Advances in Applied Mathematics 应用数学进展, 2024, 13(3), 1008-1017
Published Online March 2024 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2024.133095

基于1/t-Polyak步长的随机控制的随机梯度算法

刘晨晨

河北工业大学理学院，天津

收稿日期：2024年2月27日；录用日期：2024年3月21日；发布日期：2024年3月27日

摘　要

随机梯度下降算法已成为求解大规模有限和优化问题的流行算法，然而，由于其在迭代过程中会产生方差，导致了振荡现象。随机控制的随机梯度(SCSG)算法缩减了该方差，但SCSG算法对于步长有较强的限制。为了扩大SCSG算法的步长选择范围，基于1/t-带步长与Polyak步长，提出1/t-Polyak步长，并将其与SCSG算法结合，提出SCSGP算法。建立了SCSGP算法在强凸条件下的线性收敛性，数值实验表明SCSGP算法与其他随机梯度类算法相比有明显优势。

关键词

有限和优化，随机算法，方差缩减，1/t-带步长

1/t-Polyak Stepsize for the Stochastically Controlled Stochastic Gradient Algorithm

Chenchen Liu

School of Sciences, Hebei University of Technology, Tianjin

Received: Feb. 27th, 2024; accepted: Mar. 21st, 2024; published: Mar. 27th, 2024

Abstract

The stochastic gradient descent algorithm has become popular algorithm for solving large-scale finite-sum optimization problems. However, this algorithm leads to oscillations due to the variance in the iterative process. The stochastically controlled stochastic gradient (SCSG) algorithm reduces this variance, but the SCSG algorithm has strong limit on stepsize. To expand the range of stepsize selection of the SCSG algorithm, we propose 1/t-Polyak stepsize by combining the 1/t-band stepsize and the Polyak stepsize. Using this new stepsize for the stochastically controlled stochastic gradient (SCSG) algorithm, the SCSGP algorithm is proposed. We establish the linear convergence rate of SCSGP for strongly convex problems. Numerical experiments demonstrate a clear

文章引用: 刘晨晨. 基于 1/t-Polyak 步长的随机控制的随机梯度算法[J]. 应用数学进展, 2024, 13(3): 1008-1017.
DOI: 10.12677/aam.2024.133095

advantage of SCSGP over other stochastic gradient algorithms.

Keywords

Finite-Sum Optimization, Stochastic Algorithms, Variance Reduction, 1/t-Band Stepsize

1. 引言

考虑有限和优化问题：

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

其中分量函数 $f_i(x)$ 连续可微，假设 $f(x)$ 是强凸的。机器学习中满足条件的优化问题有很多，例如带 ℓ_2 正则项的逻辑回归问题和带 ℓ_2 正则项的最小平方回归问题等[1] [2] [3]。

当数据规模过大时，随机梯度下降(SGD)算法[4]是求解问题(1)的主流算法，即用随机梯度估计全梯度，其迭代格式为

$$x_{t+1}=x_t-\eta_t\nabla f_{i_t}(x_t),$$

其中 $\eta_t>0$ 是步长，$\nabla f_{i_t}(x_t)$ 是分量函数 $f_{i_t}(x)$ 在 x_t 处的梯度。随机梯度 $\nabla f_{i_t}(x_t)$ 与全梯度 $\nabla f(x_t)$ 之间的方差导致 SGD 即使在强凸条件下，也只能达到次线性收敛速度[5]。方差缩减梯度(SVRG)算法[6]通过内外两层循环达到缩减方差的目的，但由于其在外循环中需要计算全梯度且内循环次数较大，导致数据规模过大时计算量大。为了改善这个问题，SCSG [7]令内循环次数服从几何分布且在外循环中计算批量梯度

$$\tilde{g}=\frac{1}{|I_t|}\sum_{i\in I_t}\nabla f_i(\tilde{x}),$$

其中 $I_t\subset[n]$，$|I_t|$ 为 I_t 的批量大小，$\tilde{x}$ 为在外循环中设置的快照点。在内循环中，SCSG 用与 SVRG 相同的格式更新梯度估计量：

$$g_t=\nabla f_{i_t}(x_t)-\nabla f_{i_t}(\tilde{x})+\tilde{g}\text{。}$$

在强凸条件下，其使用固定批量可线性收敛到解的邻域。SCSG 适用于求解大规模 $n\in\left[10^4,10^9\right]$、低精度 $\varepsilon\in\left[10^{-4},10^{-2}\right]$ 的优化问题[7] [8] [9]，可以经过很少的有效循环次数收敛到上述目标精度。

步长是保证随机梯度类算法收敛的关键因素，很小的常数步长和衰减步长都会使算法收敛缓慢，并且手动调整常数步长的过程相当耗时[10] [11] [12]。Polyak 步长[13]利用迭代过程中产生的函数值和梯度自动地计算步长，避免了手动调整的过程，其计算公式为

$$\eta_t=2\frac{f(x_t)-f^*}{\left\|\nabla f(x_t)\right\|^2},$$

其中 f^* 是 $f(x)$ 的极小值。为了将 Polyak 步长与随机梯度类算法结合，Loizou 等人[14]提出 Polyak 步长的随机版本(SPS)：

$$\eta_t = 2\frac{f_{i_t}(x_t) - f_{i_t}^*}{\left\| \nabla f_{i_t}(x_t) \right\|^2},$$

其中 $f_{i_t}^*$ 是 $f_{i_t}(x)$ 的极小值。SGD 结合 SPS 步长比结合固定步长数值表现好。当 SPS 步长中 $f_{i_t}^*$ 不易求解时，可用一个下界 $\ell_{i_t}^* \le f_{i_t}^*$ 来替换[15]。最近，Wang 等人[16]介绍了 1/t-带步长，其允许步长在一定范围内扰动，具体格式为

$$\frac{m}{t} \le \eta_t \le \frac{M}{t}, \quad \forall t \ge 1,$$

其中 $m \le M$ 是正常数。显然，衰减步长 $\eta_t = \eta_0 / t$ 是 1/t-带步长的特殊情况。

受 1/t-带步长和 Polyak 步长启发提出 1/t-Polyak 步长，并将其与 SCSG 结合提出新的算法——SCSGP。在强凸光滑的条件下，SCSGP 结合变化的批量可达到线性收敛速度。数值实验结果表明 SCSGP 比 SCSG 及其他随机梯度类算法表现好。

论文其余部分概括如下：在第 2 部分中提出 1/t-Polyak 步长并描述 SCSGP 算法。收敛性分析在第 3 部分。在第 4 部分中设置了数值实验。最后在第 5 部分进行总结。

2. 1/t-Polyak 步长与 SCSGP 算法

首先，利用 Polyak 步长的随机版本并将其与 1/t-带步长结合，提出 1/t-Polyak 步长：

$$\eta_t = \begin{cases} m/t, & \bar{\eta}_t^P \le m/t; \\ \bar{\eta}_t^P, & m/t < \bar{\eta}_t^P < M/t; \\ M/t, & \bar{\eta}_t^P \ge M/t, \end{cases} \tag{2}$$

其中 $\bar{\eta}_t^P$ 形式如下：

$$\bar{\eta}_t^P = c_B \frac{f_{I_t}(\tilde{x}_{t-1}) - \ell_{I_t}^*}{\left\| \nabla f_{I_t}(\tilde{x}_{t-1}) \right\|^2},$$

其中 $\ell_{I_t}^*$ 为批量函数 $f_{I_t}(x)$ 的极小值，c_B 用于调整步长的范围，由用户给定。对于一些非负的损失函数，根据批量函数的定义可取 $\ell_{I_t}^* = 0$。结合 1/t-Polyak 步长和 SCSG 算法提出 SCSGP 算法，见算法 1。

算法 1： SCSGP

```
输入：$\tilde{x}_0, B_1, b_1, T$.
for $t = 1, 2, \cdots, T$ do
    随机选取批量 $I_t \subset [n]$，其中 $|I_t| = B_t$.
    $\tilde{v}_t = \nabla f_{I_t}(\tilde{x}_{t-1})$.
    $x_0^{(t)} = \tilde{x}_{t-1}$.
    生成 $N_t \sim Geom(\gamma_t)$.
    利用(2)计算 $\eta_t$.
    for $k = 1, 2, \cdots, N_t$ do
        随机选取小批量 $\tilde{I}_{k-1} \in [n]$，其中 $|\tilde{I}_{k-1}| = b_t$.
```

续表

```
            $v_{k-1}^{(t)} = \nabla f_{\tilde{I}_{k-1}}\left(x_{k-1}^{(t)}\right) - \nabla f_{\tilde{I}_{k-1}}\left(x_0^{(t)}\right) + \tilde{v}_t$.
            $x_k^{(t)} = x_{k-1}^{(t)} - \eta_t v_{k-1}^{(t)}$.
        end for
        $\tilde{x}_t = x_{N_t}^{(t)}$.
    end for
输出：$\tilde{x}_T$.
```

在 SCSGP 算法的第 t 次外循环中，内循环次数 $N_t \sim Geom(\gamma_t)$ 是非负的几何随机变量，其概率分布为 $P(N_t = k) = \gamma_t^k (1-\gamma_t^k)$，$\forall k = 0,1,\cdots$。值得注意的是，SCSG 中 γ_t 取固定值，但在 SCSGP 中 γ_t 随迭代次数变化。若 $\gamma_t = B_t / (B_t + b_t)$，则有

$$\mathrm{E}_{N_t \sim Geom(\gamma_t)} = \frac{\gamma_t}{1-\gamma_t} = \frac{B_t}{b_t},$$

其中 E_{N_t} 记为对 N_t 取期望。该性质在后续分析中起到重要作用。另外不难发现 $v_k^{(t)}$ 是梯度估计量 $\nabla f(x_k)$ 的有偏估计：

$$\mathrm{E}_{\tilde{I}_k} v_k^{(t)} = \nabla f\left(x_k^{(t)}\right) - \nabla f\left(\tilde{x}_{t-1}\right) + \nabla f_{I_t}\left(\tilde{x}_{t-1}\right) = \nabla f\left(x_k^{(t)}\right) + e_t, \tag{3}$$

其中 $e_t = \nabla f_{I_t}\left(\tilde{x}_{t-1}\right) - \nabla f\left(\tilde{x}_{t-1}\right)$。

3. 收敛性分析

由于几何随机变量 N_t 在收敛分析中占据重要地位，需要给出下面关键的引理。

引理 1 [8]由于 $N_t \sim Geom(\gamma_t)$，其中 $\gamma_t > 0$，则对任意满足 $\mathrm{E}\left|D_{N_t}\right| < \infty$ 的序列 $\{D_n\}$ 有

$$\mathrm{E}\left(D_{N_t} - D_{N_t+1}\right) = \left(\frac{1}{\gamma_t} - 1\right)\left(D_0 - \mathrm{E}D_{N_t}\right),$$

其中 E 记为对所有随机变量取期望。

记 $\gamma = \min_t \gamma_t$，则对任意 $t \in [T]$ 有

$$\mathrm{E}\left(D_{N_t} - D_{N_t-1}\right) \le \left(\frac{1}{\gamma} - 1\right)\left(D_0 - \mathrm{E}D_{N_t}\right)。 \tag{4}$$

为了应用(3)，需要证明用到的相关序列 $\{D_n\}$ 满足 $\mathrm{E}\left|D_{N_t}\right| < \infty$。下面引理保证了该性质。

引理 2 假设 $f_i(x)$ 是 L-光滑的，令 $\frac{ML}{t} \le \frac{1}{3}\left(\frac{b_t}{B_t}\right)^{2/3}$ 且 $B_t \ge 8b_t$，则对任意 $t \ge 1$，$\mathrm{E}\left\|\tilde{x}_t - \tilde{x}_{t-1}\right\|^2 < \infty$，$\mathrm{E}\left[f\left(\tilde{x}_t\right) - f^*\right] < \infty$，$\mathrm{E}\left\|\nabla f\left(\tilde{x}_t\right)\right\|^2 < \infty$，$\mathrm{E}\left|\left\langle e_t, \tilde{x}_t - \tilde{x}_{t-1}\right\rangle\right| < \infty$，$\mathrm{E}\left|\left\langle e_t, \nabla f\left(\tilde{x}_t\right)\right\rangle\right| < \infty$。

证明：因为 $f_i(x)$ 是 L-光滑的和(3)，可得

$$\begin{aligned} \mathrm{E}_{\tilde{I}_k} f\left(x_{k+1}^{(t)}\right) &\le f\left(x_k^{(t)}\right) - \eta_t \left\langle \mathrm{E}_{\tilde{I}_k} v_k^{(t)}, \nabla f\left(x_k^{(t)}\right)\right\rangle + \frac{L\eta_t^2}{2}\mathrm{E}_{\tilde{I}_k}\left\|v_k^{(t)}\right\|^2 \\ &= f\left(x_k^{(t)}\right) - \eta_t \left\|\nabla f\left(x_k^{(t)}\right)\right\|^2 - \eta_t \left\langle e_t, \nabla f\left(x_k^{(t)}\right)\right\rangle + \frac{L\eta_t^2}{2}\mathrm{E}_{\tilde{I}_k}\left\|v_k^{(t)}\right\|^2 \\ &\le f\left(x_k^{(t)}\right) - \eta_t\left(1 - L\eta_t\right)\left\|\nabla f\left(x_k^{(t)}\right)\right\|^2 - \eta_t \left\langle e_t, \nabla f\left(x_k^{(t)}\right)\right\rangle + \frac{L^3\eta_t^2}{2b_t}\left\|x_k^{(t)} - x_0^{(t)}\right\|^2 + L\eta_t^2\left\|e_t\right\|^2, \end{aligned} \tag{5}$$

其中最后一个不等式利用了[8]中引理 B.2。由于对任意$c>0$有$2\langle a,b\rangle \le \frac{\|a\|^2}{c}+c\|b\|^2$，令$c=2$，则有

$$\eta_t\left\langle e_t,-\nabla f\left(x_k^t\right)\right\rangle \le \frac{1}{4}\eta_t\left\|\nabla f\left(x_k^{(t)}\right)\right\|^2+\eta_t\|e_t\|^2 。\tag{6}$$

因为$\eta_t \le \frac{M}{t}$，$\frac{ML}{t}\le \frac{1}{3}\left(\frac{b_t}{B_t}\right)^{2/3}$且$B_t \ge 8b_t$，可知$\frac{3}{4}-L\eta_t>0$。由(5)和(6)得到

$$\begin{aligned}\left\|\nabla f\left(x_k^{(t)}\right)\right\|^2 \le{}& \frac{1}{\eta_t\left(\frac{3}{4}-L\eta_t\right)}\left(f\left(x_k^{(t)}\right)-\mathrm{E}_{\tilde{I}_k} f\left(x_{k+1}^{(t)}\right)\right)+\frac{1+L\eta_t}{\frac{3}{4}-L\eta_t}\|e_t\|^2\\ &+\frac{L^3\eta_t}{2b_t\left(\frac{3}{4}-L\eta_t\right)}\left\|x_k^{(t)}-x_0^{(t)}\right\|^2.\end{aligned}\tag{7}$$

注意到$x_{k+1}^{(t)}=x_k^{(t)}-\eta_t v_k^{(t)}$，用类似(5)的推导过程可得

$$\begin{aligned}\mathrm{E}_{\tilde{I}_k}\left\|x_{k+1}^{(t)}-x_0^{(t)}\right\|^2 &=\left\|x_k^{(t)}-x_0^{(t)}\right\|^2-2\eta_t\left\langle \mathrm{E}_{\tilde{I}_k} v_k^{(t)},x_k^{(t)}-x_0^{(t)}\right\rangle+\eta_t^2\mathrm{E}_{\tilde{I}_k}\left\|v_k^{(t)}\right\|^2\\ &=\left\|x_k^{(t)}-x_0^{(t)}\right\|^2-2\eta_t\left\langle\nabla f\left(x_k^{(t)}\right),x_k^{(t)}-x_0^{(t)}\right\rangle-2\eta_t\left\langle e_t,x_k^{(t)}-x_0^{(t)}\right\rangle+\eta_t^2\mathrm{E}_{\tilde{I}_k}\left\|v_k^{(t)}\right\|^2\\ &\le\left(1+\frac{\eta_t^2L^2}{b_t}\right)\left\|x_k^{(t)}-x_0^{(t)}\right\|^2-2\eta_t\left\langle\nabla f\left(x_k^{(t)}\right),x_k^{(t)}-x_0^{(t)}\right\rangle-2\eta_t\left\langle e_t,x_k^{(t)}-x_0^{(t)}\right\rangle\\ &\quad+2\eta_t^2\left\|\nabla f\left(x_k^{(t)}\right)\right\|^2+2\eta_t^2\|e_t\|^2.\end{aligned}\tag{8}$$

再次使用$2\langle a,b\rangle \le \frac{\|a\|^2}{c}+c\|b\|^2$并取$c=\frac{b_t}{8\eta_t^2B_t}$，则有

$$\left\langle -2\eta_t\nabla f\left(x_k^{(t)}\right),x_k^{(t)}-x_0^{(t)}\right\rangle \le \frac{8\eta_t^2B_t}{b_t}\left\|\nabla f\left(x_k^{(t)}\right)\right\|^2+\frac{b_t}{8B_t}\left\|x_k^{(t)}-x_0^{(t)}\right\|^2,$$

$$\left\langle -2\eta_t e_t,x_k^{(t)}-x_0^{(t)}\right\rangle \le \frac{8\eta_t^2B_t}{b_t}\|e_t\|^2+\frac{b_t}{8B_t}\left\|x_k^{(t)}-x_0^{(t)}\right\|^2 。$$

将上述不等式和(7)代入(8)得到

$$\begin{aligned}\mathrm{E}_{\tilde{I}_k}\left\|x_{k+1}^{(t)}-x_0^{(t)}\right\|^2 \le{}& \left(1+\frac{b_t}{4B_t}+\frac{3\eta_t^2L^2/2+8\eta_t^3L^3B_t/b_t}{2b_t\left(3/4-\eta_tL\right)}\right)\left\|x_k^{(t)}-x_0^{(t)}\right\|^2+\left(2\eta_t^2+\frac{8\eta_t^2B_t}{b_t}\right)\left(1+\frac{1+\eta_tL}{3/4-\eta_tL}\right)\|e_t\|^2\\ &+\frac{2\eta_t+8\eta_tB_t/b_t}{3/4-\eta_tL}\left(f\left(x_k^{(t)}\right)-\mathrm{E}_{\tilde{I}_k} f\left(x_{k+1}^{(t)}\right)\right).\end{aligned}\tag{9}$$

由$\eta_tL\le\frac{1}{3}\left(\frac{b_t}{B_t}\right)^{2/3}$和$B_t\ge 8b_t$可得

$$\begin{aligned}\frac{3\eta_t^2L^2/2+8\eta_t^3L^3B_t/b_t}{2b_t\left(3/4-\eta_tL\right)} &\le \frac{(1/6)\times\left(b_t/B_t\right)^{4/3}+(8/27)\times\left(b_t/B_t\right)}{2b_t\left(3/4-(1/3)\times\left(b_t/B_t\right)^{2/3}\right)}\\ &\le\frac{1/12B_t+8/27B_t}{2(3/4-1/12)}=\frac{41}{144B_t}\le\frac{7b_t}{24B_t}.\end{aligned}$$

结合上式和(9)有

$$\begin{aligned}\mathrm{E}_{\tilde{I}_k}\left\|x_{k+1}^{(t)}-x_0^{(t)}\right\|^2 &\le \left(1+\frac{b_t}{4B_t}+\frac{7b_t}{24B_t}\right)\left\|x_k^{(t)}-x_0^{(t)}\right\|^2+\left(2+\frac{8B_t}{b_t}\right)\left(1+\frac{1+(1/3)\times(b_t/B_t)^{2/3}}{3/4-(1/3)\times(b_t/B_t)^{2/3}}\right)\eta_t^2\left\|e_t\right\|^2\\&\quad+\frac{2\eta_t+8\eta_t B_t/b_t}{3/4-(1/3)\times(b_t/B_t)^{2/3}}\left(f\left(x_k^{(t)}\right)-\mathrm{E}_{\tilde{I}_k}f\left(x_{k+1}^{(t)}\right)\right)\\&\le\left(1+\frac{13b_t}{24B_t}\right)\left\|x_k^{(t)}-x_0^{(t)}\right\|^2+\left(\frac{21}{4}+\frac{21B_t}{b_t}\right)\eta_t^2\left\|e_t\right\|^2+\left(3+\frac{12B_t}{b_t}\right)\eta_t\left(f\left(x_k^{(t)}\right)-\mathrm{E}_{\tilde{I}_k}f\left(x_{k+1}^{(t)}\right)\right).\end{aligned}\tag{10}$$

为了证明 $\mathrm{E}\left[f\left(x_k^{(t)}\right)-f^*\right]$ 和 $\mathrm{E}\left\|x_{k+1}^{(t)}-x_0^{(t)}\right\|^2$ 的上界，记

$$G_k^{(t)}=\left(3+\frac{12B_t}{b_t}\right)\eta_t\mathrm{E}\left[f\left(x_k^{(t)}\right)-f^*\right]+\mathrm{E}\left\|x_k^{(t)}-x_0^{(t)}\right\|^2.$$

对(10)取全期望得到

$$\begin{aligned}G_{k+1}^{(t)}&\le G_k^{(t)}+\frac{13b_t}{24B_t}\mathrm{E}\left\|x_k^{(t)}-x_0^{(t)}\right\|^2+\left(\frac{21}{4}+\frac{21B_t}{b_t}\right)\eta_t^2\mathrm{E}\left\|e_t\right\|^2\\&\le\left(1+\frac{13b_t}{24B_t}\right)\left(G_k^{(t)}+\left(\frac{21}{4}+\frac{21B_t}{b_t}\right)\eta_t^2\mathrm{E}\left\|e_t\right\|^2\right)\\&\le\left(1+\frac{13b_t}{24B_t}\right)^k\left(G_0^{(t)}+\left(\frac{21}{4}+\frac{21B_t}{b_t}\right)\eta_t^2\mathrm{E}\left\|e_t\right\|^2\right).\end{aligned}$$

由 $N_t\sim Geom\left(\frac{B_t}{B_t+b_t}\right)$ 可得

$$P\left(N_t=k\right)=\frac{b_t}{B_t+b_t}\left(\frac{B_t}{B_t+b_t}\right)^k\le\left(\frac{B_t}{B_t+b_t}\right)^k,$$

$$\mathrm{E}\left(1+\frac{13b_t}{24B_t}\right)^{N_t}\le\sum_{k\ge0}\left(\frac{24B_t+13b_t}{24B_t}\right)^k\left(\frac{B_t}{B_t+b_t}\right)^k=\sum_{k\ge0}\left(\frac{24B_t+13b_t}{24B_t+24b_t}\right)^k=\frac{24B_t+24b_t}{11b_t}.$$

于是有

$$\mathrm{E}G_{N_t}^{(t)}\le\frac{24B_t+24b_t}{11b_t}\left(G_0^{(t)}+\left(\frac{21}{4}+\frac{21B_t}{b_t}\right)\eta_t^2\mathrm{E}\left\|e_t\right\|^2\right),$$

即

$$\begin{aligned}&\left(3+\frac{12B_t}{b_t}\right)\eta_t\mathrm{E}\left[f\left(x_{N_t}^{(t)}\right)-f^*\right]+\mathrm{E}\left\|x_{N_t}^{(t)}-x_0^{(t)}\right\|\\&\le\frac{24B_t+24b_t}{11b_t}\left(\left(3+\frac{12B_t}{b_t}\right)\eta_t\mathrm{E}\left[f\left(x_0^{(t)}\right)-f^*\right]+\left(\frac{21}{4}+\frac{21B_t}{b_t}\right)\eta_t^2\mathrm{E}\left\|e_t\right\|^2\right).\end{aligned}$$

分别用 $\tilde{x}_t$ 替换 $x_{N_t}^{(t)}$，用 $\tilde{x}_{t-1}$ 替换 $x_0^{(t)}$，由[8]中引理 B.3 得到 $\mathrm{E}\left\|e_t\right\|^2<\infty$，这表明 $\mathrm{E}\left\|\tilde{x}_t-\tilde{x}_{t-1}\right\|^2<\infty$ 和 $\mathrm{E}\left[f\left(\tilde{x}_t\right)-f^*\right]<\infty$。由(7)可知 $\mathrm{E}\left\|\nabla f\left(\tilde{x}_t\right)\right\|^2<\infty$。利用 $2\left\langle a,b\right\rangle\le\frac{\left\|a\right\|^2}{c}+c\left\|b\right\|^2$ 可得 $\mathrm{E}\left|\left\langle e_t,\tilde{x}_t-\tilde{x}_{t-1}\right\rangle\right|<\infty$ 和

$\mathrm{E}\left|\left\langle e_t,\nabla f\left(\tilde{x}_t\right)\right\rangle\right|<\infty$。结论得证。

现在分析强凸条件下 SCSGP 的线性收敛速度。

定理 1 假设 $f_i(x)$ 是 L-光滑的且 $f(x)$ 是 μ -强凸的，令 $b_t=t^{1/2}$， $B_t=B_0t^{3/2}$，则

$$\mathrm{E}\left[f\left(\tilde{x}_T\right)-f^*\right]\le c_0^T\Delta_f+\frac{15M\mathrm{H}^*}{8\mu mB_0}J\left(B_t<n\right),$$

其中 $c_0=\dfrac{3}{3+2\mu mB_0}$， $\Delta_f=f\left(\tilde{x}_0\right)-f^*$， $\mathrm{H}^*=\sup\limits_x\dfrac{1}{n}\sum\limits_{i=1}^n\left\|\nabla f_i(x)-\nabla f(x)\right\|$ 和 $J\left(B_t<n\right)=\begin{cases}1, & B_t<n;\\ 0, & \text{其他}.\end{cases}$

证明：由[8]中引理 B.3 和等式(20)得到

$$\begin{aligned}&\frac{\eta_tB_t}{b_t}\left(2-\frac{2b_t}{B_t}-2\eta_tL-\frac{b_t^3}{b_t^3-\eta_t^2L^2b_tB_t-\eta_t^3L^3B_t^2}\right)\mathrm{E}\left\|\nabla f\left(\tilde{x}_t\right)\right\|^2\\&\le2\mathrm{E}\left[f\left(\tilde{x}_{t-1}\right)-f\left(\tilde{x}_t\right)\right]+\frac{2\eta_t}{b_t}\left(1+\eta_tL+\frac{b_t}{B_t}\right)\mathrm{H}^*J\left(B_t<n\right).\end{aligned}\tag{11}$$

由 $\dfrac{m}{t}\le\eta_t\le\dfrac{M}{t}$， $\dfrac{ML}{t}\le\dfrac{1}{3}\left(\dfrac{b_t}{B_t}\right)^{2/3}$ 和 $B_t\ge8b_t$ 可得

$$\begin{aligned}2-\frac{2b_t}{B_t}-2\eta_tL-\frac{b_t^3}{b_t^3-\eta_t^2L^2b_tB_t-\eta_t^3L^3B_t^2}&\ge2-\frac{2b_t}{B_t}-\frac{2ML}{t}-\frac{t^3b_t^3}{t^3b_t^3-M^2tL^2b_tB_t-M^3L^3B_t^2}\\&\ge2-2\times\frac{1}{8}-2\times\frac{1}{3}\times\frac{1}{4}-\frac{1}{1-1/\left(18b_t\right)-1/\left(27b_t\right)}\ge\frac{73}{108}\ge\frac{2}{3},\end{aligned}$$

其中第三个不等式用了 $b_t\ge1$。另外，

$$1+\eta_tL+\frac{b_t}{B_t}\le1+\frac{ML}{t}+\frac{b_t}{B_t}\le1+\frac{1}{12}+\frac{1}{8}=\frac{29}{24}\le\frac{5}{4}。$$

将上述两个系数代入(11)，并再次使用 $\dfrac{m}{t}\le\eta_t\le\dfrac{M}{t}$ 得到

$$\frac{B_t}{tb_t}\mathrm{E}\left\|\nabla f\left(\tilde{x}_t\right)\right\|^2\le\frac{3}{m}\mathrm{E}\left[f\left(\tilde{x}_{t-1}\right)-f\left(\tilde{x}_t\right)\right]+\frac{15M}{4mtb_t}\mathrm{H}^*J\left(B_t<n\right).\tag{12}$$

因为 f 是 μ -强凸的，可得

$$\frac{\mu}{2}\left\|\tilde{x}_t-x^*\right\|\le\left\langle\nabla f\left(\tilde{x}_t\right),\tilde{x}_t-x^*\right\rangle+f^*-f\left(\tilde{x}_t\right)\le\frac{\left\|\nabla f\left(\tilde{x}_t\right)\right\|^2}{2\mu}+\frac{\mu\left\|\tilde{x}_t-x^*\right\|^2}{2}+f^*-f\left(\tilde{x}_t\right),$$

其中第二个不等式利用了 $2\langle a,b\rangle\le\dfrac{\|a\|^2}{c}+c\|b\|^2$。重新整理上式得

$$\left\|\nabla f\left(\tilde{x}_t\right)\right\|^2\ge2\mu\left(f\left(\tilde{x}_t\right)-f^*\right).$$

将上式代入(12)得到

$$\left(3tb_t+2\mu mB_t\right)\mathrm{E}\left[f\left(\tilde{x}_t\right)-f^*\right]\le3tb_t\mathrm{E}\left[f\left(\tilde{x}_{t-1}\right)-f^*\right]+\frac{15}{4}M\,\mathrm{H}^*J\left(B_t<n\right).$$

替换 $b_t=t^{1/2}$ 和 $B_t=B_0t^{3/2}$，然后两边同除 $3t^{3/2}+2\mu mB_0t^{3/2}$ 可得

$$
\begin{aligned}
\mathrm{E}\left[f(\tilde{x}_t)-f^*\right] &\le \left(\frac{3}{3+2\mu mB_0}\right)\mathrm{E}\left[f(\tilde{x}_{t-1})-f^*\right]+\frac{15M\,\mathrm{H}^*J(B_t<n)}{4(3+2\mu mB_0)t^{3/2}} \\
&\le \left(\frac{3}{3+2\mu mB_0}\right)\mathrm{E}\left[f(\tilde{x}_{t-1})-f^*\right]+\frac{15M\,\mathrm{H}^*J(B_t<n)}{4(3+2\mu mB_0)}.
\end{aligned}
$$

其中最后一个不等式成立是因为$t\ge1$。上式可以写为

$$\mathrm{E}\left[f(\tilde{x}_t)-f^*\right]-\frac{15M\,\mathrm{H}^*J(B_t<n)}{8\mu mB_0}\le\left(\frac{3}{3+2\mu mB_0}\right)\left(\mathrm{E}\left[f(\tilde{x}_{t-1})-f^*\right]-\frac{15M\,\mathrm{H}^*J(B_t<n)}{8\mu mB_0}\right). \qquad (13)$$

将$t=T,\cdots,1$时的(13)累加求和得到

$$\mathrm{E}\left[f(\tilde{x}_t)-f^*\right]-\frac{15M\,\mathrm{H}^*J(B_t<n)}{8\mu mB_0}\le\left(\frac{3}{3+2\mu mB_0}\right)^T\left(\mathrm{E}\left[f(\tilde{x}_0)-f^*\right]-\frac{15M\,\mathrm{H}^*J(B_t<n)}{8\mu mB_0}\right)\le c_0^T\Delta_f. \qquad (14)$$

重新整理(14)，证毕。

4. 数值实验

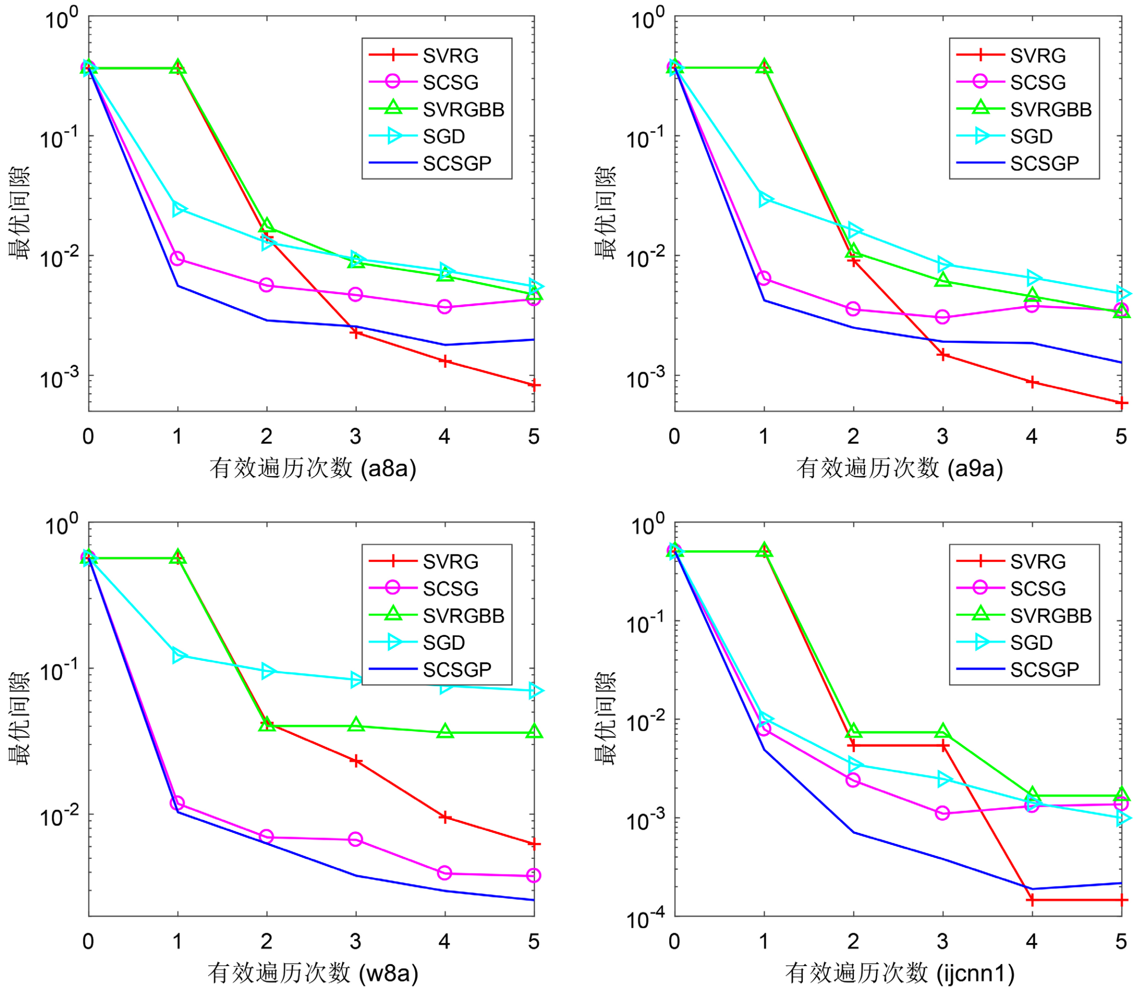

Figure 1. Comparison of different stochastic gradient algorithms
图 1. 不同随机梯度类算法的对比

考虑正则化的逻辑回归问题

$$f(x)=\frac{1}{n}\sum_{i=1}^{n}\log\left(1+\exp\left(-b_i a_i^T x\right)\right)+\frac{1}{2n}\|x\|^2,$$

其中$\left\{(a_i,b_i)\right\}_{i=1}^{n}\subset\Re^d\times\{-1,1\}$是给定的训练集。[7]中指出内循环次数$N_t$取期望值有助于增加 SCSG 算法的稳定性，且从$I_t$中选取$\tilde{I}_k$可以减小计算代价，所以在实验中设置$N_t=\lfloor B_t/b_t\rfloor$ (几何随机变量N_t的期望)且从I_t中选取$\tilde{I}_k$，其中$\lfloor\cdot\rfloor$记为向下取整。为了验证 SCSGP 的有效性，比较 SCSGP、SCSG、SVRG、SVRGBB 和 SGD。具体地，SVRG 中设置小批量$b_t\equiv 1$；SCSG 设置$B_t\equiv 0.05n$，$b_t\equiv 1$，$N_t=\lfloor B_t/b_t\rfloor$；SCSGP 设置$B_t=\lfloor B_0 t^{3/2}\wedge n\rfloor$，$b_t=\lfloor t^{1/2}\wedge n\rfloor$，$N_t=\lfloor B_t/b_t\rfloor$。表 1 给出 LIBSVM (网址：https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/)中四个标准数据集的信息。用表 2 的参数值进行对比实验，最优间隙随有效循环次数变化情况见图 1。SCSGP 明显比 SCSG 表现好，并且在前几个有效循环次数中，SCSGP 与其它随机梯度类算法相比具有更好的数值结果。

Table 1. The information of data sets
表 1. 数据集信息

数据集	n	d	L
a8a	22,696	123	3.5
a9a	32,561	123	3.5
w8a	49,749	300	28.5
ijcnn1	49,990	22	0.9842

Table 2. Parameters used for experiments
表 2. 实验中的参数设置

数据集	SGD	SVRG	SVRGBB	SCSG	SCSGP
a8a	$\eta=7/L$	$N_t=0.8n$ $\eta=0.1/L$	$N_t=0.8n$	$\eta=0.05/L$	$m=0.5/L$ $s=10$
a9a	$\eta=7/L$	$N_t=0.8n$ $\eta=0.05/L$	$N_t=0.8n$	$\eta=0.05/L$	$m=0.5/L$ $s=10$
w8a	$\eta=25/L$	$N_t=0.8n$ $\eta=2/L$	$N_t=0.8n$	$\eta=5/L$	$m=20/L$ $s=20$
ijcnn1	$\eta=80/L$	$N_t=n$ $\eta=0.1/L$	$N_t=n$	$\eta=0.1/L$	$m=1/L$ $s=15$

5. 总结

基于 Polyak 步长和 1/t-带步长提出 1/t-Polyak 步长，并将该步长与 SCSG 结合提出 SCSGP 算法。当目标函数强凸光滑时，SCSGP 线性收敛。数值实验考虑正则化的逻辑回归问题，实验结果表明在前几个有效循环次数中 SCSGP 比其他随机梯度类算法表现好。

参考文献

[1] Kasiviswanathan, S.P. and Jin, H. (2016) Efficient Private Empirical Risk Minimization for High-Dimensional Learning. *International Conference on Machine Learning*, **48**, 488-497.

[2] Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2017) Imagenet Classification with Deep Convolutional Neural Networks. *Communications of the ACM*, **60**, 84-90. https://doi.org/10.1145/3065386

[3] Sutskever, I., Martens, J., Dahl, G., *et al*. (2013) On the Importance of Initialization and Momentum in Deep Learning. *International Conference on Machine Learning*, **28**, 1139-1147.

[4] Robbins, H. and Monro, S. (1951) A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, **22**, 400-407. https://doi.org/10.1214/aoms/1177729586

[5] Bottou, L., Curtis, F.E. and Nocedal, J. (2018) Optimization Methods for Large-Scale Machine Learning. *SIAM Review*, **60**, 223-311. https://doi.org/10.1137/16M1080173

[6] Johnson, R. and Zhang, T. (2013) Accelerating Stochastic Gradient Descent Using Predictive Variance Reduction. *Advances in Neural Information Processing Systems*, **1**, 315-323.

[7] Lei, L. and Jordan, M. (2017) Less than a Single Pass: Stochastically Controlled Stochastic Gradient. *Artificial Intelligence and Statistics*, **54**, 148-156.

[8] Lei, L., Ju, C., Chen, J., *et al*. (2017) Non-Convex Finite-Sum Optimization via SCSG Methods. *Advances in Neural Information Processing Systems*, **11**, 2345-2355.

[9] Lei, L. and Jordan, M.I. (2020) On the Adaptivity of Stochastic Gradient-Based Optimization. *SIAM Journal on Optimization*, **30**, 1473-1500. https://doi.org/10.1137/19M1256919

[10] Gower, R.M., Loizou, N., Qian, X., *et al*. (2019) SGD: General Analysis and Improved Rates. *International Conference on Machine Learning*, **97**, 5200-5209.

[11] Ghadimi, S. and Lan, G. (2013) Stochastic First- and Zeroth-Order Methods for Nonconvex Stochastic Programming. *SIAM Journal on Optimization*, **23**, 2341-2368. https://doi.org/10.1137/120880811

[12] Rakhlin, A., Shamir, O. and Sridharan, K. (2011) Making Gradient Descent Optimal for Strongly Convex Stochastic Optimization. arXiv: 1109.5647.

[13] Polyak, B.T. (1987) Introduction to Optimization. Optimization Software. Publications Division, New York.

[14] Loizou, N., Vaswani, S., Laradji, I.H., *et al*. (2021) Stochastic Polyak Step-Size for SGD: An Adaptive Learning Rate for Fast Convergence. *International Conference on Artificial Intelligence and Statistics*, **130**, 1306-1314.

[15] Orvieto, A., Lacoste-Julien, S. and Loizou, N. (2022) Dynamics of SGD with Stochastic Polyak Stepsizes: Truly Adaptive Variants and Convergence to Exact Solution. *Advances in Neural Information Processing Systems*, **35**, 26943-26954.

[16] Wang, X. and Yuan, Y. (2023) On the Convergence of Stochastic Gradient Descent with Bandwidth-Based Step Size. *Journal of Machine Learning Research*, **24**, 1-49.