Advances in Applied Mathematics 应用数学进展, 2025, 14(3), 117-122
Published Online March 2025 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2025.143097

Hans汉斯

基于Polyak步长的动量方法

张欣悦[1]，张欣彤[2*]

[1]河北工业大学理学院，天津
[2]山东建筑大学理学院，山东 济南

收稿日期：2025年2月11日；录用日期：2025年2月17日；发布日期：2025年3月11日

摘　要

近年来，动量方法广泛地应用在机器学习训练中。本文基于Polyak步长和移动平均动量(MAG)方法提出了一个新的动量方法(LAGP)，并将其与随机梯度结合，提出SLAGP方法。建立了LAGP方法在半强凸条件下的线性收敛性，以及SLAGP算法在半强凸条件下的线性收敛性。数值实验表明LAGP和SLAGP与其他流行算法相比有明显优势。

关键词

机器学习，动量方法，自适应步长

Polyak Step-Size for Momentum Method

Xinyue Zhang[1], Xintong Zhang[2*]

[1]School of Sciences, Hebei University of Technology, Tianjin
[2]School of Sciences, Shandong Jianzhu University, Jinan Shandong

Received: Feb. 11th, 2025; accepted: Feb. 17th, 2025; published: Mar. 11th, 2025

Abstract

Recently, momentum methods have been widely adopted in training machine learning. In this paper, based on the Polyak step-size and the Moving Average Gradient (MAG) method, a new momentum method (LAGP) is proposed. By combining it with the stochastic gradient, the SLAGP method is developed. The linear convergence of the LAGP method under the semi-strongly convex condition, and the linear convergence of the SLAGP algorithm under the semi-strongly convex condition are

*通讯作者。

文章引用: 张欣悦, 张欣彤. 基于 Polyak 步长的动量方法[J]. 应用数学进展, 2025, 14(3): 117-122.
DOI: 10.12677/aam.2025.143097

established. Numerical experiments show that LAGP and SLAGP have significant advantages compared with other popular algorithms.

Keywords

Machine Learning, Momentum Method, Adaptive Step-Size

1. 引言

机器学习被认为是人工智能(Artificial Intelligence, AI)领域中最前沿、最能够 体现智能、发展最快速的一个分支，它是人工智能的核心，是计算机获得“智能”的根本途径[1] [2]。机器学习中的大多数问题可归结为求解如下优化问题：

$$\min_{x\in\mathbb{R}^d} f(x)=\frac{1}{n}\sum_{i=1}^{n} f_i(x),$$

其中每个 $f_i(x)$ 是损失函数。移动平均梯度(MAG)广泛用于求解上述优化问题，其迭代格式为

$$m_k=\nabla f(x_k)+\beta m_{k-1}, x_{k+1}=x_k-\eta m_k,$$

其中 $\eta>0,\beta\in(0,1)$。当前主流的自适应动量方法如 Adam、RMSProp 等，通过结合历史梯度信息动态调整学习率，显著提升了非凸优化问题的训练效率。然而，这些方法在强凸或半强凸问题中可能存在收敛速度不足或超参数敏感性问题。MAG 方法通过引入移动平均梯度有效缓解了噪声干扰，但步长固定限制了其泛化能力。

2. LAGP 和 SLAGP

首先，基于 MAG 算法提出 LAG 算法

$$m_k=\beta_1\nabla f(x_k)+\beta_2 m_{k-1}, x_{k+1}=x_k-\eta m_k,$$

其中 $\eta>0,\beta_1>0,\beta_2\in(0,1)$。借鉴 Polyak 步长[3] [4]的思想，求解

$$\min_{\eta_k}\left\|x_{k+1}(\eta_k)-x^*\right\|^2,$$

得到

$$\eta_k=\frac{\left\langle m_k, x_k-x^*\right\rangle}{\left\|m_k\right\|^2},$$

其中 x^* 是不可得的。所以，根据 f 的凸性，

$$\left\|x_{k+1}-x^*\right\|^2\le\left\|x_k-x^*\right\|^2-2\beta_1\eta_k\left[f(x_k)-f^*\right]+\eta_k^2\left\|m_k\right\|^2$$

求解上式右边的最小值得到

$$\eta_k = \beta_1 \frac{f(x_k) - f^*}{\|m_k\|^2},$$

其中 $f^* = inff(x)$。结合 LAG 和上述步长提出 LAGP 算法，见算法 1。与 Adam 通过一阶矩和二阶矩估计调整步长不同，LAGP 直接利用 Polyak 步长公式动态计算最优步长，避免了二阶矩估计带来的计算开销。此外，LAGP 的半强凸收敛性保证使其在条件数较大的问题上更具鲁棒性。

算法 1：LAGP

输入：$x_0, \beta_1 > 0, \beta_2 \in (0,1), m_0 = 0, T.$

1：**for** $t = 1, 2, \cdots, T$ **do**

2：$m_k = \beta_1 \nabla f(x_k) + \beta_2 m_{k-1}.$

3：$\eta_k = \beta_1 \frac{f(x_k) - f^*}{\|m_k\|^2},$

4：$x_{k+1} = x_k - \eta_k m_k$

5：**end for**

输出：$x_T.$

3. 收敛性分析

3.1. 确定优化

首先，给出两个关键的引理。

引理 1 假设 $m_k = \beta_1 \nabla f(x_k) + \beta_2 m_{k-1}$，$m_{-1} = 0, \beta_1 > 0, \beta_2 \in (0,1)$。那么

$$\|m_k\|^2 \le \frac{\beta_1^2}{1-\beta_2} \sum_{j=0}^{k} \beta_2^{k-j} \|\nabla f(x_j)\|^2$$

且

$$\sum_{k=0}^{T} \|m_k\|^2 \le \frac{\beta_1^2}{(1-\beta_2)^2} \sum_{k=0}^{T} \|\nabla f(x_k)\|^2。$$

引理 2 令 f 是凸的，并且 x^* 是最优解，那么对于 LAGP 生成的 $\{x_i\}_{i=0}^{k}, \eta_i \le \beta_1 (f(x_i) - f^*) / \|m_i\|^2$，有 $\langle m_{k-1}, x_k - x^* \rangle \ge 0$。

进一步，如果 f 是 L-光滑的，那么

$$\langle m_{k-1}, x_k - x^* \rangle \ge \frac{\beta_1}{2L} \sum_{j=0}^{k-1} \beta_2^{k-1-j} \|\nabla f(x_j)\|^2。$$

定理 1 假设 f 是 L-光滑和 μ-半强凸的，那么 T 步后，LAGP 的迭代满足

$$\|x_T - x^*\|^2 \le \left(1 - \frac{(1-\beta_2)\mu}{2L}\right)^T \|x_0 - x^*\|^2。$$

证明：根据引理 1 和引理 2，

$$\begin{aligned}\left\|x_{k+1}-x^*\right\|^2&=\left\|x_k-x^*\right\|^2-2\left\langle\eta_k m_k,x_k-x^*\right\rangle+\left\|\eta_k m_k\right\|^2\\&=\left\|x_k-x^*\right\|^2-2\beta_1\eta_k\left\langle\nabla f\left(x_k\right),x_k-x^*\right\rangle-2\beta_2\eta_k\left\langle m_{k-1},x_k-x^*\right\rangle+\beta_1\eta_k\left(f\left(x_k\right)-f^*\right)\\&\le\left\|x_k-x^*\right\|^2-2\beta_1\eta_k\left(f\left(x_k\right)-f^*+\frac{1}{2L}\left\|\nabla f\left(x_k\right)\right\|^2\right)\\&\quad-2\beta_2\eta_k\left\langle m_{k-1},x_k-x^*\right\rangle+\beta_1\eta_k\left(f\left(x_k\right)-f^*\right)\\&\le\left\|x_k-x^*\right\|^2-\beta_1\eta_k\left(f\left(x_k\right)-f^*\right)-\frac{\beta_1\eta_k}{L}\left\|\nabla f\left(x_k\right)\right\|^2-2\beta_2\eta_k\frac{\beta_1}{2L}\sum_{j=0}^{k-1}\beta_2^{k-1-j}\left\|\nabla f\left(x_j\right)\right\|^2\\&=\left\|x_k-x^*\right\|^2-\beta_1\eta_k\left(f\left(x_k\right)-f^*\right)-\frac{\beta_1\eta_k}{L}\sum_{j=0}^{k}\beta_2^{k-j}\left\|\nabla f\left(x_j\right)\right\|^2\\&=\left\|x_k-x^*\right\|^2-\beta_1\eta_k\left(f\left(x_k\right)-f^*\right)-\frac{\beta_1^2\left(f\left(x_k\right)-f^*\right)}{L\left\|m_k\right\|^2}\sum_{j=0}^{k}\beta_2^{k-j}\left\|\nabla f\left(x_j\right)\right\|^2\\&\le\left\|x_k-x^*\right\|^2-\left(\beta_1\eta_k+\frac{1-\beta_2}{L}\right)\left(f\left(x_k\right)-f^*\right)。\end{aligned}$$

最后，根据 f 的 μ-半强凸性，

$$\left\|x_{k+1}-x^*\right\|^2\le\left(1-\frac{\mu}{2}\left(\eta_k+\frac{1-\beta_2}{L}\right)\right)\left\|x_k-x^*\right\|^2\le\left(1-\frac{(1-\beta_2)\mu}{2L}\right)\left\|x_k-x^*\right\|^2。$$

证毕。

3.2. 随机优化

定义

$$f_{s_k}\left(x\right)=\frac{1}{\left|s_k\right|}\sum_{i\in s_k}f\left(x;\xi_i\right),\nabla f_{s_k}\left(x\right)=\frac{1}{\left|s_k\right|}\sum_{i\in s_k}\nabla f\left(x;\xi_i\right)$$

并且 $\mathbb{E}f_{s_k}\left(x\right)=f\left(x\right)$ 和 $\left\|x_{k+1}-x^*\right\|^2\le\left(1-\frac{\mu}{2}\left(\eta_k+\frac{1-\beta_2}{L}\right)\right)\left\|x_k-x^*\right\|^2\le\left(1-\frac{(1-\beta_2)\mu}{2L}\right)\left\|x_k-x^*\right\|^2$，以及 $f_{s_k}^*=inf_x f_{s_k}\left(x\right)$。

结合随机梯度提出 SLAGP，

$$m_k=\beta_1\nabla f_{s_k}\left(x_k\right)+\beta_2 m_{k-1},$$

$$\eta_k=\min\left\{\beta_1\frac{f\left(x_k\right)-f_{s_k}^*}{c\left\|m_k\right\|^2},\eta_{\max}\right\},$$

$$x_{k+1}=x_k-\eta_k m_k,$$

其中 $c>0,\eta_{\max}>0$。c 用于控制随机梯度步长的衰减速率，$\eta_{\max}$ 为避免步长过大导致发散的经验阈值。

假设 1 $\sigma^2=\mathbb{E}\left[f_{s_k}\left(x^*\right)-f_{s_k}^*\right]\le\infty$。

引理 3 对于凸函数，如果步长 $\eta_k\le\beta_1\left(f_{s_k}\left(x_k\right)-f_{s_k}^*\right)/\left\|m_k\right\|^2$，那么对于任意的 k 大于等于 1，LAGP 的迭代满足

$$\left\langle m_{k-1},x_k-x^*\right\rangle\ge\beta_1\left(1-\frac{1}{c}\right)\sum_{j=0}^{k-1}\beta_2^{k-1-j}\left(f_{s_j}\left(x_j\right)-f_{s_j}^*\right)-\beta_1\sum_{j=0}^{k-1}\beta_2^{k-1-j}\left(f_{s_j}\left(x^*\right)-f_{s_j}^*\right)。$$

定理 2 假设 f_i 是 L-光滑和 μ-半强凸的，$\eta=\max\left\{\beta_1\dfrac{f_{s_k}(x_k)-f_{s_k}^*}{\|m_k\|^2}\right\}_{k=0}^{T}<\infty$，那么 T 步后，SLAGP 的迭代满足

$$\mathbb{E}\|x_T-x^*\|^2\le(1-a_1)^T\|x_0-x^*\|^2+\frac{2\beta_1\eta\sigma^2}{a_1(1-\beta_2)},$$

其中 $a_1=\dfrac{(1-\beta_2)(c-1)\mu}{2c^2L}$。

证明：

$$\begin{aligned}\|x_{k+1}-x^*\|^2&=\|x_k-x^*\|^2-2\eta_k\langle m_k,x_k-x^*\rangle+\|\eta_k m_k\|^2\\&=\|x_k-x^*\|^2-2\eta_k\langle\beta_1\nabla f_{s_k}(x_k)+\beta_2 m_{k-1},x_k-x^*\rangle+\|\eta_k m_k\|^2\\&\le\|x_k-x^*\|^2-2\beta_1\eta_k\left(f_{s_k}(x_k)-f_{s_k}(x^*)\right)-2\beta_2\eta_k\langle m_{k-1},x_k-x^*\rangle+\|\eta_k m_k\|^2\\&\le\|x_k-x^*\|^2-\beta_1\eta_k\left(2-\frac{1}{c}\right)\left(f_{s_k}(x_k)-f_{s_k}^*\right)+2\beta_1\eta_k\left(f_{s_k}(x^*)-f_{s_k}^*\right)\\&\le\|x_k-x^*\|^2-2\beta_1\eta_k\left(1-\frac{1}{c}\right)\sum_{j=0}^{k}\beta_2^{k-j}\left(f_{s_j}(x_j)-f_{s_j}^*\right)+2\beta_1\eta_k\sum_{j=0}^{k}\beta_2^{k-j}\left(f_{s_j}(x^*)-f_{s_j}^*\right)。\end{aligned}$$

根据光滑性，

$$\sum_{j=0}^{k}\beta_2^{k-j}\left(f_{s_j}(x_j)-f_{s_j}^*\right)\ge\frac{1}{2L}\sum_{j=0}^{k}\beta_2^{k-j}\|\nabla f_{s_j}(x_j)\|^2。$$

又根据引理 3，

$$\begin{aligned}\eta_k\sum_{j=0}^{k}\beta_2^{k-j}\left(f_{s_j}(x_j)-f_{s_j}^*\right)&=\frac{\beta_1\left(f_{s_k}(x_k)-f_{s_k}^*\right)}{c\|m_k\|^2}\sum_{j=0}^{k}\beta_2^{k-j}\left(f_{s_j}(x_j)-f_{s_j}^*\right)\\&\ge\frac{(1-\beta_2)\left(f_{s_k}(x_k)-f_{s_k}^*\right)}{c\beta_1\sum_{j=0}^{k}\beta_2^{k-j}\|\nabla f(x_j)\|^2}\frac{1}{2L}\sum_{j=0}^{k}\beta_2^{k-j}\|\nabla f_{s_j}(x_j)\|^2\\&=\frac{(1-\beta_2)\left(f_{s_k}(x_k)-f_{s_k}^*\right)}{2cL\beta_1}。\end{aligned}$$

那么，

$$\begin{aligned}\|x_{k+1}-x^*\|^2&\le\|x_k-x^*\|^2-\frac{(c-1)(1-\beta_2)}{c^2L}\left(f_{s_k}(x_k)-f_{s_k}^*\right)+2\beta_1\eta_k\sum_{j=0}^{k}\beta_2^{k-j}\left(f_{s_j}(x^*)-f_{s_j}^*\right)\\&\le\left(1-\frac{(c-1)(1-\beta_2)\mu}{2c^2L}\right)\|x_k-x^*\|^2+2\beta_1\eta\sum_{j=0}^{k}\beta_2^{k-j}\left(f_{s_j}(x^*)-f_{s_j}^*\right)。\end{aligned}$$

对上述不等式取期望，

$$\begin{aligned}\mathbb{E}\|x_{k+1}-x^*\|^2&\le\left(1-\frac{(c-1)(1-\beta_2)\mu}{2c^2L}\right)\mathbb{E}\|x_k-x^*\|^2+2\beta_1\eta\sum_{j=0}^{k}\beta_2^{k-j}\sigma^2\\&\le\left(1-\frac{(c-1)(1-\beta_2)\mu}{2c^2L}\right)^{k+1}\|x_0-x^*\|^2+\frac{4c^2L\beta_1\eta\sigma^2}{(c-1)(1-\beta_2)^2\mu}。\end{aligned}$$

证毕。

4. 数值实验

4.1. 最小二乘问题

考虑最小二乘问题：

$$f(x)=\frac{1}{2}\|Ax-b\|^2$$

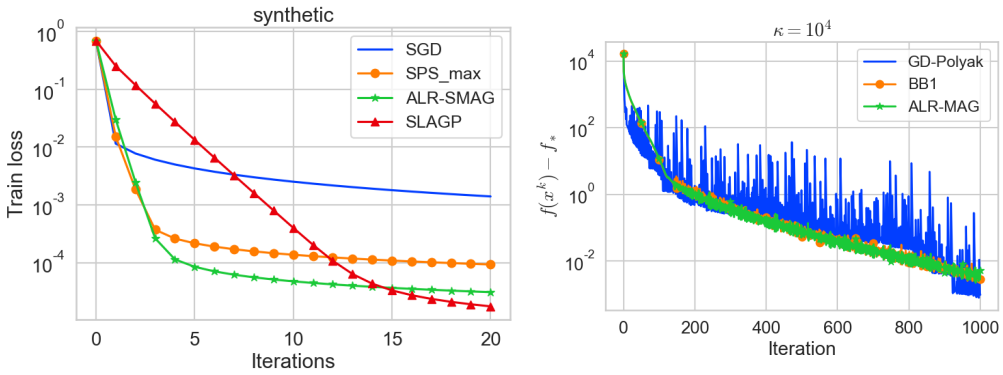
其中 $A\in\mathbb{R}^{d_1\times d}$ 是正定的， $b\in\mathbb{R}^{d_1}$ 。设置 $d_1=d=1000$ 文以及 A^TA 的条件数 κ 是 10,000。下面是与一些流行算法的对比，算法 LAGP 是明显优于其他算法的。

4.2. 逻辑回归问题

考虑逻辑回归损失：

$$f_{Log\,Reg}(w)=\frac{1}{n}\sum_{i=1}^{n}\log\left(1+\exp\left(-y_ix_i^Tw\right)\right)$$

其中 $x_i\in\mathbb{R}^d, y_i\in\{-1,+1\}$ 。下面是在合成数据集上的算法对比图。实验参数设置如下： $c=5,\eta_{\max}=100$ ，与[4]中保持一致以确保公平性。



5. 总结

基于 MAG 算法提出 LAG 算法，并基于 Polyak 步长为 LAG 设计自适应步长，提出 LAGP 算法。计算大规模机器学习问题时，借助随机梯度，应用 SLAGP 算法。分析了不同条件下的算法收敛性，数值实验证明了有效性。实验结果表明，LAGP 在最小二乘和逻辑回归任务中均显著优于一些流行的自适应算法。这表明基于 Polyak 步长的自适应策略能有效平衡方向更新与步长调整，为大规模机器学习优化提供了新思路。

参考文献

[1] Joshi, A.V. (2020) Machine Learning and Artificial Intelligence. Springer International Publishing. https://doi.org/10.1007/978-3-030-26622-6

[2] 马宪民. 人工智能的原理与方法[M]. 西安: 西北工业大学出版社, 2002.

[3] Polyak, B.T. (1987) Introduction to Optimization, Optimization Software, Inc., New York.

[4] Wang, X., Johansson, M. and Zhang, T. (2023) Generalized Polyak Step Size for First Order Optimization with Momentum. *International Conference on Machine Learning* (*PMLR*), Honolulu, HI, 35836-35863.