Gradient Analysis of Standard Softmax and Normalized Softmax

fangqing

2025 年 11 月 9 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 是否对抗的充要判据（切空间内积） 12
4.4 强保证：零对抗的半空间投影（可选） 13

1 Preface

1.1 前向传播

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

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

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

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

1.2 反向传播

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

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

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

2 Standard Softmax Gradient Analysis

2.1 雅可比矩阵

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

对任意固定的类别 ℓ，分别讨论两种情况：

(1) $\ell = j$：

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

(2) $\ell \neq j$：

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

将上面两种情况统一写为：

$$\boxed{\frac{\partial \hat{y}_j}{\partial z_\ell} = \hat{y}_j\big(\mathbb{1}_{j=\ell} - \hat{y}_\ell\big)}$$

于是雅可比矩阵（$C \times C$）为：

$$\boxed{\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z}} = \mathrm{diag}(\hat{\boldsymbol{y}}) - \hat{\boldsymbol{y}}\,\hat{\boldsymbol{y}}^{\top}}$$

2.2 海森矩阵

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

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

先看一维函数 $f(z)$，在 z_0 附近：

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

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

$$D_{\boldsymbol{v}} f(\boldsymbol{z}_0) \;=\; \lim_{t\to 0} \frac{f(\boldsymbol{z}_0 + t\boldsymbol{v}) - f(\boldsymbol{z}_0)}{t}$$

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

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

特殊地，沿坐标轴方向 $\boldsymbol{e}_i$，$D_{\boldsymbol{e}_i} f = \partial f/\partial z_i$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 $\boldsymbol{v}$ 与 ∇f 同向时取最大
设二维标量函数

$$f(x,y) = x^2 + xy$$
$$\frac{\partial f}{\partial x} = 2x + y, \qquad \frac{\partial f}{\partial y} = x$$

$\partial f/\partial x$ 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

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

∇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 $\boldsymbol{v} = (v_x, v_y)$，方向导数定义为

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

偏导是方向导数的特例：

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

取点 $(x_0, y_0) = (2,1)$：

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

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

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

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

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

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

其方向导数为

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

说明沿梯度方向的上升速率最大
从 $(2,1)$ 朝某单位方向 $\boldsymbol{v}$ 走极小步长 t：

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

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

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

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

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

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：

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

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

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

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

海森矩阵的直观含义如下：

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

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

当 $\boldsymbol{v}$ 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 $\boldsymbol{v}$ 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

$$\lambda_{\min} \;\le\; R_H(v) \;\le\; \lambda_{\max}, \qquad R_H(u_{\max}) = \lambda_{\max}, \;\; R_H(u_{\min}) = \lambda_{\min}$$

加权平均必落在最小最大特征值之间；当 $\boldsymbol{v}$ 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 $\mathbf{z} = (z_1, \ldots, z_C)^\top$，softmax 定义为

$$\hat{y}_k(\mathbf{z}) \;=\; \frac{e^{z_k}}{\sum_{j=1}^{C} e^{z_j}}, \qquad k = 1, \ldots, C$$

对任意常数 $c \in \mathbb{R}$，有

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

即

$$\mathrm{softmax}(\mathbf{z} + c\,\mathbf{1}) \;=\; \mathrm{softmax}(\mathbf{z}) \quad (\forall c)$$

令交叉熵损失

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

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

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

因此沿着 $\mathbf{1}$ 方向，L 为常数，故

$$\nabla_{\mathbf{z}} L(\mathbf{z})^{\top}\mathbf{1} \;=\; 0, \qquad H(\mathbf{z})\,\mathbf{1} \;=\; \mathbf{0}, \qquad \mathbf{1}^{\top} H(\mathbf{z})\,\mathbf{1} \;=\; 0$$

另外，已知

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

于是

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

对任意方向 $\mathbf{v} \in \mathbb{R}^C$，

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

$$\mathrm{Var}(X) = \mathbb{E}\big[(X - \mathbb{E}[X])^2\big] = \mathbb{E}\big[X^2 - 2X\,\mathbb{E}[X] + \mathbb{E}[X]^2\big]$$

$$= \mathbb{E}[X^2] - 2\,\mathbb{E}[X]\mathbb{E}[X] + \mathbb{E}[X]^2 = \mathbb{E}[X^2] - \big(\mathbb{E}[X]\big)^2.$$

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 $\boldsymbol{v} = \mathbf{1}$ 有 $\boldsymbol{v}^{\top} H \boldsymbol{v} = \mathrm{Var}(\text{常数}) = 0$，说明 H 不是正定而是半正定

2.3 Standard Softmax Analysis

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

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

真类系数是负数，把 $\boldsymbol{w}_y$ 拉向 $\boldsymbol{x}$，负类系数是正数，把 $\boldsymbol{w}_j$ 推离 $\boldsymbol{x}$，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

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

范数上界,把 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：

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

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

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 $\|W\|$（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 $\|W\|_2$ 最大放大倍数，谱范数，使

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

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

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

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

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

3 Normalized Softmax Gradient Analysis

定义

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

梯度对 $\boldsymbol{z}$ 的形式不变，但通过

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

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

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

代回可得

$$d\|x\| = \frac{1}{2}\cdot\frac{1}{\|x\|}\cdot 2\,x^\top dx = \frac{x^\top dx}{\|x\|} = \boxed{\hat{x}^\top dx}$$

$$d\hat{x} = d\Big(\frac{x}{\|x\|}\Big) = \frac{dx}{\|x\|} - x\,\frac{d\|x\|}{\|x\|^2} = \frac{1}{\|x\|}\Big(dx - \hat{x}\,\hat{x}^\top dx\Big)$$

$$= \boxed{\frac{1}{\|x\|}\big(I - \hat{x}\hat{x}^\top\big)\,dx}$$

切空间的定义

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

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

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

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 $v_{\|}$，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

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

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

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

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

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

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：

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

单位球面在 x 处的切空间与正交投影：

$$T_x\mathbb{S}^{d-1} = \{v \in \mathbb{R}^d : x^\top v = 0\}, \qquad \Pi_x := I - xx^\top.$$

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

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

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

4.1 ArcFace 的切向合力

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

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

$$\hat{w}_y = (\hat{w}_y^\top \hat{x})\,\hat{x} + \big(\hat{w}_y - (\hat{w}_y^\top \hat{x})\,\hat{x}\big) = \cos\theta_y\,\hat{x} + \Pi_x\,\hat{w}_y$$

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

$$\Rightarrow \quad \|\Pi_x \hat{w}_y\| = \sin\theta_y, \qquad \Pi_x \hat{w}_y = \sin\theta_y \, t_y, \quad t_y := \frac{\Pi_x \hat{w}_y}{\|\Pi_x \hat{w}_y\|}$$

$$\frac{\partial \cos\theta_j}{\partial x} = \frac{1}{\|x\|} \Pi_x \, \hat{w}_j, \qquad \frac{\partial \theta_y}{\partial x} = -\frac{1}{\|x\|} \, t_y$$

$$\cos\theta_j = \hat{w}_j^\top \hat{x}, \;\; \frac{\partial \hat{x}}{\partial x} = \frac{1}{\|x\|}\big(I - \hat{x}\hat{x}^\top\big) = \frac{1}{\|x\|}\Pi_x$$

$$\Rightarrow \frac{\partial \cos\theta_j}{\partial x} = \Big(\frac{\partial \hat{x}}{\partial x}\Big)^\top \hat{w}_j = \frac{1}{\|x\|}\Pi_x \hat{w}_j$$

$$\theta_y = \arccos\big(\hat{w}_y^\top \hat{x}\big) \Rightarrow \frac{\partial \theta_y}{\partial x} = \frac{d \, \arccos(u)}{du}\bigg|_{u=\cos\theta_y} \cdot \frac{\partial \cos\theta_y}{\partial x}$$

$$= -\frac{1}{\sin\theta_y} \cdot \frac{1}{\|x\|}\Pi_x \hat{w}_y$$

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

$$\frac{\partial \theta_y}{\partial x} = -\frac{1}{\|\Pi_x \hat{w}_y\|} \cdot \frac{1}{\|x\|}\Pi_x \hat{w}_y = -\frac{1}{\|x\|} \frac{\Pi_x \hat{w}_y}{\|\Pi_x \hat{w}_y\|} = -\frac{1}{\|x\|} \, t_y$$

$$\frac{\partial z_y}{\partial x} = s\frac{\partial \cos(\theta_y + m)}{\partial x} = s\big(-\sin(\theta_y + m)\big)\frac{\partial \theta_y}{\partial x} = \frac{s}{\|x\|} \, \sin(\theta_y + m) \, t_y$$

将 t_y 换回 $\Pi_x \hat{w}_y$ 得到“共线缩放”：

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

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

$$\frac{\partial z_j}{\partial x} = s\frac{\partial \cos\theta_j}{\partial x} = \frac{s}{\|x\|} \, \Pi_x \, \hat{w}_j, \qquad j \neq y$$

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

梯度完全落在 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 $\alpha(\theta_y, m)$ 调整同一方向的力度

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 $\mathcal{N}_y(x)$，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）：

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

其中 x_a^{tgt} 来自下层/EMA 的目标方向，于是

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

投影到切空间得

$$-g_{\text{pga}} \ := \ -\Pi_x \frac{\partial L_{\text{pga}}}{\partial x} \ = \ \sum_{a \in \mathcal{N}_y(x)} \alpha_a \, \Pi_x x_a^{\text{tgt}} \tag{1}$$

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 是否对抗的充要判据（切空间内积）

ArcFace 与 PGA 的是否“冲突”取决于它们在切空间的内积：

$$\langle -g_{\text{arc}},\ -g_{\text{pga}} \rangle = \Big\langle \Pi_x u,\ \sum_a \alpha_a \, \Pi_x x_a^{\text{tgt}} \Big\rangle \tag{2}$$

若该内积 ≥ 0，两者同向/弱同向，不对抗；若 < 0，则局部相悖，令

$$\theta_a := \angle\big(\Pi_x u,\ \Pi_x x_a^{\text{tgt}}\big), \qquad c_a := \cos\theta_a$$

由柯西不等式，

$$\Big\langle \Pi_x u,\ \sum_a \alpha_a \, \Pi_x x_a^{\text{tgt}} \Big\rangle = \|\Pi_x u\| \sum_a \alpha_a \, \|\Pi_x x_a^{\text{tgt}}\| \, c_a$$

$$\geq \ \|\Pi_x u\| \Big(\sum_a \alpha_a \, \|\Pi_x x_a^{\text{tgt}}\| \Big) \cdot \min_a c_a$$

因此，只要

$$\boxed{\min_{a \in \mathcal{N}_y(x)} \angle\big(\Pi_x u,\ \Pi_x x_a^{\text{tgt}}\big) \ \leq \ 90^\circ} \tag{3}$$

就有 $\langle -g_{\text{arc}}, -g_{\text{pga}} \rangle \geq 0$，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