训练稳定性与归一化

机器学习 · 深度学习基础

学习目标

- 解释前向信号与反向梯度的尺度问题怎样形成，区分传播与更新步长
- 区分初始化的对称性问题与尺度问题，解释 Xavier / He 的处理思路
- 解释输入标准化与BatchNorm的思路，跟随计算区分统计、学习参数与训练 / 推理模式
- 说明梯度裁剪改变什么，以及它不能替代哪些检查

方法作用位置：输入标准化→输入特征；Xavier / He→参数起点；BatchNorm→内部前向表示；梯度裁剪→更新前梯度。

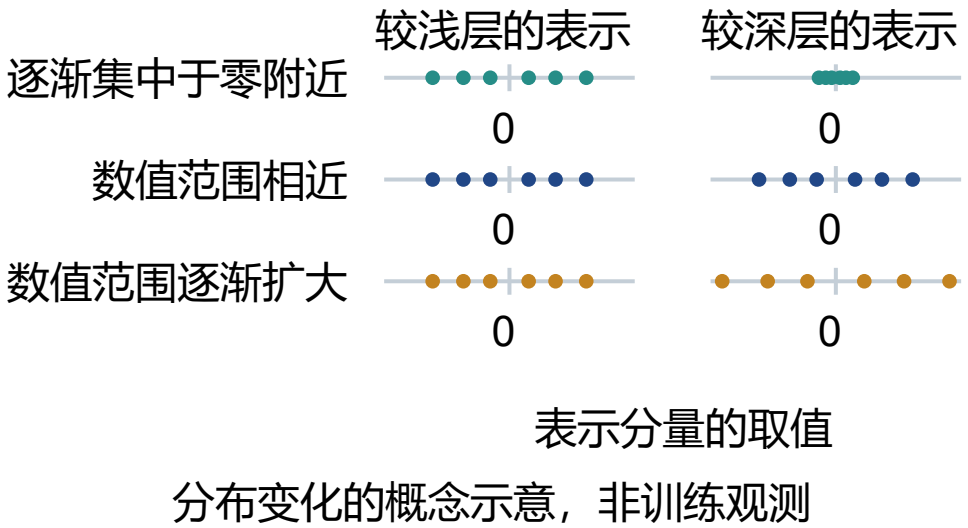
1 传播尺度与参数起点

理解传播尺度的累积与参数初始化的作用。

传播尺度的逐层累积

第4讲讨论如何更新参数。本讲先看更新所需的表示与梯度：逐层计算可能使其尺度不断缩小或放大。

同一批输入在较浅层与较深层的表示，可能出现以下变化。



分布变化的概念示意，非训练观测

用**二阶矩** $q_\ell = \mathbb{E}[h_{\ell,j}^2]$ 衡量第 ℓ 层分量的平方平均尺度。以 r_ℓ 近似描述一层带来的倍率：

$$q_\ell \approx r_\ell q_{\ell-1}$$

若各层倍率近似相同：

$$q_L \approx r^L q_0$$

- $0 < r < 1$：逐层衰减
- $r \approx 1$：尺度相近
- $r > 1$：逐层放大

初始化控制累积；激活与反向另行分析。

局部导数与梯度尺度

反向传播沿计算依赖传递敏感度，每经过一段就乘上该段的局部导数；连乘可能使尺度衰减或放大。



前向：逐层复合；反向：逐段乘局部导数；每段因子$f'_\ell = w_\ell \phi'_\ell$

对标量链 $h_\ell = f_\ell(h_{\ell-1})$，链式法则给出：

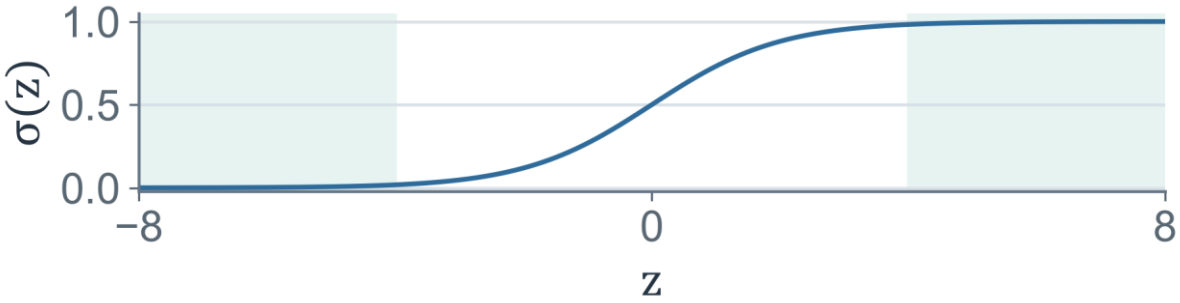
$$\frac{\partial h_L}{\partial x} = \prod_{\ell=1}^{L} f'_\ell\,(h_{\ell-1}), f'_\ell = w_\ell \phi'_\ell(w_\ell h_{\ell-1} + b_\ell)$$

- 十段局部导数均为0.5时，乘积约0.001；均为1.5时，乘积约57.7
- 跨层梯度显著衰减称为**梯度消失**，过度放大称为**梯度爆炸**
- 此式是复合映射对输入的导数，不等同所有参数的损失梯度；实际参数位移还取决于学习率和优化器

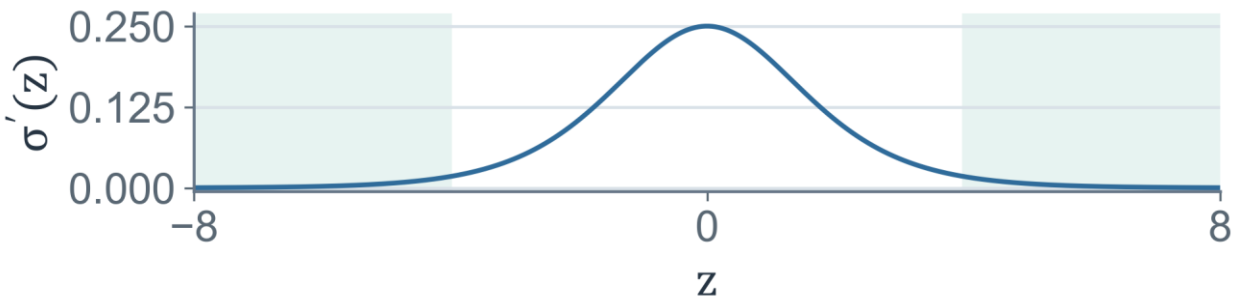
激活饱和与梯度传播

Sigmoid在两端逐渐变平：输入变化较大，输出仍变化很小，称为**激活饱和**；对应的局部导数接近零。

Sigmoid：两端变平



导数：两端接近零



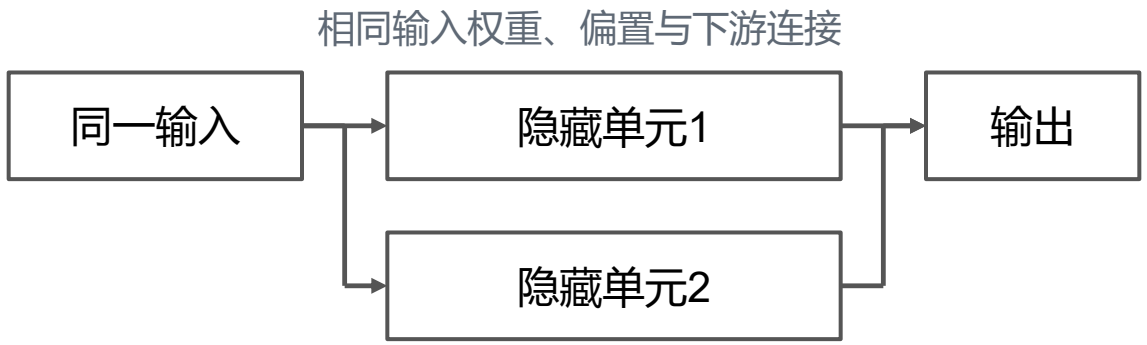
ReLU：$z < 0$导数0；$z > 0$导数1

$$\sigma(z) = \frac{1}{1+e^{-z}}, \sigma'(z) = \sigma(z)[1-\sigma(z)] \leq \frac{1}{4}$$

- 饱和区的小导数会缩小经过该激活的梯度贡献；前向数值很大，不代表反向梯度也大
- ReLU正区间不通过激活导数额外缩小梯度，负区间则阻断该局部路径
- 实际反传还含权重与路径贡献；ReLU不能保证所有单元有效或整个网络梯度尺度稳定

隐藏单元的对称性

参数初始化是在训练前指定权重、偏置等参数的初值。两个隐藏单元若连同下游联系完全对称，就可能一直执行相同计算。

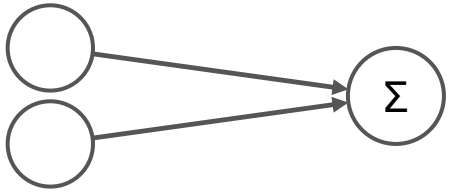


- 维持对称的条件：输入权重、偏置、下游连接均相同，且采用相同确定性更新
- 随机权重打破这种完整对称的约束，使单元不再必然学到同一表示
- 不要求所有参数都随机：不同输入权重可以配合零偏置；打破对称性与控制随机尺度是两件事

初始权重的尺度

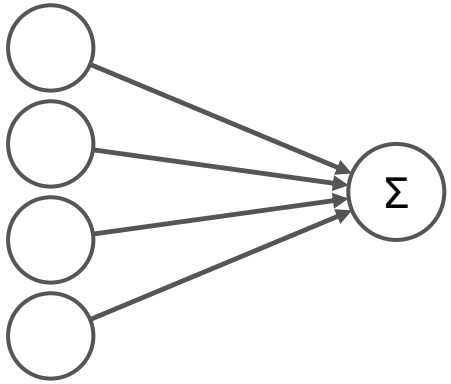
随机差异打破对称性，但同样的权重统计尺度经过更多连接累加，可能产生更大的输出尺度。

较少输入连接



两侧使用相同的
权重统计尺度

较多输入连接



独立零均值贡献累加：
连接数参与总尺度

对 $s = \sum_{j=1}^{n} w_j x_j$，权重零均值、同方差、相互独立且与输入独立，输入二阶尺度 $\mathbb{E}[x_j^2] = 1$ 时：

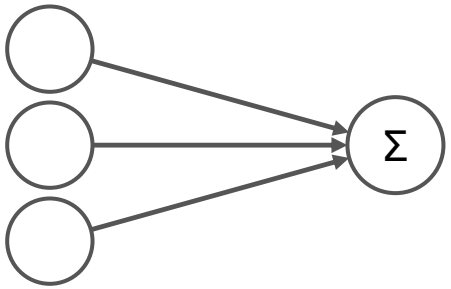
$$Var(s) \approx nVar(w)$$

- 因此权重尺度需要随输入连接数 n 调整；仅指定“小随机数”尚未确定合适尺度
- 非线性还会改变输出二阶矩，初始化也要考虑激活函数

Xavier初始化

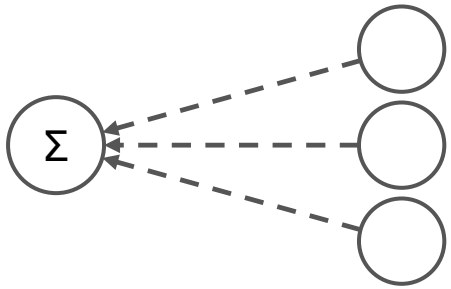
只控制前向累加，还未兼顾反向梯度贡献；两个方向涉及不同的连接数。

前向：一个输出
接收n_{in}项



希望$n_{in}Var(w) \approx 1$

反向：由后层向前层
汇总n_{out}项



希望$n_{out}Var(w) \approx 1$

前层：一个输入←后层：梯度贡献；
连接数不同时：同一方差难以同时满足两侧

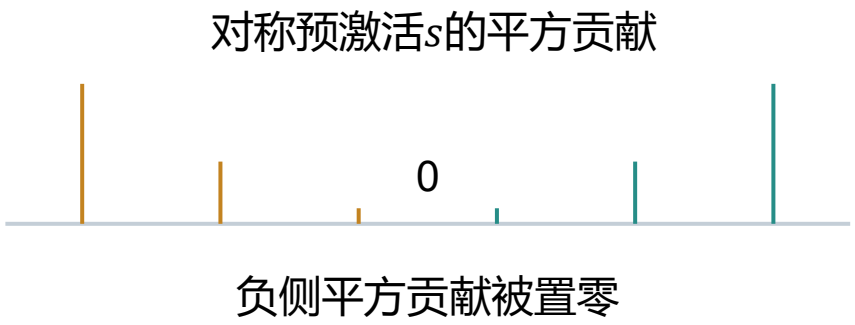
Xavier初始化采用前后向二阶尺度的折中：让两个近似倍率的平均为1，得到

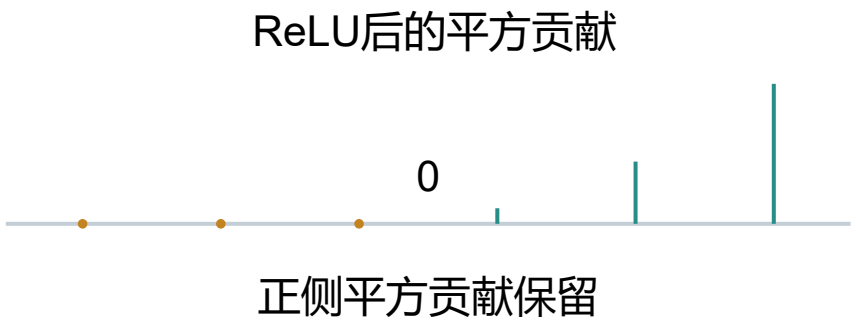
$$\frac{n_{in}+n_{out}}{2}Var(w) = 1 \Rightarrow Var(w) = \frac{2}{n_{in}+n_{out}}$$

- n_{in} 是输入单元数，n_{out} 是输出单元数，均不是批量大小
- 依据线性 / 近线性传播、零均值与独立同尺度近似；不保证任意网络的前后向尺度同时恒定，也不保证收敛

He初始化

ReLU保留正值、把负值置零。若预激活 s 关于零对称，正负两侧对平方平均贡献相同，移除负侧会使二阶矩减半。





$$\mathbb{E}[ReLU(s)^2] = \frac{1}{2}\mathbb{E}[s^2], Var(w) = \frac{2}{n_{in}}$$

- 在线性累加的独立近似下，前向二阶尺度倍率约为 $\frac{1}{2}n_{in}Var(w)$；**He初始化**用系数2补偿ReLU的影响，使倍率约为1
- 减半的是二阶矩，不是普遍的方差：ReLU后均值一般不为零
- Xavier / He建立合理起点，不保证训练期间始终保持尺度

2 特征统计与训练模式

初始化只设定参数起点，训练仍会改变内部表示；从输入标准化出发，理解BatchNorm怎样在前向计算中调整特征统计。

输入标准化调整不同特征的数值尺度

输入尺度差异会进入参数梯度，影响第4讲讨论的共同步长更新。

记样本损失为 ℓ，接收单元的加权和为

$$s = \sum_j w_j x_j + b$$

$$\frac{\partial \ell}{\partial w_j} = \delta x_j, \delta = \frac{\partial \ell}{\partial s}$$

同一样本、同一单元共享 δ。若 $\delta \neq 0$，非零输入的1000:1幅度比也传入梯度幅度比

输入标准化按特征中心化并缩放，均值 μ 和标准差 σ 由训练集估计：

$$x' = \frac{x-\mu}{\sigma+\epsilon}, \epsilon > 0$$

减少输入尺度差异，可能缓解共同步长的尺度不匹配。

- 验证、测试沿用训练统计，避免数据泄漏
- 标准化不消除特征相关性；局部幅度不代表特征重要性，批量梯度也不保持固定比例

尺度控制、数值范围与计算精度

尺度控制让信号与梯度保持适合传播和优化的幅度。

合适的尺度取决于计算过程，没有统一的 $[-1,1]$ 要求。

- 训练尺度：连乘可能使信号或梯度过度缩小、放大；输入尺度差异也影响更新。
- 浮点精度：有限有效数字带来舍入误差，极端数值可能溢出或下溢；需结合数据类型与运算处理。

标准化控制均值与标准差，不限定取值范围。 对 $1,3,5,7$，忽略 ϵ，均值为4、标准差为 $\sqrt{5}$：

$$x' = \frac{x-4}{\sqrt{5}} \Rightarrow -1.34, -0.45, 0.45, 1.34$$

舍入前均值为0、标准差为1，仍有值超过±1。固定区间不能保证精度或训练稳定。

任务要求与内部表示

监督标签约束预测结果，隐藏表示通常没有逐值指定的监督目标。

输入 x 标准化为 x'，网络参数为 θ，预测 $\hat{y}$ 与目标 y 共同决定损失：

$$x \to x' \to \hat{y} = f_\theta(x'), \ell(\hat{y}, y)$$

- 修改已训练网络：插入变换、保留其余参数，输出通常会改变；保持原计算需要满足补偿条件。
- 采用归一化结构训练：各层参数在该结构下共同学习，任务目标保持，无须逐层恢复原隐藏值。

监督目标不能保证任意变换都能恢复。输入标准化通常无需逆变换；若标准化回归目标，预测后需恢复目标单位。

表示变换与后续参数的补偿

可逆表示变换可以由紧随其后的线性层补偿，使后续计算保持一致。

设标量隐藏表示为 h，固定平移量 μ 和缩放量 $r > 0$：

$$h' = \frac{h-\mu}{r} \Leftrightarrow h = rh' + \mu$$

下一层原加权和为 $z = wh + b$。同时取 $w' = wr$、$b' = b + w\mu$，则

$$w'h' + b' = wr\frac{h-\mu}{r} + b + w\mu = wh + b = z$$

- 条件：统计固定、缩放非零，后面紧接允许相应权重和偏置的线性层。
- 这是等价参数的存在性；训练不保证找到相同预测，训练轨迹与正则化效果也可能改变。

非线性与信息恢复

变换与非线性的组合可能使不同表示变成同一个值，后续无法仅凭该值恢复原计算。

对已固定的原路径 $ReLU(h)$，在激活前插入偏移 $h-2$，其余计算不变：

原表示 h	原路径 $ReLU(h)$	改后路径 $ReLU(h-2)$
1	1	0
2	2	0

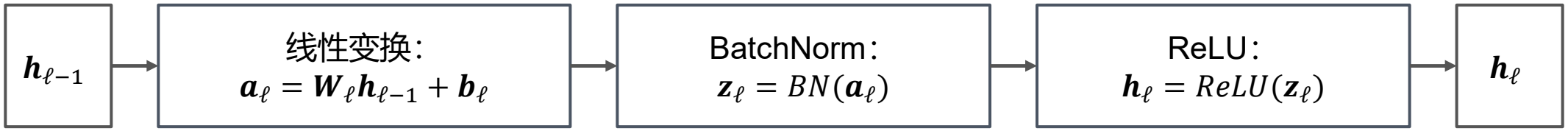
改后两个状态均为0；后续若只接收这个值，就不能同时恢复原来的1和2。

这说明修改已有网络需要检查信息与补偿条件。采用归一化结构训练时，各层可以共同学习新的表示；该反例不证明任务无法完成。

BatchNorm的作用位置

输入标准化处理进入网络的数据，初始化决定参数的起点。训练会继续改变权重和偏置，因此即使输入数据与预处理统计不变，网络内部表示的位置和尺度也可能改变。

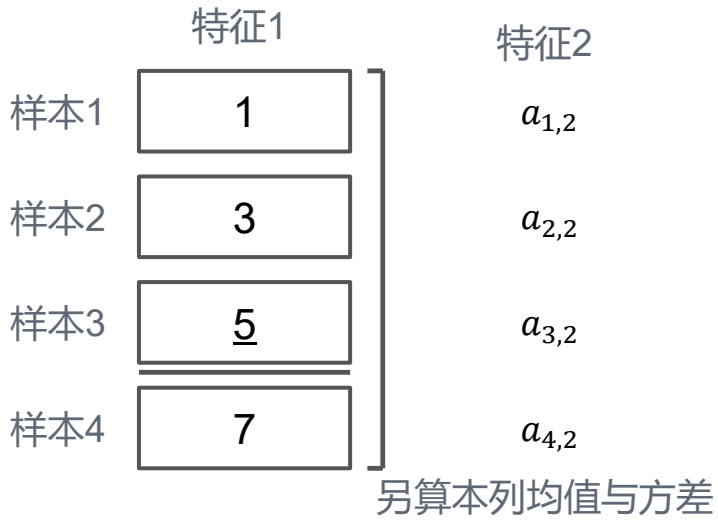
批量归一化（BatchNorm）把中心化与缩放用于内部特征：先按当前批次的统计调整尺度，再通过可学习变换选择合适的尺度和位置。本讲MLP示例把它放在线性变换之后、ReLU之前：



- $\boldsymbol{a}_\ell$ 是线性变换输出，$\boldsymbol{z}_\ell$ 是BN输出，$\boldsymbol{h}_\ell$ 是激活后表示；以下 a_i 表示同一特征在第 i 个样本中的BN输入值
- 初始化调节起点尺度；BN在后续每次训练前向中使用当前批次统计，不能用“初始化过了”代替这一计算
- 这是本讲的具体安排，不是唯一位置；内部统计变化不必然导致优化困难，BN也不保证稳定或替代合适初始化与学习率

BatchNorm的统计轴

每行一个样本，每列一个特征。选取第一列，跨四个样本统计，再归一化其中的第三个值。



$$\mu_B = \frac{1}{4}\sum_{i=1}^{4} a_i = 4$$

$$v_B = \frac{1}{4}\sum_{i=1}^{4}(a_i - 4)^2 = 5$$

$$\hat{a}_3 = \frac{5-4}{\sqrt{5+10^{-5}}} \approx 0.4472$$

跨样本统计同一特征

$m = 4, d = 2;\ a_i = a_{i,1}$

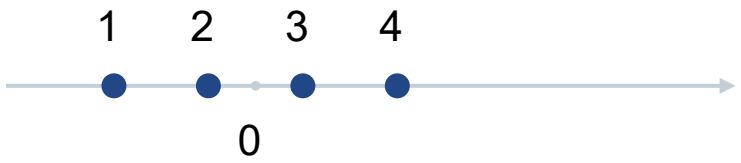
每个特征分别统计，训练前向方差分母为 $m = 4;\ \epsilon = 10^{-5}$ 防止除以零。

可学习的缩放与平移

标准化建立统计参照；同一输入 a_i 变为 $\hat{a}_i$ 后，再学习 z_i 的位置和尺度，调整进入后续激活的区间。

统计标准化

$$\hat{a}_i = (a_i - \mu_B)/\sqrt{v_B + \epsilon}$$



上页批次的标准化结果；数字为样本编号

可学习仿射变换

$$z_i = \gamma \hat{a}_i + \beta$$



$\gamma = 2$拉伸，再$\beta = 1$平移（示意）

- γ, β 按特征学习缩放与平移；均值和方差由输入计算
- 固定统计时，取 $\gamma = \sqrt{v_B + \epsilon}, \beta = \mu_B$ 可恢复原值；跨批次无统一恢复保证
- 网络共同学习适合预测的表示，无须恢复原值；训练目标也不要求 z_i 保持零均值、单位方差

BatchNorm的训练与推理

训练统计随批次组成变化；默认推理改用训练中累积的固定统计，无需为单独预测重新组批。比较同一输入 $a = 5$，取 $\gamma = 1, \beta = 0, \epsilon = 10^{-5}$。

模式与统计来源	均值 μ	方差 v	$a = 5$ 的输出，约
训练批次：1、3、5、7	4	5	0.4472
训练批次：5、7、9、11	8	5	−1.3416
推理：人为给定的固定统计	4	5	0.4472

- 默认跟踪运行统计时，训练使用当前批次统计并更新状态；eval()使用运行统计，γ, β 在两种模式下仍参与变换
- train() / eval()控制BN统计模式；no_grad()控制自动微分记录，不能代替eval()
- 表中推理统计仅为示意，并非前一批次直接产生的运行状态

BatchNorm的单值统计边界

对本讲MLP输入形状 $[m, C]$，每个特征由 m 个样本值统计，C 是特征数。每个特征仅有一个值时，统计公式退化：

$$m = 1: \mu_B = a, v_B = 0, \hat{a} = 0, z = \beta$$

- 数学上，中心化减去该值本身，输入差异被消去；γ 不能恢复被消去的差异
- 边界取决于每个特征的有效统计值数量，不只看batch size；长度轴或空间轴也可能提供多个值
- PyTorch的训练输入 $[1, C]$ 报错；默认使用固定运行统计的eval()仍允许单样本预测

BatchNorm的作用与边界

区分确定的计算机制、可观察的使用特征与对优化收益的研究解释。

层次	可以说明什么
计算机制	逐特征统计标准化，再做可学习仿射变换
使用特征	训练受批次组成影响；统计模式影响输出来源
研究解释	优化平滑性等视角提供条件性解释，不是普遍保证

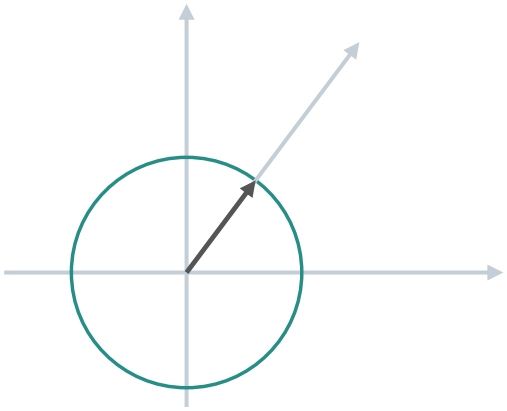
- 减少“内部协变量偏移”（训练中层输入分布变化）是原始动机之一，不是BN有效的充分解释
- 效果要结合模型、数据、批量和模式观察，不保证优化成功或泛化良好

3 梯度幅度控制

初始化与归一化不保证反向梯度始终合适；对已经形成的过大梯度，在更新前限制其范数，并区分梯度控制与参数位移。

梯度范数裁剪

L2范数 $\|\boldsymbol{g}\|_2 = (\sum_j g_j^2)^{1/2}$ 衡量整组梯度的长度。**整体范数裁剪**在优化器更新前，将超过阈值 c 的梯度统一缩短。



原梯度$\boldsymbol{g} = (3,4)$

裁剪后$\tilde{\boldsymbol{g}} = (1.2,1.6)$

范数阈值$c = 2$

同向，长度从5缩到2

反向传播→获得梯度→范数裁剪→优化器更新

$$\tilde{\boldsymbol{g}} = \boldsymbol{g} min\left(1, \frac{c}{\|\boldsymbol{g}\|_2}\right), \boldsymbol{g} \neq \boldsymbol{0}, c > 0$$

- 超过阈值时同向缩短；未超过时不变，零梯度仍为零；不同于逐分量截断
- 本例梯度乘 $2/5$；普通SGD取学习率0.1，参数位移长0.2
- 不恢复消失的梯度或修正错误导数；动量 / Adam的参数位移不直接受同一上界约束

本讲总结

先辨认问题及其来源，再选择作用于相应对象的处理思路；不同方法不能互相直接替代。

方法	作用阶段	针对的问题与处理思路
Xavier、He	参数初始化	连接数与激活改变传播尺度，按相应关系选择初始权重尺度
输入标准化	数据进入网络前	输入特征数值尺度不同，按训练统计中心化并缩放
BatchNorm	网络前向计算中	内部特征需要尺度调整时，按批次统计标准化并学习尺度与位置
梯度裁剪	梯度计算后、更新前	已形成的整体梯度过大，超过阈值时统一缩放

- 多层复合会累积尺度变化；前向二阶尺度与反向梯度分别分析，激活区间通过局部导数影响反传
- 随机差异打破对称性，合适尺度建立起点，BN与裁剪处理后续计算的不同对象
- 这些方法用于改善训练条件，不能单独保证优化成功或泛化良好