《Linear Convergence Rate in Convex Setup is Possible!》一文的问题回答

姓名学号
姓名学号
李文耀 3230102302

概述. 我们由三人组队，回答了以下三个问题。

1 问题一：论文解决的问题

1.1 更优秀的梯度下降法分析

过去的研究要么依赖于更严格的假设，如依赖于一个额外的 L-smooth 条件、要求梯度下降带 Polyak 步长；要么分析结果过于粗糙，如分析结果依赖一个可能很大的常数；要么对于收敛速率研究不够细致，例如没有研究梯度下降刚开始时的线性收敛。

论文改进了现有对 (L_0, L_1)-GD、NGD 和 Clip-GD 在凸性和 (L_0, L_1)-smoothness 假设下的收敛性分析。这些方法在初始优化阶段（即当 $\|\nabla f(x^k)\| \geq \frac{L_0}{L_1}$ 时）无需任何额外假设即可实现线性收敛；随着迭代接近最优解，收敛速度逐渐转为次线性。

特别地，对于 NGD，当 $\|\nabla f(x^k)\| \geq c$ 时，仍保持线性收敛。论文提供表格给出了 Clip-GD 线性收敛的条件：其中 $\lambda_k = 1$ 对应 GD 的情形，$\lambda_k = \frac{c}{\|\nabla f(x^k)\|}$ 对应 NGD 的情形。

1.2 对于 RCD 和 OrderRCD 方法的收敛性分析扩展

论文首次为随机坐标下降方法 RCD 及 OrderRCD 在凸性和 (L_0, L_1)-coordinate smoothness 假设下提供了收敛性分析。结果表明，这两种方法同样展现出与 full-gradient 方法类似的“先线性、后次线性”的收敛现象。

1.3 强凸条件下的推广

此外，论文还将 (L_0, L_1)-GD 的分析推广至目标函数满足 μ-strongly convex 的情形，进一步拓展了该类非标准光滑条件下的优化理论。

2 问题二：文章的假设条件

2.1 形式化表述

2.1.1 L-smoothness（Lipschitz 梯度）

函数 $f:\mathbb{R}^d\to\mathbb{R}$ 被称为 L-smooth，如果其梯度 ∇f 是 L-Lipschitz 连续的，即对任意 $x,y\in\mathbb{R}^d$ 有

$$\|\nabla f(x)-\nabla f(y)\|\le L\|x-y\|$$

若其有二阶梯度，那么有

$$\|\nabla^2 f(x)\|\le L$$

2.1.2 (L_0,L_1)-smoothness

函数 f 满足 (L_0,L_1)-smoothness，如果对任意 $x,y\in\mathbb{R}^d$，满足 $\|x-y\|\le\frac{1}{L_1}$，有

$$\|\nabla f(x)-\nabla f(y)\|\le(L_0+L_1\|\nabla f(x)\|)\|x-y\|$$

2.2 假设区别

表 1: (L_0,L_1)-smoothness 假设和 L-smoothness 假设的不同

性质	L-smoothness 假设	(L_0,L_1)-smoothness 假设
梯度变化剧烈程度	全局固定 L	依赖当前位置梯度：$L_0+L_1\|\nabla f(x)\|$
严格程度	要求梯度变化一致有界	允许梯度变化与当前位置相关
强弱关系	更强	更弱

2.3 严格程度

(L_0,L_1)-smoothness 是比 L-smoothness 更弱、更宽松的条件，因为其允许梯度在一定情况下变化可以较快，但是 L-smoothness 强制要求梯度变化必须保持缓慢。

若 f 是 L-smooth，则对于 $\|\nabla f(x)-\nabla f(y)\|\le L\|x-y\|$ 取 $L_0=L,L_1=0$，显然满足 $\|\nabla f(x)-\nabla f(y)\|\le(L_0+L_1\|\nabla f(x)\|)\|x-y\|$，即 (L_0,L_1)-smoothness。

反之，存在函数满足 (L_0,L_1)-smoothness 但不满足任何 L-smoothness。例如，考虑 $f(x)=e^x$ 在 $x\in\mathbb{R}$ 时，其梯度为 e^x，也是 e^x，无全局上界，故不存在 L 使得 $\|\nabla f(x)-\nabla f(y)\|\le L\|x-y\|$，因而它是非 L-smooth 的。但可验证其满足 $(0,1)$-smoothness，因为

$$|f'(x)-f'(y)|=|e^x-e^y|\le e^{\max(x,y)}|x-y|\le(\underbrace{0}_{L_0}+\underbrace{1}_{L_1}\cdot|f'(x)|)|x-y|.$$

3 问题三：文章的不足和前景

3.1 文章不足

3.1.1 应用前景层面

假设过于严格造成的应用范围狭窄 文中自己提到，学界和工业界大多都集中于研究更普适的非全局凸函数，但是本文的研究均是全局凸函数的情形，因此其应用范围有限。

没有讨论其他类型的优化问题 在机器学习领域常用的优化器如 adam 方法相较于文章模型更加复杂，因而本文分析只能适用于较为狭窄的研究方向。

3.1.2 分析结果层面

分析细致程度有提升空间 “线性收敛”仅在梯度超过阈值时成立，非全局性质，只在：$\|\nabla f(x_k)\| \geq L_0/L_1$ 阶段有效。分析结果中，当靠近最优解时，主导收敛部分为次线性的 $O(1/N)$，仍有进一步改进的空间。

没有解决收敛速度过于宽松的问题 收敛因子涉及 $1 - \frac{1}{4L_1R}$ 此处 R 是初始点到最优点的距离。在高维或数据规模大时，R 常非常大，导致收敛因子接近于 1，线性收敛速度非常慢。

3.2 文章意义

3.2.1 有助于解释实际现象

解释现实机器学习中“初期快速下降”的普遍现象

3.2.2 有助于梯度下降算法变式的优化

例如，如何选择 clipping 半径；clipping 如何影响收敛速度；clipping 是否阻碍早期快速下降。

3.2.3 工程运用层面

工程实现层面 对于部分对参数同时调节要求不高，对于反应速度较为看重的应用场景，文章的分析结果的梯度下降和坐标下降可以为这些垂域提供重要理论支撑。