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

姓名学号
姓名学号
姓名学号

Resumo. Este é o padrão (formato LaTeX apenas) para a submissão de trabalhos da Categoria 2 do CNMAC, destinados à divulgação de pesquisas com resultados conclusivos. *Nesta categoria, os trabalhos devem ser submetidos em Português ou Inglês, em forma de artigo com no mínimo 5 e no máximo 7 páginas, incluindo-se as referências bibliográficas.* Os *trabalhos submetidos* que *não estiverem de acordo com o formato* apresentado por esse padrão *serão rejeitados* pelo Comitê Editorial do evento, sem análise do mérito científico.

Palavras-chave. Instruções, LaTeX, Trabalhos Completos, SBMAC, CNMAC (entre 3-6 palavras-chave)

1 问题一：论文解决的问题

1.1 更优秀的梯度下降法分析

过去的论文在凸性和 (L_0, L_1)-smoothness 假设下的研究要么依赖于更严格的假设，如 Koloskova 等研究者的成果依赖于一个额外的 L-smooth 条件，Takezawa 等研究者的成果要求梯度下降带 Polyak 步长；要么分析结果过于粗糙，如 Li 等研究者的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 Gorbunov 和 Vankov 等研究者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。展示了在凸性和 (L_0, L_1)-smoothness 假设下，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。论文提供的表 1 明确给出了 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. 梯度下降（GD）：$x^{k+1} = x^k - \eta_k \nabla f(x^k)$，其中步长 $\eta_k = \frac{1}{L_0 + L_1 \|\nabla f(x^k)\|}$。当梯度范

数 $\|\nabla f(x^k)\| \geq \frac{L_0}{L_1}$ 时，算法以线性速率收敛，满足 $f(x^N) - f^* \leq \left(1 - \frac{1}{4L_1R}\right)^N F_0$ ($F_0 = f(x^0) - f^*$，$R = \|x^0 - x^*\|$)。当梯度范数 $\|\nabla f(x^k)\| < \frac{L_0}{L_1}$ 时，收敛速率退化为标准次线性，满足 $f(x^N) - f^* < \frac{4L_0R^2}{N-T}$ （T 是首次进入次线性阶段的迭代次数)。

2. 归一化梯度下降（NGD)：$x^{k+1} = x^k - \eta_k \cdot \frac{\nabla f(x^k)}{\|\nabla f(x^k)\|}$，步长 $\eta_k = \eta \leq \frac{c}{L_0 + L_1 c}$ （c 是梯度范数下界，即 $\|\nabla f(x^k)\| \geq c$)。可以实现线性收敛 $f(x^N) - f^* \leq \left(1 - \frac{\eta}{2R}\right)^N F_0$。只要梯度范数不小于 c（前期迭代满足)，就能保持线性收敛；若 $c < \frac{L_0}{L_1}$，则收敛速率会下降，且最终精度受 c 限制

3. 裁剪梯度下降 (Clip-GD)：$x^{k+1} = x^k - \eta_k \cdot clip_c(\nabla f(x^k))$，其中裁剪操作 $clip_c(\nabla f(x^k)) = \min\left\{1, \frac{c}{\|\nabla f(x^k)\|}\right\} \nabla f(x^k)$，步长 $\eta_k = \frac{1}{L_0 + L_1 \cdot \min\{\|\nabla f(x^k)\|, c\}}$。收敛速率满足 $f(x^N) - f^* = \mathcal{O}\left(\min\left\{\frac{L_0R^2}{N-T}, \left(1 - \frac{\rho}{R}\right)^T F_0\right\}\right)$ ($\rho = \frac{c}{\max\{L_0, L_1 c\}}$)。梯度大于 c 时，按 NGD 模式；梯度小于 c 时，按 GD 模式。若 $c \geq \frac{L_0}{L_1}$，前期线性收敛速率与 GD 一致；若 $c < \frac{L_0}{L_1}$，前期线性收敛速率略低。

1.2 对于 RCD 和 OrderRCD 方法的收敛性分析扩展

过去的论文对于坐标下降类方法 RCD 的分析集中于 L-smoothness 假设，而对于 OrderRCD 方法，则没有给出任何分析。

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. RCD：当多数坐标的梯度范数 $\geq L_0/L_1$ 时，期望误差满足 $\mathbb{E}[f(x^N)] - f^* = \mathcal{O}\left(\left(1 - \frac{\rho}{dR}\right)^N F_0\right)$ ($\rho = 1/(4\sqrt{2}L_1)$，d 为维度，$F_0 = f(x^0) - f^*$) 。当梯度接近最优解时，误差退化为 $\mathcal{O}\left(\frac{dL_0R^2}{N}\right)$。当 $L_0 = 0$（强增长条件）时，可线性收敛到任意精度，迭代次数仅需 $O(dL_1R \log \frac{F_0}{\varepsilon})$，显著优于传统坐标下降在固定光滑性下的复杂度。

2. OrderRCD：具有与 RCD 完全一致的收敛速率——$\mathbb{E}[f(x^N)] - f^* = \mathcal{O}\left(\max\left\{\left(1 - \frac{\rho}{dR}\right)^N F_0, \frac{dL_0R^2}{N}\right\}\right)$

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，以下是一个详细证明。

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

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

因此，L-smoothness $\Rightarrow (L,0)$-smoothness，但逆命题不成立。所以，(L_0,L_1)-smoothness 是比 L-smoothness 更弱、更宽松的条件。

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。

4 模板留档

4.1 表格示例

表 2: Categorias dos trabalhos.

Categoria do trabalho	Número de páginas	Tipo do trabalho
1	2	A, B e C
2	entre 5 e 7	apenas C

图 1: Exemplo de imagem. Fonte: indicar.