Batch size

1. what

从数据集中取 `m` 个样本进行训练，每组称之为**batch**，样本数 `m` 为 `batch size` 。

2. why

1. **大**数据集下，样本无法一次性在内存内完成计算。
2. 使用batch能帮助训练。

基于*Gradient Descent*的训练，*full batch*（即使用全部训练数据时）更新权重，更能接近训练数据本身的“曲线”；*mini-batch*则由于抽样的原因，方差大，权重的更新与训练数据本身存在偏差。

3. how to choose batch size

`batch size` 影响了什么?

> - 训练速度
> - 收敛/Loss
> - 泛化性能

batch 较大：

- 发挥硬件性能。内存利用率、大矩阵乘法的并行化效率提高；
- 一次 `epoch` 的 `iteration` 更少，相对更快训练；
- 与样本本身“曲线”接近，震荡小；

batch 较小：

- 引入噪声，导致震荡；
- 更多 `iteration` ；

tradeoff

- 研究发现[3], 大batch容易陷入**sharp minimum**，小batch能越过去，反而接近**flat minimum**。

- 推荐：
  硬件允许下先用大batch避免震荡，快速收敛；下降到某一阈值后用小batch微调。

4. more?

- 基于2阶导来训练？待了解！

> “对于强大的二阶优化算法如共轭梯度法、L-BFGS来说，如果估计不好一阶导数，那么对二阶导数的估计会有更大的误差，这对于这些算法来说是致命的。”

- batch的延申研究：*Batch Normalization*，很好的介绍在这[4]。
- GPU 显存与batch关系。[5]

reference：

1. **https://www.zhihu.com/question/61607442/answer/440944387**
2. **https://www.zhihu.com/question/32673260**
3. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima
4. **https://zhuanlan.zhihu.com/p/34879333**
5. **https://zhuanlan.zhihu.com/p/31558973**