Noname manuscript No.
(will be inserted by the editor)

Convergence Rate of Projected Subgradient Method with Time-varying Step-sizes

Zhihan Zhu · Yanhao Zhang · Yong Xia

Received: date / Accepted: date

Abstract We establish the optimal ergodic convergence rate for the classical projected subgradient method with time-varying step-sizes. This convergence rate remains the same even if we slightly increase the weight of the most recent points, thereby relaxing the ergodic sense.

Keywords Subgradient method · step-size · ergodic convergence rate · nonsmooth convex optimization

Mathematics Subject Classification (2000) 90C25, 90C30

1 Introduction

Consider the nonsmooth convex optimization problem

$$x^* \in \mathrm{argmin}_{x\in\mathcal{X}} f(x),$$

where $\mathcal{X} \subset \mathbb{R}^n$ is a compact convex set that is contained within the Euclidean ball $B(x^*, R)$, and f is (possibly nonsmooth) convex and Lipschitz on $\mathcal{X}$, i.e., there is an $L > 0$ such that $\|g\| \leq L$ for any $g \in \partial f(x) \neq \emptyset$ and $x \in \mathcal{X}$.

The classical projected subgradient method (PSG) iterates the following equations for $t \geq 1$:

$$\begin{cases} y_{t+1} = x_t - \eta_t g_t, \text{ where } g_t \in \partial f(x_t), \\ x_{t+1} = \mathrm{argmin}_{x\in\mathcal{X}} \|x - y_{t+1}\|. \end{cases}$$

Utilizing the following constant step size

$$\eta_s \equiv \frac{R}{L\sqrt{t}},\ s = 1, \cdots, t, \tag{1}$$

Zhihan Zhu · Yanhao Zhang · Yong Xia (corresponding author)
School of Mathematical Sciences, Beihang University, Beijing 100191, People's Republic of China. E-mail: {zhihanzhu, yanhaozhang, yxia}@buaa.edu.cn

PSG achieves an optimal ergodic convergence rate (see, for example, [1–3]) expressed as [1]

$$f\left(\frac{\sum_{s=1}^{t} x_s}{t}\right) - f(x^*) \le \frac{RL}{\sqrt{t}}. \tag{2}$$

A more practical approach is to use a step-size that monotonically decreases towards 0. One such time-varying step-size, motivated by (1) and suggested in [1, 2], is expressed as follows:

$$\eta_s = \frac{R}{L\sqrt{s}},\ s = 1, \cdots, t. \tag{3}$$

However, using this step size results in sub-optimal ergodic convergence rate [1, 2], as it adds an additional $\log(t)$ factor compared to the right-hand side of (2).

The contribution of this note is to demonstrate that PSG with the time-varying step-size (3) indeed achieves the following optimal convergence rate.

Theorem 1 *PSG with the time-varying step-size* (3) *satisfies*

$$f\left(\frac{\sum_{s=1}^{t} x_s}{t}\right) - f(x^*) \le \frac{3RL}{2\sqrt{t}}.$$

In Section 2, we present a more generalized convergence analysis, allowing us to provide some insightful observations.

2 Analysis

Consider PSG with a general step-size η_s, which is assumed to be positive and non-increasing. We can establish the following convergence rate.

Theorem 2 *For any fixed $k \ge -1$, PSG with a positive and non-increasing step-size sequence $\{\eta_s\}$ satisfies*

$$f\left(\frac{\sum_{s=1}^{t} \frac{1}{\eta_s^k} x_s}{\sum_{s=1}^{t} \frac{1}{\eta_s^k}}\right) - f(x^*) \le \frac{\frac{R^2}{\eta_t^{k+1}} + L^2 \sum_{s=1}^{t} \frac{1}{\eta_s^{k-1}}}{2\sum_{s=1}^{t} \frac{1}{\eta_s^k}}. \tag{4}$$

Proof According to the definition of subgradient, we have

$$\begin{aligned}
f(x_s) - f(x^*) &\le g_s^T (x_s - x^*) \\
&= \frac{1}{\eta_s}(x_s - y_{s+1})^T (x_s - x^*) \\
&= \frac{1}{2\eta_s}(\|x_s - y_{s+1}\|^2 + \|x_s - x^*\|^2 - \|y_{s+1} - x^*\|^2) \qquad (5) \\
&= \frac{1}{2\eta_s}(\|x_s - x^*\|^2 - \|y_{s+1} - x^*\|^2) + \frac{\eta_s}{2}\|g_s\|^2 \\
&\le \frac{1}{2\eta_s}(\|x_s - x^*\|^2 - \|x_{s+1} - x^*\|^2) + \frac{\eta_s}{2}L^2, \qquad (6)
\end{aligned}$$

[1] $f((\sum_{s=1}^{t} x_s)/t)$ can be replaced with $\min_{s=1,\cdots,t} f(x_s)$ based on similar analysis.

where (5) follows from the elementary identity $2a^Tb = \|a\|^2 + \|b\|^2 - \|a-b\|^2$, and (6) holds since $\|g_s\| \leq L$ and

$$\|y_{s+1} - x^*\|^2 \geq \|x_{s+1} - x^*\|^2,$$

which is implied by the projection theorem.

Consequently, we have

$$\begin{aligned}
&\left(\sum_{s=1}^{t} \frac{1}{\eta_s^k}\right)\left(f\left(\frac{\sum_{s=1}^{t} \frac{1}{\eta_s^k} x_s}{\sum_{s=1}^{t} \frac{1}{\eta_s^k}}\right) - f(x^*)\right) \\
&\leq \sum_{s=1}^{t} \frac{1}{\eta_s^k}(f(x_s) - f(x^*)) \quad \text{(since } f \text{ is convex)} \\
&\leq \sum_{s=1}^{t} \frac{1}{2\eta_s^{k+1}}(\|x_s - x^*\|^2 - \|x_{s+1} - x^*\|^2) + \sum_{s=1}^{t} \frac{1}{2\eta_s^{k-1}} L^2 \qquad (7) \\
&= \frac{1}{2\eta_1^{k+1}}\|x_1 - x^*\|^2 + \sum_{s=2}^{t}\left(\frac{1}{2\eta_s^{k+1}} - \frac{1}{2\eta_{s-1}^{k+1}}\right)\|x_s - x^*\|^2 \\
&\quad - \frac{1}{2\eta_t^{k+1}}\|x_{t+1} - x^*\|^2 + \sum_{s=1}^{t} \frac{1}{2\eta_s^{k-1}} L^2 \\
&\leq \frac{R^2}{2\eta_1^{k+1}} + \frac{R^2}{2}\sum_{s=2}^{t}\left(\frac{1}{\eta_s^{k+1}} - \frac{1}{\eta_{s-1}^{k+1}}\right) + \sum_{s=1}^{t} \frac{1}{2\eta_s^{k-1}} L^2 \qquad (8) \\
&= \frac{R^2}{2\eta_t^{k+1}} + \sum_{s=1}^{t} \frac{1}{2\eta_s^{k-1}} L^2,
\end{aligned}$$

where (7) follows from (6), and (8) holds since $1/\eta_s^{k+1} - 1/\eta_{s-1}^{k+1} \geq 0$ when $k \geq -1$. The proof is complete. □

Remark 1 *By setting $k = -1$ in Theorem 2, the upper bound on the right-hand side* (4) *simplifies to*

$$\frac{R^2 + L^2 \sum_{s=1}^{t} \eta_s^2}{2\sum_{s=1}^{t} \eta_s},$$

which is exactly the same as the result presented in [1]. Then PSG with the time-varying step-size (3) *satisfies*

$$f\left(\frac{\sum_{s=1}^{t} \frac{1}{\sqrt{s}} x_s}{\sum_{s=1}^{t} \frac{1}{\sqrt{s}}}\right) - f(x^*) \leq \frac{2RL + RL \log t}{4(\sqrt{t+1} - 1)}, \qquad (9)$$

which is sub-optimal. Note that computing the weighted average of the iterates in the second half of the sequence yields the optimal convergence rate [4, Corollary 3.2].

Remark 2 *By setting $k = 0$ in Theorem 2, we can immediately obtain the optimal convergence rate, as presented in Theorem 1.*

Remark 3 *By setting any k such that $k > -1$ in Theorem 2, the convergence rate of $f((\sum_{s=1}^{t} x_s/\eta_s^k)/(\sum_{s=1}^{t} 1/\eta_s^k)) - f(x^*)$ will be $\mathcal{O}(1/\sqrt{t})$, without the presence of a $\log t$ factor. In comparison with the sub-optimal case* (9)*, when $k > 0$, the weighting scheme $(\sum_{s=1}^{t} x_s/\eta_s^k)/(\sum_{s=1}^{t} 1/\eta_s^k)$ assigns smaller weights to the initial points and larger weights to the most recent points. This new result can be referred to as the "weak" ergodic convergence rate.*

Remark 4 *We can apply the same proof techniques to extend the conclusion of weak ergodic convergence to mirror descent, Nesterov's dual averaging, and other schemes with time-varying step sizes for solving nonsmooth convex optimization, see [1,2].*

Funding

This research was supported by National Key R&D Program of China under grant 2021YFA1003300, the National Natural Science Foundation of China under grant 12171021, and the Fundamental Research Funds for the Central Universities.

Data Availability

The manuscript has no associated data.

References

1. Y. Nesterov, Lectures on convex optimization, volume 137, Springer, 2018.
2. S. Bubeck, Convex optimization: Algorithms and complexity. Foundations and Trends Trends® in Machine Learning, 8(3-4): 231-357, 2015.
3. P. Rigollet, K. Li, Convex optimization for machine learning. Massachusetts Institute of Technology: MIT OpenCouseWare, https://ocw.mit.edu/, Oct. 2015.
4. G. Lan, First-order and Stochastic Optimization Methods for Machine Learning, Springer-Nature, 2020