

Gradient Analysis of Standard Softmax and Normalized Softmax

fangqing

2025 年 11 月 12 日

目录

1	Preface	2
1.1	前向传播	2
1.2	反向传播	2
2	Standard Softmax Gradient Analysis	3
2.1	雅可比矩阵	3
2.2	海森矩阵	3
2.3	Standard Softmax Analysis	7
3	Normalized Softmax Gradient Analysis	8
4	PGA	10
4.1	ArcFace 的切向合力	10
4.2	PGA K/Z 对齐的切向合力	12
4.3	切空间内积判断对抗性	12

1 Preface

1.1 前向传播

feature: $\mathbf{x} \in \mathbb{R}^D$, target: $y \in \{1, 2, \dots, C\}$, dataset: $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^N$

$$\mathbf{z} = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_C \end{bmatrix} = \underbrace{\begin{bmatrix} w_{11} & w_{21} & \cdots & w_{D1} \\ w_{12} & w_{22} & \cdots & w_{D2} \\ \vdots & \vdots & \ddots & \vdots \\ w_{1C} & w_{2C} & \cdots & w_{DC} \end{bmatrix}}_{\mathbf{W}^\top \in \mathbb{R}^{C \times D}} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_C \end{bmatrix} = \mathbf{W}^\top \mathbf{x} + \mathbf{b}$$

$$\hat{\mathbf{y}} = \text{softmax}(\mathbf{z}) = \begin{bmatrix} e^{z_1} / \sum_{k=1}^C e^{z_k} \\ e^{z_2} / \sum_{k=1}^C e^{z_k} \\ \vdots \\ e^{z_C} / \sum_{k=1}^C e^{z_k} \end{bmatrix}, \quad L(\mathbf{W}, \mathbf{b}) = - \sum_{k=1}^C y_k \log \hat{y}_k$$

$$J(\mathbf{W}, \mathbf{b}) = \frac{1}{N} \sum_{i=1}^N L^{(i)} = - \frac{1}{N} \sum_{i=1}^N \sum_{k=1}^C y_k^{(i)} \log \hat{y}_k^{(i)}$$

1.2 反向传播

$$\frac{\partial J}{\partial w_{nm}} = \sum_{i=1}^N \frac{\partial J}{\partial L^{(i)}} \left(\sum_{j=1}^C \frac{\partial L^{(i)}}{\partial \hat{y}_j^{(i)}} \frac{\partial \hat{y}_j^{(i)}}{\partial z_n^{(i)}} \right) \frac{\partial z_n^{(i)}}{\partial w_{nm}}$$

$$J = \frac{1}{N} \sum_{i=1}^N L^{(i)}, \quad \frac{\partial J}{\partial L^{(i)}} = \frac{1}{N}$$

$$\frac{\partial L}{\partial \hat{y}_j} = - \frac{y_j}{\hat{y}_j}, \quad \mathbf{z} = \mathbf{W}^\top \mathbf{x} + \mathbf{b}, \quad z_n = \sum_{m=1}^D w_{mn} x_m + b_n, \quad \frac{\partial z_n}{\partial w_{nm}} = x_m$$

2 Standard Softmax Gradient Analysis

2.1 雅可比矩阵

$$s = \sum_{k=1}^C e^{z_k}, \quad \hat{y}_j = \frac{e^{z_j}}{s}$$

对任意固定的类别 ℓ ，分别讨论两种情况：

(1) $\ell = j$ ：

$$\frac{\partial \hat{y}_j}{\partial z_j} = \frac{\partial}{\partial z_j} \left(\frac{e^{z_j}}{s} \right) = \frac{e^{z_j}}{s} - \frac{e^{z_j}}{s^2} \frac{\partial s}{\partial z_j} = \hat{y}_j - \hat{y}_j \cdot \frac{e^{z_j}}{s} = \hat{y}_j(1 - \hat{y}_j)$$

(2) $\ell \neq j$ ：

$$\frac{\partial \hat{y}_j}{\partial z_\ell} = \frac{\partial}{\partial z_\ell} \left(\frac{e^{z_j}}{s} \right) = -\frac{e^{z_j}}{s^2} \frac{\partial s}{\partial z_\ell} = -\frac{e^{z_j}}{s^2} e^{z_\ell} = -\hat{y}_j \hat{y}_\ell$$

将上面两种情况统一写为：

$$\frac{\partial \hat{y}_j}{\partial z_\ell} = \hat{y}_j (\mathbb{I}_{j=\ell} - \hat{y}_\ell)$$

于是雅可比矩阵 ($C \times C$) 为：

$$\frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{z}} = \text{diag}(\hat{\mathbf{y}}) - \hat{\mathbf{y}} \hat{\mathbf{y}}^\top$$

2.2 海森矩阵

$$\frac{\partial L}{\partial z_j} = \sum_{k=1}^C \frac{\partial L}{\partial \hat{y}_k} \frac{\partial \hat{y}_k}{\partial z_j} = \sum_{k=1}^C \left(-\frac{y_k}{\hat{y}_k} \right) \hat{y}_k (\mathbb{I}_{k=j} - \hat{y}_j) = \hat{y}_j - \mathbb{I}_{j=y}$$

$$\nabla_z^2 L = \frac{\partial}{\partial \mathbf{z}} (\hat{\mathbf{y}} - \mathbf{y}) = \underbrace{\frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{z}}}_{\text{Jacobian of softmax}} - \underbrace{\frac{\partial \mathbf{y}}{\partial \mathbf{z}}}_{=0} = \text{diag}(\hat{\mathbf{y}}) - \hat{\mathbf{y}} \hat{\mathbf{y}}^\top$$

先看一维函数 $f(z)$, 在 z_0 附近:

$$f(z_0 + \Delta) \approx f(z_0) + f'(z_0) \Delta + \frac{1}{2} f''(z_0) \Delta^2$$

这里 $f'(z_0)$ 是斜率, $f''(z_0)$ 是弯曲程度, 多维函数 $f(\mathbf{z})$, ($\mathbf{z} \in \mathbb{R}^d$) 方向导数定义为沿单位方向 \mathbf{v} 的瞬时变化率:

$$D_{\mathbf{v}}f(\mathbf{z}_0) = \lim_{t \rightarrow 0} \frac{f(\mathbf{z}_0 + t\mathbf{v}) - f(\mathbf{z}_0)}{t}$$

若 f 可微, 梯度 $\nabla f(\mathbf{z}_0) \in \mathbb{R}^d$ 满足

$$\boxed{D_{\mathbf{v}}f(\mathbf{z}_0) = \nabla f(\mathbf{z}_0) \cdot \mathbf{v}} \quad (\text{任意单位向量 } \mathbf{v}, \text{ 梯度在 } \mathbf{v} \text{ 上的投影})$$

特殊地, 沿坐标轴方向 \mathbf{e}_i , $D_{\mathbf{e}_i}f = \partial f / \partial z_i$ 偏导数就是方向导数的特例, 最陡上升方向正是 ∇f 本身, 点积在 \mathbf{v} 与 ∇f 同向时取最大
设二维标量函数

$$f(x, y) = x^2 + xy$$

$$\frac{\partial f}{\partial x} = 2x + y, \quad \frac{\partial f}{\partial y} = x$$

$\partial f / \partial x$ 是沿 $(1, 0)$ 方向的瞬时变化率, $\partial f / \partial y$ 是沿 $(0, 1)$ 方向的瞬时变化率

$$\nabla f(x, y) = (2x + y, x)$$

∇f 指向函数上升最快的方向, $\|\nabla f\|$ 等于最大上升速率, 对任意单位向量 $\mathbf{v} = (v_x, v_y)$, 方向导数定义为

$$D_{\mathbf{v}}f(x, y) = \nabla f(x, y) \cdot \mathbf{v} = (2x + y) v_x + x v_y.$$

偏导是方向导数的特例:

$$D_{(1,0)}f = \frac{\partial f}{\partial x}, \quad D_{(0,1)}f = \frac{\partial f}{\partial y}.$$

取点 $(x_0, y_0) = (2, 1)$:

$$\frac{\partial f}{\partial x}(2, 1) = 2 \cdot 2 + 1 = 5, \quad \frac{\partial f}{\partial y}(2, 1) = 2.$$

$$\nabla f(2, 1) = (5, 2).$$

沿 45° 方向（单位向量 $\mathbf{v} = \frac{1}{\sqrt{2}}(1, 1)$ ）的方向导数：

$$D_{\mathbf{v}}f(2, 1) = \nabla f(2, 1) \cdot \mathbf{v} = (5, 2) \cdot \frac{1}{\sqrt{2}}(1, 1) = \frac{7}{\sqrt{2}} \approx 4.95$$

从 $(2, 1)$ 朝 45° 方向迈一个单位小步， f 大约增加 4.95，沿梯度方向，最陡上升方向的单位向量

$$\hat{\mathbf{g}} = \frac{\nabla f(2, 1)}{\|\nabla f(2, 1)\|} = \frac{(5, 2)}{\sqrt{5^2 + 2^2}} = \frac{(5, 2)}{\sqrt{29}}$$

其方向导数为

$$D_{\hat{\mathbf{g}}}f(2, 1) = \nabla f(2, 1) \cdot \hat{\mathbf{g}} = \|\nabla f(2, 1)\| = \sqrt{29} \approx 5.385$$

说明沿梯度方向的上升速率最大

从 $(2, 1)$ 朝某单位方向 \mathbf{v} 走极小步长 t ：

$$f(2, 1 + t\mathbf{v}) \approx f(2, 1) + t D_{\mathbf{v}}f(2, 1) = f(2, 1) + t \nabla f(2, 1) \cdot \mathbf{v}$$

例如取 $\mathbf{v} = \frac{1}{\sqrt{2}}(1, 1)$ ，有

$$f(2, 1 + t\mathbf{v}) \approx f(2, 1) + t \cdot \frac{7}{\sqrt{2}}$$

若 f 二次可微，海森矩阵（Hessian） $H(\mathbf{z}_0) = \nabla^2 f(\mathbf{z}_0) \in \mathbb{R}^{d \times d}$ 收集了所有二阶偏导。沿方向 \mathbf{v} 的二阶方向导数为，通过二次型 $\mathbf{v}^\top H \mathbf{v}$ 投影到某个方向

$$\boxed{D_{\mathbf{v}}^2 f(\mathbf{z}_0) = \mathbf{v}^\top H(\mathbf{z}_0) \mathbf{v}}$$

从一维函数的角度看高维函数的一二阶方向导函数的证明：

$$g(t) = f(\mathbf{z}_0 + t\mathbf{v})$$

$$g'(t) = \nabla f(\mathbf{z}_0 + t\mathbf{v}) \cdot \mathbf{v} \Rightarrow D_{\mathbf{v}}f(\mathbf{z}_0) = g'(0) = \nabla f(\mathbf{z}_0) \cdot \mathbf{v}$$

$$g''(t) = \frac{d}{dt}(\nabla f(\mathbf{z}_0 + t\mathbf{v}) \cdot \mathbf{v})$$

$$= (\nabla^2 f(\mathbf{z}_0 + t\mathbf{v}) \mathbf{v}) \cdot \mathbf{v} \Rightarrow \boxed{D_{\mathbf{v}}^2 f(\mathbf{z}_0) = \mathbf{v}^\top \nabla^2 f(\mathbf{z}_0) \mathbf{v}}.$$

海森矩阵的直观含义如下：

$$v^\top H v = v^\top Q \Lambda Q^\top v = (Q^\top v)^\top \Lambda (Q^\top v) = \sum_{i=1}^d \lambda_i \alpha_i^2.$$

$$\|v\| = 1 \Rightarrow \sum_i \alpha_i^2 = 1 \Rightarrow v^\top H v = \sum_{i=1}^d \lambda_i \underbrace{\alpha_i^2}_{\text{weights}}.$$

当 v 不是单位向量的时候：

$$R_H(v) := \frac{v^\top H v}{v^\top v} = \frac{\sum_i \lambda_i \alpha_i^2}{\sum_i \alpha_i^2} = \sum_{i=1}^d \lambda_i \underbrace{\frac{\alpha_i^2}{\sum_j \alpha_j^2}}_{\text{normalize weights}}$$

α_i^2 是 v 在主方向 u_i 上的投影长度平方，归一化后各权重之和为 1，因此 $R_H(v)$ 就是把各主曲率 λ_i 按能量占比做加权平均

$$\lambda_{\min} \leq R_H(v) \leq \lambda_{\max}, \quad R_H(u_{\max}) = \lambda_{\max}, \quad R_H(u_{\min}) = \lambda_{\min}$$

加权平均必落在最小最大特征值之间；当 v 与 u_{\min} （或 u_{\max} ）对齐时，全部权重集中在该主方向上，Rayleigh 商取到上下界

设 logits 向量 $z = (z_1, \dots, z_C)^\top$ ，softmax 定义为

$$\hat{y}_k(z) = \frac{e^{z_k}}{\sum_{j=1}^C e^{z_j}}, \quad k = 1, \dots, C$$

对任意常数 $c \in \mathbb{R}$ ，有

$$\hat{y}_k(z + c \mathbf{1}) = \frac{e^{z_k + c}}{\sum_{j=1}^C e^{z_j + c}} = \frac{e^{z_k} e^c}{e^c \sum_{j=1}^C e^{z_j}} = \hat{y}_k(z),$$

即

$$\text{softmax}(z + c \mathbf{1}) = \text{softmax}(z) \quad (\forall c)$$

令交叉熵损失

$$L(z) = - \sum_{k=1}^C y_k \log \hat{y}_k(z)$$

（其中 y 为目标分布，one-hot 时 $\sum_k y_k = 1$ ）。由于 L 只依赖 \hat{y} ，

$$L(z + c \mathbf{1}) = L(z) \quad (\forall c)$$

因此沿着 $\mathbf{1}$ 方向, L 为常数, 故

$$\nabla_z L(\mathbf{z})^\top \mathbf{1} = 0, \quad H(\mathbf{z}) \mathbf{1} = \mathbf{0}, \quad \mathbf{1}^\top H(\mathbf{z}) \mathbf{1} = 0$$

另外, 已知

$$\nabla_z L(\mathbf{z}) = \hat{\mathbf{y}}(\mathbf{z}) - \mathbf{y},$$

于是

$$\nabla_z L(\mathbf{z})^\top \mathbf{1} = \sum_{k=1}^C (\hat{y}_k - y_k) = \underbrace{\sum_{k=1}^C \hat{y}_k}_{=1} - \underbrace{\sum_{k=1}^C y_k}_{=1} = 0$$

对任意方向 $\mathbf{v} \in \mathbb{R}^C$,

$$\mathbf{v}^\top H(\mathbf{z}) \mathbf{v} = \sum_{k=1}^C \hat{y}_k v_k^2 - \left(\sum_{k=1}^C \hat{y}_k v_k \right)^2 = \text{Var}_{k \sim \hat{\mathbf{y}}}(v_k) \geq 0$$

$$\begin{aligned} \text{Var}(X) &= \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2 - 2X \mathbb{E}[X] + \mathbb{E}[X]^2] \\ &= \mathbb{E}[X^2] - 2\mathbb{E}[X]\mathbb{E}[X] + \mathbb{E}[X]^2 = \mathbb{E}[X^2] - (\mathbb{E}[X])^2. \end{aligned}$$

因此 $H(\mathbf{z}) \succeq 0$, 当 \mathbf{v} 为常数向量 (如 $\mathbf{v} = \mathbf{1}$) 时, 方差为 0, 故该方向的二阶导为 0, 说明 L 为凸但非严格凸, 并且由于 $\sum_k \hat{y}_k = 1$, 取 $\mathbf{v} = \mathbf{1}$ 有 $\mathbf{v}^\top H \mathbf{v} = \text{Var}(\text{常数}) = 0$, 说明 H 不是正定而是半正定

2.3 Standard Softmax Analysis

由 $z_j = \mathbf{w}_j^\top \mathbf{x} + b_j$ 得

$$\boxed{\frac{\partial L}{\partial \mathbf{x}} = \sum_{j=1}^C (\hat{y}_j - \mathbb{1}_{j=y}) \mathbf{w}_j = \underbrace{\sum_{j \neq y} \hat{y}_j \mathbf{w}_j}_{\text{负类排斥合力}} - \underbrace{\mathbf{w}_y}_{\text{正类吸引}}}$$

真类系数是负数, 把 \mathbf{w}_y 拉向 \mathbf{x} , 负类系数是正数, 把 \mathbf{w}_j 推离 \mathbf{x} , 每看一个样本, 就把正类中心拉向样本, 把负类中心推离样本

$$\boxed{\frac{\partial L}{\partial \mathbf{w}_j} = (\hat{y}_j - \mathbb{1}_{j=y}) \mathbf{x}, \quad \frac{\partial L}{\partial b_j} = \hat{y}_j - \mathbb{1}_{j=y}}$$

范数上界, 把 \mathbf{W} 看成一个线性放大器, 这个最大放大倍数就叫谱范数也等于最大奇异值:

$$\hat{\mathbf{y}} \approx \mathbf{e}_k \ (k \neq y) \Rightarrow \hat{\mathbf{y}} - \mathbf{e}_y \approx \mathbf{e}_k - \mathbf{e}_y, \quad \|\mathbf{e}_k - \mathbf{e}_y\|_2 = \sqrt{1^2 + (-1)^2} = \sqrt{2}.$$

$$\left\| \frac{\partial L}{\partial \mathbf{x}} \right\| = \|\mathbf{W}(\hat{\mathbf{y}} - \mathbf{e}_y)\| \leq \|\mathbf{W}\|_2 \|\hat{\mathbf{y}} - \mathbf{e}_y\|_2 \leq \sqrt{2} \|\mathbf{W}\|_2, \quad \left\| \frac{\partial L}{\partial \mathbf{w}_j} \right\| \leq \|\mathbf{x}\|.$$

在标准形式 $z_j = \mathbf{w}_j^\top \mathbf{x} + b_j$ 下, 若缺少归一化/正则, 训练倾向通过增大 $\|\mathbf{W}\|$ (有时也增大 $\|\mathbf{x}\|$) 来放大 $z_y - z_{k \neq y}$, 而非优化角度; 这会提升 $\|\mathbf{W}\|_2$ 最大放大倍数, 谱范数, 使

$$\left\| \frac{\partial L}{\partial \mathbf{x}} \right\| = \|\mathbf{W}(\hat{\mathbf{y}} - \mathbf{e}_y)\| \leq \|\mathbf{W}\|_2 \|\hat{\mathbf{y}} - \mathbf{e}_y\| \leq \sqrt{2} \|\mathbf{W}\|_2$$

的上界变大, 链式反传更易爆/抖, 若未做 log-sum-exp 等数值稳定, 过大 logits 也会造成前向溢出, 结果是模型更依赖强度 (模长) 而非方向 (角度/判别边界), 判别边界变差, 泛化与鲁棒性下降

$$\log \sum_i e^{z_i} = m + \log \sum_i e^{z_i - m} \quad (\text{任取 } m).$$

取 $m = \max_i z_i$ 时, $z_i - m \leq 0$ (所有指数 ≤ 1), 数值稳定:

$$\log \sum_i e^{z_i} = \max_i z_i + \log \sum_i e^{z_i - \max_j z_j}$$

3 Normalized Softmax Gradient Analysis

定义

$$\hat{\mathbf{x}} = \frac{\mathbf{x}}{\|\mathbf{x}\|}, \quad \hat{\mathbf{w}}_j = \frac{\mathbf{w}_j}{\|\mathbf{w}_j\|}, \quad z_j = s \hat{\mathbf{w}}_j^\top \hat{\mathbf{x}} = s \cos \theta_j, \quad z_j \in [-s, s]$$

梯度对 \mathbf{z} 的形式不变, 但通过

$$\frac{\partial \hat{\mathbf{x}}}{\partial \mathbf{x}} = \frac{1}{\|\mathbf{x}\|} (I - \hat{\mathbf{x}} \hat{\mathbf{x}}^\top), \quad \frac{\partial \hat{\mathbf{w}}_j}{\partial \mathbf{w}_j} = \frac{1}{\|\mathbf{w}_j\|} (I - \hat{\mathbf{w}}_j \hat{\mathbf{w}}_j^\top)$$

$$\|\mathbf{x}\| = (x^\top x)^{1/2}. \quad \text{因此} \quad d\|\mathbf{x}\| = \frac{1}{2} (x^\top x)^{-1/2} d(x^\top x)$$

$$\text{而 } d(x^\top x) = d\left(\sum_{i=1}^d x_i^2\right) = \sum_{i=1}^d 2x_i dx_i = 2 \sum_{i=1}^d x_i dx_i = \boxed{2x^\top dx}$$

代回可得

$$\begin{aligned} d\|x\| &= \frac{1}{2} \cdot \frac{1}{\|x\|} \cdot 2x^\top dx = \frac{x^\top dx}{\|x\|} = \boxed{\hat{x}^\top dx} \\ d\hat{x} &= d\left(\frac{x}{\|x\|}\right) = \frac{dx}{\|x\|} - x \frac{d\|x\|}{\|x\|^2} = \frac{1}{\|x\|} \left(dx - \hat{x} \hat{x}^\top dx\right) \\ &= \boxed{\frac{1}{\|x\|} (I - \hat{x} \hat{x}^\top) dx} \end{aligned}$$

切空间的定义

$$T_{\hat{x}} S^{d-1} = \{v \in \mathbb{R}^d : \hat{x}^\top v = 0\}$$

球面上的某点，沿球表的方向就是切空间，指向地心/离心的是径向

$$v = v_{\parallel} + v_{\perp}, \quad v_{\parallel} = (\hat{x}^\top v) \hat{x} \quad (\text{沿半径}), \quad v_{\perp} = v - (\hat{x}^\top v) \hat{x} \quad (\text{贴着球面})$$

把 v 丢进

$$P := I - \hat{x} \hat{x}^\top$$

得到

$$Pv = (I - \hat{x} \hat{x}^\top)v = v - (\hat{x}^\top v) \hat{x} = v_{\perp}$$

P 会去掉径向分量 v_{\parallel} ，仅保留切向分量 v_{\perp}

$$P\hat{x} = (I - \hat{x} \hat{x}^\top)\hat{x} = \hat{x} - (\hat{x}^\top \hat{x})\hat{x} = 0$$

该式子表示半径方向被完全杀掉，与 $T_{\hat{x}} S^{d-1} = \hat{x}^\perp$ 一致

$$d\hat{x} = \frac{1}{\|x\|} (I - \hat{x} \hat{x}^\top) dx = \frac{1}{\|x\|} P dx$$

上面式子告诉我们，对于 Normalized Softmax，方向的变化 $d\hat{x}$ 来自切分量
径向分量控制长度的一阶变化（这里不存在，被分解掉了）；切向分量控制
方向的一阶变化；且方向变化幅度与 $1/\|x\|$ 成正比 $\|x\|$ 越大，方向越不敏
感

$$\frac{\partial L}{\partial x} = \sum_j \frac{\partial L}{\partial z_j} \frac{\partial z_j}{\partial x} = \sum_j (p_j - \mathbb{1}_{j=y}) s \frac{1}{\|x\|} (I - \hat{x} \hat{x}^\top) \hat{w}_j$$

$$= \frac{s}{\|x\|} (I - \hat{x}\hat{x}^\top) \sum_j (p_j - \mathbb{1}_{j=y}) \hat{w}_j$$

$$\frac{\partial L}{\partial w_j} = (p_j - \mathbb{1}_{j=y}) s \frac{1}{\|w_j\|} (I - \hat{w}_j \hat{w}_j^\top) \hat{x}$$

两个投影算子把更新限制在各自的切空间内, 只改角度不鼓励改模长, logits 有界于 $[-s, s]$ 提升数值稳定, 系数 $s/\|\cdot\|$ 决定更新尺度, $\|x\|$ 或 $\|w_j\|$ 越大, 单位变化对方向的影响越小, 也就更加稳定

4 PGA

特征与类别权重均作单位化:

$$\mathbf{x} \in \mathbb{S}^{d-1}, \quad \hat{\mathbf{w}}_j = \frac{\mathbf{w}_j}{\|\mathbf{w}_j\|} \in \mathbb{S}^{d-1}, \quad z_j = s \hat{\mathbf{w}}_j^\top \mathbf{x}, \quad p_j = \text{softmax}_j(z).$$

单位球面在 \mathbf{x} 处的切空间与正交投影:

$$T_{\mathbf{x}}\mathbb{S}^{d-1} = \{v \in \mathbb{R}^d : \mathbf{x}^\top v = 0\}, \quad \Pi_{\mathbf{x}} := I - \mathbf{x}\mathbf{x}^\top.$$

注意对任意由 logits 组合得到的损失 $L(\{z_j\})$, 链式法则给出:

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{x}} &= \sum_j \frac{\partial L}{\partial z_j} \frac{\partial z_j}{\partial \mathbf{x}} = \sum_j (p_j - \mathbb{1}_{j=y}) s \frac{1}{\|\mathbf{x}\|} (I - \hat{\mathbf{x}}\hat{\mathbf{x}}^\top) \hat{\mathbf{w}}_j \\ &= \frac{s}{\|\mathbf{x}\|} \Pi_{\mathbf{x}} \sum_j (p_j - \mathbb{1}_{j=y}) \hat{\mathbf{w}}_j \end{aligned}$$

4.1 ArcFace 的切向合力

$$\hat{\mathbf{x}} = \frac{\mathbf{x}}{\|\mathbf{x}\|}, \quad \hat{\mathbf{w}}_j = \frac{\mathbf{w}_j}{\|\mathbf{w}_j\|}, \quad z_y = s \cos(\theta_y + m), \quad z_j = s \cos \theta_j \quad (j \neq y)$$

$$L = -\log \frac{e^{z_y}}{\sum_k e^{z_k}}, \quad \frac{\partial L}{\partial z_j} = p_j - \mathbb{1}_{j=y}, \quad p_j = \frac{e^{z_j}}{\sum_k e^{z_k}}$$

$$\hat{\mathbf{w}}_y = (\hat{\mathbf{w}}_y^\top \hat{\mathbf{x}}) \hat{\mathbf{x}} + (\hat{\mathbf{w}}_y - (\hat{\mathbf{w}}_y^\top \hat{\mathbf{x}}) \hat{\mathbf{x}}) = \cos \theta_y \hat{\mathbf{x}} + \Pi_{\mathbf{x}} \hat{\mathbf{w}}_y$$

$$\|\Pi_{\mathbf{x}} \hat{\mathbf{w}}_y\|^2 = \|\hat{\mathbf{w}}_y - \cos \theta_y \hat{\mathbf{x}}\|^2 = 1 + \cos^2 \theta_y - 2 \cos \theta_y (\hat{\mathbf{w}}_y^\top \hat{\mathbf{x}})$$

$$\begin{aligned}
&= 1 - \cos^2 \theta_y = \sin^2 \theta_y \\
\Rightarrow \quad &\|\Pi_x \hat{\mathbf{w}}_y\| = \sin \theta_y, \quad \Pi_x \hat{\mathbf{w}}_y = \sin \theta_y \mathbf{t}_y, \quad \mathbf{t}_y := \frac{\Pi_x \hat{\mathbf{w}}_y}{\|\Pi_x \hat{\mathbf{w}}_y\|} \\
&\frac{\partial \cos \theta_j}{\partial \mathbf{x}} = \frac{1}{\|\mathbf{x}\|} \Pi_x \hat{\mathbf{w}}_j, \quad \frac{\partial \theta_y}{\partial \mathbf{x}} = -\frac{1}{\|\mathbf{x}\|} \mathbf{t}_y \\
&\cos \theta_j = \hat{\mathbf{w}}_j^\top \hat{\mathbf{x}}, \quad \frac{\partial \hat{\mathbf{x}}}{\partial \mathbf{x}} = \frac{1}{\|\mathbf{x}\|} (I - \hat{\mathbf{x}} \hat{\mathbf{x}}^\top) = \frac{1}{\|\mathbf{x}\|} \Pi_x \\
&\Rightarrow \frac{\partial \cos \theta_j}{\partial \mathbf{x}} = \left(\frac{\partial \hat{\mathbf{x}}}{\partial \mathbf{x}} \right)^\top \hat{\mathbf{w}}_j = \frac{1}{\|\mathbf{x}\|} \Pi_x \hat{\mathbf{w}}_j \\
&\theta_y = \arccos(\hat{\mathbf{w}}_y^\top \hat{\mathbf{x}}) \Rightarrow \frac{\partial \theta_y}{\partial \mathbf{x}} = \frac{d \arccos(u)}{du} \Big|_{u=\cos \theta_y} \cdot \frac{\partial \cos \theta_y}{\partial \mathbf{x}} \\
&= -\frac{1}{\sin \theta_y} \cdot \frac{1}{\|\mathbf{x}\|} \Pi_x \hat{\mathbf{w}}_y
\end{aligned}$$

又因为 $\sin \theta_y = \sqrt{1 - (\hat{\mathbf{w}}_y^\top \hat{\mathbf{x}})^2} = \|\Pi_x \hat{\mathbf{w}}_y\|$, 故:

$$\begin{aligned}
\frac{\partial \theta_y}{\partial \mathbf{x}} &= -\frac{1}{\|\Pi_x \hat{\mathbf{w}}_y\|} \cdot \frac{1}{\|\mathbf{x}\|} \Pi_x \hat{\mathbf{w}}_y = -\frac{1}{\|\mathbf{x}\|} \frac{\Pi_x \hat{\mathbf{w}}_y}{\|\Pi_x \hat{\mathbf{w}}_y\|} = -\frac{1}{\|\mathbf{x}\|} \mathbf{t}_y \\
\frac{\partial z_y}{\partial \mathbf{x}} &= s \frac{\partial \cos(\theta_y + m)}{\partial \mathbf{x}} = s(-\sin(\theta_y + m)) \frac{\partial \theta_y}{\partial \mathbf{x}} = \frac{s}{\|\mathbf{x}\|} \sin(\theta_y + m) \mathbf{t}_y
\end{aligned}$$

将 \mathbf{t}_y 换回 $\Pi_x \hat{\mathbf{w}}_y$ 得到 “共线缩放”:

$$\boxed{\frac{\partial z_y}{\partial \mathbf{x}} = \frac{s}{\|\mathbf{x}\|} \underbrace{\frac{\sin(\theta_y + m)}{\sin \theta_y}}_{:=\alpha(\theta_y, m)} \Pi_x \hat{\mathbf{w}}_y}$$

可见方向与 $\Pi_x \hat{\mathbf{w}}_y$ 共线, 仅被 $\alpha(\theta_y, m)$ 缩放; 当 $m = 0$ 时 $\alpha = 1$

$$\begin{aligned}
\frac{\partial z_j}{\partial \mathbf{x}} &= s \frac{\partial \cos \theta_j}{\partial \mathbf{x}} = \frac{s}{\|\mathbf{x}\|} \Pi_x \hat{\mathbf{w}}_j, \quad j \neq y \\
\frac{\partial L}{\partial \mathbf{x}} &= \sum_j \frac{\partial L}{\partial z_j} \frac{\partial z_j}{\partial \mathbf{x}} \\
&= (p_y - 1) \frac{s}{\|\mathbf{x}\|} \alpha(\theta_y, m) \Pi_x \hat{\mathbf{w}}_y + \sum_{j \neq y} p_j \frac{s}{\|\mathbf{x}\|} \Pi_x \hat{\mathbf{w}}_j \\
&= \frac{s}{\|\mathbf{x}\|} \Pi_x \left(\sum_{j \neq y} p_j \hat{\mathbf{w}}_j - \alpha(\theta_y, m) \hat{\mathbf{w}}_y \right)
\end{aligned}$$

梯度完全落在 \mathbf{x} 的切空间内 (被 Π_x 投影), “真类吸引” 与 “切向方向” 严格共线, margin 仅通过 $\alpha(\theta_y, m)$ 调整同一方向的力量

4.2 PGA K/Z 对齐的切向合力

对同类且可靠的邻居（经掩码与 EMA 平滑）集合 $\mathcal{N}_y(\mathbf{x})$ ，K/Z 对齐的一个统一写法是最大化方向一致性（或最小化某相似度损失）：

$$L_{\text{pga}} = \sum_{a \in \mathcal{N}_y(\mathbf{x})} \beta_a \phi(\langle \mathbf{x}, \mathbf{x}_a^{\text{tgt}} \rangle), \quad \beta_a \geq 0, \quad \phi'(\cdot) \leq 0$$

其中 $\mathbf{x}_a^{\text{tgt}}$ 来自下层/EMA 的目标方向，于是

$$\frac{\partial L_{\text{pga}}}{\partial \mathbf{x}} = \sum_a \beta_a \phi'(\langle \mathbf{x}, \mathbf{x}_a^{\text{tgt}} \rangle) \mathbf{x}_a^{\text{tgt}} = - \sum_a \alpha_a \mathbf{x}_a^{\text{tgt}} \quad (\alpha_a := \beta_a |\phi'| \geq 0)$$

投影到切空间得

$$-\mathbf{g}_{\text{pga}} := -\Pi_{\mathbf{x}} \frac{\partial L_{\text{pga}}}{\partial \mathbf{x}} = \sum_{a \in \mathcal{N}_y(\mathbf{x})} \alpha_a \Pi_{\mathbf{x}} \mathbf{x}_a^{\text{tgt}} \quad (1)$$

即 PGA 的下降方向是同类目标邻居在切空间的加权均值

4.3 切空间内积判断对抗性

ArcFace 与 PGA 的是否“冲突”取决于它们在切空间的内积：

$$\langle -\mathbf{g}_{\text{arc}}, -\mathbf{g}_{\text{pga}} \rangle = \left\langle \Pi_{\mathbf{x}} \mathbf{u}, \sum_a \alpha_a \Pi_{\mathbf{x}} \mathbf{x}_a^{\text{tgt}} \right\rangle \quad (2)$$

若该内积 ≥ 0 ，两者同向/弱同向，不对抗；若 < 0 ，则局部相悖，令

$$\theta_a := \angle(\Pi_{\mathbf{x}} \mathbf{u}, \Pi_{\mathbf{x}} \mathbf{x}_a^{\text{tgt}}), \quad c_a := \cos \theta_a$$

由柯西不等式，

$$\begin{aligned} \left\langle \Pi_{\mathbf{x}} \mathbf{u}, \sum_a \alpha_a \Pi_{\mathbf{x}} \mathbf{x}_a^{\text{tgt}} \right\rangle &= \|\Pi_{\mathbf{x}} \mathbf{u}\| \sum_a \alpha_a \|\Pi_{\mathbf{x}} \mathbf{x}_a^{\text{tgt}}\| c_a \\ &\geq \|\Pi_{\mathbf{x}} \mathbf{u}\| \left(\sum_a \alpha_a \|\Pi_{\mathbf{x}} \mathbf{x}_a^{\text{tgt}}\| \right) \cdot \min_a c_a \end{aligned}$$

因此，只要

$$\min_{a \in \mathcal{N}_y(\mathbf{x})} \angle(\Pi_{\mathbf{x}} \mathbf{u}, \Pi_{\mathbf{x}} \mathbf{x}_a^{\text{tgt}}) \leq 90^\circ \quad (3)$$

就有 $\langle -\mathbf{g}_{\text{arc}}, -\mathbf{g}_{\text{pga}} \rangle \geq 0$ ，充分保证两者不对抗，这与实践中同类掩码加上可靠边相吻合，稳定邻居与真类方向通常形成锐角