arXiv:2601.00457v1 [cs.LG] 1 Jan 2026

混合专家模型中的几何正则化：权重与激活之间的脱节

Hyunjun Kim韩国科学技术院（KAIST）
hyunjun1121@kaist.ac.kr

摘要

混合专家模型（MoE）通过稀疏激活实现效率，但几何正则化在专家专门化中的作用仍不明确。我们应用正交性损失来增强专家多样性，却发现它在多个方面未能奏效：它并未减少权重空间重叠（MSO 实际上**最多增加了**114%），**激活空间**重叠无论是否进行正则化都仍然很高（∼0.6），而且对性能的影响**并不一致**——在 WikiText-103 上略有提升（-0.9%），在 TinyStories 上略有下降（$+0.9\%$），在 PTB 上的结果则差异很大（标准差 > 1.0）。我们对 7 种正则化强度的分析发现，权重正交性与激活正交性之间不存在显著相关性（$r = -0.293, p = 0.523$）。这些发现表明，权重空间正则化既无法实现其几何目标，也不能可靠地提升性能，因此不适用于提高混合专家模型的多样性。

1 引言

混合专家模型（MoE）通过在每次输入时仅激活一部分参数，实现高效扩展（Shazeer 等人，2017年；Fedus 等人，2022年）。一种常见假设是，专家表征应当正交，以尽量减少干扰（Chen 等人，2022年，2023年）。这种直觉源于线性代数：正交向量具有最大的可区分性，组合时其输出不会相互干扰。

假设。正交正则化应能提升专家多样性并降低困惑度。

发现。事实并非如此，而且这种方法并不可靠。在三个数据集（TinyStories、WikiText-103、PTB）上，几何正则化的结果不一致：在 WikiText-103 上略有改善（-0.9%），在 TinyStories 上略有退化（$+0.9\%$），而在 PTB 上则**高方差**（标准差 > 1.0）。

为何如此？ 我们发现了**权重-激活差距**：权重空间正交性（MSO $\approx 10^{-4}$）并不会转化为激活空间正交性（MSO ≈ 0.6）。在 7 种正则化强度下，我们发现权重重叠与激活重叠之间不存在显著相关性（$r = -0.293, p = 0.523$），这表明权重几何结构与功能正交性在很大程度上彼此独立。

贡献。

1. 我们证明，正交正则化无法降低权重 MSO——实际上还会使其增加多达 114%——并且对损失产生不一致的影响：在 WikiText-103 上略有改善（-0.9%），在 TinyStories 上略有退化（$+0.9\%$），在 PTB 上则呈现高方差（标准差 > 1.0）。

2. 我们发现了权重-激活脱节：激活重叠比权重重叠高出 $\sim 1000\times$，且二者不存在显著相关性（$r = -0.293$，$p = 0.523$，n=7）。

3. 我们证明，权重空间正则化是一个不可靠的优化目标——它既无法实现几何目标，也不能可靠地提升性能。

2 方法

正交损失。对于专家权重矩阵$\{W_i\}_{i=1}^{N}$，我们定义：

$$\mathcal{L}_{\text{orth}} = \sum_{i<j} |\langle \tilde{W}_i, \tilde{W}_j \rangle|^2 \tag{1}$$

其中，$\tilde{W}_i = \text{vec}(W_i)/\|\text{vec}(W_i)\|$ 是归一化后的展平权重向量。该损失鼓励专家表征之间保持正交性，并以权重 λ 加入语言建模目标。

均方重叠（MSO）。我们使用以下方法衡量几何多样性：

$$\text{MSO} = \frac{2}{N(N-1)} \sum_{i<j} |\langle \tilde{W}_i, \tilde{W}_j \rangle|^2 \quad (2)$$

MSO 越低，表示专家越正交（越多样）。我们分别计算权重 和激活的 MSO。

激活 MSO。 对于共同激活且产生输出 $\{h_i\}$ 的专家，我们计算：

$$\text{MSO}_{\text{act}} = \mathbb{E}_x \left[\frac{2}{k(k-1)} \sum_{i<j \in \mathcal{S}(x)} \left(\frac{\langle h_i, h_j \rangle}{\|h_i\| \|h_j\|} \right)^2 \right] \quad (3)$$

其中， $\mathcal{S}(x)$ 是输入 x 对应的 k 个选定专家的集合。这衡量了专家输出在实际输入上的功能相似性。

3 实验

实验设置。 我们在 TinyStories 上训练 NanoGPT-MoE（∼130M 个参数、8 个专家、6 层、top-2 路由）（Eldan 和 Li，2023年），使用 AdamW（Loshchilov 和 Hutter，2019年）（学习率=5×10^{-4}， β_1=0.9， β_2=0.95，权重衰减=0.1）进行 10K 次迭代。每个混合专家层包含 8 个专家，隐藏维度为 512，中间维度为 2048。TinyStories 实验使用 5 个随机种子（42、123、456、789、1337）。

实现细节。 我们对每个专家的**上投影权重**($W_{\text{up}} \in \mathbb{R}^{d_{\text{ffn}} \times d_{\text{model}}}$)进行正则化。计算两两内积前，会将每个权重矩阵展平并进行 L2 归一化。 λ扫描使用 7 个值：$\{0, 0.001, 0.005, 0.01, 0.05, 0.1, 0.2\}$。MSO 在**每一层**分别计算，并在全部 6 个混合专家层上**取平均**。激活 MSO 是针对**门控后**的专家输出计算的，涉及得分排名前 2 的选定专家，且**不按**门控得分加权。我们不使用辅助负载均衡损失。

3.1 正则化下的权重 MSO

表1揭示了一个令人惊讶的发现：正交正则化并未降低权重 MSO，反而**增加**了它。尽管损失明确惩罚专家重叠，但最终的权重 MSO 随着 λ上升，这表明正则化干扰了自然训练动态，而非强制实现正交性。

方法	权重 MSO	Δ
基线	5.43×10^{-4}	—
+正交（λ=0.001）	7.52×10^{-4}	+39%
+正交（λ=0.01）	1.16×10^{-3}	+114%

表1：正交性正则化会增加权重 MSO，这与其预期效果相反。基线权重已经接近正交。

方法	验证集困惑度	Δ%	p值
基线	5.94 ± 0.08	—	—
+正交 (λ=0.001)	6.00 ± 0.32	+0.9%	0.727

表 2：正交性正则化并未改善困惑度（$p = 0.727$，配对 t 检验， n=5个随机种子）。方差从 0.08 增至 0.32。

3.2 困惑度未改善

尽管正则化目标是明确的，困惑度的改善**在统计上并不显著**（表2）。

为何 p 值偏高? 较高的 p 值（p =0.727）反映了效应量极小且方差增加。基线的标准差较低（0.08），而 λ=0.001 会将方差提高至 0.32——方差 4× 增加会使训练不稳定。困惑度略微上升（+0.9%），但与这一方差相比微不足道，这表明正则化引入了噪声，却没有带来任何好处。

3.3 权重—激活间隙

表3揭示了核心发现：权重与激活的几何结构从根本上彼此脱节。

相关性分析。图1直观展示了这种脱节：随着 λ 增加，权重 MSO 上升（施加了正则化），但激活 MSO 仍保持在 ∼0.57 不变。在 7 种正则化强度下，我们发现皮尔逊 $r = -0.293$ ($p = 0.523$, 95% 置信区间：$[-0.857, 0.590]$)—结果在统计上不显著。这证实权重与激活的几何结构彼此独立。

3.4 跨数据集验证

为检验我们的发现能否推广到 TinyStories 之外，我们在 WikiText-103（Merity 等人，2016年）和 PennTreebank（Marcus 等人，1993年）这两个特征各异的标准基准数据集上评估正交正则化（表4）。

依赖数据集的效应。 多随机种子实验显示，效果不一。在 WikiText-103（1.18 亿词元）上，正交正则化带来了

λ	权重 MSO	激活 MSO	比率
0（基线）	5.43×10^{-4}	0.572	1053×
0.001	7.52×10^{-4}	0.581	773×
0.01	1.16×10^{-3}	0.577	496×
0.1	2.04×10^{-3}	0.593	290×
0.2	2.78×10^{-3}	0.564	203×

表 3：激活 MSO（∼0.57）保持不变，而权重 MSO 随着 λ增加。皮尔逊 $r = -0.293$, $p = 0.523$（n=7），表明不存在显著相关性。

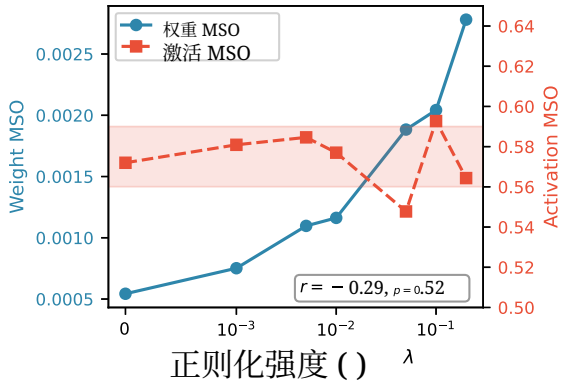

图 1：权重-激活差距。权重 MSO 会响应正则化；激活 MSO 则不会。不存在显著相关性（$r = -0.293$, $p = 0.523$）。

小幅且稳定的提升（−0.9%）。然而，在 PTB（120 万词元）上，不同随机种子下的结果波动很大（标准差 ∼ 1.0），因此难以得出可靠结论。这种高方差表明，几何正则化的效果可能取决于数据集与随机种子的交互作用，而不仅仅取决于数据集特征。

解读。PTB 上的高方差（标准差∼1.0）与 WikiText-103 上的（标准差 ∼0.05）相比，可能反映了数据集规模效应。较小的数据集可能表现出更多依赖随机种子的专家专门化模式，从而导致结果不稳定。无论方向如何，这种不一致本身都会削弱权重空间正则化的可靠性——一种仅在特定随机种子与数据集组合下有效的方法并不实用。

4 分析：差距为何存在?

非线性的作用。 现代混合专家采用非线性激活函数（SiLU/Swish）（Ramachandran 等人，2017年）和层归一化（Ba 等人，2016年）。考虑专家计算：

$$h_i = \text{LayerNorm}(\text{SiLU}(W_i \cdot x)) \qquad (4)$$

数据集	词元	Base	Orth	Δ
WikiText-103	118M	$3.76_{\pm.02}$	$3.73_{\pm.05}$	−0.9%
TinyStories	2.1M	$5.94_{\pm.08}$	$6.00_{\pm.32}$	+0.9%
PTB	1.2M	$6.17_{\pm.95}$	$5.74_{\pm1.11}$	−7.0%*

表 4：跨数据集验证（WikiText-103/PTB 使用 n=3 个随机种子，TinyStories 使用 n=5 个，λ=0.001）。*PTB 显示出高方差；效果方向并不可靠。

即使 $W_i \perp W_j$，非线性也会以非平凡的方式改变几何结构。SiLU 会根据激活幅值进行逐元素门控，而层归一化则会对激活重新居中并缩放，使其具有单位方差。这些操作可以压缩专家输出之间的角度差异，无论权重几何如何，都会增加激活重叠。

输入分布的影响。 权重正交性约束的是静态参数，但激活正交性取决于输入分布。如果输入 x 在不同的权重子空间上的投影相似（例如，由于输入分布具有低秩结构），那么即使 $W_i \perp W_j$，所得激活 $W_i \cdot x$ 和 $W_j \cdot x$ 仍会相关。自然语言输入可能因语义规律而呈现此类结构。

数学直觉。 考虑两个权重矩阵 W_1，W_2 具有弗罗贝尼乌斯正交性$\langle W_1, W_2 \rangle_F = \text{tr}(W_1^T W_2) = 0$。这一约束仅确保 $W_1^T W_2$ 中的元素总和 为零——它并不 意味着 $W_1^T W_2 = 0$。对于输入 x，激活内积为 $\langle z_1, z_2 \rangle = x^T W_1^T W_2 x$。由于 $W_1^T W_2$ 可以具有任意的非零结构（只有它的迹受到约束），对于典型输入，这个二次型通常不为零。因此，权重的弗罗贝尼乌斯正交性无法保证激活正交性。

5 相关工作

MoE扩展与架构。 现代混合专家范式源自 Shazeer 等人(2017年)，他们通过稀疏门控展示了如何扩展至万亿参数。 Fedus 等人 (2022年) 在 Switch Transformers 中采用 top-1路由简化了这一范式，而 GShard（Lepikhin 等人，2021年）则实现了高效的分布式训练。近期研究探索了更细粒度的专家分解：DeepSeek-MoE（Dai 等人，2024年）在每层使用 64 个细粒度专家，而 Mixtral（Jiang 等人，2024年）则通过 top-2 路由使用 8 个专家，实现了出色的性能。这些架构假设专家多样性

会在训练过程中自然形成。

专家专门化与多样性。 多种方法通过改进路由来提升专家多样性。X-MoE (Chi 等人，2022年) 采用超球面路由和余弦归一化门控，以缓解表征坍塌。HyperRouter (Nguyen 等人，2024年) 通过超网络动态生成路由器参数。S2MoE (Do 等人，2025年)采用结合高斯噪声的随机学习，以防止专家特征重叠。Re-MoE (Fang 等人，2025年)提出结合 L1 正则化的 ReLU 路由，以实现可微的专家选择。SMoE-Dropout (Chen 等人，2023年)采用随机路由以防止专家坍塌。CompeteSMoE (Pham 等人，2024年)采用基于竞争的路由来解决表征坍塌问题。Lo 等人 (2025年)最近的分析发现，专家多样性会随层深度增加，但同时得出结论，专家专门化的程度“仍值得怀疑”。Liu 等人 (2023年)报告称，即使在表现良好的混合专家模型中，专家表征的相似度也可能高达 99%，并提出用正交优化器作为解决方案。我们的结果进一步印证了这种质疑：即使权重空间度量显示出多样性，激活空间重叠依然存在。关键在于，这些方法主要针对路由或优化方面的多样性，而我们直接分析权重空间几何约束。

深度学习中的几何分析。 神经坍缩（Papyan 等人，2020年；Zhu 等人，2021年）表明，在训练终末阶段，分类器表征会收敛为等角紧框架（ETFs）。这种几何结构——类别表征之间的分离达到最大——启发了我们的假设：专家表征同样应能从正交性中受益。然而，分类器头是将特征映射到逻辑值的线性映射，而混合专家则是具有内部结构的非线性变换。我们的负面结果表明，神经坍缩的类比并不适用：非线性会破坏这种几何结构。

6 讨论

我们的结果对这一隐含假设提出了质疑：权重空间正交性会带来功能多样性。几何正则化不仅无效，而且不可靠——在 PTB 这类较小的数据集上表现出较高方差（标准差 >1.0）。

对混合专家设计的启示。

- **权重空间正则化并不可靠**。其效果并不一致：在 WikiText-103 上略有改善（-0.9%），在 TinyStories 上有所退化（$+0.9\%$），在 PTB 上则呈现高方差。这种不可预测性使其不适合作为通用策略。

- **激活空间正则化可能更为合适**。在训练过程中直接约束 MSOact，或许能够在不受权重-激活脱节影响的情况下实现功能多样性。

- **自然训练已经实现了较低的权重 MSO**。基线权重 MSO（$\sim 10^{-4}$）已经接近正交，这表明梯度下降会隐式地对专家几何进行正则化（Neyshabur，2017年）。

- **数据集规模很重要**。 较小的数据集表现出更高的依赖随机种子的方差，使正则化效果变得难以预测。

7 结论

我们研究了对混合专家模型的专家权重进行几何正则化能否提升模型性能，发现这种做法不仅无效，而且不可靠。正交损失未能降低权重 MSO （反而最多增加了 114% ），且对下游性能的影响不一致：在 WikiText-103 上略有提升（-0.9%），在 TinyStories 上略有下降（$+0.9\%$），在 PTB 上则呈现高方差（标准差 >1.0）。我们发现，权重正交性与激活正交性之间存在根本性的脱节：激活 MSO 仍 $\sim 1000\times$ 高于权重 MSO，且二者之间没有显著相关性（$r = -0.293$， $p = 0.523$， n=7）。这种差距源于非线性变换（SiLU、层归一化）和输入分布的影响。我们的分析表明，几何正则化既未能实现其几何目标，也无法可靠地提升性能，因此不适合作为提高混合专家模型多样性的策略。未来的工作应探索激活空间正则化，或直接针对功能行为的替代多样性度量。

局限性

规模。我们的实验采用 NanoGPT-MoE（∼130 M 个参数）。权重与激活之间的差距是否会在更大规模（1B+个参数）或不同架构（例如 Mixtral、DeepSeek-MoE）中持续存在，仍是一个悬而未决的问题。先前的研究报告了更大规模下路由器层面的增益（Wang 等人，2024年）；我们的发现可能仅适用于特定设置。

统计功效。 TinyStories 实验使用 5 个随机种子进行统计检验。WikiText-103 实验使用 3 个随机种子，结果显示出一致的改进（3.76 ± 0.02 基线与 3.73 ± 0.05正交）。PTB 实验在不同随机种子间表现出高方差（6.17 ± 0.95 基线与 5.74 ± 1.11正交），因此关于 PTB 的结论并不可靠。报告性能下降的单随机种子结果可能并不具有代表性。

作用机制。 我们发现了权重-激活差距，但尚未充分解释其成因。尽管我们推测非线性和输入结构可能发挥作用，但 SiLU 和层归一化如何改变几何关系，仍有待未来进行严谨的数学分析。

基线。 我们没有与 SMoE-丢弃法(Chen 等人，2023年)或无损失均衡(Wang 等人，2024年)进行比较；这些方法可能通过其他机制（随机路由、无辅助损失偏置更新）取得不同结果。

层间变化。 权重-激活差距在不同层之间有所不同：早期层（0–3）呈现较大 的差距（∼3000–5600×），而后期层（4– 5）呈现较小的差距（∼300–450×）。这种模式可能是因为后期层的权重 MSO 更高（∼10×高于早期层），从而减小了比率的分母。后期层为何会形成正交性较低的权重，仍有待未来研究全面分析。

正则化强度。 我们的跨数据集实验使用固定的 $\lambda = 0.001$。最佳正则化强度可能因数据集而异；全面的超参数搜索或许能揭示几何正则化在哪些条件下有益。

未来方向。值得探索的替代方法包括：(1) 梯度空间正交性，约束专家梯度指向不同方向；(2) 路由多样性损失，以最大化专家选择熵；(3) 对比目标，使专家输出而非权重彼此远离；或 (4) 激活空间正则化，在训练期间直接惩罚 MSO_{act}。

参考文献

Jimmy Lei Ba、Jamie Ryan Kiros 和 GeoffreyE Hin- ton。2016年。层归一化。*arXiv* 预印本 *arXiv:1607.06450*。

Tianlong Chen、Zhenyu Zhang、Ajay Jaiswal、Shiwei Liu 和 Zhangyang Wang。2023年。稀疏混合专家作为新型丢弃法：扩展稠密和自瘦身 Transformer。载于第十一届国际学习表征会议。

Zixiang Chen、Yihe Deng、Yue Wu、Quanquan Gu 和 Yuanzhi Li。2022年。探索深度学习中专家混合的理解。载于神经信息处理系统进展。

Zewen Chi、Shaohan Huang、Li Dong、Shuming Ma 和 Furu Wei。2022年。X-MoE：采用细粒度自适应专家选择的稀疏模型。*arXiv* 预印本 *arXiv:2204.09179*。

Damai Dai、Chengqi Deng、Chenggang Zhao、R X Xu、Huazuo Gao、Deli Chen、Jiashi Li、Wangding Zeng、Xingkai Yu、Y Wu、Zhenda Xie、Xiao Ruan、Xinghan Liu、Chong Ruan、Wenfeng Liang 和 Guanting Shen。2024年。DeepSeekMoE：迈向专家混合语言模型中的终极专家专门化。*arXiv* 预印本 *arXiv:2401.06066*。

Giang Do 等 1 人。2025年。S2MoE：通过随机学习实现鲁棒稀疏专家混合。*arXiv* 预印本 *arXiv:2503.23007*。

Ronen Eldan 和 Yuanzhi Li。2023年。TinyStories：语言模型可以有多小，同时仍能说出连贯英语？*arXiv* 预印本 *arXiv:2305.07759*。

Ziteng Fang、Yifan He、Zhi Liu、Zhen Liu、Duo Zhang 和 Shuicheng Zhai。2025年。ReMoE：采用 ReLU 路由的完全可微专家混合。*ICLR* 会议论文集。

William Fedus、Barret Zoph 和 Noam Shazeer。2022年。Switch transformers：通过简单高效的稀疏性扩展至万亿参数模型。机器学习研究杂志，23(120):1–39。

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bam- ford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lenber, GuillaumeMilber, Guillaume Lample, Lucile Saudis, Lelio Lavaud, Teven Le Scao, Timothée Lechatiel, Thibaut Lavril, Thomas Wang，以及另外 2 位作者。2024年。Mixtral专家混合。*arXiv* 预印本 *arXiv:2401.04088*。

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer 和 Zhifeng Chen。2021年。GShard：通过条件计算和自动分片扩展巨型模型。发表于国际学习表征会议。

Zhidong Liu, Weixing Wang, Xiaofeng Li, Qunyang Li, Zhaopeng Wen, Hai-Tao Guo, 和 Ziyang Zheng。2023年。利用正交优化器丰富语言模型的专家混合表示。*arXiv* 预印本 *arXiv:2310.09762*。

Ka Man Lo、Zeyu Huang、Zihan Qiu、Zili Wang 和 Jie Fu。2025年。深入探究大语言模型中的专家混合。载于计算语言学协会：*NAACL 2025* 论文集，第4427–4447页。

Ilya Loshchilov 和 Frank Hutter。2019年。解耦权重衰减正则化。载于国际学习表征会议。

Mitchell P Marcus、Beatrice Santorini 和 Mary Ann Marcinkiewicz。1993年。构建大型英语标注语料库：宾州树库。《计算语言学》，19(2):313–330。

Stephen Merity、Caiming Xiong、James Bradbury 和 Richard Socher。2016年。指针哨兵混合模型。*arXiv* 预印本 *arXiv:1609.07843*。

Behnam Neyshabur。2017年。深度学习中的隐式正则化。*arXiv* 预印本 *arXiv:1709.01953*。

Truong Son Nguyen、Nguyen Minh Ngo、Nguyen Duc Nguyen、Ngan Nguyen、The Quan Tran 和 Quang Pham Le。2024年。HyperRouter：迈向稀疏专家混合的高效训练与推理。*arXiv* 预印本 *arXiv:2410.14574*。

Vardan Papyan、XY Han 和 David L Donoho。2020年。深度学习训练终末阶段神经坍缩的普遍性。载于《美国国家科学院院刊》，第117卷，第24652–24663页。

Quang Pham、Giang Do、Huy Nguyen、TrungTin Nguyen、Chenghao Liu、Mina Sartipi、Binh T Nguyen、Savitha Ramasamy、Xiaoli Li、Steven Hoi 和 Nhat Ho。2024年。CompeteSMoE：通过竞争实现稀疏专家混合的有效训练。载于第*62*届计算语言学协会年会论文集。

PrajitRamachandran、Barret Zoph 和 Quoc V Le。2017年。搜索激活函数。*arXiv* 预印本 *arXiv:1710.05941*。

Noam Shazeer、AzaliaMirhoseini、Krzysztof Maziarz、Andy Davis、Quoc Le、Geoffrey Hinton 和 Jeff Dean。2017年。《规模惊人的神经网络：稀疏门控专家混合层》。*arXiv* 预印本 *arXiv:1701.06538*。

Lean Wang、Huazuo Mao、Haohan He、Deli Chen 和 Guanting Shen。2024 年。《专家混合的无辅助损失负载均衡策略》。*arXiv* 预印本 *arXiv:2408.15664*。

Zhihui Zhu、Tianyu Ding、Jinxin Zhou、Xiao Li、Chong You、Jeremias Sulam 和 Qing Qu。2021年。无约束特征下神经坍缩的几何分析。神经信息处理系统进展，34。