第 8 章：非凸优化算法

部分 2：随机梯度法

授课教师：曹语

课程主页：https://yucaoyc.github.io/math3806

目录

背景

上一部分中，我们先用“全部训练样本作为一个 batch”看清楚训练循环。每次更新本质上是在执行

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \alpha_k \cdot \text{当前数据给出的梯度}.$$

现在把它抽象成如下优化问题：

$$\min_{\boldsymbol{\theta}} \; f(\boldsymbol{\theta}) = \frac{1}{M}\sum_{i=1}^{M} f_i(\boldsymbol{\theta}), \qquad f_i(\boldsymbol{\theta}) = \mathcal{L}\big(p_{\boldsymbol{\theta}}(\boldsymbol{x}_i), y_i\big).$$

问题：如果 M 很大，为什么训练时不一定每一步都计算完整梯度 $\nabla f(\boldsymbol{\theta})$？

本节核心： 随机梯度是完整梯度的便宜估计，但会引入噪声。

学习目标

完成本节后，希望能够：

- 区分全梯度下降、随机梯度下降（Stochastic Gradient Descent, SGD）和小批量 SGD；
- 解释单样本梯度为什么是完整梯度的无偏估计；
- 从图像上理解随机噪声如何帮助探索非凸目标；
- 说清楚非凸 SGD 理论通常保证什么；
- 解释 batch size 如何影响梯度噪声和每步计算成本。

课前回顾：完整梯度

若

$$f(\boldsymbol{\theta}) = \frac{1}{M}\sum_{i=1}^{M} f_i(\boldsymbol{\theta}),$$

则完整梯度为

$$\nabla f(\boldsymbol{\theta}) = \frac{1}{M}\sum_{i=1}^{M} \nabla f_i(\boldsymbol{\theta}).$$

问题：当 M 非常大时，完整梯度下降的主要瓶颈是什么？

课前回顾：完整梯度

若

$$f(\boldsymbol{\theta}) = \frac{1}{M}\sum_{i=1}^{M} f_i(\boldsymbol{\theta}),$$

则完整梯度为

$$\nabla f(\boldsymbol{\theta}) = \frac{1}{M}\sum_{i=1}^{M} \nabla f_i(\boldsymbol{\theta}).$$

问题：当 M 非常大时，完整梯度下降的主要瓶颈是什么？

答案：每一步都要扫描全部样本，计算成本和数据读取成本都可能很高。

目录

两个熟悉例子

例子 1： Logistic 回归中，令

$$p_i = p_{\boldsymbol{a},b}(\boldsymbol{x}_i) = \sigma(\boldsymbol{a}^\top \boldsymbol{x}_i + b), \qquad \ell_i(\boldsymbol{a}, b) = -\big[y_i \log p_i + (1 - y_i) \log(1 - p_i)\big].$$

因此训练目标是

$$\min_{\boldsymbol{a},b} \ \frac{1}{M} \sum_{i=1}^{M} \ell_i(\boldsymbol{a}, b) = -\frac{1}{M} \sum_{i=1}^{M} \big[y_i \log p_i + (1 - y_i) \log(1 - p_i)\big].$$

例子 2： 最小二乘中，若 $\boldsymbol{a}_i^\top$ 是 $\boldsymbol{A}$ 的第 i 行，令 $\ell_i(\boldsymbol{\theta}) = \big|\boldsymbol{a}_i^\top \boldsymbol{\theta} - y_i\big|^2$. 因此训练目标是

$$\min_{\boldsymbol{\theta}} \ \frac{1}{M} \sum_{i=1}^{M} \ell_i(\boldsymbol{\theta}) = \frac{1}{M} \sum_{i=1}^{M} \Big|\boldsymbol{a}_i^\top \boldsymbol{\theta} - y_i\Big|^2.$$

把参数统一记为 $\boldsymbol{\theta}$，它们都具有共同形式：$\min_{\boldsymbol{\theta}} \ \frac{1}{M} \sum_{i=1}^{M} f_i(\boldsymbol{\theta})$.

挑战：大数据 + 非凸

$$\min_{\boldsymbol{\theta}} \; f(\boldsymbol{\theta}) = \frac{1}{M} \sum_{i=1}^{M} f_i(\boldsymbol{\theta}).$$

- **大数据**：全部样本同时参与一次更新可能太贵；
- **非凸性**：神经网络参数化后，$f_i(\boldsymbol{\theta})$ 通常不是凸函数；
- **实际需求**：希望用较低成本做很多次参数更新。

问题：只用一个样本的梯度，会不会完全没有数学依据?

参与推导：随机梯度的无偏性

在给定当前点 $\boldsymbol{\theta}^{(k)}$ 后，随机均匀选择 $i_k \in \{1, 2, \ldots, M\}$，并令

$$\boldsymbol{g}_k = \nabla f_{i_k}(\boldsymbol{\theta}^{(k)}).$$

问题：计算条件期望 $\mathbb{E}[\boldsymbol{g}_k \mid \boldsymbol{\theta}^{(k)}]$。它和完整梯度有什么关系？

参与推导：随机梯度的无偏性

在给定当前点 $\boldsymbol{\theta}^{(k)}$ 后，随机均匀选择 $i_k \in \{1, 2, \ldots, M\}$，并令

$$\boldsymbol{g}_k = \nabla f_{i_k}(\boldsymbol{\theta}^{(k)}).$$

问题：计算条件期望 $\mathbb{E}[\boldsymbol{g}_k \mid \boldsymbol{\theta}^{(k)}]$。它和完整梯度有什么关系？

答案：

$$\mathbb{E}[\boldsymbol{g}_k \mid \boldsymbol{\theta}^{(k)}] = \frac{1}{M} \sum_{i=1}^{M} \nabla f_i(\boldsymbol{\theta}^{(k)}) = \nabla f(\boldsymbol{\theta}^{(k)}).$$

给定当前迭代点时，单样本梯度是完整梯度的无偏估计，但有方差。

三种更新方式

$$\min_{\boldsymbol{\theta}} \ f(\boldsymbol{\theta}) = \frac{1}{M}\sum_{i=1}^{M} f_i(\boldsymbol{\theta})$$

- **全梯度下降**

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \alpha_k \nabla f(\boldsymbol{\theta}^{(k)}).$$

- **SGD**

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \alpha_k \nabla f_{i_k}(\boldsymbol{\theta}^{(k)}).$$

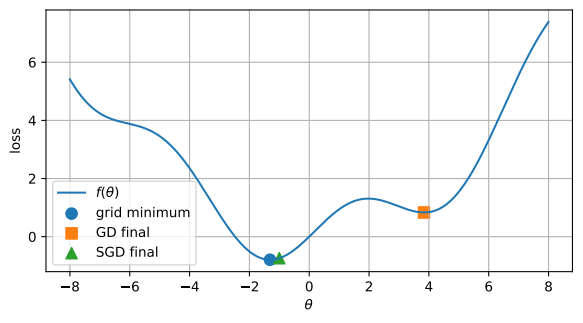
- **Mini-batch SGD**

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \alpha_k \frac{1}{|S_k|}\sum_{i\in S_k} \nabla f_i(\boldsymbol{\theta}^{(k)}).$$

目录

一维实验：随机性会带来什么？

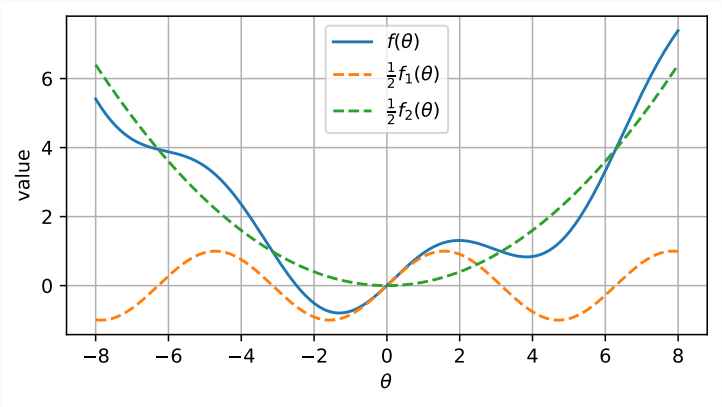
考虑非凸函数 $f(\theta) = \sin(\theta) + (0.1) \cdot \theta^2 = \frac{1}{2} f_1(\theta) + \frac{1}{2} f_2(\theta)$，其中 $f_1(\theta) = 2\sin(\theta)$，$f_2(\theta) = (0.2) \cdot \theta^2$。从相同初值 $\theta_0 = 5$ 出发，全梯度下降和 SGD 可能到达不同区域。



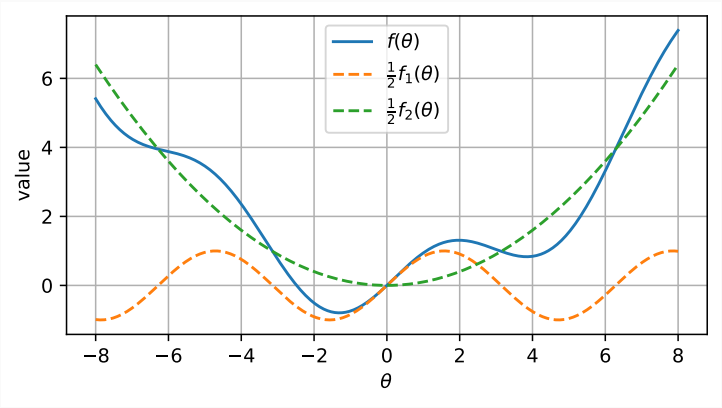
问题：图中只标出最终落点。为什么 SGD 的随机更新有时可能把迭代带到另一个局部区域？

直观解释：噪声帮助探索

问题：基于下图，为什么 SGD 有机会离开某些局部区域?



直观解释：噪声帮助探索



在局部区域附近，完整梯度可能很小；但随机选中的某个 f_i 仍可能给出较大的更新方向。这个方向不一定让 f 每一步都下降，却可能把点推离当前局部区域。

Mini-batch 的直觉

方法	每次更新成本	轨迹特点
Full-batch GD	高，需要全部样本	较平滑，但每步较贵
Mini-batch SGD	中等，只用一小批样本	有噪声，更新更频繁
单样本 SGD	低，只用一个样本	噪声最大，曲线波动明显

目录

两个基本假设：平滑性与噪声大小

为了解释 SGD，先保留两个常见假设。

假设 1： f 是 L-smooth，即对任意 $\boldsymbol{x}, \boldsymbol{y}$，

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \le L\|\boldsymbol{x} - \boldsymbol{y}\|,$$

等价地，若 f 二阶可微，则 $-L\boldsymbol{I} \preceq \nabla^2 f(\boldsymbol{\theta}) \preceq L\boldsymbol{I}$。且随机梯度无偏：

$$\mathbb{E}[\nabla f_{i_k}(\boldsymbol{\theta}) \mid \boldsymbol{\theta}] = \nabla f(\boldsymbol{\theta}).$$

假设 2： 随机梯度噪声的二阶矩有界：

$$\mathbb{E}[\|\nabla f_{i_k}(\boldsymbol{\theta})\|^2 \mid \boldsymbol{\theta}] \le \sigma^2 + \|\nabla f(\boldsymbol{\theta})\|^2.$$

直观理解： 在无偏性下，这等价于噪声方差不超过 σ^2；σ^2 越大，随机梯度越吵。

一次更新的下降估计

由平滑性可得，对 SGD 更新

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \alpha_k \nabla f_{i_k}(\boldsymbol{\theta}^{(k)}),$$

有

$$\mathbb{E}[f(\boldsymbol{\theta}^{(k+1)}) \mid \boldsymbol{\theta}^{(k)}] \leq f(\boldsymbol{\theta}^{(k)}) - \alpha_k \left\| \nabla f(\boldsymbol{\theta}^{(k)}) \right\|^2 + \frac{L\alpha_k^2}{2} \mathbb{E}\left[\left\| \nabla f_{i_k}(\boldsymbol{\theta}^{(k)}) \right\|^2 \mid \boldsymbol{\theta}^{(k)} \right].$$

进一步由假设 2，

$$\mathbb{E}[f(\boldsymbol{\theta}^{(k+1)}) \mid \boldsymbol{\theta}^{(k)}] \leq f(\boldsymbol{\theta}^{(k)}) - \left(\alpha_k - \frac{L\alpha_k^2}{2} \right) \left\| \nabla f(\boldsymbol{\theta}^{(k)}) \right\|^2 + \frac{L\alpha_k^2}{2} \sigma^2.$$

非凸情形：理论保证什么？

定理

在假设 *1* 和假设 *2* 下，若 $p^\star = \inf_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) > -\infty$，取固定步长 $\alpha_k = \alpha \in (0, \frac{1}{L}]$，则

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=0}^{K-1}\left\|\nabla f(\boldsymbol{\theta}^{(k)})\right\|^2\right] \le \alpha L\sigma^2 + \frac{2\big(f(\boldsymbol{\theta}^{(0)}) - p^\star\big)}{\alpha K}.$$

- 非凸情形通常不保证找到全局最优解；
- 这里控制的是历史迭代点的平均梯度范数平方，而不是最后一个点；
- 右侧有两项：噪声平台 $\alpha L\sigma^2$ 和随 K 下降的优化误差。

Batch size 如何进入理论?

若 mini-batch 大小为 n_b，并且样本独立均匀抽取（有放回），则随机梯度平均后方差降低。这里 σ^2 表示单样本随机梯度噪声的方差上界：

$$\mathbb{E}\left\| \frac{1}{n_b} \sum_{j=1}^{n_b} \nabla f_{i_{k,j}}(\boldsymbol{\theta}^{(k)}) \right\|^2 \leq \frac{\sigma^2}{n_b} + \left\| \nabla f(\boldsymbol{\theta}^{(k)}) \right\|^2.$$

若是不放回抽样，也有类似结论，只是方差项会带有限总体校正因子。

问题：增大 n_b 一定更好吗?

Batch size 如何进入理论?

若 mini-batch 大小为 n_b，并且样本独立均匀抽取（有放回），则随机梯度平均后方差降低。这里 σ^2 表示单样本随机梯度噪声的方差上界：

$$\mathbb{E}\left\| \frac{1}{n_b} \sum_{j=1}^{n_b} \nabla f_{i_{k,j}}(\boldsymbol{\theta}^{(k)}) \right\|^2 \leq \frac{\sigma^2}{n_b} + \left\| \nabla f(\boldsymbol{\theta}^{(k)}) \right\|^2.$$

若是不放回抽样，也有类似结论，只是方差项会带有限总体校正因子。

问题：增大 n_b 一定更好吗?

答案：增大 n_b 会降低梯度噪声，使更新更稳定；但每次更新更贵，单位时间内可完成的更新次数可能减少。所以 batch size 是稳定性和计算成本之间的折中。

可选对照：强凸时的更强结论

若额外假设 f 是 μ-强凸的，则可以讨论函数值误差 $f(\boldsymbol{\theta}^{(K)}) - p^\star$。

固定步长 $\alpha \in (0, \frac{1}{L}]$ 时，有典型形式

$$\mathbb{E}\big[f(\boldsymbol{\theta}^{(K)}) - p^\star\big] \le \frac{\alpha L \sigma^2}{2\mu} + (1 - \alpha\mu)^K \left(f(\boldsymbol{\theta}^{(0)}) - p^\star - \frac{\alpha L \sigma^2}{2\mu} \right).$$

若使用递减步长 $\alpha_k = \frac{\beta}{k+\gamma}$，可得到 $O(1/K)$ 量级的误差界。【具体可参见 Royer 笔记的 Theorem 2.3.2】

目录

课堂快速检查

1. 若 $M = 10^6$，为什么 full-batch GD 每一步可能很慢?
2. SGD 中的随机梯度为什么可以看作完整梯度的无偏估计?
3. 在非凸 SGD 定理中，$\frac{1}{K}\sum_{k=0}^{K-1}\left\|\nabla f(\boldsymbol{\theta}^{(k)})\right\|^2$ 变小说明什么?

课堂快速检查

1. 若 $M = 10^6$，为什么 full-batch GD 每一步可能很慢?
2. SGD 中的随机梯度为什么可以看作完整梯度的无偏估计?
3. 在非凸 SGD 定理中，$\frac{1}{K}\sum_{k=0}^{K-1}\left\|\nabla f(\boldsymbol{\theta}^{(k)})\right\|^2$ 变小说明什么?

答案： 1. 每步要扫描全部样本；2. 给定当前点并均匀采样时，单样本梯度的条件期望等于完整梯度；3. 说明平均意义下接近一阶驻点附近；固定步长时通常停在噪声平台，不等于保证全局最优。

总结

本节的主线：

- 数据驱动优化常写成 $f = \frac{1}{M}\sum_i f_i$；
- SGD 用随机样本梯度近似完整梯度，降低每步成本；
- mini-batch 通过平均多个样本梯度降低噪声；
- 非凸 SGD 理论通常保证平均梯度范数变小；
- batch size 是稳定性和计算成本之间的折中。

阅读材料

- Clement W. Royer 的随机梯度笔记：Lecture Notes on Stochastic Gradient Methods