arXiv:2603.20587v1 [cs.LG] 21 Mar 2026

正交多面体区域中的神经坍塌

James Alcala* Rayna Andreeva† Vladimir A. Kobzar‡ Dustin G. Mixon‡§
Sanghoon Na¶ Shashank Sule¶ Yangxinyu Xie‖

摘要

训练用于分类的神经网络时，已知训练集的特征向量会坍塌到正则单纯形的顶点，前提是特征空间的维数 d和类别数 n 满足条件 $n \leq d+1$。这一现象称为 $neuralcollapse$。对于语言模型等其他应用，人们则采用 $n \gg d$。此时仍会出现神经坍塌现象，但涌现出的几何图形有所不同。我们刻画了 $orthoplex\ regime$ 中的这些几何图形，其中 $d+2 \leq n \leq 2d$。我们的分析主要运用了 Radon 定理和凸性。

1 引言

深度神经网络 $N\colon\ \mathbb{R}^r \to \mathbb{R}^n$ 可以分解为一个特征映射 $F\colon\ \mathbb{R}^r \to \mathbb{R}^d$，其后接一个仿射线性映射 $L\colon\ \mathbb{R}^d \to \mathbb{R}^n$，再接 softmax函数 $\mathbb{R}^n \to \mathbb{R}^n$:

$$N \;=\; \mathrm{softmax} \circ L \circ F.$$

在为分类任务进行训练时， n表示需要区分的类别数量。理想的神经网络会将 $x \mapsto e_k$映射为 $e_k \in \mathbb{R}^n$ ，其中 $e_k \in \mathbb{R}^n$ 表示对应于目标分类 $k \in [n]$ 的标准基元素（又称独热编码）：$=\{1,\ldots,n\}$ 数据点 $x \in \mathbb{R}^r$。为了构建这样的神经网络，我们先收集大量由 x及其对应的 k组成的训练集，然后训练用于定义 F和 L 的参数，直到训练集中的每个（x， k）都满足 N（x） $\approx e_k$。

2020年，Papyan、Han 和 Donoho [9] 发现了一种几何现象，称为神经坍塌；当训练此类神经网络直至训练损失降为零时，这种现象便会出现。在这个“训练终止阶段”，有几件事会同时发生。首先，对于训练集中任意一对具有相同目标分类的点 x和 x' ， k相应的特征会发生坍塌：$F(x) \approx F(x')$。特别地，特征空间 h_1，.....， h_n 中会出现一些特殊点 $\mathbb{R}^d$，使得

$$F(x) \approx h_k$$

对于训练集中的每个 (x,k)。此外，若训练集中每个类别的样本数相同，那么这些特殊点 $\{h_k\}_{k\in[n]}$ 构成一个正则单纯形的顶点。（当然，这隐含地假设 $d \geq n-1$ ，这样 $\mathbb{R}^d$ 才有可能包含一个 n顶点单纯形；稍后会进一步讨论。）接下来，如果仿射线性映射 $L\colon\ \mathbb{R}^d \to \mathbb{R}^n$ 的第 k 个分量为

$$L(h)_k = \langle w_k, h\rangle + b_k,$$

那么，我们训练过程得到的权重向量 $\{w_k\}_{k\in[n]}$ 在 $\mathbb{R}^d$ 中构成了一个以原点为中心的正则单纯形的顶点，使得

$$w_k \propto h_k - \frac{1}{n}\sum_{i=1}^{n} h_i.$$

*南加州大学数学系，洛杉矶，加利福尼亚州
†爱丁堡大学信息学院，爱丁堡，苏格兰
‡俄亥俄州立大学数学系，哥伦布，俄亥俄州§俄亥俄州立大学转化数据分析研究所，哥伦布，俄亥俄州
¶马里兰大学数学系，大学公园，马里兰州
‖宾夕法尼亚大学统计与数据科学系，费城，宾夕法尼亚州

简而言之，仿射线性映射 L 最终在几何上与训练集的特征向量 $F(x)$ 对齐。

自神经坍塌被经验发现以来，研究者已在一种无约束特征模型 [7]下对其进行了理论分析。该模型将分析与 F 的架构解耦，将训练集的特征向量视为自由变量。（这种建模选择是合理的，因为在实际应用中，F 通常过度参数化。）迄今为止，几乎所有关于神经坍塌的研究都聚焦于 $n \le d+1$ 这一情形，此时特征域能够容纳一个 n顶点单纯形。但在许多应用场景（例如语言模型）中，$n \gg d$。那么，当 $n > d+1$时，会出现哪些几何图形?

当 d 固定且 $n \to \infty$时，[6] 证明涌现出的图形在球面上均匀分布。[4] 中的较新研究对 $n > d+1$ 情形进行了非渐近分析。该研究从每个类别中取 m个样本，并将对应的无约束特征向量记为 $H = \{h_{k,i}\}_{k\in[n],\ i\in[m]}$。与之前一样，每个类别的权重向量为 $W = \{w_k\}_{k\in[n]}$。最后，我们通过从 L 中移除偏置 b_k，并进一步将每个 $h_{k,\ i}$ 和 w_k 限制在 $\mathbb{R}^d$的单位球面上，避免平移和尺度歧义带来的麻烦。给定固定的温度 $\tau > 0$，我们希望最小化交叉熵损失：

$$\begin{aligned} \text{minimize} \qquad & \mathcal{L}_{\mathrm{CE}}^{(\tau)}(W,H) := \frac{1}{mn}\sum_{k=1}^{n}\sum_{i=1}^{m} -\log\left(\frac{\exp(\langle w_k, h_{k,i}\rangle/\tau)}{\sum_j \exp(\langle w_j, h_{k,i}\rangle/\tau)}\right) \\ \text{subject to} \qquad & \|h_{k,i}\| = \|w_k\| = 1 \quad \text{for all } k\in[n],\ i\in[m]. \end{aligned}$$

当 $\tau \to 0$ 时，[4] 中的引理 3.1 表明，$\mathcal{L}^{(\tau)}$ 交叉熵 的极小化解会收敛到所谓的硬最大值 问题的极小化解：

$$\begin{aligned} \text{minimize} \qquad & \mathcal{L}_{\mathrm{HM}}(W,H) := \max_{k\in[n]}\max_{i\in[m]}\max_{k'\neq k}\langle w_k - w_{k'}, h_{k,i}\rangle \\ \text{subject to} \qquad & \|h_{k,i}\| = \|w_k\| = 1 \quad \text{for all } k\in[n],\ i\in[m]. \end{aligned}$$

如果 (W,H) 是 $\mathcal{L}_{\mathrm{HM}}(W,H)$ 的极小化解，我们就称其为一个**hardmax 码**。在这种 hardmax 设定下，[4] 观察到神经坍塌现象的一种推广，其中涌现出的几何对象是正则单纯形的推广。为明确说明这一点，给定一个球面构型 $X = \{x_k\}_{k\in[n]}$，记

$$\delta(X) := \min_{j\in[n]} \operatorname{dist}(x_j, \operatorname{conv}\{x_i\}_{i\in[n]\setminus\{j\}}).$$

如果 X 最大化 $\delta(X)$，我们就称其为一个**softmax 码**。事实证明，对于每个 hardmax 码 (W,H)，W 都是一个 softmax 码；反过来，每个 softmax 码 W 都可以扩展为一个 hardmax 码 (W,H)；参见 [4] 中的定理 C.7。此外，通常可以对每个 $k\in[n]$和 $i\in[m]$ 取 $h_{k,i} = w_k$（例如，当 $d=2$ 或 $n \le d+1$ 时），并且人们猜想这总是可行的。

迄今为止，softmax 编码已在两种情形下得到刻画：$d=2$ 和 $n \le d+1$。在这两种情形下，softmax 编码分别构成以原点为中心的正则多边形或单纯形的顶点。我们尤其关注**正交多面体区域**，其中 $d+2 \le n \le 2d$。在此区域内，已知任意 n 个 $2d$ 点 $\pm e_1$，……，$\pm e_d \in \mathbb{R}^d$ 都会构成 softmax 码。有趣的是，这些都是球面码的例子，也就是说，它们是塔姆斯问题的解：在球面上寻找 n个点，使最小两两距离最大化。

本文刻画了正交多面体区域中的所有 softmax 编码。图 1 展示了一个示意图，说明我们的结果彼此之间（以及与 [4] 的工作）有何关联。与$d=2$ 和 $n \le d+1$ 区域的情形非常相似，我们证明，softmax 编码恰好就是正交多面体区域中的球面码；第 2 节将讨论这一刻画。接下来，在第 3 节中，我们证明，对于正交多面体区域中的每个 softmax 码 W，取 $H = W$ 即可得到一个 hardmax 码 (W,H)。在第 4 节中，我们发现，对于正温度 τ，某些 softmax 编码 W 的交叉熵 CE(W,W) 值比其他编码更小，这意味着 softmax 编码并非都同样出色。有趣的是，当 τ 足够小时，最佳 softmax 编码 W 具有“低熵”；而当 τ 足够大时，则具有“高熵”。最后，我们在第 5 节进行讨论。

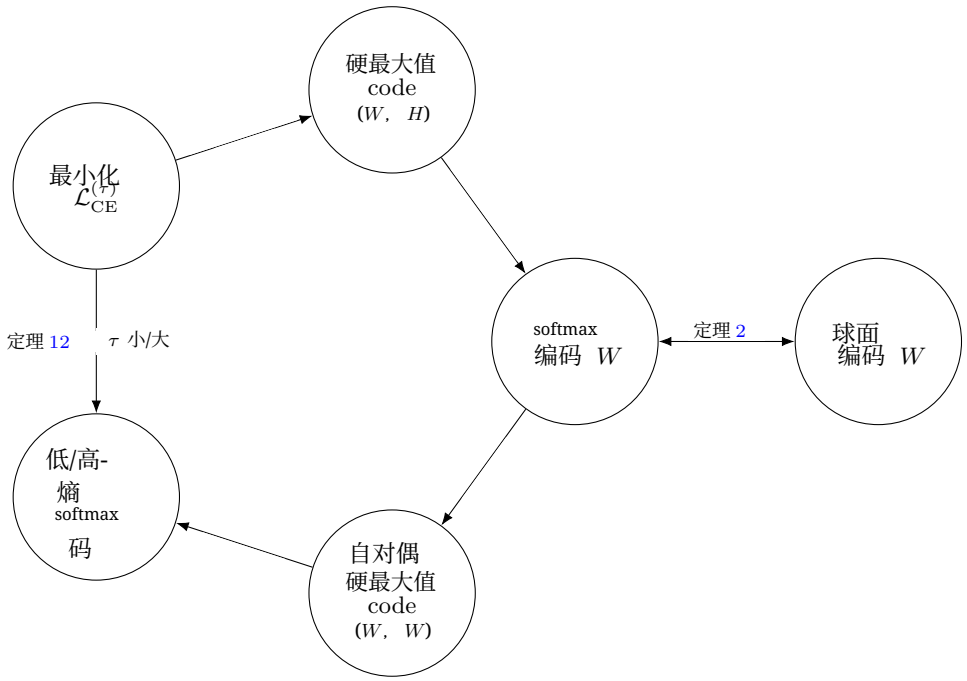

图 1：正交多面体区域内神经坍塌主要结果的示意图。定理 2、5 和 12 分别见第2、3和4节。

2 正交多面体区域中的 softmax 码

给定一个球面构型 $X = \{x_i\}_{i\in[n]}$，我们记

$$\alpha(X) := \max_{\substack{i,j\in[n]\\ i\neq j}} \langle x_i, x_j\rangle.$$

若 X 最小化 $\alpha(X)$，我们就称 X 为球面码。由于球面码可用于模拟编码，长期以来一直受到研究。由于 α 是非凸的，且存在许多局部极小化解，因此球面码通常难以找到；不过，在正交多面体区域中，球面码已得到全面理解：

命题 1（Rankin 正交多面体界；参见 [11],，参见 [8]）。固定 $d \geq 2$ 和 $n \geq d+2$。对于 S^{d-1} 中的每个构型 $X = \{x_i\}_{i\in[n]}$，均有

$$\alpha(X) \geq 0.$$

此外，当且仅当 $n \leq 2d$ 时，等号成立。

接下来给出本节的主要结果，它将正交多面体区域中的球面码与 softmax 码联系起来：

定理 2。固定 $d \geq 2$ 和 $n \geq d+2$。对于 S^{d-1}中的每个构型 $X = \{x_i\}_{i\in[n]}$，都有

$$\delta(X) \leq 1.$$

此外，$\delta(X) = 1$ 当且仅当 $\alpha(X) = 0$。也就是说，在正交多面体区域中，*softmax* 编码恰好就是球面码。

在证明定理 2之前，我们先证明几个引理。

引理 3。固定 $d \geq 2$ 和 $n \geq d+2$。给定 S^{d-1}中的一个构型 $X = \{x_i\}_{i\in[n]}$，假设存在一个划分 $[n] = A \sqcup B$，其中 A 和 B 均非空，且交集

$$K := \operatorname{conv}\{x_i\}_{i\in A} \cap \operatorname{conv}\{x_i\}_{i\in B}$$

非空，且满足 $0 \notin K$。然后 $\delta(X) < 1$。

证明。令 v 表示 0 在 K 上的投影。（值得注意的是，v 非零，因为 $0 \notin K$。）由于 $v \in K \subseteq$属于 $\mathrm{conv}\{x_i\}_{i\in[n]}$，我们可以将 v 表示为凸组合：

$$v = \sum_{i=1}^{n} a_i x_i.$$

由此可得

$$\|v\|^2 = \langle v, v\rangle = \left\langle \sum_{i=1}^{n} a_i x_i, v \right\rangle = \sum_{i=1}^{n} a_i \langle x_i, v\rangle \le \max_{i\in[n]} \langle x_i, v\rangle,$$

也就是说，存在 $j \in [n]$ 使得 $\langle x_j, v\rangle \ge \|v\|^2$。这进而意味着

$$\|x_j - v\|^2 = 1 - 2\langle x_j, v\rangle + \|v\|^2 \le 1 - 2\|v\|^2 + \|v\|^2 = 1 - \|v\|^2.$$

不失一般性，设 $j \in A$，因此

$$v \in K \subseteq \mathrm{conv}\{x_i\}_{i\in B} \subseteq \mathrm{conv}\{x_i\}_{i\in[n]\setminus\{j\}}.$$

因此，

$$\delta(X) \le \mathrm{dist}(x_j, \mathrm{conv}\{x_i\}_{i\in[n]\setminus\{j\}}) \le \|x_j - v\| \le \sqrt{1 - \|v\|^2} < 1,$$

其中最后一步利用了 v 非零这一事实。 □

引理4。 固定 $d \ge 2$ 和 $n \ge d+2$。对于 S^{d-1}中的每个构型 $X = \{x_i\}_{i\in[n]}$，都有 t

$$\alpha(X) > 0 \qquad \Longrightarrow \qquad \delta(X) < 1.$$

证明。 根据 Radon 定理 [10], , 存在一个划分 $[n] = A \sqcup B$，使得 A 和 B 均非空，且其交集

$$K := \mathrm{conv}\{x_i\}_{i\in A} \cap \mathrm{conv}\{x_i\}_{i\in B}$$

也非空。根据引理 3，不妨设 $0 \in K$。固定 j，$k \in [n]$ 使得 $\langle x_j, x_k\rangle > 0$，并记 $\theta :=$ $\arccos\langle x_j, x_k\rangle \in [0, \frac{\pi}{2})$。不失一般性，设$^{-}j \in A$，因此

$$0 \in K \subseteq \mathrm{conv}\{x_i\}_{i\in B} \subseteq \mathrm{conv}\{x_i\}_{i\in[n]\setminus\{j\}}.$$

由于 $\mathrm{conv}\{x_i\}_{i\in[n]\setminus\{j\}}$ 包含 x_k 和 0，根据凸性，它也包含整条线段 $\mathrm{conv}\{x_k, 0\}$。因此，

$$\delta(X) \le \mathrm{dist}(x_j, \mathrm{conv}\{x_i\}_{i\in[n]\setminus\{j\}}) \le \mathrm{dist}(x_j, \mathrm{conv}\{x_k, 0\}) = \sin\theta < 1.$$ □

定理 2的证明。 界 $\delta(X) \le 1$ 由 [4]中的引理C.9给出。同时，“此外”陈述的“仅当”方向可由 Rankin 正交多胞体界（命题 1）与引理 4结合得到。剩下需要证明“若”方向。为此，假设 $\alpha(X) = 0$。那么对于每个 $j \in [n]$，每个满足 x_i 的 $i \ne j$ 都位于半空间

$$H_j := \{z \in \mathbb{R}^d : \langle z, x_j\rangle \le 0\}.$$

由于 H_j 是凸的，因此可得 $\mathrm{conv}\{x_i\}_{i\in[n]\setminus\{j\}} \subseteq H_j$，从而

$$\mathrm{dist}(x_j, \mathrm{conv}\{x_i\}_{i\in[n]\setminus\{j\}}) \ge \mathrm{dist}(x_j, H_j) = 1,$$

其中最后一步由投影定理得出。接着对 $j \in [n]$ 求最小值，得到 $\delta(X) \ge 1$，但由于一般而言 $\delta(X) \le 1$，我们得出 $\delta(X) = 1$。 □

3 从缺乏松动点到自对偶性

在本节中，我们将说明，正交多面体区域中的 hardmax 码具有一种自对偶性：

定理 5。 正交多面体区域中的每个 *hardmax* 码$(W = \{w_k\}_{k\in[n]}, H = \{h_{k,i}\}_{k\in[n],\ i\in[m]})$ 均满足 $h_{k,i} = w_k$，对任意 $k \in [n]$ 和 $i \in [m]$均成立。

我们的证明运用了 [4], 中的一项重要结果，我们将在给出必要定义后陈述该结果。

定义 6。

(a) $X = \{x_i\}_{i\in[n]}$ 的一个**softmax 松动点** 是满足下述条件的指标 $j \in [n]$：

$$\operatorname{dist}(x_j, \operatorname{conv}\{x_i\}_{i\in[n]\setminus\{j\}}) > \delta(X).$$

(b) $X = \{x_i\}_{i\in[n]}$ 的一个**Tammes 松动点** 是满足下述条件的指标 $j \in [n]$：

$$\max_{i\in[n]\setminus\{j\}} \langle x_i, x_j\rangle < \alpha(X).$$

现在，我们可以揭示定理 5的证明技巧了。该技巧通过以下命题来实现：

命题 7 ([4]中的定理 3.7)。 固定 d 和 n，使得 S^{d-1} 中的每个大小为-n 的 *softmax* 码都没有 *softmax* 松动点，并且每个大小为-n 的球面码都没有 *Tammes* 松动点。则以下陈述等价：

(a) 每个硬最大值码 (W, H) 都满足 $h_{k,\ i} = w_k$，对任意 $k \in [n]$ 和 $i \in [m]$均成立。

(b) S^{d-1} 中的大小为-n的 *softmax* 码恰好是 S^{d-1}中的大小为-n 的球面码。

在 [4], 中已经证明，在 $d = 2$ 和 $n \le d+1$区域内，softmax 码与球面码完全相同，并且在两种意义下都无摇摆点。因此，命题 7 表明这些码是自对偶的。我们将沿用这一方法，证明正交多面体区域中的自对偶性。接下来仍需证明这些码在两种意义下都无摇摆点。

引理 8. 正交多面体区域中的每个 *softmax* 码都没有 *softmax* 松动点。

证明。 固定一个正交多面体区域中的 softmax 码 $X = \{x_i\}_{i\in[n]}$，即 $\delta(X) = 1$。根据 Radon 定理 [10], ，存在一个划分 $[n] = A \sqcup B$，其中 A 和 B 均非空，使得交集

$$K := \operatorname{conv}\{x_i\}_{i\in A} \cap \operatorname{conv}\{x_i\}_{i\in B}$$

也非空。根据引理 3，我们有 $0 \in K$。

考虑任意一个 $j \in [n]$。不失一般性，我们有 $j \in A$，因此

$$0 \in K \subseteq \operatorname{conv}\{x_i\}_{i\in B} \subseteq \operatorname{conv}\{x_i\}_{i\in[n]\setminus\{j\}}.$$

由此可得

$$1 = \delta(X) \le \operatorname{dist}(x_j, \operatorname{conv}\{x_i\}_{i\in[n]\setminus\{j\}}) \le \|x_j - 0\| = 1,$$

因此

$$\operatorname{dist}(x_j, \operatorname{conv}\{x_i\}_{i\in[n]\setminus\{j\}}) = \delta(X).$$

由于 j 是任意的，我们得出 X 没有 softmax 松动点。 □

引理 9. 正交多面体区域中的每个球面码都没有 *Tammes* 松动点。

我们的证明 引理 9利用了对正交多面体区域中球面码的分类：

命题 10 （ [5] 中的定理 2）。 固定 $d \ge 2$ 和 $n \in [d+2, 2d]$, 并考虑 S^{d-1} 中任意一个达到 *Rankin* 正交多面体界等号的构型$\{x_i\}_{i\in[n]}$。那么存在一个可能为空的子集 $S_0 \subseteq [n]$ 以及一个划分 $S_1 \sqcup \cdots \sqcup S_l = [n] \setminus S_0$，满足 $l \ge n - d$，使得

(a) $|S_0| =$ 维张成空间$\{x_i\}_{i\in S_0}$,

(b) $|S_j| =$ 维张成空间$\{x_i\}_{i\in S_j} + 1$ 对于每个 $j \in [l]$，且

(c) 当 $j \neq j'$ 时，张成空间$\{x_i\}_{i\in S_j} \perp$ 张成空间$\{x_i\}_{i\in S_{j'}}$。

引理的证明 *9.* 固定一个正交多面体区域内的球面码 $X = \{x_i\}_{i\in[n]}$，即 $\alpha(X) = 0$。任取 $j \in [n]$。根据命题 10，存在 $i \in [n] \setminus \{j\}$ 使得 $\langle x_i, x_j\rangle = 0 = \alpha(X)$。由于$j$ 是任意的，我们得出 X 没有 Tammes 松动点。

□

定理的证明 *5.* 固定 d 和 n，使得 $d + 2 \leq n \leq 2d$。根据引理 8，n 中的每个大小为 S^{d-1} 的 softmax 码都没有 softmax 松动点；根据引理 9，每个大小为n的球面码都没有 Tammes 松动点。根据定理 2，n 中的大小为 S^{d-1} 的 softmax 码，恰好是n中的大小为 S^{d-1}的球面码。结论由命题 7得出。

□

4 并非所有 softmax 码都生而平等

至此，我们已经了解正交多面体区域中的所有 softmax 码，并且知道它们都是自对偶的，也就是说，每个 softmax 码都能唯一扩展为一个 hardmax 码。本节中，我们回到原本的非零温度情形。事实证明，在这种情形下，某些 softmax 码比其他码更受青睐，而这种偏好由温度决定。我们先给出一个定义。

定义 11。

(a) 若一个 softmax 码由一个p点正则单纯形的顶点和一个正交多胞体组成，二者彼此正交且共同张成环境空间，则称该 softmax 码关于参数 p 为**低熵**。(b) 若一个 softmax 码由若干两两正交的 p 点或（$p + 1$）点正则单纯形的顶点组成，且这些单纯形共同张成环境空间，则称该 softmax 码关于参数 p 为**高熵**。

图 2展示了低熵和高熵 softmax 码的示例。每个正交多胞体的顶点（其中 $n = 2d$）同时关于 $p = 2$ 呈低熵，关于 $p = 2$呈高熵。类似地，三角双锥的顶点（其中 $d = 3$ 且 $n = 5$）同时关于 $p = 3$呈低熵，关于 $p = 2$呈高熵。若要举出这两种概念不重合的例子，就需要考虑 $d = 4$ 和 $n = 6$。这里，一个低熵 softmax 码（其中 $p = 4$）由三维子空间中一个四面体的顶点以及正交补中的一对对跖点构成。另一方面，一个高熵 softmax 码（其中 $p = 2$）则是在两个正交平面中各放置一个等边三角形的顶点。

下文中，我们用 SMC(d, n) 表示 S^{d-1} 中所有 n 点 softmax 码的集合。现在可以陈述本节的主要结果了。

定理 12. 固定满足 $d + 2 \leq n \leq 2d$ 的 d 和 n。存在阈值

$$0 < \tau_-(n) < \tau_+(n) < \infty$$

使得对于每个温度 $\tau < \tau_-(n)$（分别地，$\tau > \tau_+(n)$），在$W \in \mathrm{SMC}(d, n)$ 约束下，$\mathcal{L}^{(\tau)}$ 交叉熵(W, W) 的极小化解是低熵（分别地，高熵） *softmax* 码。

根据定理2，我们可以将命题10 视为对可行域 SMC(d, n) 的一种分解。采用命题10的记号，我们使用划分 $S_0 \sqcup S_1 \sqcup \cdots \sqcup S_l = [n]$，使得批次 $\{w_k\}_{k\in S_i}$对于 $i \in \{0, 1, \ldots, l\}$ 是相互正交的。于是，需要最小化的量为

$$\begin{aligned}
\mathcal{L}^{(\tau)}_{\mathrm{CE}}(W, W) &= \frac{1}{n}\sum_{k=1}^{n} -\log\left(\frac{\exp(\langle w_k, w_k\rangle/\tau)}{\sum_j \exp(\langle w_j, w_k\rangle/\tau)}\right) \\
&= \frac{1}{n}\sum_{i=0}^{l}\sum_{k\in S_i} \log\left(\sum_{j=1}^{n}\exp(\langle w_j, w_k\rangle/\tau)\right) - \frac{1}{\tau} \\
&= \frac{1}{n}\sum_{i=0}^{l}\sum_{k\in S_i} \log\left(n - |S_i| + e^{1/\tau} + \sum_{\substack{j\in S_i \\ j\neq k}}\exp(\langle w_j, w_k\rangle/\tau)\right) - \frac{1}{\tau}.
\end{aligned}$$

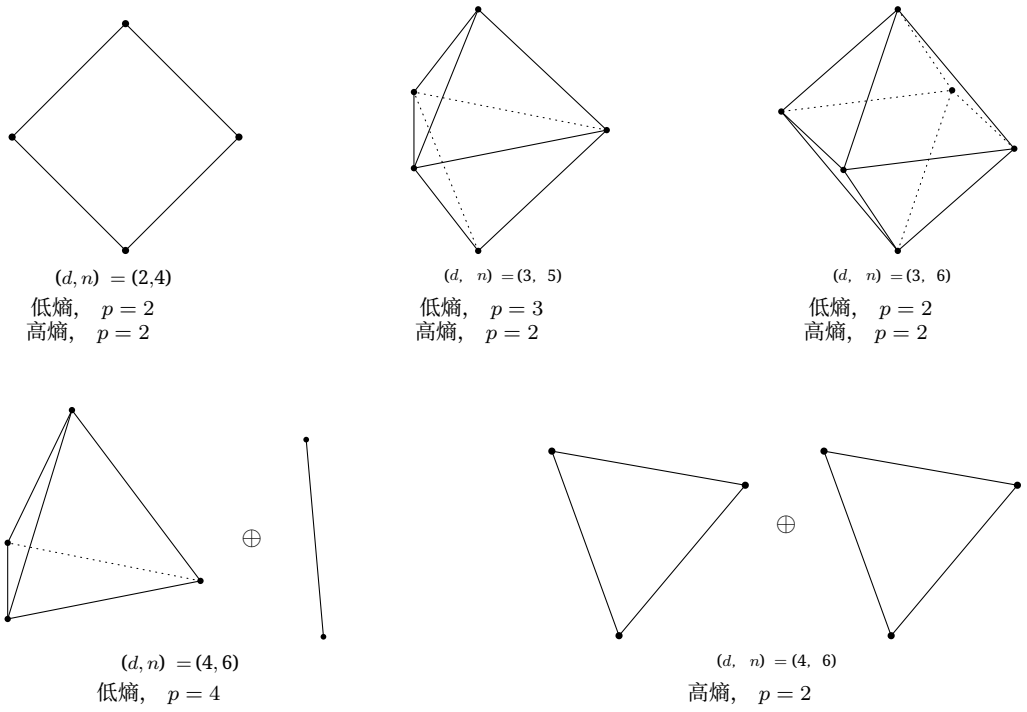

图 2：低熵和高熵 softmax 码的最小示例。

特别地，通过分解为相互正交的批次，我们的目标函数按批次拆分为以下形式的项

$$\mathcal{L}_{\tau,c}(X) := \sum_k \log\left(c + \sum_{j\neq k} \exp(\langle x_j, x_k\rangle/\tau)\right),$$

其中 $c > 0$ 由 $n - |S_i| + e^{1/\tau}$ 在应用于第 i个批次时给出。受 [6], 的启发，我们证明 $\mathcal{L}_{\tau,c}(X)$ 在以原点为中心的正则单纯形的顶点处取得最小值。

引理 13。固定 d 和 n，使得 $2 \le n \le d+1$，以及 $\tau, c > 0$。使 $\mathcal{L}_{\tau,c}(X)$ 最小的 S^{d-1} 中的 n 点构型$X = \{x_i\}_{i\in[n]}$ 满足

$$\sum_{i=1}^n x_i = 0 \qquad and \qquad \langle x_i, x_j\rangle = -\frac{1}{n-1} \quad \forall\, i, j \in [n],\, i \neq j.$$

证明。令 $s := \sum_i x_i$，且 $\beta := 1/\tau$。于是，对每个 $k \in [n]$，由詹森不等式可得

$$\frac{1}{n-1}\sum_{j\neq k} \exp(\langle x_j, x_k\rangle/\tau) \ge \exp\left(\frac{1}{n-1}\sum_{j\neq k}\langle x_j, x_k\rangle/\tau\right) = \exp\left(\frac{\beta}{n-1}\big(\langle s, x_k\rangle - 1\big)\right).$$

由对数函数的单调性可知

$$\mathcal{L}_{\tau,c}(X) \ge \sum_{k=1}^n \log\left(c + (n-1)\exp\left(\frac{\beta}{n-1}\big(\langle s, x_k\rangle - 1\big)\right)\right) = \sum_{k=1}^n \log\left(c + ae^{b\langle s, x_k\rangle}\right),$$

其中 $a = (n-1)e^{-\beta/(n-1)}$ 且 $b = \beta/(n-1)$。由于函数 $t \mapsto \log(c + ae^{bt})$ 是凸且单调的，再次应用詹森不等式可得

$$\frac{1}{n}\mathcal{L}_{\tau,c}(X) \ge \frac{1}{n}\sum_{k=1}^n \log\left(c + ae^{b\langle s, x_k\rangle}\right) \ge \log\left(c + ae^{b\frac{1}{n}\sum_{k=1}^n \langle s, x_k\rangle}\right) = \log\left(c + ae^{b\|s\|^2/n}\right) \ge \log(c + a).$$

最后，我们刻画 X 何时在此下界中取等号。最后一步取等要求 $s = 0$，这进而保证倒数第二个不等式取等。剩下的不等式当且仅当每个 k 都具有一个常数 c_k，使得 $\langle x_j, x_k\rangle = c_k$ 对每个 $j \neq k$均成立时达到等号。考虑 $0 = \langle s, x_k\rangle = 1 + (n-1)c_k$，我们必然有 $c_k = -1/(n-1)$。 □

因此，对于满足 $W \in \mathrm{SMC}(d,\ n)$ 的 $\mathcal{L}^{(\tau)}$ 交叉熵$(W,\ W)$ 的每个极小化解，批次 S_1、.....、S_l 对应以原点为中心的正则单纯形的顶点。结果表明，剩下的批次 S_0 为空，此外还有 $l = n - d$:

引理 14。固定 d 和 n，使得 $d + 2 \le n \le 2d$，以及 $\tau > 0$。考虑以下问题的任意一个极小化解：$\mathcal{L}^{(\tau)}$ 交叉熵$(W,\ W)$ 满足约束 $W \in \mathrm{SMC}(d,\ n)$，并考虑任意一个划分 $S_0 \sqcup S_1 \sqcup \cdots \sqcup S_l = [n]$，如命题 *10*所述。那么 S_0 为空，W 张成 $\mathbb{R}^d$，且 $l = n - d$。

证明。假设 S_0 非空。若 S_0 是单元素集，则将引理13 应用于球面构型 $X := \{w_k\}_{k \in S_0 \cup S_1}$ 在张成空间 X 中可得，W不是以下问题的极小化解：$\mathcal{L}^{(\tau)}$ 交叉熵$(W,\ W)$，满足约束$W \in \mathrm{SMC}(d,\ n)$。否则，我们可以将引理 13 应用于球面构型 $X := \{w_k\}_{k \in S_0}$在张成空间 X 中，从而得到相同的结论。

接下来，假设 W 不张成 $\mathbb{R}^d$。那么，将引理13 应用于位于张成空间 X +(span W)中的球面构型$X := \{w_k\}_{k \in S_0 \cup S_1}$，可知在满足$W \in \mathrm{SMC}(d, n)$的条件下，$W$并非 $\mathcal{L}^{(\tau)}$ 交叉熵(W, W)的极小化解。

最后，对于每个 $i \in [l]$，令 $d_i := |S_i| - 1$ 表示 $\{w_k\}_{k \in S_i}$的张成空间的维数。那么

$$d = \sum_{i=1}^{l} d_i \qquad \text{and} \qquad n = \sum_{i=1}^{l} (d_i + 1) = d + l,$$

也就是说，$l = n - d$。 □

总之，每个极小化解都由两两正交且以原点为中心的正则单纯形的顶点构成。接下来，我们确定最佳的单纯形大小元组。对于每个 $i \in [l]$，令 $d_i := |S_i| - 1$ 表示 $\{w_k\}_{k \in S_i}$的张成空间的维数。我们的任务是确定正整数 $d_1, \ldots, d_l$，它们的和均为 d （由引理 14），使得下式最小化

$$\begin{aligned}
\mathcal{L}_{\mathrm{CE}}^{(\tau)}(W, W) &= \frac{1}{n} \sum_{i=0}^{l} \sum_{k \in S_i} \log \left(n - |S_i| + e^{1/\tau} + \sum_{\substack{j \in S_i \\ j \neq k}} \exp(\langle w_j, w_k \rangle / \tau) \right) - \frac{1}{\tau} \\
&= \frac{1}{n} \sum_{i=1}^{l} (d_i + 1) \log \left(n - (d_i + 1) + e^{1/\tau} + d_i e^{-1/(\tau d_i)} \right) - \frac{1}{\tau} \\
&= \frac{1}{n} \sum_{i=1}^{l} f_{n,\tau}(d_i) - \frac{1}{\tau},
\end{aligned}$$

其中，第一个等式来自我们对 $\mathcal{L}^{(\tau)}$ 交叉熵(W, W) 的原始化简，第二个等式应用了引理 13 和14，最后一个等式则来自如下定义 $f_{n,\tau}$：$[1, n-1] \to \mathbb{R}$

$$f_{n,\tau}(x) = (x + 1) \log \left(n - x - 1 + e^{1/\tau} + x e^{-1/(\tau x)} \right).$$

图 3 展示了在特殊情形 $n = 10$ 下我们离散优化问题的示意图。一般而言，当 τ 足够小时，$f_{n,\tau}$ 是凹的（相应地，当其足够大时则是凸的），这进而决定了 $(d_1, \ldots, d_l)$ 的哪种取值能使 $\mathcal{L}^{(\tau)}$ 交叉熵(W, W) 最小。

引理 15. 对于每个整数 $n > 1$，存在阈值

$$0 < \tau_-(n) < \tau_+(n) < \infty$$

使得对于每个 $\tau < \tau_-(n)$ （相应地，$\tau > \tau_+(n)$），$f_{n,\tau}$ 都是严格凹的（相应地，严格凸的）。

引理 15 的证明是一个技术性的微积分论证，我们将其推迟到本节末尾。目前，我们用这个引理来证明本节的主要结果。

定理*12*的证明。假设 W是以下目标函数的极小化解：$\mathcal{L}^{(\tau)}$ 交叉熵$(W,\ W)$，并满足 $W \in \mathrm{SMC}(d,\ n)$。那么，如上所述，引理13 和14 表明 W 由以下集合的顶点构成 $l := n - d$ 两两正交且以原点为中心的正则单纯形，其中第 i个单纯形的维数为 d_i，且满足$\sum_{i=1}^{l} d_i = d$；此外，

$$\mathcal{L}_{\mathrm{CE}}^{(\tau)}(W, W) = \frac{1}{n} \sum_{i=1}^{l} f_{n,\tau}(d_i) - \frac{1}{\tau}.$$

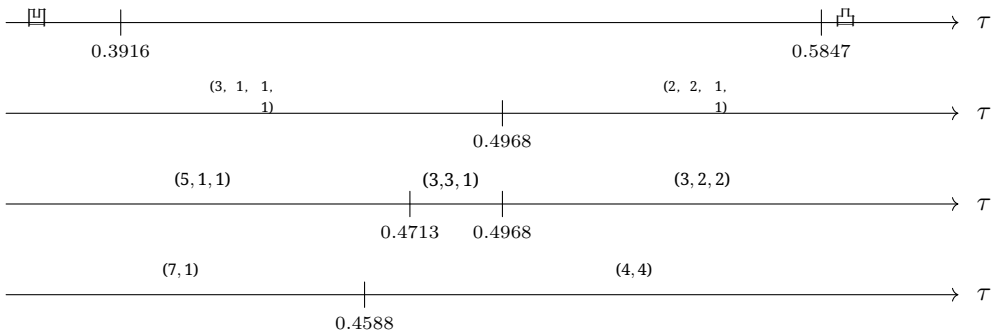

图 3：固定 $n = 10$。第一个图展示了 $f_{n,\tau}$ 何时为凹或凸。接下来的三个图展示了 $(d_1, \ldots, d_l)$ 随温度 τ 变化时的最佳选择。这里，我们分别取 d 为 6、7 和 8。我们省略 $d = 5$ 情况，因为相应的 softmax 码在旋转意义下唯一。

因此，只需确定正整数元组（d_1，……，d_l），使其和为 d，并使上式右端最小。考虑引理15中给出的阈值 (n) 和$(n_\circ$

情形 I： $\tau < \tau_{-}(n)$。我们声称，最优的 $d_1 \geq \cdots \geq d_l$ 具有 $d_2 = 1$，因此相应的 softmax 码是参数为 $p = d_1 + 1$ 的低熵码。为证明这一点，我们将证明$d_2 > 1$ 意味着（d_1，……，d_l）并非最优。只需验证由 $d_1' := d_1 + 1$、$d_2' := d_2 - 1$ 和 $d_i' := d_i$ 定义的（d_1'，……，d_l'）对所有 $i > 2$ 都是一种改进。首先，

$$\sum_{i=1}^{l} f_{n,\tau}(d_i) - \sum_{i=1}^{l} f_{n,\tau}(d_i') = \big(f_{n,\tau}(d_2) - f_{n,\tau}(d_2 - 1)\big) - \big(f_{n,\tau}(d_1 + 1) - f_{n,\tau}(d_1)\big).$$

根据中值定理，存在满足条件的 c_1 和 c_2

$$d_2 - 1 < c_2 < d_2 \leq d_1 < c_1 < d_1 + 1$$

使得

$$f_{n,\tau}'(c_1) = f_{n,\tau}(d_1 + 1) - f_{n,\tau}(d_1), \qquad f_{n,\tau}'(c_2) = f_{n,\tau}(d_2) - f_{n,\tau}(d_2 - 1).$$

由此可得

$$\sum_{i=1}^{l} f_{n,\tau}(d_i) - \sum_{i=1}^{l} f_{n,\tau}(d_i') = f_{n,\tau}'(c_2) - f_{n,\tau}'(c_1) > 0,$$

其中，最后一步是根据以下事实得出的：$f_{n,\ \tau}$ 由引理 15 可知是严格凹函数。

情形 II： $\tau > \tau_{+}(n)$。我们声称，最优的 $d_1 \geq \cdots \geq d_l$ 具有 $d_1 - d_l \leq 1$，因此相应的 softmax 码是参数为 $p = d_l + 1$ 的高熵码。为证明这一点，我们将证明 $d_1 - d_l > 1$ 意味着（d_1，……，d_l）并非最优。只需验证由 $d_1' := d_1 - 1$、$d_l' := d_l + 1$ 和 $d_i' = d_i$ 定义的（d_1'，……，d_l'）对所有 $1 < i < l$ 都是一种改进。首先，

$$\sum_{i=1}^{l} f_{n,\tau}(d_i) - \sum_{i=1}^{l} f_{n,\tau}(d_i') = \big(f_{n,\tau}(d_1) - f_{n,\tau}(d_1 - 1)\big) - \big(f_{n,\tau}(d_l + 1) - f_{n,\tau}(d_l)\big).$$

根据中值定理，存在满足以下条件的 c_1 和 c_l

$$d_l < c_l < d_l + 1 \leq d_1 - 1 < c_1 < d_1$$

使得

$$f_{n,\tau}'(c_1) = f_{n,\tau}(d_1) - f_{n,\tau}(d_1 - 1), \qquad f_{n,\tau}'(c_l) = f_{n,\tau}(d_l + 1) - f_{n,\tau}(d_l).$$

由此可得

$$\sum_{i=1}^{l} f_{n,\tau}(d_i) - \sum_{i=1}^{l} f_{n,\tau}(d_i') = f_{n,\tau}'(c_1) - f_{n,\tau}'(c_l) > 0,$$

其中最后一步是根据以下事实得出的：$f_{n,\ \tau}$ 由引理15可知是严格凸的。 □

引理15的证明。为方便起见，我们记

$$g(x) := n - x - 1 + e^{\beta} + xe^{-\beta/x},$$

其中 $\beta := 1/\tau$。值得注意的是，对每个 $x \in [1, n-1]$，$g(x) > 0$。我们有

$$f''_{n,\tau}(x) = \frac{1}{g(x)}\left((x+1)g''(x) + 2g'(x) - (x+1)\frac{g'(x)^2}{g(x)}\right).$$

有时，写作 $u := 1/(\tau x) = \beta/x$ 会更方便。我们有

$$g'(x) = -1 + (1+u)e^{-u}, \qquad g''(x) = \frac{u^2 e^{-u}}{x}.$$

接下来，我们将直接使用以下基本结果，不再另行解释：

$$1 \le x \le n-1\,, \qquad 0 \le u \le \beta\,.$$

要证明凹性，只需证明

$$Q := (x+1)g''(x) + 2g'(x) - (x+1)\frac{g'(x)^2}{g(x)} < 0.$$

对于前两项，我们有

$$(x+1)g''(x) = (1+\tfrac{1}{x})u^2 e^{-u} \le 2\beta^2 e^{-\beta/n},$$
$$2g'(x) = 2\big(-1 + (1+u)e^{-u}\big) \le 2\big(-1 + (1+\beta)e^{-\beta/n}\big),$$

并且由于

$$g(x) = n - x - 1 + e^{\beta} + xe^{-\beta/x} \ge e^{\beta} + xe^{-\beta/x} > 0,$$

第三项满足

$$-(x+1)\frac{g'(x)^2}{g(x)} \le 0.$$

综合以上各项，

$$Q \le 2\big(-1 + (1+\beta+\beta^2)e^{-\beta/n}\big),$$

对于所有足够大的 β，该式均为负；也就是说，只需对某个足够小的 $\tau_-(n)$ 取 $\tau < \tau_-(n)$ 即可。

对于凸性结论，只需证明 $Q > 0$。对于前两项，我们有

$$(x+1)g''(x) = (1+\tfrac{1}{x})u^2 e^{-u} \ge (1+\tfrac{1}{n})u^2 e^{-u}, \qquad 2g'(x) = 2\big(-1 + (1+u)e^{-u}\big).$$

对于第三项，我们有

$$g'(x)^2 = \big(-1 + (1+u)e^{-u}\big)^2, \qquad \frac{g(x)}{x+1} = \frac{n+e^{\beta}}{x+1} - 1 + \frac{x}{x+1}e^{-u} \ge \frac{1}{n} + \frac{1}{2}e^{-u}.$$

综合以上各项，

$$Q \ge (1+\tfrac{1}{n})u^2 e^{-u} + 2\big(-1 + (1+u)e^{-u}\big) - \frac{\big(-1 + (1+u)e^{-u}\big)^2}{\frac{1}{n} + \frac{1}{2}e^{-u}}$$
$$= \tfrac{1}{n}u^2 + (\tfrac{1}{3} + \tfrac{1}{n})u^3 + \cdots,$$

其中最后一步是关于 u、以 0 为中心的泰勒级数展开。根据泰勒定理可知，对于所有充分小的 u，都有 $Q > 0$。由于 $u = 1/(\tau x) \le 1/\tau$，因此只需对某个充分大的 τ_+ 取 $\tau > \tau_+(n)$。

□

5 讨论

本文刻画了正交多面体区域中的 softmax 编码，证明了它们始终是自对偶的，并确定了其中哪些编码在极低温度和极高温度下更受交叉熵损失偏好。仍有几个问题尚待解决。首先，softmax 编码是否始终是球面码？已知当 $d = 2$ 或 $n \leq d + 1$时，这一吻合关系成立；现在我们知道，当 $d+2 \leq n \leq 2d$时，这一关系同样成立。对于 (d, n)，仅剩少数几个例子，其对应的球面码已为人所知 [3],，因此这些自然成为进一步评估的下一批对象。是否存在普适最优性理论的推广 [2]，能够解释这种吻合关系？更一般地说，是否有希望确定具有 $n > 2d$的 softmax 编码？最后，若能更好地理解具有正温度 $\tau > 0$的交叉熵损失极小化解，以及优化器如何随 τ演化，将会很有意义；这在定性上类似于 [1]中所研究的问题。

致谢

本研究始于2024年美国数学学会数学研究社区研讨会“对抗式、可解释及可说明人工智能的数学”。DGM 获得了美国国家科学基金会 DMS 2220304 项目的资助。本文中的所有图均借助 ChatGPT 制作。

参考文献

[1] X. Chen, V. Gonzalez, E. Goodman, S. Kang, K. A. Okoudjou, 普适最优构形：关于 p-框架势能，《计算数学进展》46 (2020年) 4。[2] H. Cohn, A. Kumar，球面点的普适最优分布，《美国数学会杂志》20 (2007) 99–148。[3] T. Ericson, V. Zinoviev，《欧几里得球面上的码》，爱思唯尔，2001。[4] J. Jiang, J. Zhou, P. Wang, Q. Qu, D. G. Mixon, C. You, Z. Zhu，适用于大量类别的广义神经坍塌，2024年国际机器学习大会，22010–22041。[5] E. J. King, D. G. Mixon, H. Parshall, C. Wells，低复杂度的唯一最优码具有对称性，《实验数学杂志》2 (2026) 81–113。[6] J. Lu, S. Steinerberger，交叉熵损失下的神经坍塌，《应用与计算调和分析》59 (2022)，224–241。[7] D. G. Mixon, H. Parshall, J. Pi，无约束特征下的神经坍塌，《采样理论、信号处理与数据分析》20 (2022) 11。[8]J. O’Rourke，在 $\mathbb{R}^d$ 中打包钝角向量，mathoverflow.net/questions/208484/packing-obtuse-vectors-in-mathbbrd/。[9] V. Papyan, X. Y. Han, D. L. Donoho，深度学习训练终末阶段神经坍塌的普遍性，《美国国家科学院院刊》117 (2020年) 24652–24663。[10] J. Radon，含有公共点的凸集族，Math. Ann. 83 (1921) 113–115。[11] R. A. Rankin，球冠在 n 维空间中的最密堆积，《格拉斯哥数学杂志》2 (1955) 139–144。