$



(a) 两位专家　(b) $r < 0$　(c) $r > 1$

图 2：**定理 3.2 的直观示意。当** $r < 0$ **时，密度差异显著的点会导致聚合失败；当** $r > 1$ **时，共识点也会导致失败。**我们考察公式 (20) 中的差距在 r 各个范围内的位置。“平均对数”表示各个对数似然的平均值（公式 (20) 右侧）。左图 (a)：我们考虑两个高斯分布$\mathcal{N}(\pm 2.5,\ 1)$。中图 (b)：对于 $r < 0$，定理 3.1 的不等式在 $x = 2.5$处成立。我们发现，差距并不局限于某处，并且会随 x 增大而扩大。右图 (c)：对于 $r \geq 1$，算术平均（$r = 1$）在全局上优于平均对数密度，而所有 $r > 1$ 在 $x = 0$附近均呈现非常局部化（但非孤立点）的负面效应。

我们通过针对**两类不同的反例**来证明定理 3.2，这两类反例取决于$\mathbf{x}$ 相对于区间 $[0, 1]$的**位置** r。更确切地说，当 $r < 0$时，公式 (15) 在分歧点处失效；而当$r > 1$ 时，则在共识点处失效。作为支持这一观点和定理 3.2 的示例，我们考虑如下定义的两个对称高斯密度：

$$p^{(1)}(x) = \mathcal{N}(x; -m, 1), \quad \text{and} \quad p^{(2)}(x) = \mathcal{N}(x; m, 1). \tag{21}$$

聚合在分歧点处会失效。当 $r < 0, x = m$ 满足公式 (20) 时。该点对应着显著的不平衡，因为当 m 较大时

$$p^{(2)}(m) \gg p^{(1)}(m) \approx 0, \tag{22}$$

也就是说，该位置的密度存在强烈分歧。实际上，二者之间的差距呈指数级增长，如下所示

$$\frac{p^{(2)}(m)}{p^{(1)}(m)} = e^{2m^2}, \tag{23}$$

而图 2a 展示了这种情况，其中 $x = m$对应于 $p^{(2)}$ 的众数和 $p^{(1)}$的尾部。因此，当 $r < 0$时，该点处的平均对数似然高于对数幂平均，如图 2b 所示。这种现象的出现，是因为悲观（类似最小值）的幂平均会对至少一个密度赋予近乎零概率质量的区域施以重罚。附录 C.1 给出了详细证明，其中我们表明，当 m 较大时，这种惩罚效应会使公式 (17) 中的詹森不等式反转。

聚合在共识点处失效。当 $r > 1$时，我们相反地表明 $x = 0$ 满足公式 (20)。这对应于一个正面共识点，即两个密度相一致并赋予非零概率质量的位置。事实上，在该点，

$$p^{(1)}(0) = p^{(2)}(0) > 0, \tag{24}$$

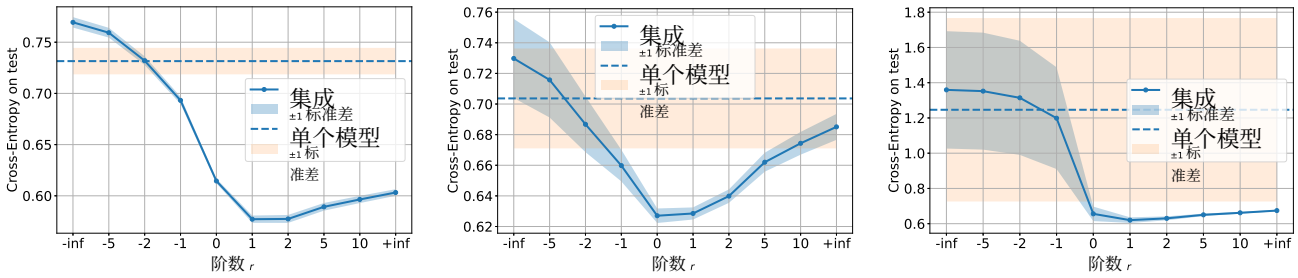
而两个密度在其他地方则不同。图 2a 展示了这种情况。如图 2c 所示，当 $r > 1$时，平均对数似然在该点超过 $\mathring{p}_{2,r}$。与前一种情况不同，这种一致性是局部的，从而导致不等式出现高度局部的失效。该失效源于归一化：尽管类最大值聚合偏向较大的值，归一化会将概率质量重新分配到某一密度占优的区域，从而削弱共识点的贡献。附录 C.1 给出了详细证明，我们在其中明确表明，尽管詹森不等式以等号成立，$Z_{2,r}$ 仍使公式 (20) 成立。

非连续反例。 在极端区域中，这一现象更加明显，并且可以通过离散（分类）情形加以说明。我们可以证明，对于 $r = -\infty$ （最小值），两个在 $y \in \{0, 1\}$ 上、真实标签为 $y = 0$的分类器，分别为 $p^{(1)} = (0.9, 0.1)$和 $p^{(2)} = (0.01, 0.99)$，构成一个反例。同样，我们可以证明，对于 $r = +\infty$ （最大值），两个在 $y \in \{0, 1, 2\}$ 上、真实标签为 $y = 0$的不同分类器，分别为 $p^{(1)} = (0.9, 0.03, 0.07)$和 $p^{(2)} = (0.9, 0.07, 0.03)$，也会违反该不等式。推导过程见附录 C.2。

4 实验

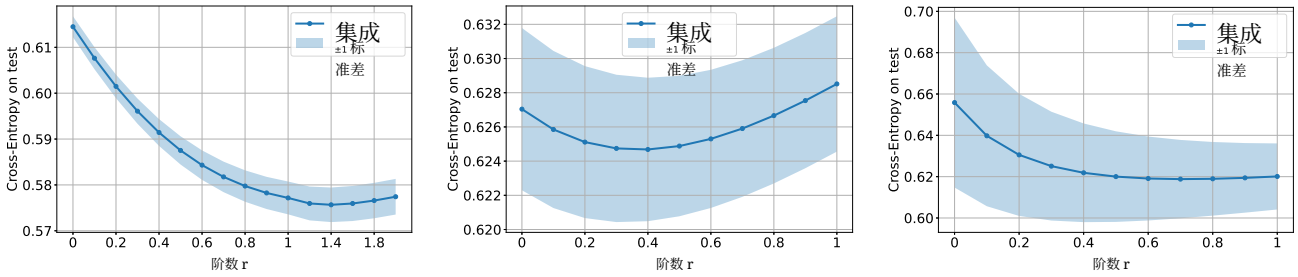
在本节中，我们通过实证检验理论安全区间 $r \in [0, 1]$（来自定理 3.1）内的广义均值聚合，是否能在所考察的所有模型和任务上实际改善负对数似然，并分析超出该范围后性能如何退化。

如何构建用于集成的模型。 为了在实践中研究聚合行为，我们使用不同的随机初始化和数据打乱，在相同的数据集和损失上独立训练多个具有相同架构的神经网络。该设置对应于深度集成 （DE）



(a) CIFAR-100 (b) MedMNIST (c) IMDb

图 3：**归一化幂平均似然的交叉熵行为示意图**$\mathring{p}_{k,\ r}$ **在三种分类设置下的交叉熵表现**。所有图均呈现 U 形性能曲线：极端聚合会放大模型分歧，而最优值位于 $[0,1]$。与定理 3.1 一致，$r \in [0,1]$ 区域的表现始终稳定优于单个模型不确定性带。相比之下，负阶数（$r<0$）表现较差，这可能是由于对主导类别存在分歧，正如第 3.3 节所述。



(a) CIFAR-100 (b) MedMNIST (c) IMDb

图 4：**归一化幂平均似然的交叉熵表现示意图**$\mathring{p}_{k,\ r}$**，重点关注** r **附近的** $[0,1]$ **取值（图 3 的放大视图）**。在 MedMNIST (b) 和 IMDb (c) 上，最优似然位于可靠性区间 $[0,1]$ 内（定理 3.1）；而在 CIFAR-100 (a) 上，它略微超出该区间（≈ 1.4）。这表明，尽管区间 $[0,1]$ 提供了稳定可靠的范围，但在实际应用中，略微超出该范围的温和乐观仍可能有益。

框架由 Lakshminarayanan 等人（2017年）和 Fort 等人（2019年）提出。在这种情况下，每个预测分布 $p^{(i)}$ 由独立训练的分类器的 softmax 输出表示，并在多个此类模型上进行聚合。对于每个实验，我们构建规模为 10 的集成，并在 5 次独立的集成抽样中重复该过程以降低方差。聚合在推理时使用阶数为 r 的广义平均进行（定义 2.1）。

数据集与架构。我们在视觉、医学影像和自然语言处理领域评估聚合，以检验这一现象的普遍性。对于多类分类，我们使用 CIFAR-100（Krizhevsky, 2009），这是一个基准数据集，包含 60,000 张低分辨率彩色图像，涵盖 100 个类别，每类仅有 600 张图像。庞大的标签空间和有限的每类数据增加了分类难度，也提高了模型间分歧的可能性。在医学领域，我们考察 DermaMNIST（Yang等人, 2023），其中包含涵盖多个诊断类别的皮肤镜皮损图像。该数据集严重不平衡，良性黑素细胞痣约占样本的三分之二的样本，而若干类别仅占 1–5%。我们为这些视觉任务训练标准的卷积残差网络（He等人, 2016; Zagoruyko和 Komodakis, 2016）。最后，为了评估自然语言处理中的聚合，我们在 IMDb（Maas等人, 2011）上训练 Transformer（Vaswani等人, 2017）。IMDb 是一个均衡的二分类情感分类数据集，包含 50,000 条电影评论。更多详情请参见附录A。

评估。我们使用测试集交叉熵评估性能，它对应于预测分布下真实标签的负对数似然。我们衡量其随 r 的变化情况，并特别关注预测的安全区间 $r \in [0,1]$。

4.1 中间阶比极端阶更可靠

通过分析整个范围 $r \in \mathbb{R}$ 内的聚合，我们在图 3中表明，性能呈现出典型的 U 形曲线，而可靠的改进仅限于 $r \in [0,1]$ 这一范围。

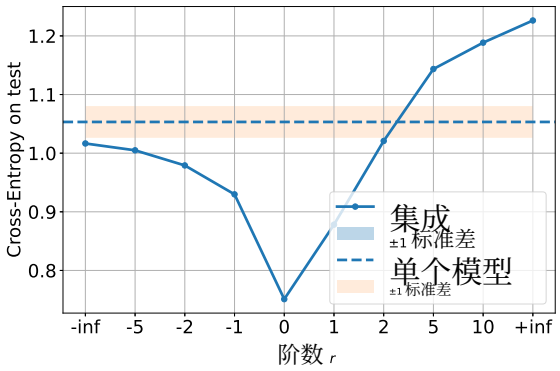


图 5：**在受控的近共识状态下，归一化幂平均似然在 CIFAR-100 上的交叉熵表现示意图**$\mathring{p}_{k,\,r}$。这一趋势与图 3 相反：过度乐观的聚合会产生负面影响，而最大值算子的表现最差。

在全部三个数据集上，极端阶数都会导致性能下降，而中间值，尤其是 $[0,1]$（参见定理 3.1），始终优于单个模型。即使在模型间不确定性很强的情况下，这一模式仍然成立（图 3c）：所有 $r \in [0,1]$ 均低于不确定性带，从而支持该区间的可靠性。

在极端情况下，$r = -\infty$ 的交叉熵比单个模型更差，这可能是因为分歧被放大了（第 3.3 节），尤其是在 CIFAR-100 上，较大的标签空间会增加冲突预测。相反，在 DermaMNIST 上，$r = +\infty$ 达到了单个模型不确定性区域，这表明类最大值聚合可能会放大共同的高置信度错误，而这可能是类别不平衡所致。

最后，对于 r的所有取值，集成方差都比单个模型有所降低。此外，$r < 0$对应的方差仍然较大，而 $r \geq 0$ 对应的方差则有所下降，这表明较大的取值（尤其是中间取值）能使 r 实现更强的方差缩减，同时获得最佳的整体对数似然，从而支持“群体智慧”效应。

4.2 什么是最优阶数 r?

通过对每张图进行更细致的扫描（图 4），我们发现最优 r 通常处于中间位置，而非极端取值；但尽管理论上可靠的区域是 $I = [0,1]$，经验上最优的 r 取值有时仍会落在该区域之外。

在图4b和图4c中，经验最优值分别位于$[0.3, 0.5] \subset I$ 和 $[0.6, 1] \subset I$。相比之下，图4a表明，温和乐观（$r \in [1.2, 1.6] \not\subseteq I$）在实践中仍可提高模型似然。

这表明，区间 $[0,1]$ 并不能保证最优性，因为当 r 的取值大于 1 时，经验表现可能更好。Safont等人（2019年）的相关工作直接从分类数据中学习最优聚合参数，采用的是一类由 α-散度导出的替代聚合规则。

我们注意到，规范值 $r = 0$和 $r = 1$之间不存在普适的阶数关系，因为正如图 4 所示，在不同数据集上，两者都无法始终占优。有趣的是，在类别不平衡数据集上，算术平均数（$r = 1$）似乎表现得不那么理想，这与 Buchanan 等人（2023）在预测场景中的观察一致。

4.3 近共识使乐观有害

我们在图 5 中表明，当模型对某一类别赋予几乎相同的概率、但在其他方面存在差异时，聚合行为可能会与图 3a 至图 3c 中的情况相反：极端乐观规则会产生危害，而最大值型聚合的表现最差。这支持了第 3.3 节中的反例。

对于独立训练的预测器而言，这种结构化的近共识很难实现，因为预测器之间的一致性通常会随类别和输入而波动。为隔离这一效应并生成图 5，我们构建了一个可控的 CIFAR-100 实验：复制一个已训练的预测器，然后对其 logits 添加小幅高斯扰动，但不扰动真实类别的 logit（同时确保概率之和为一）。

5 结论

我们研究了通过对密度应用阶数为 r 的归一化广义平均来聚合概率分布的影响。该方法定义了一个连续族，并将经典混合与几何聚合规则作为特例纳入其中。我们从理论上证明，$r \in [0,1]$ 是确保对数似然系统性优于单个模型的区间，而在该区间之外，不同阶数可能因不同机制而无法实现这一点。简单的高斯场景以及分类任务上的实验验证均支持这些结果。值得注意的是，我们证实了几何平均和算术平均数的特殊作用；它们的广泛应用与其界定理论安全区间 $[0,1]$ 相吻合。

我们的工作通过对数似然评估密度运算，而以往研究侧重于准确率。例如，Buchanan 等人（2023）比较了归一化 logits 的平均值与 Softmax 输出（对数池化与线性池化），并观察到其表现因数据集而异：在长尾数据集上，对数池化（$r = 0$）更受青睐。这证实了数据在聚合性能中的作用。
虽然我们的结果确定了区间 $r \in [0,1]$ 是一个有理论依据的可靠范围，但实验表明，经验上的最优值可能取决于数据或模型的属性。因此，未来研究的一个自然方向是更好地刻画最优聚合阶数 r。

致谢

本研究由法国政府通过法国2030计划资助，该计划由法国国家研究署管理。PAM 也获得了法国政府的资助，该资助由法国国家研究署管理，项目编号为“ANR-23-IACL-0001”；RS 获得的资助项目编号为“ANR-15-IDEX-01”；DG 则通过 NIM-ML 项目“ANR-21-CE23-0005-01”获得资助。

参考文献

Shun-ichi Amari。通过最小化 α-散度整合随机模型。*Neural Computation*，19(10):2780–2796，2007年。

Shun-ichi Amari。信息几何及其应用。施普林格，2016年。Shun-ichi Amari 和 Hiroshi Nagaoka。信息几何方法，第 191 卷。美国数学学会，2000年。Christopher M. Bishop。模式识别与机器学习，第 4 卷。施普林格，2006年。E. Kelly Buchanan、Geoff Pleiss 和 John P. Cunningham。集成对长尾数据的影响。载于*NeurIPS* 机器学习中的重尾研讨会，2023年。Weilin Cai、Juyong Jiang、Fan Wang、Jing Tang、Sunghun Kim 和 Jiayi Huang。大语言模型中混合专家的综述。*IEEE*知识与数据工程汇刊，2025年。Yanshuai Cao 和 David J. Fleet。用于自动且有原则地融合高斯过程预测的广义专家乘积。载于 *NeurIPS* 现代非参数方法 *3*：学习流程自动化研讨会，2014。Luiz M. Carvalho、Daniel A. M. Villela、Flavio C. Coelho 和 Leonardo S. Bastos。对数池化权重的贝叶斯推断。贝叶斯分析，18(1)：223–251，2023年。doi：10.1214/22-BA1311。Herman Chernoff。基于观测值总和的假设检验渐近效率度量。《数学统计年刊》，493–507页，1952年。Robert T. Clemen 和 Robert L. Winkler。风险分析中如何结合专家提供的概率分布。风险分析，19(2)：187–203，1999年。Samuel Cohen、Rendani Mbuvha、Tshilidzi Marwala 和 Marc P. Deisenroth。高斯过程专家乘积的修复。国际机器学习会议，2020年。

Roger Cooke。不确定性专家：科学中的意见与主观概率。牛津大学出版社，1991年。

José Manuel Corcuera 和 Federica Giummolè。用于预测的广义贝叶斯规则。斯堪的纳维亚统计杂志，26(2):265–279，1999年。

Miguel De Carvalho。平均值，你究竟是什么意思？《美国统计学家》，70(3):270–274，2016年。

Yilun Du、Conor Durkan、Robin Strudel、Joshua B. Tenen- baum、Sander Dieleman、Rob Fergus、Jascha Sohl- Dickstein、Arnaud Doucet 和 Will Sussman Grathwohl。减少、重用、回收：利用基于能量的扩散模型和马尔可夫链蒙特卡罗进行组合生成。载于国际机器学习会议，第202卷，第8489–8510页。PMLR，2023。Yousef El-Laham、Niccolò Dalmasso、Svitlana Vyetrenko、Vamsi K Potluru 和 Manuela Veloso。Mixup 正则化：一种概率视角。*arXiv* 预印本 *arXiv:2502.13825*，2025。Lee Elkin 和 Richard Pettigrew。意见汇集。剑桥大学出版社，2025。Stanislav Fort、Huiyi Hu 和 Balaji Lakshminarayanan。深度集成：一种损失景观视角。*arXiv* 预印本 *arXiv:1912.02757*，2019。Simon French。群体共识概率分布：批判性综述。曼彻斯特大学决策理论系，1983。Christian Genest 和 James V. Zidek。概率分布的组合：一篇批评性评论及附注书目。《统计科学》，1(1):114–135，1986。Christian Genest、Kevin J. McConway 和 Mark J. Schervish。外部贝叶斯汇集算子的特征刻画。《统计年鉴》，第487–501页，1986。Ian Goodfellow、Yoshua Bengio、Aaron Courville 和 Yoshua Bengio。深度学习，第1卷。麻省理工学院出版社，剑桥，2016。Godfrey H. Hardy、John Edensor Littlewood 和 George Pólya。不等式。剑桥大学出版社，1952年。Trevor Hastie、Robert Tibshirani 和 Jerome Friedman。统计学习要素。施普林格统计学丛书，纽约，2009。何恺明、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论文集，第770–778页，2016年。

Geoffrey E. Hinton。通过最小化对比散度来训练专家乘积。*NeuralComputation*，14(8):1771–1800，2002。

Robert A. Jacobs、Michael I. Jordan、Steven J. Nowlan 和 Geoffrey E. Hinton。局部专家自适应混合。*Neural Computation*，3(1):79–87，1991年。

Michael I. Jordan 和 Robert A. Jacobs。专家的分层混合与期望最大化算法。*Neural Computation*，6(2):181–214，1994。

Alex Krizhevsky。从微小图像中学习多层特征。2009。

Ludmila I. Kuncheva。组合模式分类器：方法与算法。约翰·威立父子出版公司，2014。

Robert Laddaga。Lehrer 与共识提案。综合，36(4):473–477，1977。

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。神经信息处理系统进展，30:6402–6413，2017年。

Aixin Liu、Bei Feng、Bing Xue、Bingxuan Wang、Bochao Wu、Chengda Lu、Chenggang Zhao、Chengqi Deng、ChenyuZhang、Chong Ruan 等。Deepseek-v3 技术报告。*arXiv* 预印本 *arXiv:2412.19437*，2024。

Haitao Liu、Jianfei Cai、Yi Wang 和 Yew Soon Ong。面向大规模高斯过程回归的广义稳健贝叶斯委员会机。发表于国际机器学习会议，第 3131–3140 页。PMLR，2018。

Nan Liu、Shuang Li、Yilun Du、Antonio Torralba 和 Joshua B. Tenenbaum。利用可组合扩散模型进行组合式视觉生成。发表于 欧洲计算机视觉会议，第 423–439 页。施普林格，2022。

Andrew Maas、Raymond E. Daly、Peter T. Pham、Dan Huang、Andrew Y. Ng 和 Christopher Potts。用于情感分析的词向量学习。发表于 计算语言学协会第 *49* 届年会论文集：人类语言技术，第 142–150 页，2011。

Pierre-Alexandre Mattei 和 Damien Garreau。集成模型是否一直在变得更好？机器学习研究期刊，26(201):1–46，2025年。

Adam Paszke、Sam Gross、Francisco Massa、Adam Lerer、James Bradbury、Gregory Chanan、Trevor Killeen、Zem- ing Lin、Natalia Gimelshein、Luca Antiga 等。PyTorch：一种命令式风格、高性能的深度学习库。神经信息处理系统进展，32，2019年。

Raphaël Razafindralambo、Rémy Sun、Frédéric Precioso、Damien Garreau 和 Pierre-Alexandre Mattei。两个分数何时胜过一个？扩散模型集成研究。机器学习研究汇刊，2026年。

Gonzalo Safont、Addisson Salazar 和 Luis Vergara。多分类器得分的多类 α 积分。*Neural Computation*，31(4):806–825，2019年。

Mervyn Stone。意见池。数理统计年刊，第1339–1342页，1961年。

James Surowiecki。群体的智慧：为什么多数人比少数人更聪明，以及集体智慧如何塑造商业、经济、社会和国家。Doubleday，2004年。

Lucas Theis、Aäron van den Oord 和 Matthias Bethge。关于生成模型评估的说明。*arXiv* 预印本 *arXiv:1511.01844*，2015年。

Philipp Tschandl、Cliff Rosendahl 和 Harald Kittler。HAM10000 数据集：常见色素性皮肤病变的大规模多来源皮肤镜图像集。《科学数据》，5(1):1–9，2018。

Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N. Gomez、Łukasz Kaiser 和 Illia Polosukhin。注意力就是你所需要的一切。《神经信息处理系统进展》，30，2017年。

Robert L. Winkler。主观概率分布的共识。《管理科学》，15(2):B–61，1968。

Jiancheng Yang、Rui Shi、Donglai Wei、Zequan Liu、Lin Zhao、Bilian Ke、Hanspeter Pfister 和 Bingbing Ni。MedMNIST v2：面向二维与三维生物医学图像分类的大规模轻量级基准。《科学数据》，10(1):41，2023。

Sergey Zagoruyko与Nikos Komodakis。宽残差网络。载于英国机器视觉会议（*BMVC*）论文集，第87号，第1–46页，2016年。

超越集成聚合中的混合与乘积：广义平均的似然视角（补充材料）

Raphaël Razafindralambo[1] **Rémy Sun**[1] **Frédéric Precioso**[1] **Damien Garreau**[2] **Pierre-Alexandre Mattei**[1]

[1]蔚蓝海岸大学、Inria、法国国家科学研究中心、I3S/LJAD、Maasai，尼斯，法国[2]维尔茨堡尤利乌斯-马克西米利安大学，计算机科学研究所 / CAIDAS，维尔茨堡，德国

本附录通过方法细节、补充示意图和额外的理论推导，对正文进行补充：

- 附录A介绍与第4节实验相对应的详细实验设置。
- 附录B探讨高斯设定下广义均值更广泛的行为，以扩展图1中的示例。
- 附录C给出第 3.3 节中提出的“群体智慧”反例的理论推导。
- 附录D推导了具有 $k \geq 1$ 密度的 $\mathbb{R}^d$ 高斯情形下归一化常数的解析表达式。

A 实验详情

我们对第4节中的实验进行更详细的描述，涵盖实验方案以及数值稳定的计算流程。录用后我们将提供代码。

A.1 数据集和模型

我们采用了三个数据集，并分别为每个数据集配备了适合相应分类任务的架构 .

我们使用 CIFAR-100（Krizhevsky, 2009），这是一个包含 60,000 张尺寸为 32×32 的彩色图像的基准数据集，涵盖 100 个物体类别，每个类别仅有 600 张图像。这些类别涵盖种类繁多的自然物体，包括动物、交通工具、家居用品和日常场景。我们的训练集包含 50k 张图像。我们训练 WideResNet-28-10 模型（Zagoruyko和Komodakis, 2016），其中 28 表示网络深度，10 表示控制通道数量的加宽因子。这是适用于小分辨率图像分类任务的常用架构。

我们还评估 DermaMNIST（Yang等人, 2023），该数据集源自 HAM10000 集合（Tschandl 等人, 2018），这是一个由皮肤镜皮损图像构成的医学影像数据集，图像被分为七种诊断类别。该数据集规模相对较小，且类别严重不平衡：良性黑素细胞痣约占样本的三分之二，而有几类仅占 1–5%。我们采用官方数据集划分（≈ 7k 训练 / 1k 验证 / 2k 测试）。我们使用标准的 28×28 分辨率版本，并采用 ResNet-18 架构，遵循 MedMNIST基准（Yang等人, 2023）推荐的实验方案。

最后，我们在 IMDb电影评论数据集（Maas等人, 2011）上进行评估。该数据集是一个大规模自然语言语料库，包含来自互联网电影数据库的 50,000 条用户撰写的评论，任务是将评论分类为正面或负面。我们训练轻量级的基于Transformer的文本分类器，其结构由词元嵌入、正弦位置编码和一个小型Transformer编码器组成，后接均值池化和线性分类头。我们采用 35k/7.5k/7.5k 的训练/验证/测试集划分。

A.2 训练集成

每项实验中，我们训练五个相互独立的深度集成（Lakshminarayanan 等人, 2017），每个集成都包含十个使用交叉熵损失训练的模型。每个集成中的模型均使用不同的随机初始化和数据打乱方式独立训练。我们根据验证性能选择最佳检查点。报告的均值和标准差是基于五个集成计算的，用于评估统计稳定性。

A.3 评估用于分类的广义平均

本节将说明如何对 softmax 输出应用广义均值聚合，以进行分类中的似然评估。

在这种设置下，每个模型都会生成一个定义在 L 类别上的预测分布。给定 k 个预测器，我们将其记为

$$p^{(i)}(y = \ell \mid x), \qquad i = 1, \dots, k, \tag{25}$$

输入 x 时，分配给类别 ℓ 的概率。在这种情况下，对似然取平均相当于聚合预测的类别概率。因此，我们对各模型的预测采用阶数为 r 的广义（幂）均值。对于每个类别 y，我们定义

$$M_{k,r}(y \mid x) = \begin{cases} \left(\frac{1}{k} \sum_{i=1}^{k} \left(p^{(i)}(y \mid x) \right)^r \right)^{\frac{1}{r}}, & r \neq 0, \\ \exp\left(\frac{1}{k} \sum_{i=1}^{k} \log p^{(i)}(y \mid x) \right), & r = 0. \end{cases} \tag{26}$$

由于此操作对每个类别独立执行，所得分数的总和不一定为 1。因此，我们按需对各类别进行归一化，以获得有效的预测分布，

$$\mathring{p}_{k,r}(y \mid x) = \frac{M_{k,r}(y \mid x)}{\sum_{y=1}^{L} M_{k,r}(y \mid x)}. \tag{27}$$

稳定的对数域实现。直接在概率空间中计算可能会导致数值不稳定性，因为 Softmax 概率极小，且归一化步骤涉及指数和。因此，我们使用数值稳定的 logsumexp 运算在对数域中进行聚合（例如，torch.logsumexp[1]）。

对于有限的 $r \neq 0$，预测的广义平均写作

$$\log M_{k,r}(y \mid x) = \frac{1}{r} \left(\log \sum_{i=1}^{k} \exp\big(r \log p^{(i)}(y \mid x) \big) - \log k \right), \tag{28}$$

这可以通过对模型执行 logsumexp 运算来高效实现。

对于 $r = 0$，几何平均数写作

$$\log M_{k,0}(y \mid x) = \frac{1}{k} \sum_{i=1}^{k} \log p^{(i)}(y \mid x). \tag{29}$$

最后，通过取指数并在类别维度上进行归一化，得到预测分布，

$$\mathring{p}_{k,r}(y \mid x) = \exp\big(\log M_{k,r}(y \mid x) - \log Z(x) \big), \tag{30}$$

其中归一化常数为

$$\log Z(x) = \log \sum_{y} \exp\big(\log M_{k,r}(y \mid x) \big). \tag{31}$$

[1]使用 PyTorch 库实现（Paszke 等，2019年）。

B 高斯设定下的可视化实验

本节中，我们分析连续情形（对应于回归），以说明专家配置和数据分散度如何共同影响 $\overset{\circ}{p}_{k,\,r}$ 的幂参数 r 对负对数似然的作用。

我们考虑对称的高斯专家 $p^{(1)}$ 和 $p^{(2)}$，其均值分别为 μ_1 和 $\mu_2 = -\mu_1$，标准差固定为1.8，并在从标准差为σ 的中心高斯分布中抽取的样本上评估聚合性能（其中 $\sigma = 4$ 对应于图1）。对于每种设定，负对数似然均在数万个样本上估计。通过改变 σ 以及专家之间的分离程度，我们阐明了支配不同聚合机制的机理。

B.1 数据集中如何塑造聚合性能

在本节中，我们表明，数据在一致点附近的集中可能导致最大值聚合表现不佳，而数据在专家的支撑集上的分散则会加剧最小值型聚合的缺陷。

在图6中，减小 σ （相对于图1中的 $\sigma = 4$ 配置）首先会加剧U形行为。对于 σ 的中间值（$\{3, 2, 1\}$），所有阶数 r 均优于个体基线（虚线）。当 σ 变得非常小（σ =0.2）时，趋势发生逆转：曲线保持相似结构，但排序相反，且最小值型聚合的表现不如单个专家似然。这一效应的产生是因为 σ 决定了样本主要落在不一致区域（从而惩罚最小值型聚合）还是一致区域（从而在$p^{(1)} \neq p^{(2)}$时惩罚最大值型聚合），而中间值则提供了平衡权衡。

B.2 专家间距如何影响聚合性能

本节将说明，与图1相比，让专家彼此更加接近（其中 $\mu_1 = 3.5$）会减弱幂平均的影响，因为 $p^{(1)} \approx p^{(2)}$；而增大专家之间的间距，则可以解决图1中遇到的 r 负值问题。

在图7中，当 μ_1 和 μ_2 接近时，广义平均的影响有限，且在 r 较小时略呈负面，因为测试样本主要落在不一致区域。增大 μ_1 和 μ_2 之间的间距，可以缩小这些不一致区域，并使样本集中在归一化广义平均的众数附近。

C “群体智慧”的失败：反例

我们在连续设定下（使用高斯密度）和离散情形下（见第 3.3 节）都给出了反例。前者考察可靠性区间（$r < 0$ 和 $r > 1$）之外的一般情形，后者则聚焦于对应最小值和最大值算子的极端聚合情形。

C.1 对称高斯密度

本节中，我们证明定理 3.2。我们证明可以找到两个正密度 $p^{(1)},\ p^{(2)} \in \mathcal{P}$ 满足 $p^{(1)} \neq p^{(2)}$，以及一点 $\mathbf{x}$，使得

$$\log \overset{\circ}{p}_{2,r}(\mathbf{x}) < \frac{1}{2} \sum_{i=1}^{2} \log p^{(i)}(\mathbf{x}) \tag{32}$$

当 $r \notin [0, 1]$ 时。

证明。我们考虑两个对称高斯分布。令 $\varphi(x) : x \mapsto (2\pi)^{-1/2}\exp(-x^2/2)$ 为标准正态密度。令 $m > 0$，并考虑

$$p^{(1)}(x) = \varphi(x + m), \qquad p^{(2)}(x) = \varphi(x - m). \tag{33}$$

它们分别对应于 $\mathcal{N}(-m, 1)$ 和 $\mathcal{N}(m, 1)$。

为了构造 $r > 1$ 和 $r < 0$ 的反例，我们考虑两个关注点：分别是 $x = 0$，即两个密度重合的唯一点，以及 $x = m$，它对应于 $p^{(2)}$ 的众数，并位于 $p^{(1)}$ 的尾部。

在分歧点处，聚合会失效 $r < 0$。我们证明，若 $r < 0$ 且 $x = m$，只要 m 足够大，公式 (17) 中的詹森不等式（因此公式 (15)）就可以反转。图 2b 展示了负面的 r 如何在 $|x|$ 和 m 足够大时反转公式 (15)。我们有

$$p^{(2)}(m) = \varphi(0), \qquad p^{(1)}(m) = \varphi(2m) = \varphi(0)e^{-2m^2}. \tag{34}$$

对于任意 $r < 0$, $x = m$处的幂平均满足

$$M_{2,r}(m) = \left(\tfrac{1}{2}\big(p^{(1)}(m)^r + p^{(2)}(m)^r\big)\right)^{1/r} \tag{35}$$

$$= \varphi(0)\left(\tfrac{1}{2}(1 + e^{-2rm^2})\right)^{1/r}. \tag{36}$$

取对数并与平均对数密度比较，得到

$$\begin{aligned} &\log M_{2,r}(m) - \tfrac{1}{2}\Big(\log p^{(1)}(m) + \log p^{(2)}(m)\Big) \\ &\quad = \frac{1}{r}\log\Big(\tfrac{1}{2}(1 + e^{-2rm^2})\Big) + m^2. \end{aligned} \tag{37}$$

由于 $r < 0$，项 e^{-2rm^2}随 m呈指数增长，因此当 $m \to +\infty$时，$\frac{1}{r}\log\Big(\frac{1}{2}(1 + e^{-2rm^2})\Big) + m^2 \sim \frac{1}{r}\log(e^{-2rm^2}) + m^2 = -2m^2 + m^2 = -m^2$，所以右侧在 $m \to +\infty$时趋于 $-\infty$。因此，当 m 足够大时，

$$\log M_{2,r}(m) < \tfrac{1}{2}\Big(\log p^{(1)}(m) + \log p^{(2)}(m)\Big), \tag{38}$$

这与公式 (17) 和公式 (15) 相矛盾。

广义均值无法改善一致点，当 $r > 1$。这一次我们证明，公式 (15) 在 $x = 0$ 处不成立，并非源于反转公式 (17)；事实上，公式 (17) 在此处将取等号，而是因为归一化将概率质量从该点重新分配到了别处。图 2c 展示了这一现象，该点附近的性能出现了明显退化。利用引理 2.1，对于所有 $r > 1$

$$Z_{2,r} = \int_{\mathbb{R}} M_{2,r}(x)\,\mathrm{d}x > \int_{\mathbb{R}} M_{2,1}(x)\,\mathrm{d}x \tag{39}$$

$$= \int_{\mathbb{R}} \frac{p^{(1)}(x) + p^{(2)}(x)}{2}\,\mathrm{d}x = 1. \tag{40}$$

这里公式 (39) 成立，因为 $p^{(1)}$ 与 $p^{(2)}$ 不同的集合具有严格为正的测度。现在考虑 $x = 0$，即关于两个均值 $\pm m$ 的对称点。我们有

$$p^{(1)}(0) = p^{(2)}(0) = \varphi(m) > 0. \tag{41}$$

因此，由公式 (39) 和 (41)，复合密度满足

$$\mathring{p}_{2,r}(0) = \frac{M_{2,r}(0)}{Z_{2,r}} = \frac{\varphi(m)}{Z_{2,r}} \tag{42}$$

$$\implies \log \mathring{p}_{2,r}(0) = \log\varphi(m) - \log Z_{2,r} < \log\varphi(m). \tag{43}$$

由于

$$\frac{1}{2}\sum_{i=1}^{2}\log p^{(i)}(0) = \frac{1}{2}\sum_{i=1}^{2}\log\varphi(m) = \log\varphi(m), \tag{44}$$

这与公式 (15) 相矛盾，并验证了公式 (32)。 □

C.2 极端聚合机制下的反例

我们给出了离散情形下的明确反例，表明针对 $r \in [0,1]$ 得出的似然不等式并不适用于极端聚合机制 $r=-\infty$（最小聚合）和 $r=+\infty$（最大值聚合）。我们考察分类问题，其中每个预测器都会输出一个类别上的概率向量，并通过逐分量聚合后再进行归一化。

在 $r=-\infty$ 分歧点处失效。 考虑标签为 $y \in \{0,1\}$、真实标签为 $y=0$ 的二分类问题。设两个分类器给出

$$p^{(1)} = (0.9,\, 0.1), \qquad p^{(2)} = (0.01,\, 0.99). \tag{45}$$

最小聚合为每个类别选取所分配的最小概率：

$$M_{2,-\infty} = \big(\min(0.9, 0.01), \min(0.1, 0.99)\big) = (0.01,\, 0.1). \tag{46}$$

归一化后，

$$\overset{\circ}{p}_{2,-\infty} = \frac{(0.01,\, 0.1)}{0.01 + 0.1} = (0.0909,\, 0.9091). \tag{47}$$

因此，真实标签的聚合对数似然为

$$\log \overset{\circ}{p}_{2,-\infty}(y=0) = \log(0.0909) \approx -2.40. \tag{48}$$

个体平均对数似然等于

$$\frac{1}{2}\sum_{i=1}^{2} \log p^{(i)}(y=0) = \frac{1}{2}\big(\log 0.9 + \log 0.01\big) \approx -2.36. \tag{49}$$

因此，

$$\log \overset{\circ}{p}_{2,-\infty}(0) < \frac{1}{2}\sum_{i=1}^{2} \log p^{(i)}(0), \tag{50}$$

这与公式 (15) 中的“群体智慧”原则相矛盾。

对 $r=+\infty$ 在一致点处的失效。 考虑标签为 $y \in \{0,1,2\}$、真实标签为$y=0$的三分类。设两个分类器产生

$$p^{(1)} = (0.9,\, 0.03,\, 0.07), \qquad p^{(2)} = (0.9,\, 0.07,\, 0.03). \tag{51}$$

最大值聚合会选择分配给每个类别的最大概率：

$$M_{2,+\infty} = \big(\max(0.9, 0.9), \max(0.03, 0.07), \max(0.07, 0.03)\big) = (0.9,\, 0.07,\, 0.07). \tag{52}$$

归一化之后，

$$\overset{\circ}{p}_{2,+\infty} = \frac{(0.9,\, 0.07,\, 0.07)}{0.9 + 0.07 + 0.07} = (0.865,\, 0.067,\, 0.067). \tag{53}$$

因此，真实标签的聚合对数似然为

$$\log \overset{\circ}{p}_{2,+\infty}(y=0) = \log(0.865) \approx -0.145. \tag{54}$$

个体平均对数似然等于

$$\frac{1}{2}\sum_{i=1}^{2} \log p^{(i)}(y=0) = \frac{1}{2}\big(\log 0.9 + \log 0.9\big) = \log 0.9 \approx -0.105. \tag{55}$$

因此，

$$\log \overset{\circ}{p}_{2,+\infty}(0) < \frac{1}{2}\sum_{i=1}^{2} \log p^{(i)}(0), \tag{56}$$

这再次与公式 (15) 中的对数似然不等式相矛盾。

D 积分的解析表达式：高斯分布的广义均值

在本节中，我们推导广义平均在 $r \in \mathbb{R}$ 情形下归一化常数的解析表达式。我们表明，在可靠性区间 $[0,1]$（第 3.2 节）内可以得到闭式公式，即针对 $r=0$ 以及 r 为正整数的倒数的取值。需要注意的是，情形 $r=1$ 是平凡的，因为归一化常数等于一；而对于情形 $r=0$，所得密度已被广泛推导（Cao 与 Fleet，2014；Cohen 等人，2020），尤其表明加权几何均值仍保持为高斯分布。另一方面，我们表明，对于 r 的其他取值，包括位于 $[0,1]$，之外的取值，该积分并不具有有限的闭式表达式，尽管我们知道它是有限的（命题 3.1）。这进一步强调，区间 $[0,1]$ 不仅构成理论上的可靠性机制（定理 3.1），也是某些聚合仍保持解析可处理性的唯一区域，而超出该区间并不能提供可比的解析或实际优势。

回归中的动机。 在概率回归中，高斯密度通常被用作似然函数。因此，研究高斯密度的广义平均具有实际意义：若相应的归一化常数可解析处理，则聚合可直接导出有效的基于高斯的回归模型以及可处理的模型似然，而无需数值积分或近似方案，除非可能在预测时刻计算均值时仍涉及积分。

记号与目标。 设 $p^{(1)}\dots,p^{(k)}$ 为 $\mathcal{X}=\mathbb{R}^d$ 中的高斯密度，且 $d\geq 1$。形式上，我们对所有$\mathbf{x}\in\mathcal{X}$

$$p^{(i)}(\mathbf{x})=\frac{1}{(2\pi)^{d/2}\,|\Sigma_i|^{1/2}}\exp\!\left(-\frac{1}{2}(\mathbf{x}-\mu_i)^\top\Sigma_i^{-1}(\mathbf{x}-\mu_i)\right),\qquad i=1,\dots,k, \tag{57}$$

其中 $\mu_i\in\mathbb{R}^d$ 表示均值向量，$\Sigma_i\in\mathbb{R}^{d\times d}$ 为对称正定协方差矩阵（我们记作$\Sigma_i\succ 0$）。

目标是确定

$$Z_{k,r}=\int\left(\frac{1}{k}\sum_{i=1}^{k}p^{(i)}(\mathbf{x})^r\right)^{\frac{1}{r}}\mathrm{d}\mathbf{x}. \tag{58}$$

我们区分两种情况。首先，我们考虑可处理积分，它们对应于 $r=0$（几何平均数）以及 $r=1/n$，且满足 $n\in\mathbb{N}\setminus\{0\}$，涵盖安全区间 $[0,1]$ 中的一个可数子集。然后，我们考察不满足上述条件且不对应可处理积分的值 $r\in\mathbb{R}$。

D.1 可处理积分

假设 $n\in\mathbb{N}\setminus\{0\}$。所关注的积分可写为

$$Z_{k,r}=\int\left(\frac{1}{k}\sum_{i=1}^{k}p^{(i)}(\mathbf{x})^{\frac{1}{n}}\right)^{n}\mathrm{d}\mathbf{x}. \tag{59}$$

Thi这种表述使我们能够直接应用以下多项式定理，定义如下 s.

引理 D.1 (多项式定理)。 设 k 为正整数，n 为非负整数。设 $a_1,\dots,a_k\in F$，其中 F 是一个域（例如，$F=\mathbb{R}$）。那么，

$$(a_1+\cdots+a_k)^n=\sum_{\substack{n_1+\cdots+n_k=n\\ n_i\geq 0}}\binom{n}{n_1,\dots,n_k}\prod_{i=1}^{k}a_i^{n_i}, \tag{60}$$

其中，多项式系数定义为

$$\binom{n}{n_1,\dots,n_k}=\frac{n!}{n_1!\cdots n_k!}. \tag{61}$$

主要根据公式（60），我们可以推导出以下命题。

命题 D.1。设 $p^{(1)}$、...、$p^{(k)}$ 是参数为(μ_i, Σ_i)的 $k \geq 1$ 高斯密度。若 $r = \frac{1}{n}$ 且 $n \in \mathbb{N} \setminus \{0\}$，则有

$$\begin{aligned} Z_{k,r} = \frac{1}{k^n} \sum_{\substack{n_1+\cdots+n_k=n \\ n_i \geq 0}} \frac{n!}{n_1! \cdots n_k!} (2\pi)^{\frac{d}{2}} \Big| \sum_{i=1}^k \frac{n_i}{n} \Sigma_i^{-1} \Big|^{-\frac{1}{2}} \prod_{i=1}^k (2\pi)^{-\frac{dn_i}{2n}} |\Sigma_i|^{-\frac{n_i}{2n}} \\ \cdot \exp\left(\frac{1}{2} \Big(\sum_{i=1}^k \frac{n_i}{n} \mu_i^\top \Sigma_i^{-1} \Big) \Big(\sum_{i=1}^k \frac{n_i}{n} \Sigma_i^{-1} \Big) \Big(\sum_{i=1}^k \frac{n_i}{n} \Sigma_i^{-1} \mu_i \Big) - \frac{1}{2} \sum_{i=1}^k \frac{n_i}{n} \mu_i^\top \Sigma_i^{-1} \mu_i \right). \end{aligned} \tag{62}$$

若 $r = 0$,

$$\begin{aligned} Z_{k,r} =& (2\pi)^{\frac{d}{2}} \Big| \sum_{i=1}^k \frac{1}{k} \Sigma_i^{-1} \Big|^{-\frac{1}{2}} \prod_{i=1}^k (2\pi)^{-\frac{d}{2k}} |\Sigma_i|^{-\frac{1}{2k}} \\ & \cdot \exp\left(\frac{1}{2} \Big(\sum_{i=1}^k \frac{1}{k} \mu_i^\top \Sigma_i^{-1} \Big) \Big(\sum_{i=1}^k \frac{1}{k} \Sigma_i^{-1} \Big) \Big(\sum_{i=1}^k \frac{1}{k} \Sigma_i^{-1} \mu_i \Big) - \frac{1}{2} \sum_{i=1}^k \frac{1}{k} \mu_i^\top \Sigma_i^{-1} \mu_i \right). \end{aligned} \tag{63}$$

证明。由引理 D.1 可得

$$Z_{k,r} = \frac{1}{k^n} \sum_{\substack{n_1+\cdots+n_k=n \\ n_i \geq 0}} \binom{n}{n_1, \ldots, n_k} \underbrace{\int \prod_{i=1}^k p_i(\mathbf{x})^{\frac{n_i}{n}}}_{I_k}. \tag{64}$$

注意，I_k 是加权几何均值的积分，因为 $\sum_{i=1}^k \frac{n_i}{n} = 1$。令 $w_i = \frac{n_i}{n}$表示所有 i对应的权重。对于所有$\mathbf{x} \in \mathcal{X}$，有

$$\prod_{i=1}^k p_i(\mathbf{x})^{w_i} = \prod_{i=1}^k (2\pi)^{-\frac{dw_i}{2}} |\Sigma_i|^{-\frac{w_i}{2}} \exp\left(-\frac{1}{2} \sum_{i=1}^k w_i (\mathbf{x} - \mu_i)^\top \Sigma_i^{-1} (\mathbf{x} - \mu_i) \right) \tag{65}$$

$$= C_k \exp\left(-\frac{1}{2} \big(\mathbf{x}^\top Q_k \mathbf{x} - 2 b_k^\top \mathbf{x} + c_k \big) \right) \tag{66}$$

其中

$$C_k = \prod_{i=1}^k (2\pi)^{-\frac{dw_i}{2}} |\Sigma_i|^{-\frac{w_i}{2}}, \quad Q_k = \sum_{i=1}^k w_i \Sigma_i^{-1} \succ 0, \quad b_k = \sum_{i=1}^k w_i \Sigma_i^{-1} \mu_i, \quad c_k = \sum_{i=1}^k w_i \mu_i^\top \Sigma_i^{-1} \mu_i. \tag{67}$$

通过配方法，我们可以提取出一项，其积分有显式表达式。事实上，

$$\prod_{i=1}^k p_i(x)^{w_i} = C_k \exp\left(-\frac{1}{2} \big(\mathbf{x}^\top Q_k \mathbf{x} - 2 b_k^\top \mathbf{x} \big) \right) \exp\left(-\frac{1}{2} c_k \right) \tag{68}$$

$$= C_k \exp\left(-\frac{1}{2} \big(\mathbf{x}^\top Q_k \mathbf{x} - 2 (Q_k^{-1} b_k)^\top Q_k \mathbf{x} \big) \right) \exp\left(-\frac{1}{2} c_k \right) \tag{69}$$

$$= C_k \exp\left(-\frac{1}{2} \big(\mathbf{x}^\top Q_k \mathbf{x} - 2 (Q_k^{-1} b_k)^\top Q_k \mathbf{x} \big) + \underbrace{(Q_k^{-1} b_k)^\top Q_k (Q_k^{-1} b_k)}_{b_k^\top Q_k b_k} \right) \exp\left(\frac{1}{2} b_k^\top Q_k b_k - \frac{1}{2} c_k \right) \tag{70}$$

$$= C_k \exp\left(-\frac{1}{2} \big(\mathbf{x} - Q_k^{-1} b_k \big)^\top Q_k \big(\mathbf{x} - Q_k^{-1} b_k \big) \big) \right) \exp\left(\frac{1}{2} b_k^\top Q_k b_k - \frac{1}{2} c_k \right) \tag{71}$$

$$= C_k (2\pi)^{\frac{d}{2}} |Q_k|^{-\frac{1}{2}} \, \mathcal{N}(\mathbf{x};\, Q_k^{-1} b_k, Q_k^{-1}) \exp\left(\frac{1}{2} b_k^\top Q_k b_k - \frac{1}{2} c_k \right) \tag{72}$$

其中，方程（69）源于 Q_k 可逆且对称这一事实。利用 $\int \mathcal{N}(\mathbf{x};\, Q_k^{-1} b_k, Q_k^{-1}) \mathrm{d}\mathbf{x} = 1$，我们得到

$$\int \prod_{i=1}^k p_i(\mathbf{x})^{w_i} \mathrm{d}\mathbf{x} = (2\pi)^{\frac{d}{2}} |Q_k|^{-\frac{1}{2}} \prod_{i=1}^k (2\pi)^{-\frac{dw_i}{2}} |\Sigma_i|^{-\frac{w_i}{2}} \exp\left(\frac{1}{2} b_k^\top Q_k b_k - \frac{1}{2} c_k \right) \tag{73}$$

其中， Q_k, b_k, c_k 在方程（67）中定义。最后，我们使用方程（64）证明命题 D 的第一个结论 .1. 对于 $r = 0$ 这种情况，我们只需对 $w_i = \frac{1}{k}$ 运用方程（65）推导方程（73）时所采用的相同推理。 – □

D.2 非平凡积分

一旦 r 不属于 D.1 节所研究的情形，尤其是当 $r \notin [0, 1]$ 时，我们就无法像之前那样用有限和导出显式积分$Z_{k,\ r}$。

为此，我们考虑简单的情况 $k = 2$，并假设 α 是任意实数（或复数）。它可以是负数。无需额外限制，引理 D.1 可通过二项式级数推广，如下引理所述

引理 D.2（**二项式级数**）。设 a_1， $a_2 \in \mathbb{R}$。那么，若 $|a| < |b|$,

$$(a + b)^\alpha = \sum_{i=0}^{\infty} \binom{\alpha}{i} a^k b^{\alpha-k}. \tag{74}$$

其中，广义二项式系数定义为

$$\binom{\alpha}{i} = \frac{\alpha(\alpha - 1)(\alpha - 2) \dots (\alpha - i - 1)}{i!} \tag{75}$$

根据该引理，$Z_{k,r}$ 对于一般的 $r \in \mathbb{R} \setminus \{0\}$ 可表示为涉及二项式级数的积分，方法与D.1 节中的做法相同。为确保级数展开收敛，该积分必须分解为对应于情形 $p^{(1)} < p^{(2)}$ 与 $p^{(2)} < p^{(1)}$的两个区域。然而，除非 r^{-1} 为正整数（如D.1 节中那样），所得二项式展开便不再是有限和。因此，求值$Z_{k,r}$ 通常需要近似或数值方法，既可用于计算该积分本身，也可用于截断相应的级数表示。

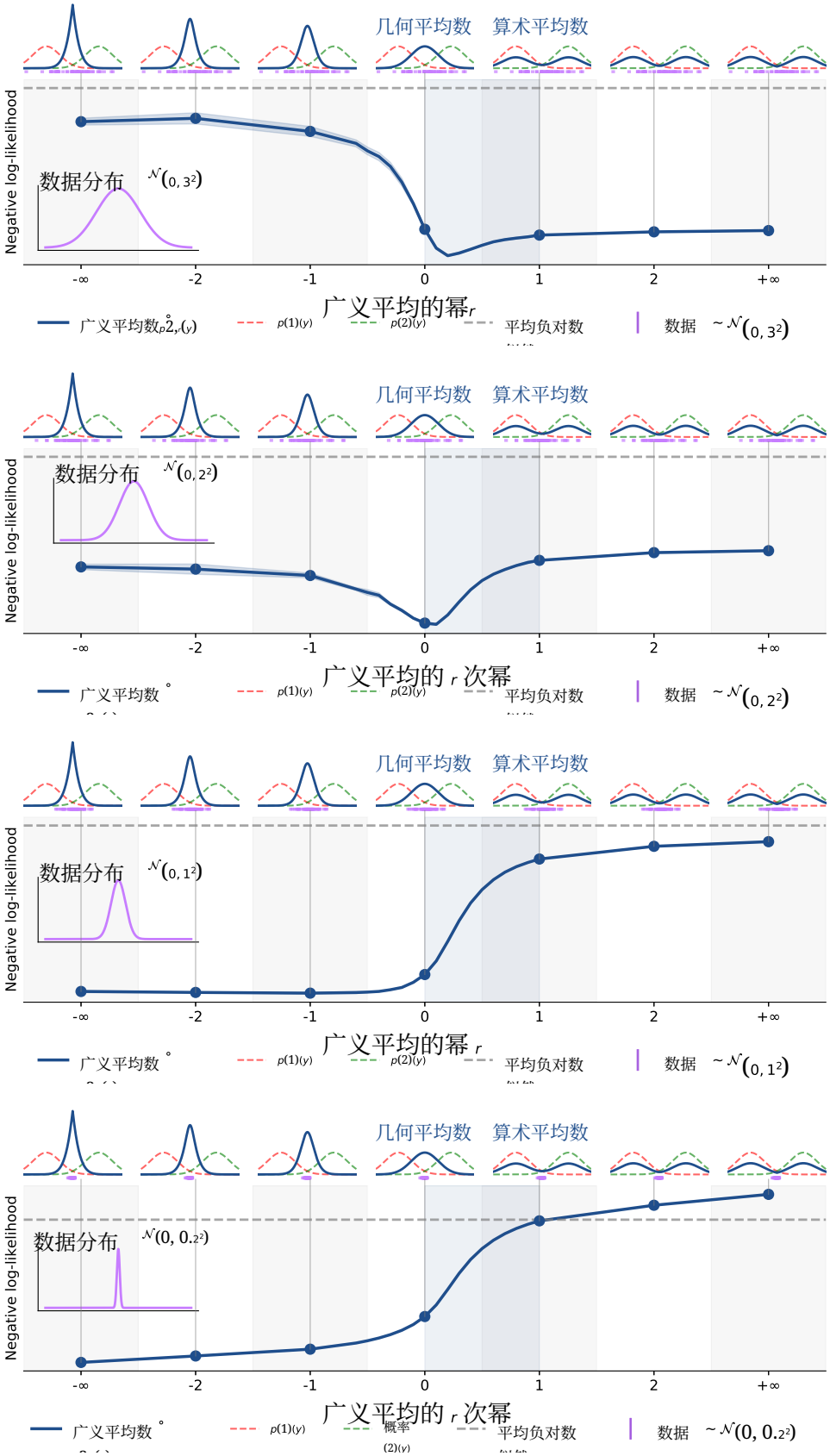


图6：**减小 σ 对聚合行为的影响**。 从上到下依次为：大（$\sigma = 3$）、中等（$\sigma = 2$）、小（$\sigma = 1$）和非常小 σ （$\sigma = 0.2$）。U形行为先变得更加明显，随后在中等 σ 时，所有 r 均优于个体基线；而当 σ 非常小时，趋势则反转。

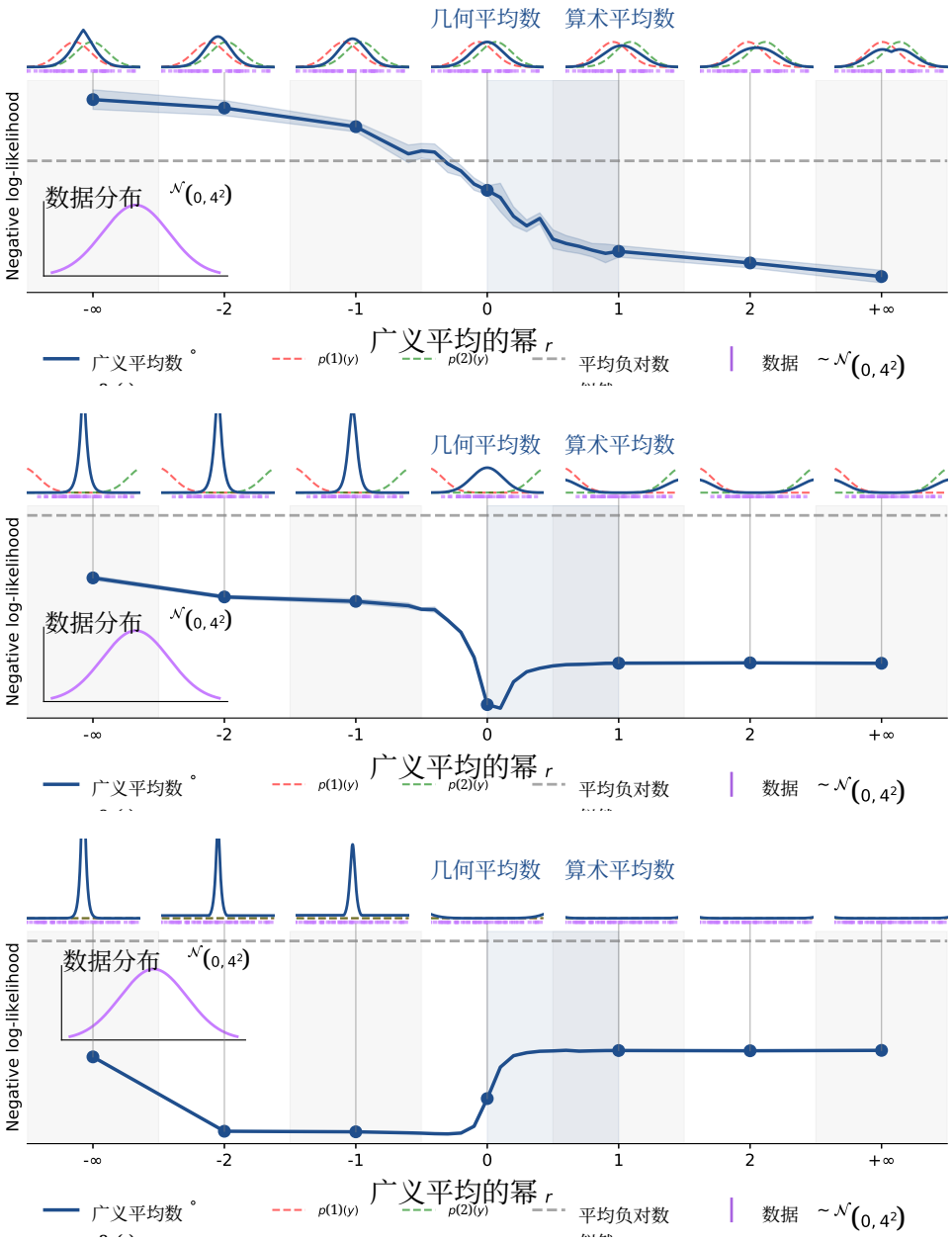


图7：**对 μ_1 和 μ_2 对聚合性能的影响**。从上到下：合并的专家、中度分离的专家和高度分离的专家。当 μ_1 和 μ_2 接近时，不一致区域频繁出现，这解释了类似最小值的聚合表现不佳的原因。随着均值彼此分离，测试样本逐渐集中于广义平均的众数上，这种影响也随之减弱。