优化算法与学习率

机器学习 · 深度学习基础

学习目标

- 解释小批量梯度、共同学习率与优化轨迹之间的关系
- 说明动量如何利用历史趋势，RMSProp如何调整坐标尺度
- 解释Adam的方向 / 幅度组合、零初始修正及全局学习率作用
- 说明学习率调度的动机，并在明确配置与预算下比较优化进展

1 梯度估计与步长

批次决定本次梯度，学习率与目标曲率共同影响参数位移。

训练目标与局部更新

优化算法利用已计算的梯度，寻找使训练集平均损失降低的参数：

$$L(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} \ell_i(\boldsymbol{\theta})$$

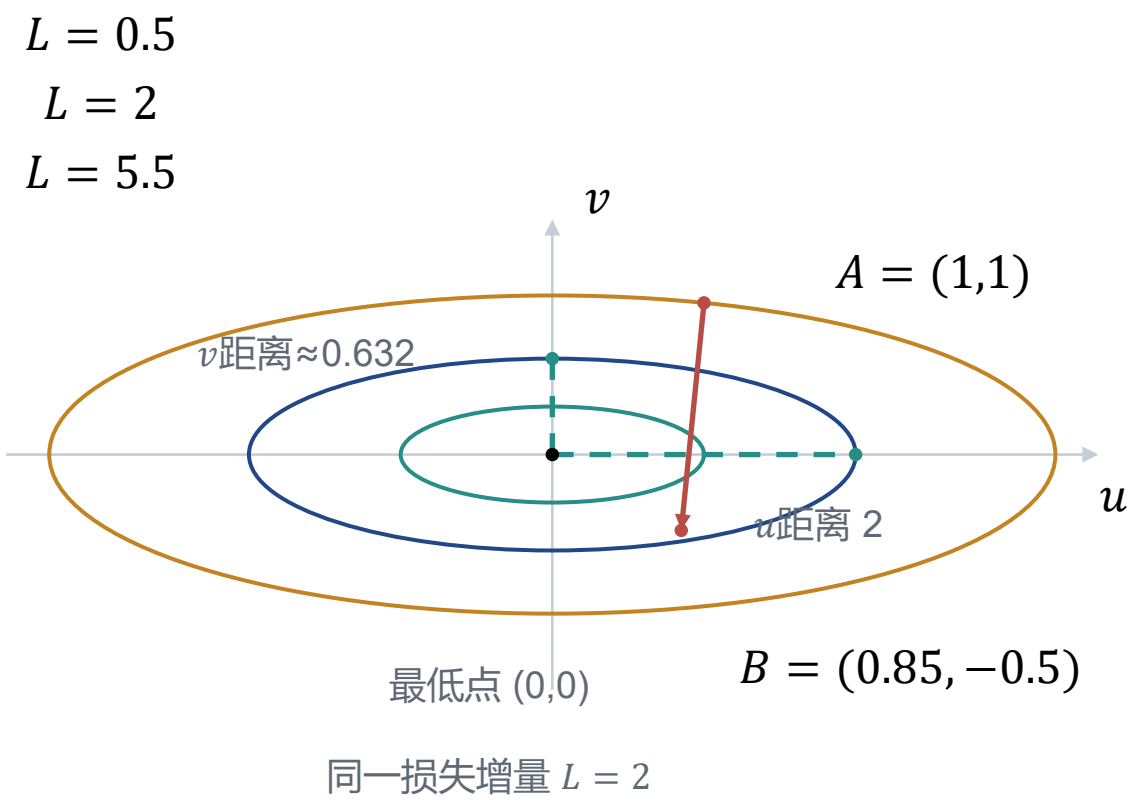
- N是样本数，ℓ_i是单样本损失，$\boldsymbol{\theta}$是展开的参数向量
- 在损失可微且梯度非零处，沿负梯度走足够小的一步可以降低损失；实际位移还取决于梯度估计与学习率
- 深层网络目标通常非凸；局部下降或零梯度均不能保证全局最优
  - 局部极小、鞍点和平坦区域只用于说明局部信息的边界
- 降低训练损失与改善泛化是不同问题；后者由验证和测试判断

曲率差异与共同学习率

等高线连接同损失位置；陡峭 v 方向达到同样损失增量所需距离更短。

$L(u,v)=\frac{1}{2}u^2+5v^2, \nabla L=(u,10v)$

- 同一 $L=2$：沿 u 距离为2，沿 v 约 0.632
- 为照顾 v 缩小共同学习率，也会减慢 u 方向的进展
- 教学函数的一步更新 $\eta=0.15$：$A\to B$ 中 v 跨零，u 仅小幅前进



同一参数处，不同批次提供不同梯度

设单参数模型预测 w，四个目标值为 0、2、4、6，单样本损失为 $\ell_i(w) = \frac{1}{2}(w - y_i)^2$。在固定 $w = 5$ 处：

目标 y	0	2	4	6
梯度 $w - y$	5	3	1	−1

- 同一参数处，全体平均梯度为2；前两样本平均为4，后两样本平均为0。批次不同，估计就可能不同
- 小批量梯度是整体梯度的估计；固定参数处均匀抽取同大小批次，所有可能批次的梯度平均等于整体梯度，某一次却可以偏离。这种偏离称为**梯度噪声**

SGD 用当前批次平均梯度更新

小批量随机梯度下降（mini-batch SGD）用一批样本的平均梯度进行一次更新：

$$\boldsymbol{g}_t = \frac{1}{|\mathcal{B}_t|}\sum_{i\in\mathcal{B}_t} \nabla \ell_i(\boldsymbol{\theta}_{t-1}), \boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \eta_t \boldsymbol{g}_t$$

- $\mathcal{B}_t$ 是当前批次；t 从1开始，梯度在更新前的参数处计算。d 是展开后的参数个数，$\boldsymbol{\theta}$ 与 $\boldsymbol{g}$ 均有 d 个分量
- $\eta_t > 0$ 是本次更新的**学习率**，与梯度一起决定参数位移的大小
- 本例从 $w = 5$、$\eta = 0.1$ 出发，使用平均梯度 4 更新到 4.6；若使用另一批的 0，则仍为 5
- 批次梯度可能偏离全体梯度，使更新后的整体损失暂时上升；若每步还换用当前批次计算损失，曲线又包含样本组成变化的影响
- 批量增大通常减小估计噪声，但每次计算更大；实际速度还受设备与数据处理影响

一轮、一次更新与处理样本量是不同单位

epoch 是完整遍历一次训练集；**step** 是进行一次参数更新。

$N = 100$，批量均整除样本数	batch size	每 epoch 更新次数	每 epoch 处理样本
小批次	10	10	100
大批次	50	2	100

- 同样 10 次更新，前者处理 100 个样本，后者处理 500 个；横轴相同步数不代表相同数据量
- 比较批次或优化器时，明确固定的是epoch数、处理样本量、更新次数还是用时；它们描述不同的训练投入

学习率与收敛轨迹

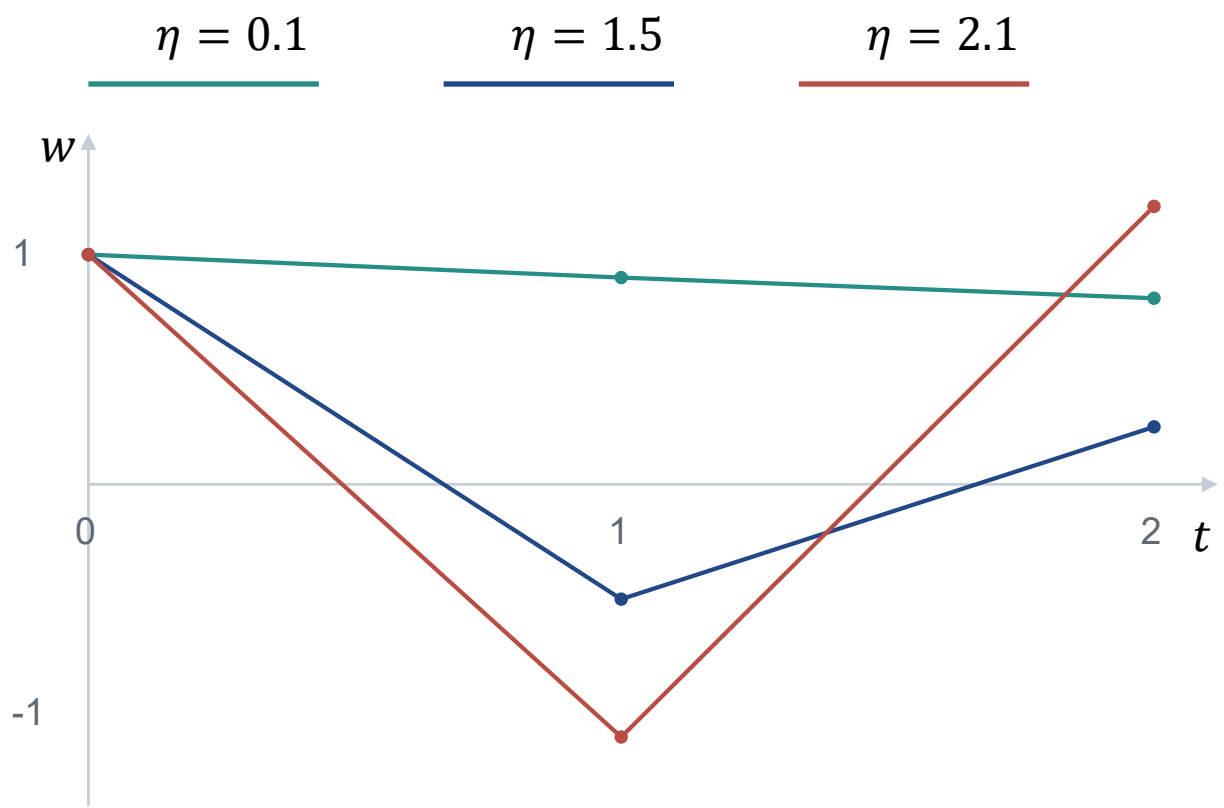
回到狭长曲面的步长问题，先隔离一个方向：
设 $L(w) = w^2/2$，梯度为 w。

$w_t = (1-\eta)w_{t-1}$

从 $w_0 = 1$ 出发：

η	w_1	w_2	轨迹含义
0.1	0.9	0.81	同侧缓慢接近
1.5	−0.5	0.25	两侧摆动且幅度缩小
2.1	−1.1	1.21	摆动且幅度增大

- 仅对此二次函数，$0 < \eta < 2$使$|1-\eta| < 1$，参数逐渐接近0



曲率与共同学习率

二次函数 $L(w) = aw^2/2$、$a > 0$ 中，**曲率** a 表示梯度随参数变化的快慢。更新乘数为 $1 - \eta a$；要让任意初值的参数逐渐接近0，这个乘数的绝对值必须小于1。

$$w_t = (1 - \eta a)w_{t-1}, |1 - \eta a| < 1 \Leftrightarrow 0 < \eta < \frac{2}{a}$$

- 曲率越大，允许的学习率上界越小。回到狭长曲面：u 方向 $a = 1$，v 方向 $a = 10$，共同正学习率需满足更紧的 $\eta < 0.2$
- 这个限制照顾了较陡方向，却可能使较缓方向进展很慢
- 区间只适用于上述二次函数；网络中的批次噪声、参数位置及训练阶段还会影响轨迹

2 历史与逐参数尺度

方向平均保留历史趋势，平方平均提供幅度，两者组合成Adam更新。

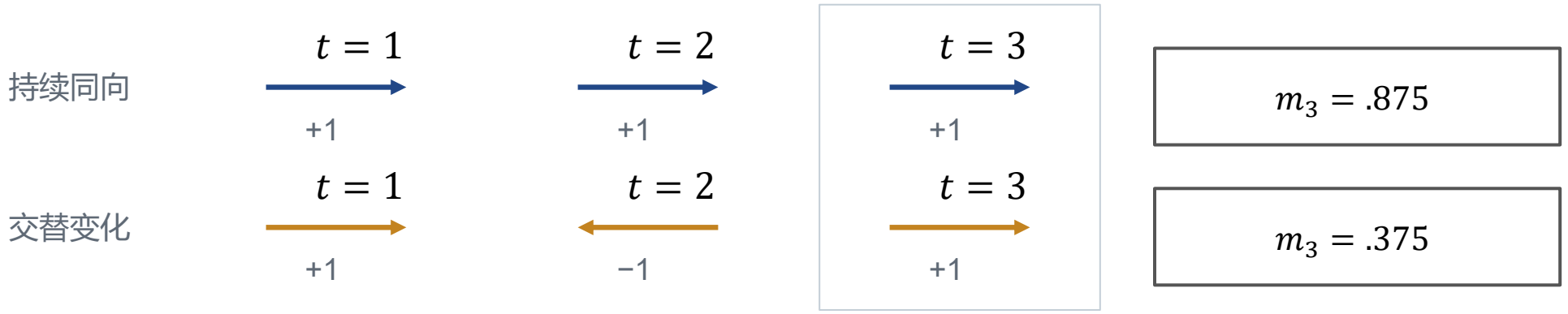
梯度历史与动量

动量（Momentum）用衰减的历史平均决定更新：持续同向的信息留下来，来回交替的作用部分抵消。

- 两组当前梯度都是+1，历史平均却不同；动量改变的是更新所使用的信息
- **指数加权平均**保留部分旧历史，再加入当前梯度；越早的信息留下得越少

$m_t = \beta m_{t-1} + (1-\beta) g_t, m_0 = 0, 0 \le \beta < 1$

β 越大，对较早历史保留越久；本例取 $\beta = 0.5$。图中箭头表示梯度符号，参数位移方向与它相反。



历史改变当前更新量

当前位置的梯度描述当前坡度，历史平均还保留前几步的趋势。本例每次梯度幅度相同，同向历史得到较大的平均量，交替历史得到较小的平均量；参数沿平均梯度的反方向移动。

$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \eta_t \boldsymbol{m}_t$

$\beta = 0.5$、$m_0 = 0$；下表只验证前页两种历史在同一平均规则下的差异：

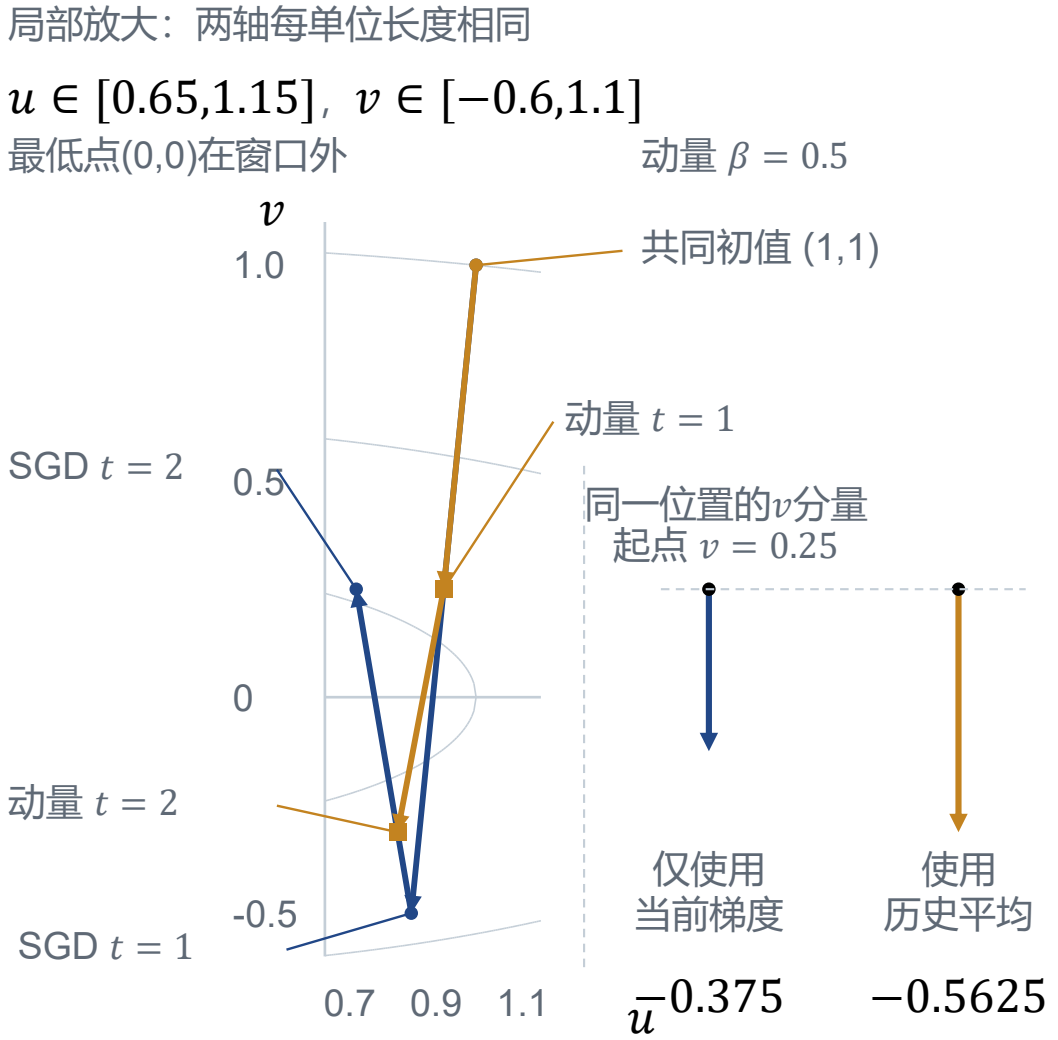
梯度历史	m_1	m_2	m_3
+1、+1、+1	0.5	0.75	0.875
+1、−1、+1	0.5	−0.25	0.375

- 例如同向历史取 $\eta = 0.2$，第三步位移为 $-0.2\times0.875 = -0.175$；交替历史的末步平均更小
- 每个坐标执行相同递推。本课采用带 $(1-\beta)$ 的平均形式，其他速度形式需换算状态及学习率
- 这是指定梯度输入，尚未绑定某个函数；函数上的行为由下一页解释

动量的历史作用与滞后

沿用狭长曲面 $L(u,v)=u^2/2+5v^2$，从(1,1)出发，$\eta=0.15$。动量采用 $\beta=0.5$、零初始状态。

- SGD使用当前梯度；动量还保留历史
- 第二步：当前 $g_v=2.5$，历史 $m_v=3.75$
- 历史保留趋势，也会延迟响应；旧历史仍推动 v 越过0
- 仅画两步；相同 η 不代表同等调优，不能据此排名



指数平均的记忆与边界

前面的趋势保留与响应滞后，来自同一份有记忆的状态。将递推展开，可以看清每次梯度留下了多少作用：

$$m_t = (1-\beta)\sum_{k=0}^{t-1}\beta^k g_{t-k}$$

- k 是从当前步向前回溯的步数：$k=0$ 对应当前梯度，权重为 $1-\beta$；较早的梯度对应更大的 k，在 $0<\beta<1$ 时权重更小
- 零初始且 $0<\beta<1$ 时，这些权重的和还不足1；在Adam偏差修正中继续使用这一结果
- β、η 和初始状态共同影响轨迹；历史平均不能完全消除噪声
- 动量不保证每步下降、全局最优或最佳泛化
- 接下来区分另一类历史：保留梯度的幅度，而不保留正负符号

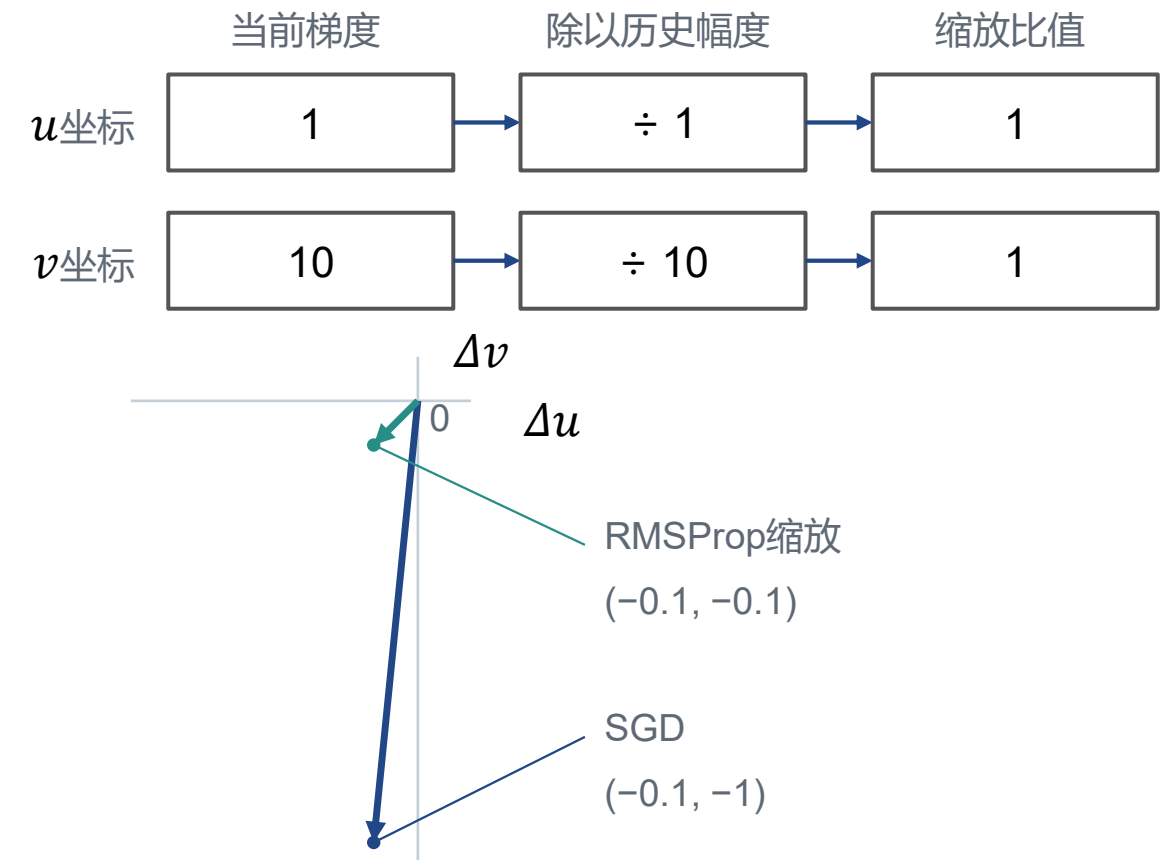
RMSProp的逐坐标缩放

RMSProp按各坐标的历史梯度幅度分别缩放，既改变位移大小，也可能改变更新方向。

$$v_t = \rho v_{t-1} + (1-\rho) g_t^2, \Delta\theta_t = -\eta_t \frac{g_t}{\sqrt{v_t}+\epsilon}$$

- 平方防止正负抵消，平均记录幅度，开方后按坐标相除
- 相同 η 下，两个坐标按不同比例缩放，方向随之改变
- 图取恒定梯度极限、忽略 ϵ，不表示第一步或收敛优势
- 逐坐标运算；$0 \le \rho < 1$、$v_0 = 0$、$\epsilon > 0$ 且在根号外。v 不是曲率或严格方差

同一梯度 (1,10)；历史幅度 (1,10)



$\eta = 0.1$；同一坐标比例，忽略ϵ

Adam的历史信息组合

Adam组合两类历史：分子保留动量的有符号趋势，分母按RMSProp的思路调整各坐标尺度，η 控制整体位移。有符号历史同时含方向和幅度。

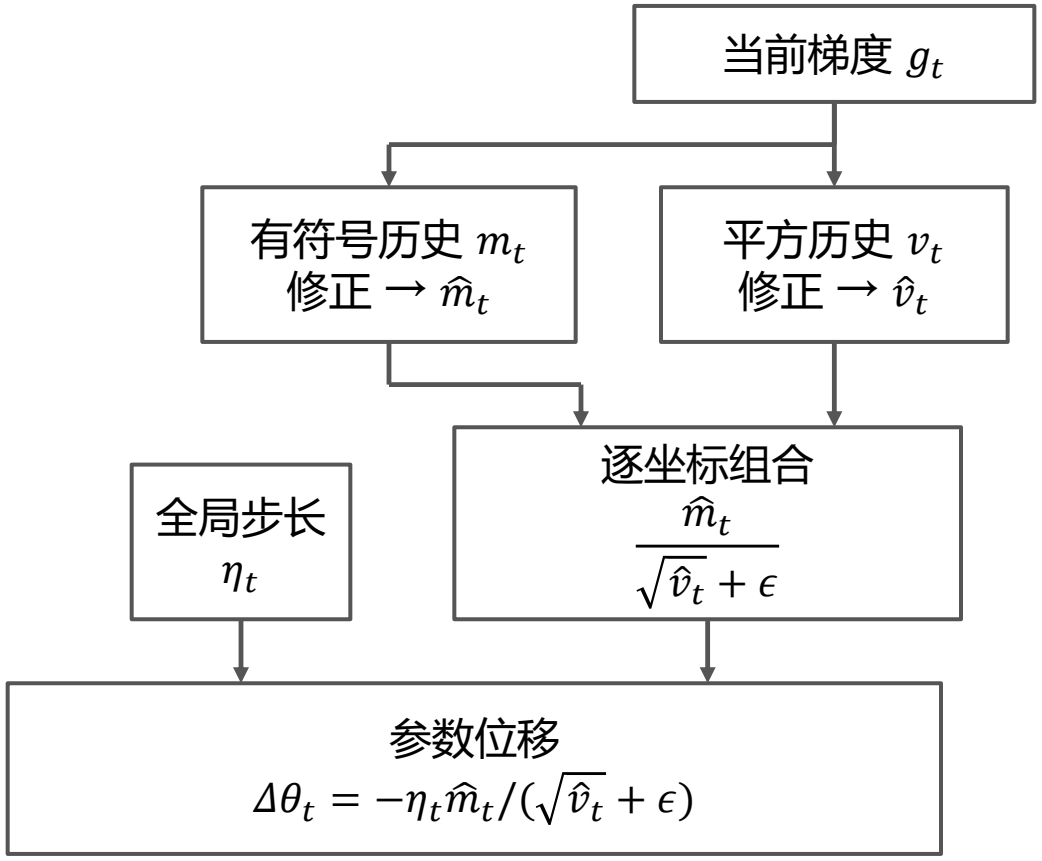
$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2$$

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t}, \hat{v}_t = \frac{v_t}{1-\beta_2^t}$$

$$\theta_t = \theta_{t-1} - \eta_t \frac{\hat{m}_t}{\sqrt{\hat{v}_t}+\epsilon}$$

- 逐坐标运算；在 θ_{t-1} 处求 g_t，$t \geq 1$
- $m_0 = v_0 = 0$，$0 \leq \beta_1, \beta_2 < 1$；$\epsilon > 0$ 在根号外
- 两种 β 控制记忆，带帽状态补偿零初值的影响；η 仍需选择



$$\theta_t = \theta_{t-1} + \Delta\theta_t$$

零初始状态与偏差修正

沿用动量的历史展开式，零初始时实际梯度的累计权重为：

$$(1-\beta)\sum_{k=0}^{t-1}\beta^k = 1-\beta^t$$

在常用的 $0<\beta<1$ 时，这个和小于1，早期状态因而受到零初值的影响。**偏差修正**除以累计权重，使实际历史的权重和归一化为1；$\beta=0$ 时无需修正。

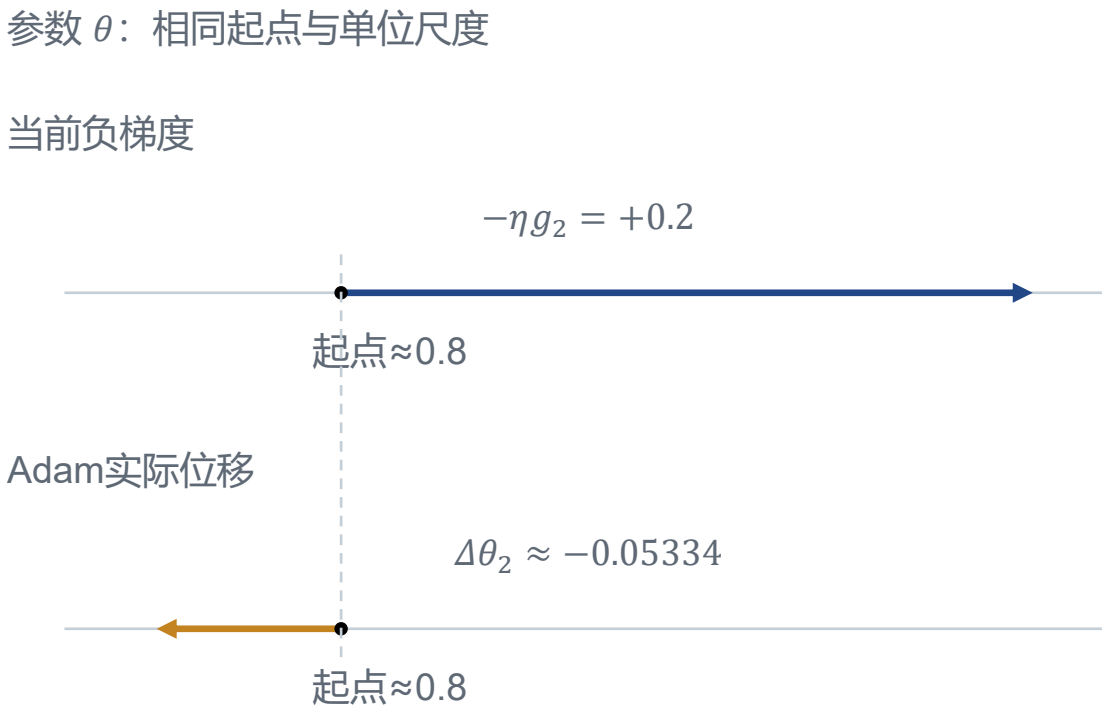
$$\hat{m}_t = \frac{m_t}{1-\beta_1^t}, \hat{v}_t = \frac{v_t}{1-\beta_2^t}$$

- 一次简单验证：$g_1=2$、$\beta_1=0.9$，则 $m_1=0.2$；归一化后 $\hat{m}_1=0.2/0.1=2$
- 两类状态各用自己的 β，因此分别除以 $1-\beta_1^t$ 与 $1-\beta_2^t$
- 二阶状态按同一思路修正平方梯度的平均
  - 梯度均值及平方梯度均值分别稳定时，可补偿相应平均的零初始偏差
- 修正补偿零初值，不消除随机波动；后续更新仍受实际梯度历史与 η 影响

Adam的历史方向与当前梯度

当梯度变号，历史方向未必立即变号。指定输入 $g_1 = 2$、$g_2 = -1$，从 $\theta_0 = 1$ 出发；$m_0 = v_0 = 0$，$\beta_1 = 0.9$、$\beta_2 = 0.99$、$\eta = 0.2$、$\epsilon = 10^{-8}$。

- $g_2 < 0$，当前负梯度向右
- $m_2 = 0.08$，$\hat{m}_2 \approx 0.421 > 0$，历史仍为正
- Adam位移向左；幅度状态只缩放大小，符号由 $-\hat{m}_2$ 决定
- 历史会使更新滞后于当前梯度
- 指定输入不是函数轨迹，不能据此判断损失变化



3 学习率调度与结果比较

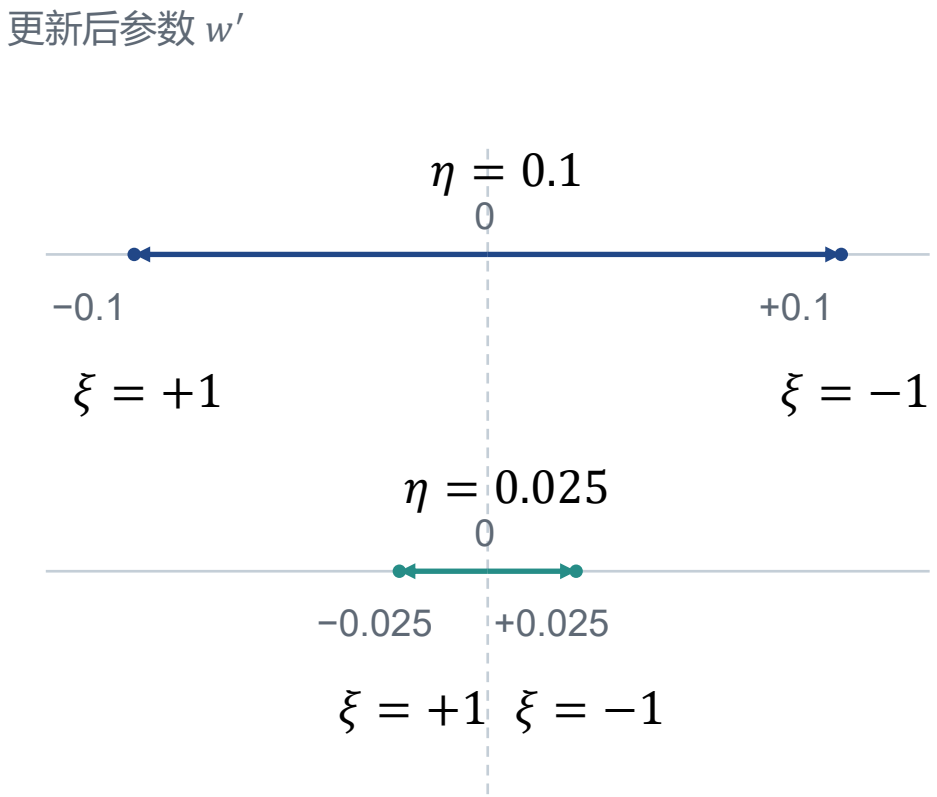
解释阶段步长变化，再在明确配置、指标和预算下比较进展。

梯度噪声与学习率

沿用 $L(w) = w^2/2$，最优点 $w = 0$。构造一个无偏但有波动的梯度 $g = w + \xi$，其中 ξ 等概率取 $+1$ 或 -1；平均梯度为 w，单次梯度不一定为零。

从同一个 $w = 0$ 做一步 $w' = w - \eta g$：

学习率 η	$\xi = +1$ 时的 w'	$\xi = -1$ 时的 w'	两种情况的损失 $L(w')$
0.1	−0.1	+0.1	0.005
0.025	−0.025	+0.025	0.0003125



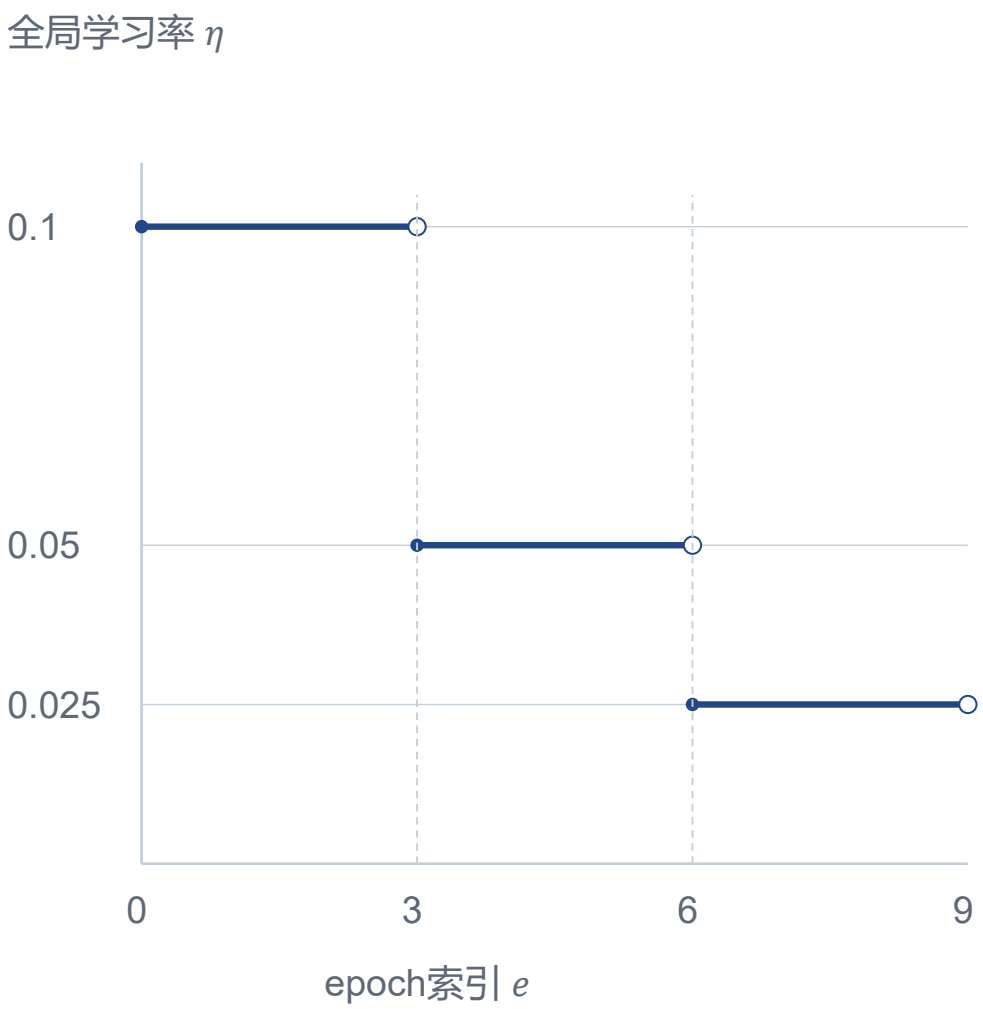
- 此处更新由噪声推动；η 减为四分之一，偏移减为四分之一，二次损失增量减为十六分之一
- 远离最优点时，η 也缩放真实梯度进展；η 过小会走得慢，衰减时机仍需结合进展判断

学习率调度与训练阶段

学习率调度使不同训练阶段采用不同的全局更新尺度。阶梯衰减是代表性方案：每完成 K 个epoch，将 η 乘以 $0 < \gamma < 1$。

$$\eta(e) = \eta_0 \gamma^{\lfloor e/K \rfloor}$$

- 图中 $\eta_0 = 0.1$、$K = 3$、$\gamma = 0.5$，epoch索引 e 从0开始
- 较小 η 减小噪声位移，也减慢真实进展；初值与时机需结合训练和验证判断
- 调度改变 η，不改变损失定义，不重置动量或Adam历史
- 计数单位是epoch，区别于step；阶梯并非唯一调度方式

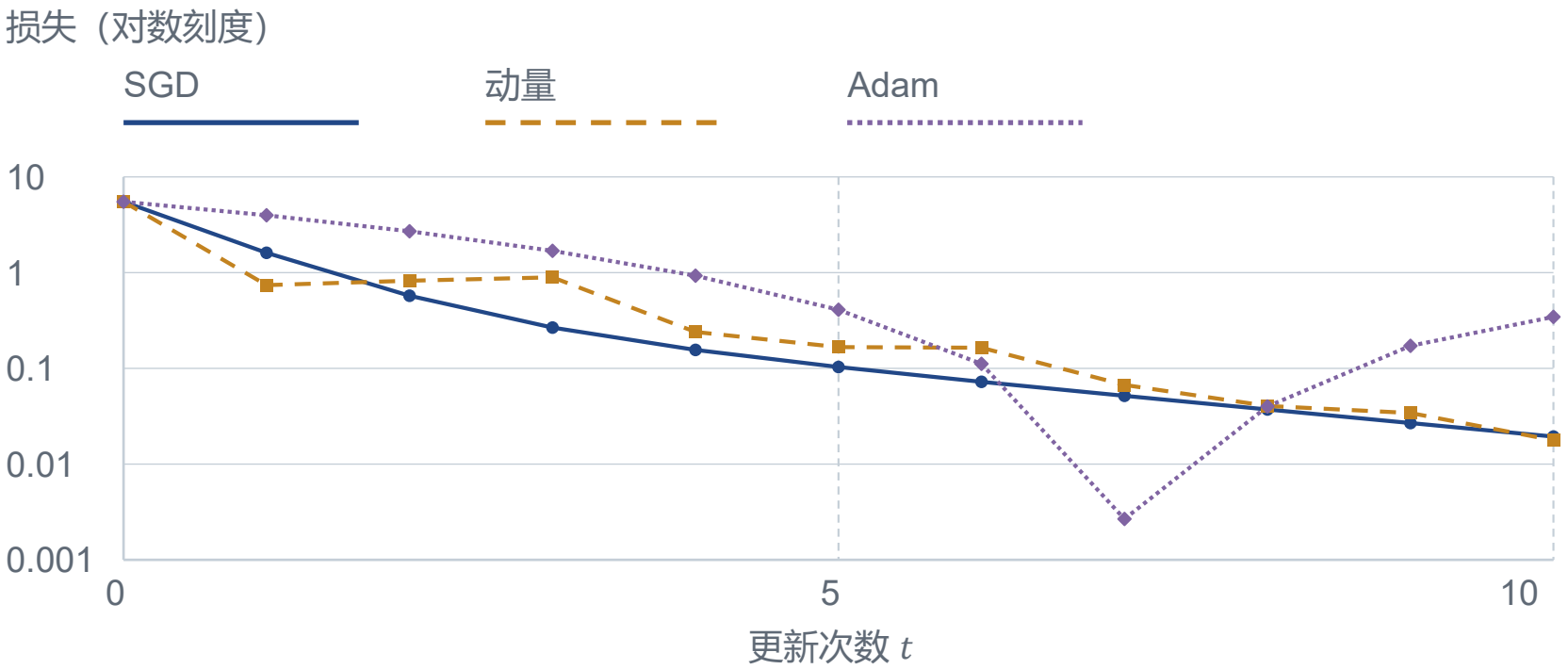


优化器轨迹比较

复用 $L(u,v)=u^2/2+5v^2$，初值(1,1)，每步在各自更新前参数处重算全梯度；没有随机批次。

- $\eta=0.15$ 固定，历史初值零；动量 $\beta=0.5$，Adam $\beta_1=0.9$、$\beta_2=0.99$、$\epsilon=10^{-8}$

配置	$t=5$	$t=10$
SGD	0.103	0.0194
动量	0.167	0.0177
Adam	0.410	0.345



$t=5$：SGD较低

$t=10$：动量较低

指定配置的计算结果；损失纵轴为对数刻度，相邻主刻度相差10倍。

优化器比较的条件

前页回答的是：这些配置在这个函数、这些观察时点下取得了什么优化进展。

- **观察到的结果**：第5步与第10步，SGD和动量的损失排序不同；比较结论依赖观察时点与指标
  - Adam的历史也持续影响轨迹，首次更新的幅度性质不能推广到每一步
- **能够支持的判断**：本例比较指定配置在相同全梯度计算次数下的优化进展
  - 相同 η 不意味着各方法得到同等程度的调优，不能据此判断算法普遍优劣
  - 未测量用时，也没有验证或测试结果；效率与泛化表现仍需各自的证据

真实训练的比较条件

比较前固定任务、数据划分、模型及初始化条件，明确各方法配置。不同目标需要不同证据：

比较目标	评价依据与对齐条件
参数优化进展	同一完整训练集、损失定义与归约方式；明确样本处理量或更新预算
计算效率	同一计算平台，记录用时；比较明确时间预算内的进展
泛化表现	验证集选择配置，独立测试集报告；注明数据与训练条件

- 不同批量的相同step数可能处理不同样本量；相同 η 也不代表同等调优
- 前例只比较同梯度次数的优化进展；第6讲进一步讨论验证选择与训练诊断

本讲总结

不同方法改变的是更新所使用的信息，学习率调度改变的是训练阶段的全局尺度。

方法	更新使用的核心信息
SGD	当前批次梯度
Momentum	当前梯度与有符号的历史趋势
RMSProp	当前梯度与历史幅度
Adam	有符号历史梯度与历史幅度
学习率调度	不同训练阶段的全局学习率

- 历史方向也包含幅度；按坐标缩放仍需全局学习率，历史也可能产生滞后
- 零初始修正补偿早期权重；比较须说明配置、指标与预算，优化进展不等同于泛化