Parameter-free Clipped Gradient Descent Meets Polyak

Yuki Takezawa[1,2], Han Bao[1,2], Ryoma Sato[3], Kenta Niwa[4], Makoto Yamada[2]

[1]Kyoto University, [2]OIST, [3]NII, [4]NTT Communication Science Laboratories

Background

▶ **Gradient Descent**

$$x_{t+1} = x_t - \eta_t \nabla f(x_t). \tag{1}$$

Assumption (L-smoothness)

There exists a constant $L > 0$ such that it holds that for all $x, y \in \mathbb{R}^d$,

$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|. \tag{2}$$

Theorem (Gradient Descent)

Assume that f is convex and L-smooth. Then, gradient descent with $\eta_t = \frac{1}{L}$ satisfies

$$f(\bar{x}) - f(x^\star) \leq \mathcal{O}\left(\frac{L\|x_0 - x^\star\|^2}{T}\right). \tag{3}$$

where $\bar{x} := \frac{1}{T}\sum_{t=0}^{T-1} x_t$.

▶ **Clipped Gradient Descent**

$$x_{t+1} = x_t - \eta_t \min\left\{1, \frac{c}{\|\nabla f(x_t)\|}\right\} \nabla f(x_t). \tag{4}$$

Assumption ((L_0, L_1)-smoothness)

There exists constants $L_0 > 0$ and $L_1 > 0$ that satisfies

$$\|\nabla f(x) - \nabla f(y)\| \leq (L_0 + L_1 \|\nabla f(x)\|) \|x - y\|, \tag{5}$$

for all $x, y \in \mathbb{R}^d$ with $\|x - y\| \leq \frac{1}{L_1}$.

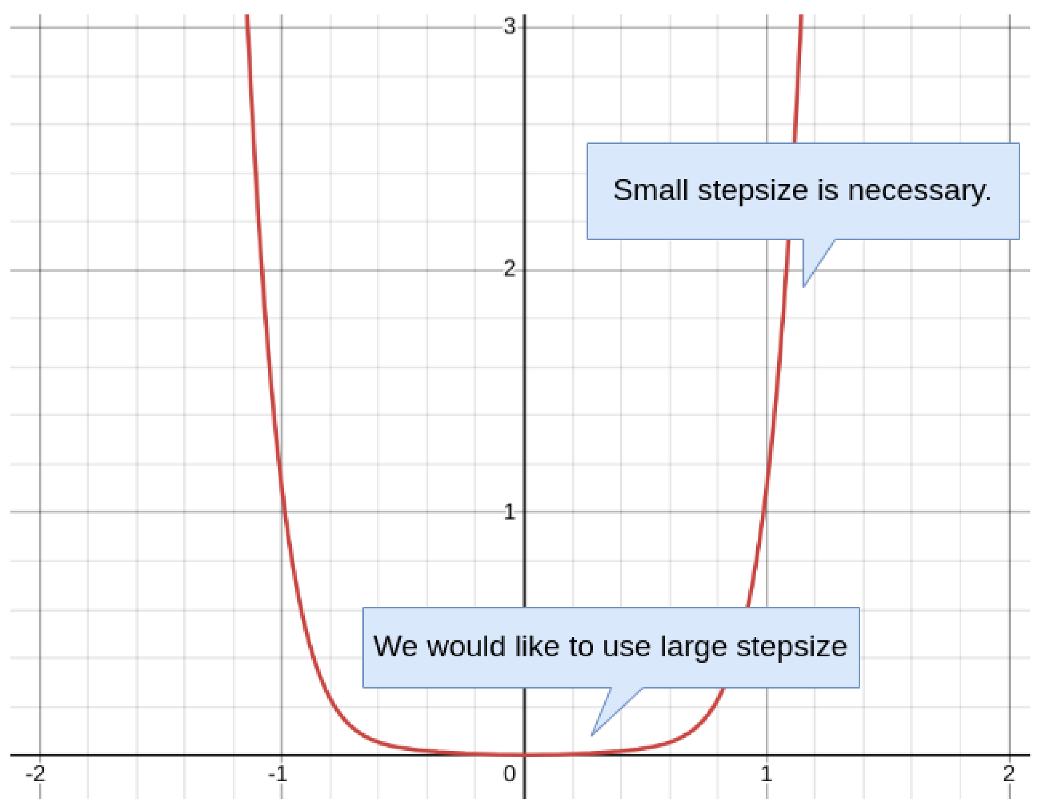
Theorem (Clipped Gradient Descent)

Assume that f is convex, L-smooth, and (L_0, L_1)-smooth. Then, clipped gradient descent with $\eta_t = \frac{1}{L_0}$ and $c = \frac{L_0}{L_1}$ satisfies

$$f(\bar{x}) - f(x^\star) \leq \mathcal{O}\left(\frac{L_0\|x_0 - x^\star\|^2}{T} + \frac{L L_1^2 \|x_0 - x^\star\|^4}{T^2}\right), \tag{6}$$

where $\bar{x} := \frac{1}{T}\sum_{t=0}^{T-1} x_t$.

▷ Since $L_0 \ll L$ in practice, clipped gradient descent can converge faster than gradient descent.



Contribution

Q: Can we develop a parameter-free method whose convergence rate is asymptotically independent of L under (L_0, L_1)-smoothness?

▷ We discover that Polyak stepsize can converge as fast as clipped gradient descent.

▷ We make Polyak stepsize parameter-free without losing the asymptotic independence of L by proposing Inexact Polyak Stepsize.

New Convergence Result of Polyak Stepsize

▶ **Polyak Stepsize**

▷ It is well-known that Polyak stepsize allows gradient descent to converge as fast as the optimal stepsize.

$$\eta_t = \frac{f(x_t) - f(x^\star)}{\|\nabla f(x_t)\|^2}. \tag{7}$$

▶ **Analysis of Polyak Stepsize under (L_0, L_1)-smoothness**

▷ Polyak stepsize can also achieve the same convergence rate as clipped gradient descent.

Theorem (Polyak Stepsize)

Assume that f is convex, L-smooth, and (L_0, L_1)-smooth. Then, gradient descent with Polyak stepsize satisfies

$$f(x_\tau) - f(x^\star) \leq \mathcal{O}\left(\frac{L_0\|x_0 - x^\star\|^2}{T} + \frac{L L_1^2 \|x_0 - x^\star\|^4}{T^2}\right),$$

where $\tau := \arg\min_{0 \leq t \leq T} f(x_t)$.

▶ **Several existing papers proposed parameter-free versions of Polyak stepsize, while they lost the fruitful property under (L_0, L_1)-smoothness.**

▷ They proposed to use the lower bound $l^\star$ instead of $f(x^\star)$.

▷ To prevent the stepsize from becoming too large, they make the stepsize monotonically decreasing.

$$\eta_t = \frac{f(x_t) - l^\star}{\|\nabla f(x_t)\|^2}.$$

Proposed Method

Algorithm 1 Inexact Polyak Stepsize

1: **Input:** The number of iterations T and lower bound $l^\star$.
2: $f^{\text{best}}, x^{\text{best}} \leftarrow f(x_0), x_0$.
3: **for** $t = 0, 1, \cdots, T-1$ **do**
4: $\quad x_{t+1} \leftarrow x_t - \frac{f(x_t) - l^\star}{\sqrt{T}\|\nabla f(x_t)\|^2}\nabla f(x_t)$.
5: $\quad$ **if** $f(x_{t+1}) \leq f^{\text{best}}$ **then**
6: $\quad\quad f^{\text{best}}, x^{\text{best}} \leftarrow f(x_{t+1}), x_{t+1}$.
7: **return** x^{best}.

Theorem (Inexact Polyak Stepsize)

Assume that f is convex, L-smooth, and (L_0, L_1)-smooth. Then, gradient descent with Inexact Polyak stepsize satisfies

$$f(x^{\text{best}}) - f(x^\star) \leq \mathcal{O}\left(\frac{L_0\|x_0 - x^\star\|^2 + \sigma^2}{\sqrt{T}} + \frac{L L_1^2 \|x_0 - x^\star\|^4}{T} + \frac{L_1^2 L \sigma^4}{L_0^2 T}\right),$$

where $\sigma^2 := f(x^\star) - l^\star$.

▷ The convergence rate of Inexact Polyak Stepsize is asymptotically independent of L.

▷ The convergence rates of DecSPS and AdaSPS depend on $D_T (:= \max_{0 \leq t \leq T} \|x_t - x^\star\|)$, while the rate of Inexact Polyak Stepsize depends on $\|x_0 - x^\star\|$.

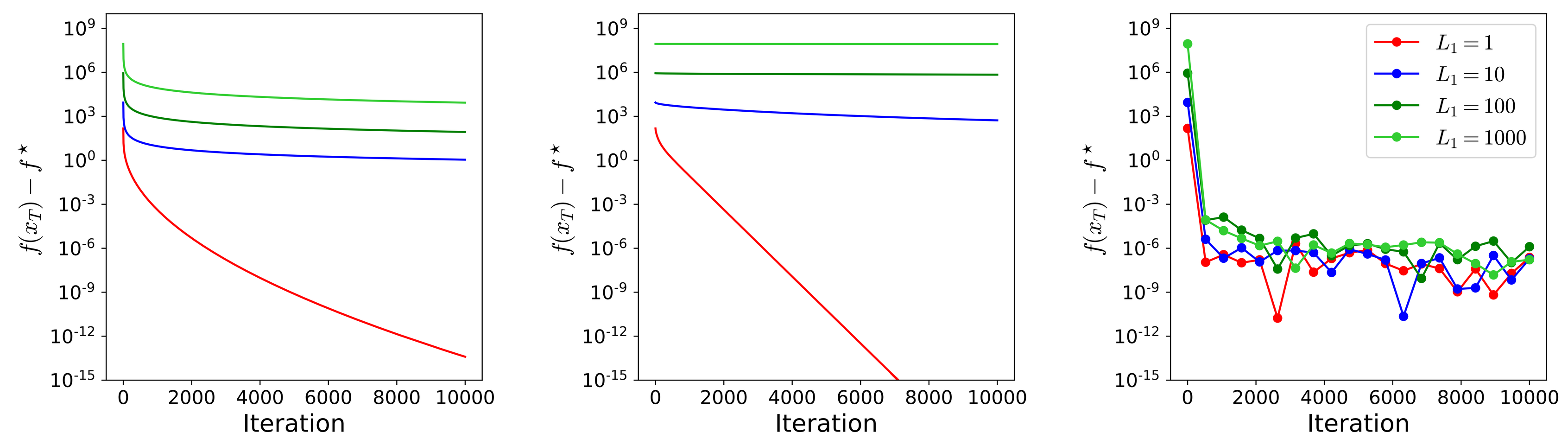
Table: Summary of convergence rates of parameter-free methods based on Polyak stepsize.

Algorithm	Convergence Rate
DecSPS (Orvieto et al., 2022)	$\mathcal{O}\left(\frac{\max\{L, \eta_0^{-1}\} D_T^2 + \sigma^2}{\sqrt{T}}\right)$
AdaSPS (Jiang et al., 2023)	$\mathcal{O}\left(\frac{L D_T^2 \sigma}{\sqrt{T}} + \frac{L^2 D_T^4}{T}\right)$
Inexact Polyak Stepsize (Ours)	$\mathcal{O}\left(\frac{L_0\|x_0 - x^\star\|^2 + \sigma^2}{\sqrt{T}} + \frac{L L_1^2 \|x_0 - x^\star\|^4}{T} + \frac{L_1^2 L \sigma^4}{L_0^2 T}\right)$

Numerical Results

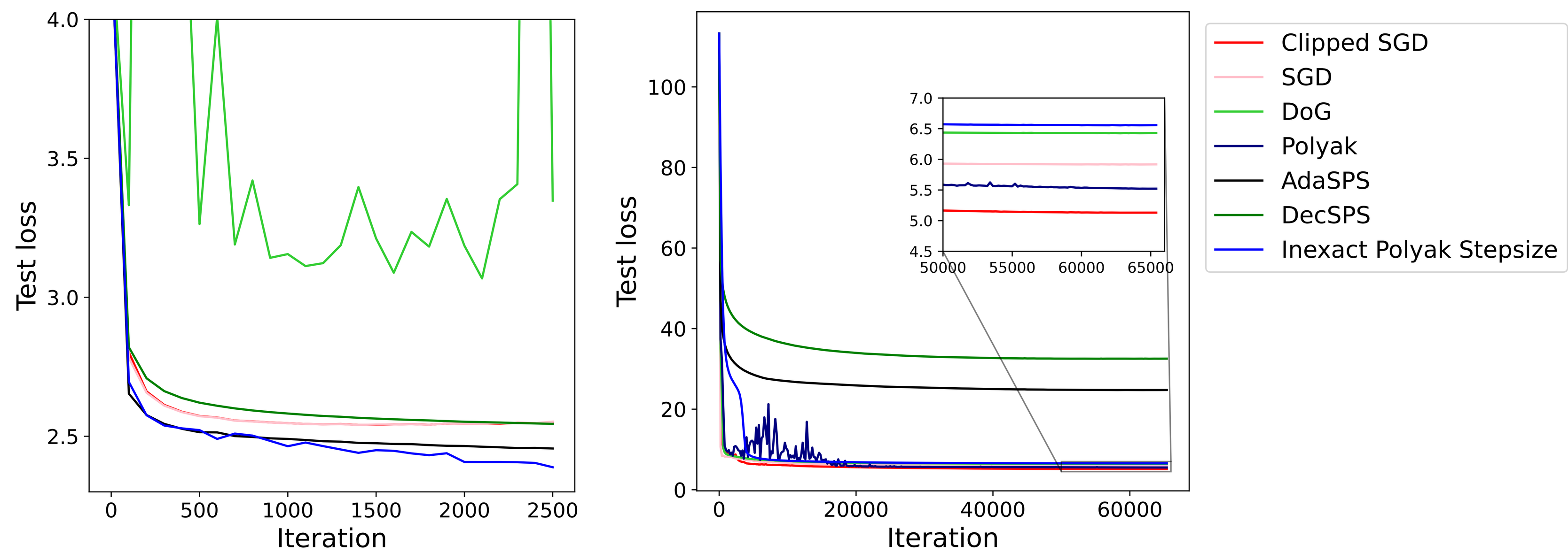
▶ **Synthetic Function**

▷ The convergence behavior of Inexact Polyak stepsize is almost the same for all L_1.



(a) DecSPS (b) AdaSPS (c) Inexact Polyak Stepsize

▶ **Neural Networks**



(a) Nano-GPT (b) T5