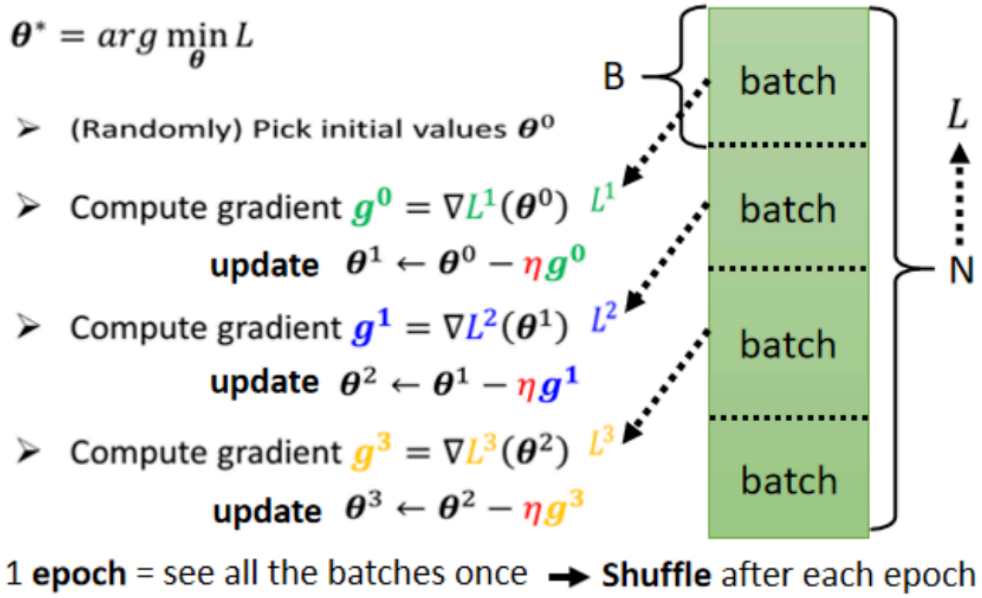
Batch & Momentum

1. Batch

1.1 为什么要用Batch



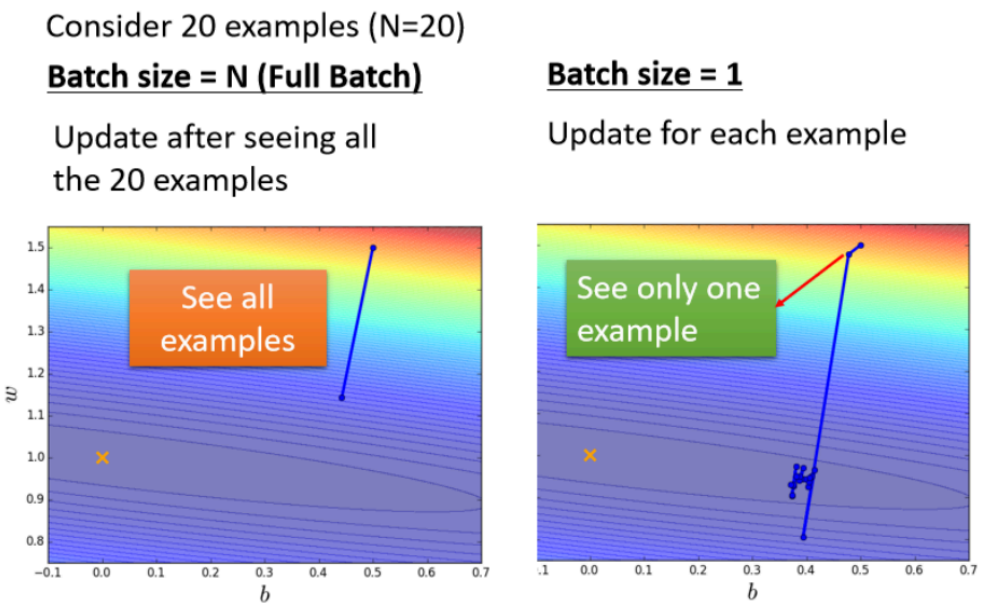
训练时不会每次用全部资料计算梯度，而是把训练集拆成 **多个 batch**：

- 每个 batch 大小 = B（又称 mini-batch）
- 每看完整个训练集一次 = **1 epoch**
- 每个 epoch 开始前通常会 **shuffle** 数据，让 batch 组合不同

目的：

- 降低一次梯度计算成本
- 引入随机性，改善训练dynamics
- 加快一次 epoch 的计算速度（借助 GPU 并行）

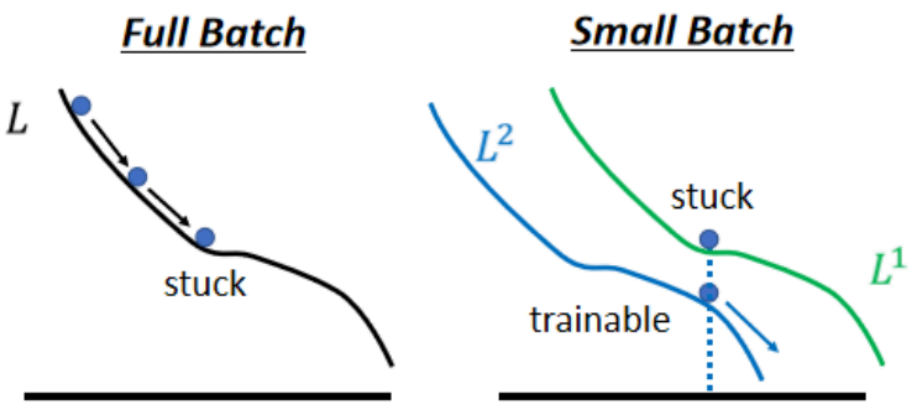
1.2 Small Batch vs Large Batch



- **Full Batch（B = 全部数据）**
  - 梯度 **完全准确、非常稳定**
  - 但每次更新要看全部数据 → 更新频率低
  - 容易被 local minima 或 saddle point 卡住
- **Small Batch（B 很小，如 1 或 32）**
  - 梯度 **noisy**, 每次更新方向不一样
  - 更新频率高
  - 引入随机性 → 较容易跳脱差的区域

1.3 为什么 small batch 训练更好?

核心原因：Noisy Gradient



- 假设两个 batch 对应两个 loss：$L_1(\theta)$、$L_2(\theta)$，它们的梯度方向不同
  - 若 L_1 在某点让梯度变 0（卡住），换 batch 变成 L_2 后梯度又不会是 0，因此继续训练
  - 因此 small batch 较不容易在 saddle point / flat region 停住

理解纠正：每个batch有一个loss function

> **换 batch 之后 Loss 会变化，是因为每个 batch 定义的是一个不同的子损失函数。即使参数不变，Loss 也会改变**

在训练过程中，我们真正想优化的完整损失函数是：

$$L(\theta) = \frac{1}{N}\sum_{i=1}^{N} \ell(x_i, y_i; \theta)$$

每个 batch 定义了自己的"子损失函数"：

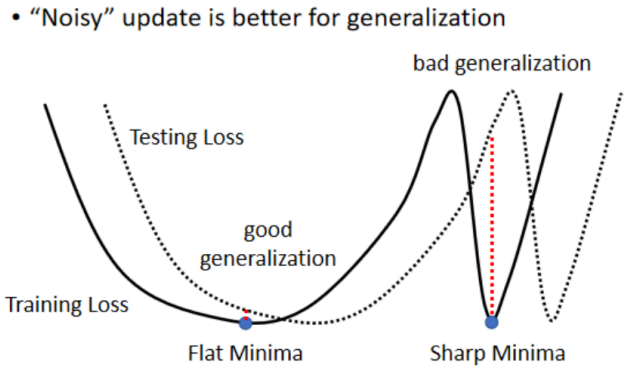
$$L_B(\theta) = \frac{1}{|B|}\sum_{i \in B} \ell(x_i, y_i; \theta)$$

1.4 Small Batch 对 Testing（泛化）更好

论文《On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima》发现：

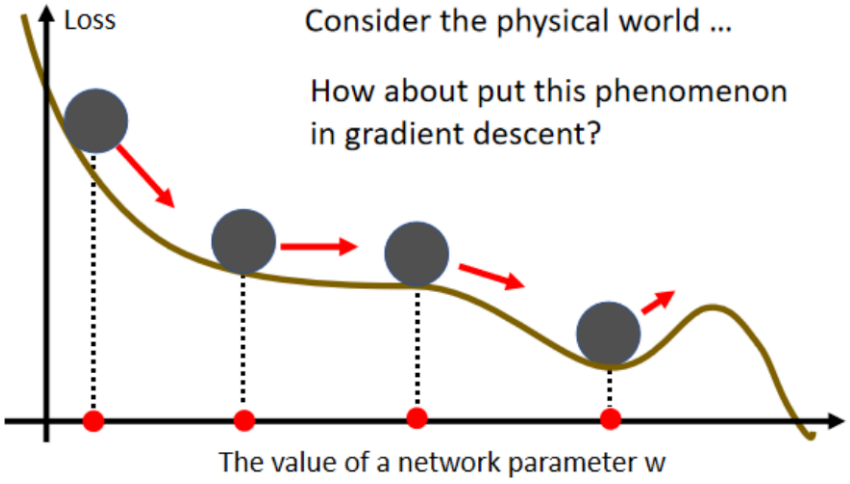
- 就算调参让 large batch 与 small batch 在 training 上 equally good，**small batch 的 testing accuracy 更高**

解释：



- training loss landscape 有 **sharp minima**（尖窄）与 **flat minima**（宽平）
- large batch 容易落入 sharp minima → 对 testing 敏感（泛化差）
- small batch 更新方向随机 → 容易跳出 narrow region → 落入 flat minima → 泛化更好

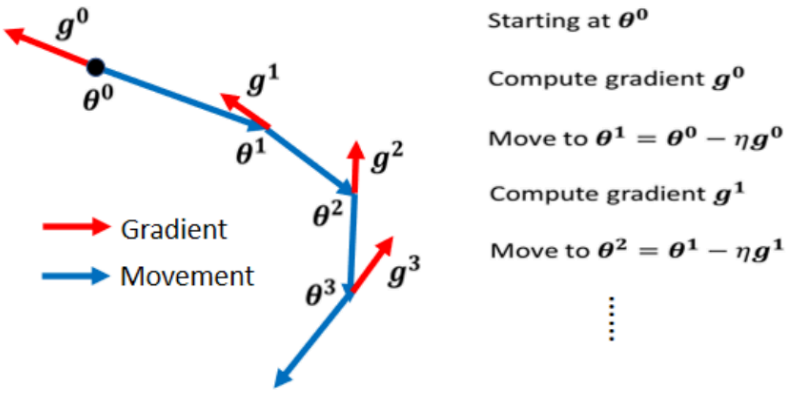
2. Momentum：解决训练卡住问题



Momentum 模拟“惯性”：

- 把 **前一次的更新方向** 加入当前梯度
- 遇到 small gradient（如 saddle point）仍能继续前进
- 有机会跨过 local minima 的小坑，找到更好解

2.1 Vanilla Gradient Descent

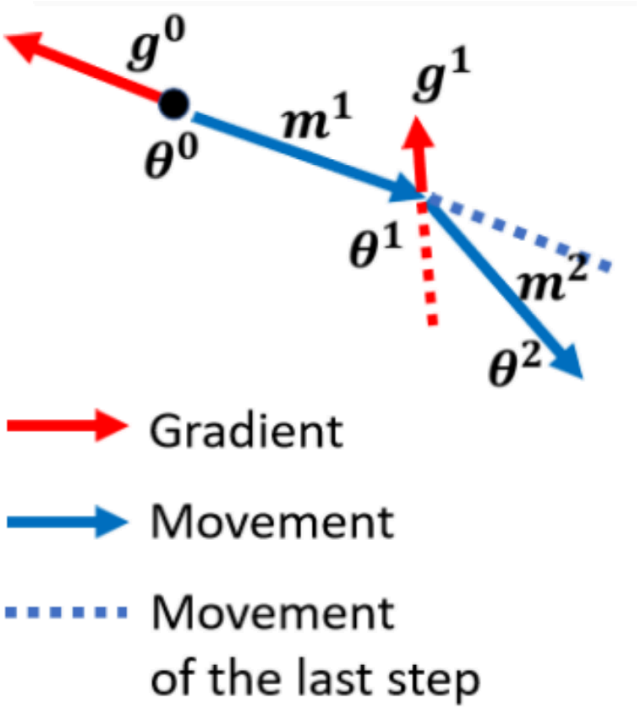


更新规则：$\theta^{t+1} = \theta^t - \eta g^t$

遇到：

- gradient = 0 → 停住
- local minima / saddle point → 无法继续前进

2.2 Momentum Gradient Descent



定义动量变量m_t：$m_t = \lambda m_{t-1} - \eta g_t$

更新参数：

其中：

- η = learning rate
- λ = momentum 系数（通常 0.9）

也就是：$m_t = -\eta(g_t + \lambda g_{t-1} + \lambda^2 g_{t-2} + \cdots)$

- Momentum 等于 **过去所有梯度的衰减加权和**
- 有“惯性”