正则化与训练诊断

机器学习 · 深度学习基础

学习目标

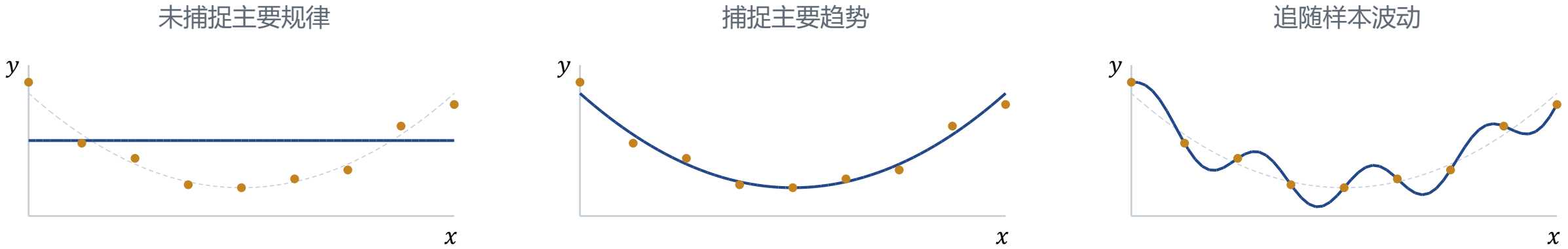
- 区分训练拟合、验证选择与独立测试，结合条件判断泛化问题
- 说明L2、权重衰减、Dropout和早停针对的问题、处理思路及各自改变的对象
- 从现象提出候选，选择有区分力的检查，说明干预后如何再验证
- 说明搜索尺度、预算和验证指标如何支持配置选择

1 泛化判断与正则化方法

数据分工、方法作用对象与验证选择。

拟合现象与诊断候选

训练拟合描述模型对训练样本的适应，泛化表现描述它在未用于参数学习的新数据上的表现。三种拟合示意用于理解二者的区别，实际判断仍需训练与验证记录。



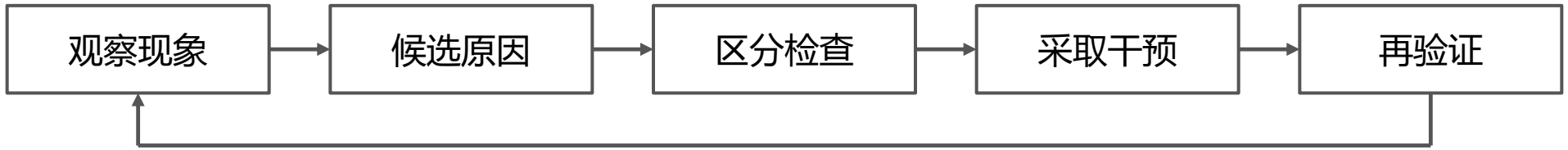
同一组指定样本；灰虚线为假设规律，蓝线为拟合示意，不是训练或验证成绩

- 欠拟合：尚未充分学到可学习规律；训练表现差仍需区分容量、优化与实现
- 过拟合：训练拟合很好，却未同样适用于新数据；差距只是线索，须核对划分、分布与评价模式
- 容量、数据量、噪声和训练条件共同影响结果；观察之后先提候选，再检查原因

数据职责与诊断路径

数据	主要用途	选择责任
训练集	更新参数，拟合预处理统计	训练成绩不能代替新数据评价
验证集	选择配置、时刻和检查方向	不作为日常参数梯度的同一批样本
测试集	对已选方案作独立评价	不反复用于挑选配置

- 先核对划分、损失计算与评价模式；数据损失、含正则的总目标和准确率分开记录



结果进入下一轮观察；单次检查不保证唯一定位

- 训练好、验证差时先查划分、预处理与模式；证据支持过拟合后，再用验证结果比较正则化、增加数据或早停

正则化与权重幅度

模型有时会通过较大权重响应训练样本中的局部变化，降低训练损失；这可能包含对偶然变化的适应。**正则化**在拟合之外加入偏好或约束，限制权重幅度是一种尝试。

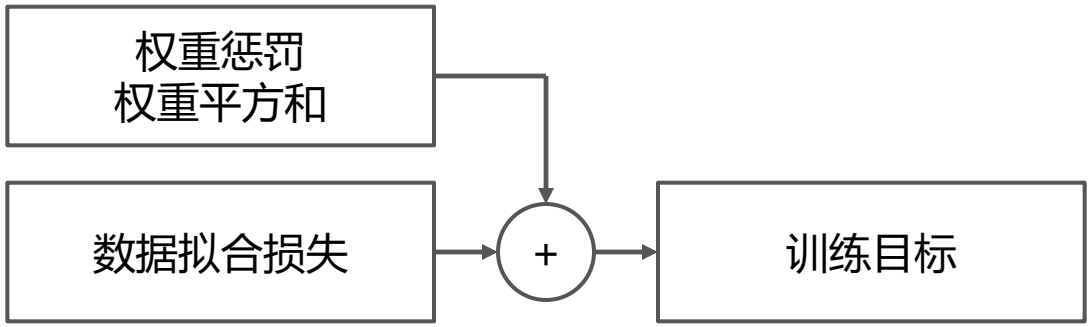
在线性函数 $f(x) = wx$ 中，输入扰动δ产生：

$$f(x+\delta) - f(x) = w\delta$$

- 输入单位相同时，$\delta = 0.1$：$w = 2$使响应变化0.2，$w = 0.5$使其变化0.05
- 若希望减少对这类变化的敏感性，可把控制权重幅度加入训练目标，与数据拟合共同权衡
- 这只是线性敏感性的动机支点，不是深层网络过拟合的统一原因；限制过强会损害拟合，泛化效果仍需验证

L2 在目标中增加权重大小的惩罚

L2 正则化用权重平方和把限制权重的偏好写入训练目标，优化器同时权衡数据拟合与权重幅度。



所选权重展开为$\boldsymbol{w}$，偏置不默认包含；$\|\boldsymbol{w}\|_2^2 = \sum_j w_j^2$。

$J(\boldsymbol{w}) = L_{data}(\boldsymbol{w}) + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2, \nabla J = \boldsymbol{g}_{data} + \lambda \boldsymbol{w}$

- L_{data}为平均数据损失，$\boldsymbol{g}_{data} = \nabla L_{data}$；$\lambda \geq 0$控制惩罚强度
- 惩罚梯度为$\lambda \boldsymbol{w}$，沿其负方向把权重往零拉；$\lambda = 0$时该项为零
- 同一参数下，总目标不小于数据损失，不代表预测变差；预测表现应比较同样计算的数据损失与验证指标

普通SGD下的L2与衰减

权重衰减直接缩放旧权重；普通SGD中，L2的额外梯度可等价写成乘法衰减：

$$\boldsymbol{w}_{new} = \boldsymbol{w} - \eta(\boldsymbol{g}_{data} + \lambda \boldsymbol{w}) = (1 - \eta\lambda)\boldsymbol{w} - \eta \boldsymbol{g}_{data}$$

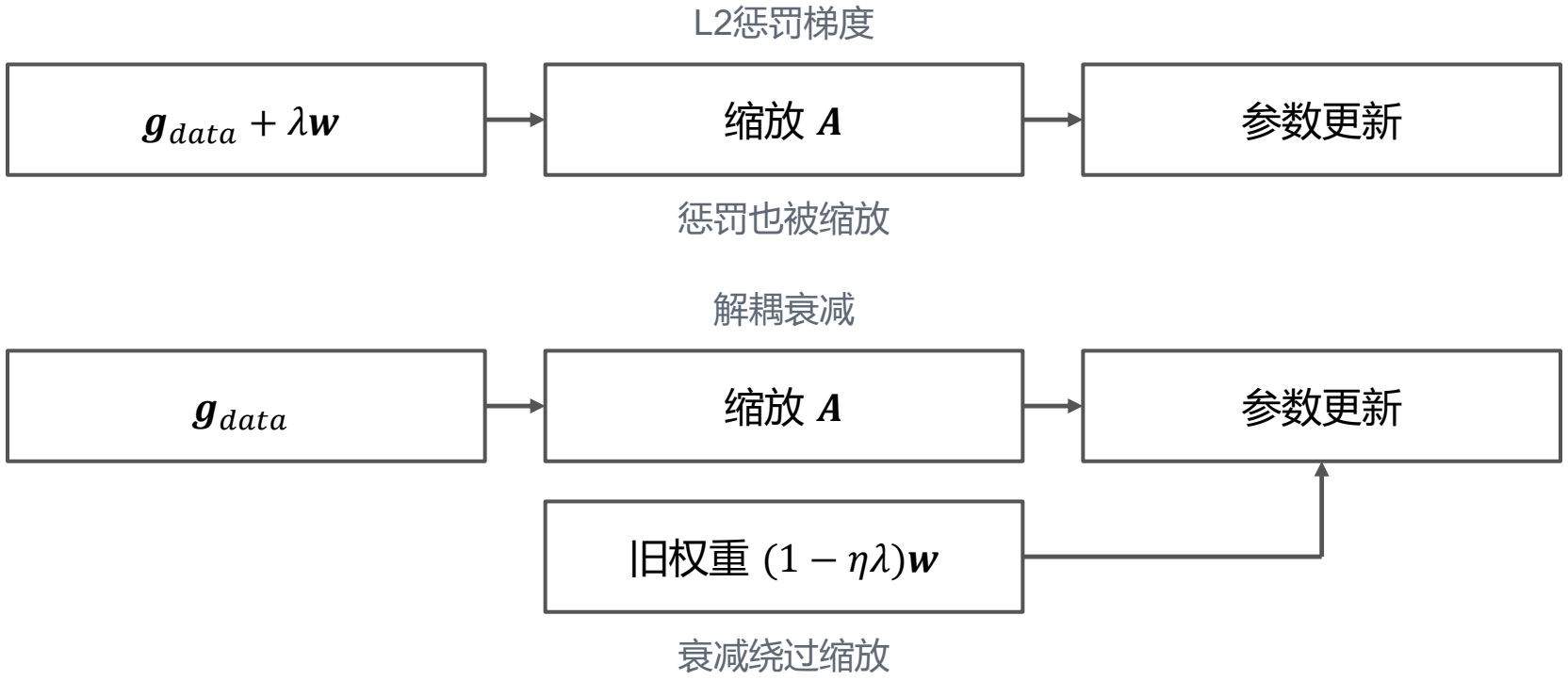
旧权重衰减	–	数据梯度更新
$(1 - \eta\lambda)\boldsymbol{w}$		$\eta \boldsymbol{g}_{data}$

普通SGD：额外梯度对应旧权重的乘法衰减

- $\eta > 0$、$\lambda \geq 0$；当$0 \leq \eta\lambda \leq 1$，旧权重项幅度保留或缩小，但数据梯度仍可能增大总权重
- 等价依赖普通SGD、同一权重集合及相同的η、λ定义；不适用于所有优化器
- 实际配置须核对受惩罚权重，偏置不默认包含

自适应更新中的两条路径

关键区别是权重惩罚是否经过自适应缩放：L2随数据梯度一起缩放，**解耦衰减**独立作用于参数。令A为当前逐坐标对角缩放算子，对照两条路径：

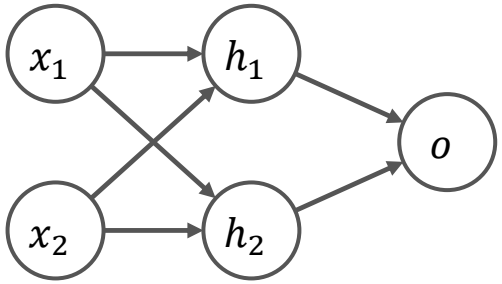


- A非单位算子时，两种更新一般不同；**AdamW**采用解耦衰减
- 固定A仅作缩放位置对照，非完整Adam；真实Adam的L2项还进入梯度历史

Dropout改变训练中的激活组合

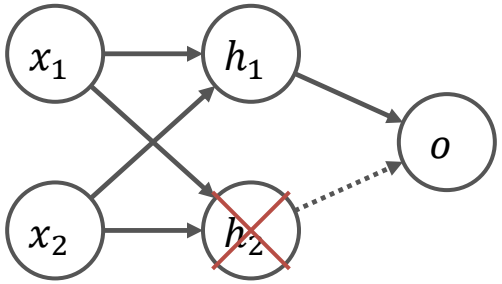
预测若依赖少数激活的固定组合，组合缺失就可能影响预测。**Dropout**在训练时随机屏蔽激活，让模型在变化的组合下仍按损失学习，以减轻固定依赖。

完整激活
输入层 → 隐藏层 φ → 得分层



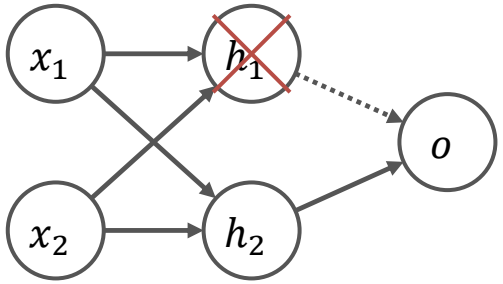
$\boldsymbol{z} = (1,1)$

训练掩码A
输入层 → 隐藏层 φ → 得分层



$\boldsymbol{z} = (1,0)$
输出置零

训练掩码B
输入层 → 隐藏层 φ → 得分层



$\boldsymbol{z} = (0,1)$
输出置零

隐藏层：$\boldsymbol{W}_1, \boldsymbol{b}_1, \phi$；得分层：$\boldsymbol{W}_2, b_2$；屏蔽激活输出，不删除权重

- 三图在同一参数状态下比较掩码；屏蔽的是激活输出，未屏蔽路径仍正常训练
- 训练仍更新权重，不永久删除参数；效果需验证，下一页补偿屏蔽引起的尺度变化

倒置缩放与推理行为

随机置零会改变激活的平均尺度。取激活$h_j = 2$，保留概率$q = 0.5$，先比较是否补偿尺度：



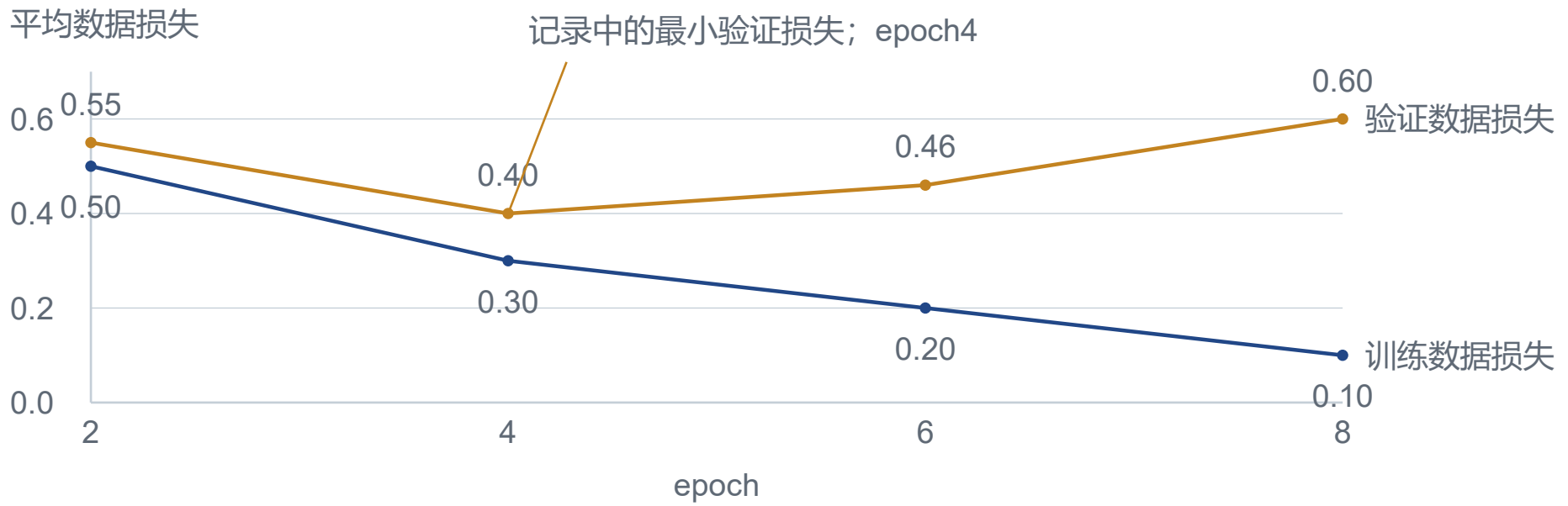
倒置Dropout训练时把保留激活除以q；$0 < q \leq 1$，各单元独立采样，丢弃率为$1 - q$：

$$z_j \sim Bernoulli(q), h'_j = \frac{z_j h_j}{q}$$

- 固定h_j时，$\mathbb{E}[h'_j] = qh_j/q = h_j$，保持局部激活期望；不保证非线性网络预测的平均相同或损失不变
- 推理直接使用完整$\boldsymbol{h}$，不采样、不再乘q；与BN一起检查训练 / 推理模式
- Dropout不直接收缩权重，其作用与L2不同；效果仍需验证

继续降低训练损失，验证损失可能回升

下图为教学构造，记录各时刻在推理模式下计算的平均数据损失：



- 验证表现回落，因此按验证指标选择训练时刻。依记录中的最小验证损失，应选epoch4，尽管epoch8的训练损失最低
- **早停**依据验证指标及给定停止规则确定何时停止训练，再恢复选定的模型状态；观察到验证损失回升不能唯一确定原因

早停是模型选择，也需要一致规则

- 训练中保存候选模型状态及验证指标，停止后恢复选定状态。模型状态包括参数，以及BN运行均值 / 方差等预测所需状态；只换回权重、却保留后续时刻的BN统计，可能无法恢复当时的预测
- 是否等待若干次未改善、监测损失还是其他指标，是需提前说明的配置，不存在通用的固定等待轮数
- 早停根据验证数据作选择，独立测试保留到方案确定之后；本讲不把测试曲线用于挑时刻

2 训练检查与诊断证据

检查、监测与追加结果共同支持诊断。

初始损失先核对任务和接口是否相符

采用哪项干预，需要先检查当前困难发生在哪个环节。初始损失可先用于核对任务与计算接口。

沿用前文交叉熵：二分类硬标签任务若每个样本的两类预测概率均为0.5，单样本数据损失为：

$$-\ln 0.5 = \ln 2 \approx 0.6931$$

- 先核对输出是得分还是概率、标签编码、损失的输入要求及平均方式；初始数值有明确条件才能作检查支点
- 若总目标含L2惩罚，不能要求它也恰为0.6931；纯数据损失和总目标要分开记录
- 必要时复用第二讲的有限差分核对导数，控制随机计算并避开不适合差分的点，不重新把它作为训练算法

小样本拟合检查学习管线能否工作

完整数据上训练不佳时，可以先缩小到一个小且标签一致、模型有表达能力的样本集，暂时减弱正则化。这使检查集中于学习管线能否拟合少量样本，再观察数据损失是否显著降低。

- 若仍无法拟合，检查标签 / 输入、损失接口、梯度是否到达参数、优化器是否包含应更新参数及实际参数位移
- 梯度出现不等于参数真的更新；反向、清零和更新时机需要与完整训练循环核对
- 小样本拟合成功只排除部分问题；即使存在数据处理错误，模型也可能记住少量样本，不能据此证明完整实现正确或能够泛化

观测证据与待检查问题

每种记录回答不同问题。先说明它能区分什么，再决定下一项检查。

观察证据	能支持的问题	仍需检查
数据损失、总目标、训练 / 验证指标	比较条件是否一致，拟合是否改善	划分、评价模式与损失定义
梯度范数与实际参数位移	有梯度之后是否真的更新	步长、更新调用、参数绑定
分层激活 / 梯度及BN状态	传播尺度或模式是否异常	输入、初始化与训练 / 推理设置

- 批次损失与完整数据平均分开；训练 / 验证比较需注明相同的计算与模式条件
- 某项统计正常只能排除部分候选，不能证明所有环节都正确

有梯度不等于参数有效更新

反向计算产生梯度，优化器执行更新；两者分别检查。**实际更新比例**衡量参数本步的实际位移相对原参数有多大：

$$\Delta \boldsymbol{w} = \boldsymbol{w}_{new} - \boldsymbol{w}, \rho = \frac{\|\Delta \boldsymbol{w}\|_2}{\|\boldsymbol{w}\|_2}$$

- 梯度非零而位移为零，只能确认本步参数未更新；仍需区分实际步长、更新调用与参数绑定
- 同一参数组且参数范数非零才计算ρ；分母为零时看绝对位移，不设通用正常阈值
- 动量、Adam或裁剪下，原始梯度乘学习率通常不能替代实际位移
- 分层激活饱和并伴随早层小梯度时，回查输入尺度与初始化；这是传播线索，区别于零位移的更新线索

训练与验证损失只能限定诊断候选

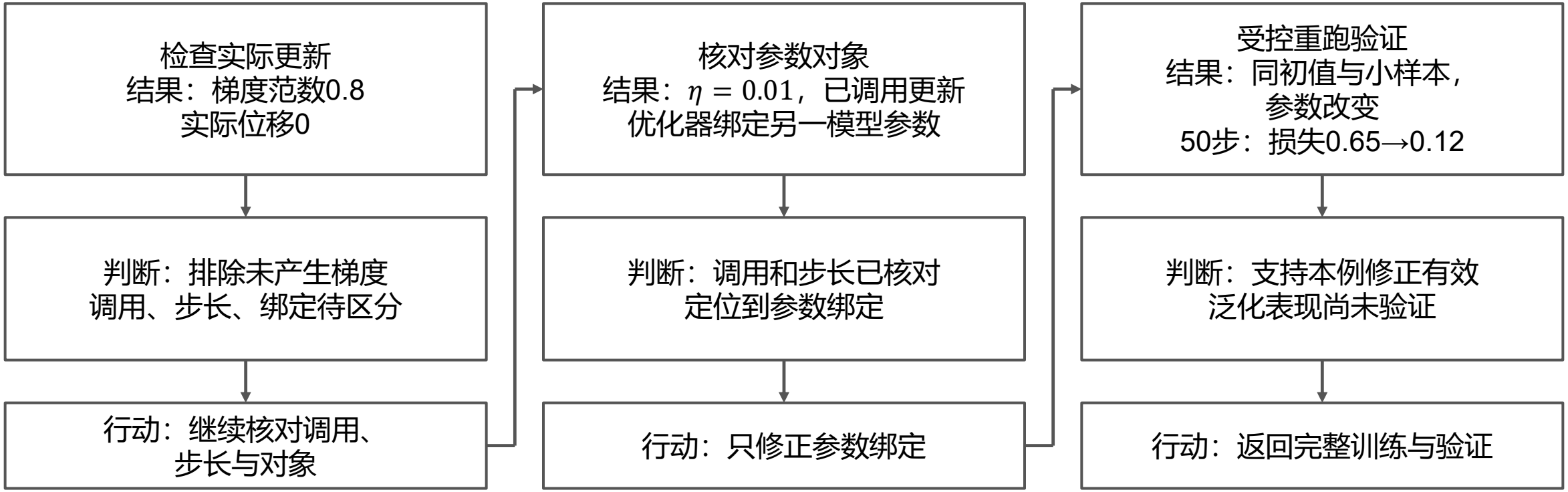
以下是采用相同损失计算方法的教学构造，比较的是平均数据损失：

现象	训练	验证	仍需区分的候选
两者接近初始基线	0.65	0.67	更新未有效进行、优化困难、容量或数据限制
训练损失低，验证损失较高	0.08	0.65	拟合未推广、分布或预处理差异、评价模式不一致

- 这些损失值能缩小问题范围，却不能唯一确定原因；先核对损失计算、训练 / 推理模式及数据划分条件
- 第一个现象不自动意味着欠拟合只需更大模型；第二个现象不自动意味着只需增加正则化

证据逐步定位参数绑定

教学构造，未运行真实训练：同一干净小样本、可表达模型，损失起点0.65。候选包括未更新、步长为零和参数未被优化器管理。



诊断通过新增证据逐步缩小原因范围，再采取对应修正并复验；当前没有增加容量或正则化的依据。

参数更新后，仍需区分优化、数据与泛化问题

前例处理的是参数没有被更新的情况。若实际位移非零，训练损失却仍几乎不改善，就需要继续区分优化、数据与容量等候选。

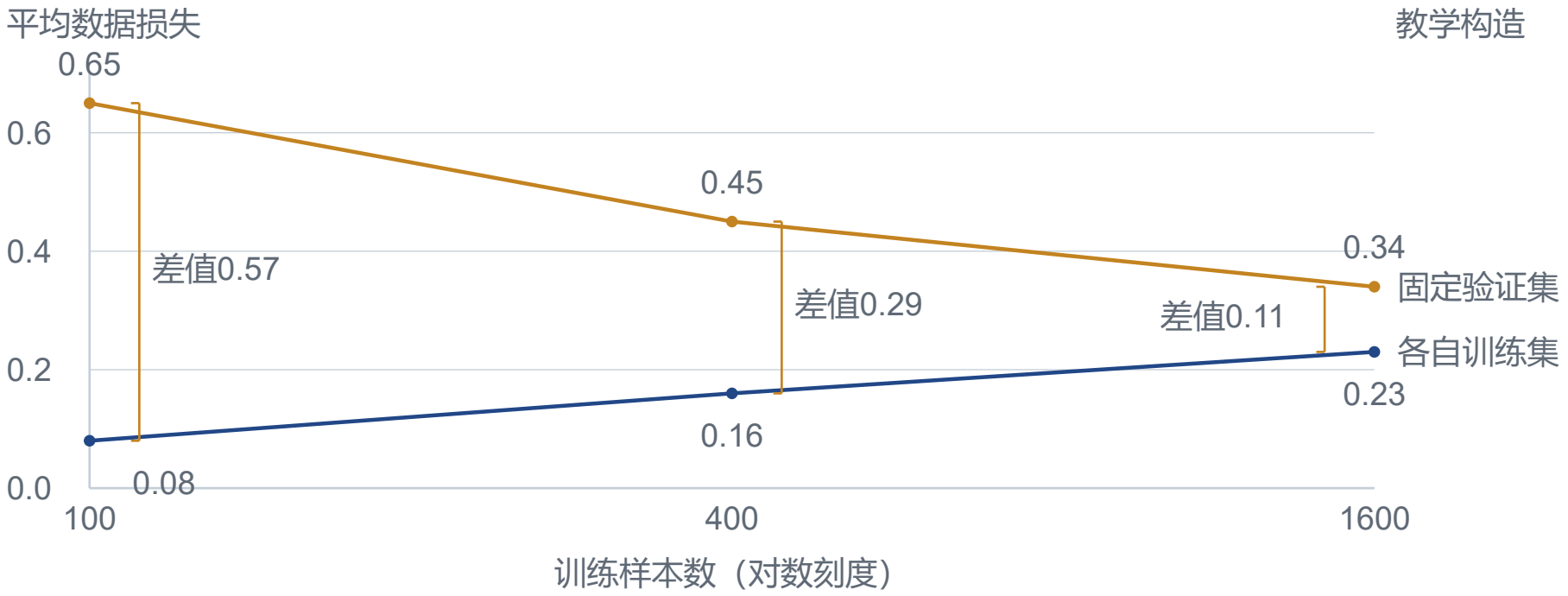
- 可用小样本拟合及梯度 / 更新尺度检查学习管线和优化条件；若这些局部检查未发现问题，再结合任务与数据考察容量。任何单项通过仍保留其证据边界
- 对“训练损失低、验证损失高”，先核对划分、预处理统计、标签编码和BN / Dropout模式；这些条件一致后，再选择数据规模或正则强度作受控变化，观察新数据表现是否改善
- 每次记录改了什么、预期区分什么以及结果支持什么；学习率、初始化和模式的机制使用前两讲已有定义

改变训练样本量，观察新数据表现如何变化

训练拟合好、验证表现差时，可检查增加数据是否改善新数据表现。以下为教学构造。

同一训练池取100、400、1600个样本，固定模型、初值和正则设置，分别训练到指定稳定条件；真实实验需记录各规模训练量。

蓝线比较各自训练集的平均损失；黄线始终使用同一验证集。均在推理模式下计算，不含惩罚。



本例验证损失改善，训练损失却升高

增加训练数据，可能使训练拟合看起来不如原来，却改善新数据表现；这是本例的指定趋势，不是普遍规律。

比较对象	实验条件	本例能支持的判断
训练损失升高	来自不同训练集	不能把损失差异直接归为优化程度变化
验证损失下降	验证集始终相同	为新数据表现改善提供比较依据

- 可将增加同来源数据作为调整候选；仍需重复抽样 / 初始化，检查优化是否充分、数据分布和标签质量
  - 各规模按指定稳定条件训练，计算消耗可能不同；不声称等预算优势，不用测试集挑选规模

平均指标之外，还要看错误发生在哪里

总体指标可能掩盖不同群体的错误。发现评价疑点后，进一步检查错误分布；这与改变训练量或继续训练回答不同问题。

指定100个分类结果，真实正类40个、负类60个：

真实类别	预测正确	预测错误	该类错误比例
正类	32	8	$8/40 = 20\%$
负类	55	5	$5/60 \approx 8.33\%$

- 总准确率为87%；这个平均数没有显示正类错误比例更高，应查看该类的代表样本与相关条件
- 类别分组和错误样本提供进一步线索，不证明标签质量、模型机制或数据分布的唯一原因；不要仅挑几个成功样本判断整体表现

3 超参数搜索与方案选择

用范围、预算和验证指标选择配置。

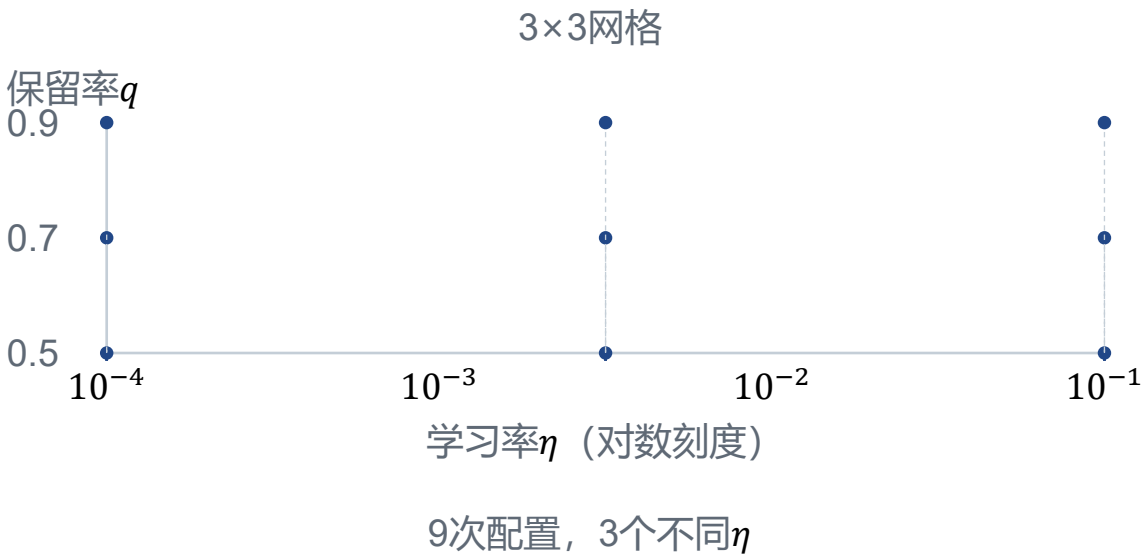
搜索尺度与预算

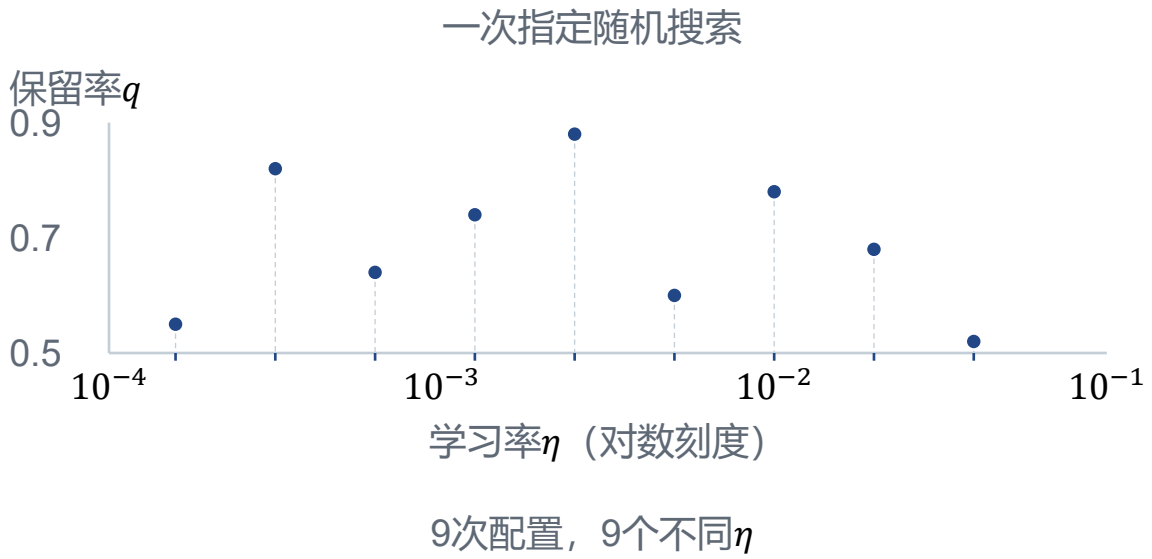
诊断提出干预方向后，还需选择调整的强度和范围。有限预算下，搜索要兼顾参数尺度与候选覆盖，再用验证结果比较配置。学习率、正则强度、Dropout保留率等**超参数**由训练者设置，区别于梯度更新的模型权重。

- 搜索尺度应匹配参数性质：学习率从0.001到0.002是加倍，从0.1到0.101只增加约1%；正学习率和正则强度可按对数量级搜索
  - 如η在10^{-4}到10^{-1}先宽后细；$\lambda = 0$单独保留，不能对零取对数
  - 保留率q有界，应使用合适区间，不机械套用对数尺度
- **随机搜索**在指定范围和尺度中抽取配置；同样9次投入的取值覆盖由下一页对照
- 覆盖更广不保证成绩更好；最终候选仍按同一预算及验证规则比较，测试不参与反复选择

本例随机搜索探索更多不同学习率

设同时搜索学习率η与Dropout保留率q，每个配置的训练预算相同，总共只做9次。先假设这个任务对η更敏感、对给定范围内的q不太敏感：





- 在这个假设下，网格把一部分预算用于同一η下改变较不敏感的q；随机搜索可把更多试验放在不同η上，这说明其作为起步方法的理由
- 哪些参数重要因任务而异；这组点阵解释预算分配，不是9次随机搜索必定取得更好验证成绩的证明

短训练用于初筛，范围边界仍需检查

粗筛和精查是投入不同的搜索阶段；每轮候选采用该轮约定的训练预算，短训练结果不能直接替代最终模型选择。

- 先在宽范围作短训练，排除明显不学习或发散的配置，再围绕有效区域细化；短训练只能初筛，最终候选仍需足够训练并按同一规则比较
- 若最好结果落在搜索范围边界，检查是否需要向该方向扩展范围；“当前范围内最好”不说明范围之外没有更好取值

配置由验证成绩选择，再做独立评价

假定三配置采用同一数据划分、模型及相同处理样本预算，表中数据损失均按同一评价模式计算，数值为教学构造：

配置	训练损失	验证损失
A	0.12	0.48
B	0.20	0.42
C	0.10	0.46

- 若规则为最低验证损失，应选B；它不是训练损失最低的C。结论只适用于这些候选和给定记录
- 保存配置、模型状态、预算和指标；状态应包含预测所需的参数及BN运行统计。反复从大量候选中选择，也可能使选出的方案过多适合验证集，最终测试保留为独立评价

本讲总结

面对训练结果，先明确问题和调整依据，再选择处理思路并检验结果。



结果进入下一轮观察；单次检查不保证唯一定位

- 观察使用相同方式计算的损失、实际更新、数据量对照及错误分布，单个现象不唯一确定原因
- 检查实现、优化、数据与模式，再有依据地调整：L2在目标中惩罚权重，衰减在更新中缩放权重，Dropout改变激活组合，早停按验证选时刻
- 搜索限定范围、尺度和预算，用验证集选择配置 / 时刻，恢复所选模型状态；测试只作独立评价
- 干预后受控再验证并记录结果；训练损失下降或局部检查通过，都不能单独证明泛化