Methods for Convex (L_0, L_1)-Smooth Optimization:

- Clipping
- Acceleration
- Adaptivity

Mohamed bin Zayed University of Artificial Intelligence

Eduard Gorbunov*
Nazarii Tupitsa*
Sayantan Choudhury
Alen Aliev
Peter Richtárik
Samuel Horváth
Martin Takáč (*equal contribution)

Problem

$$\min_{x\in\mathbb{R}^d} f(x)$$

Standard L-smoothness: $\|\nabla^2 f(x)\|_2 \le L$

- Equivalent definition: $\|\nabla f(x) - \nabla f(y)\| \le L\|x-y\|$
- ✗ Does not hold on $\mathbb{R}^d$ in Deep Learning
- ✗ Even if satisfied on a compact, constant L can be large

(L_0, L_1)-smoothness: $\|\nabla^2 f(x)\|_2 \le L_0 + L_1\|\nabla f(x)\|$

- ✓ Proposed and empirically validated by Zhang et al. (2020) [7]

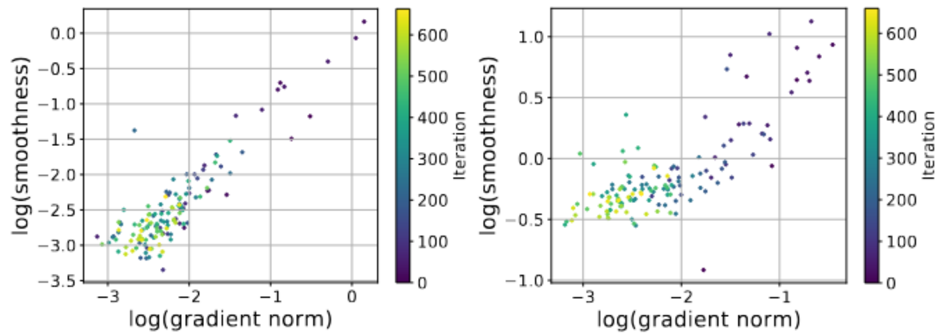

Figure 1: Estimated smoothness constant for two DL tasks (Zhang et al., 2020) [7]

- ✓ Generalized by Zhang et al. (2020) [6] and Chen et al. (2023) [1] to

$$\|\nabla f(x) - \nabla f(y)\| \le \left(L_0 + L_1 \sup_{u\in[x,y]} \|\nabla f(u)\|\right)\|x-y\|$$

- ✓ Chen et al. (2023) [1] proved that (L_0, L_1)-smoothness is equivalent to

$$\|\nabla f(x) - \nabla f(y)\| \le (L_0 + L_1\|\nabla f(y)\|)\exp\left(L_1\|x-y\|\right)\|x-y\|.$$

Existing convergence results in the convex case

Let f be **convex** and (L_0, L_1)-smooth. **Goal:** bound the number of iterations $N(\varepsilon)$ needed to reach $f(x^N) - f(x^*) \le \varepsilon$ for a given method, $x^* \in \arg\min_{x\in\mathbb{R}^d} f(x)$.

- Li et al. (2024) [3]: $N = \mathcal{O}\left(\frac{L_0(1+L_1R_0\exp(L_1R_0))R_0^2}{\varepsilon}\right)$
  - ✗ Exponentially large factor of $R_0 = \|x^0 - x^*\|$
- Koloskova et al. (2023) [2], Takezawa et al. (2024) [5]: $N = \mathcal{O}\left(\max\left\{\frac{L_0R_0^2}{\varepsilon}, \sqrt{\frac{R_0^4 L L_1^2}{\varepsilon}}\right\}\right)$
  - ✗ Derived under additional L-smoothness assumption
- 🤔 Is it possible to derive better results without extra assumptions? Yes!

References

1. Chen et al. *Generalized-smooth nonconvex optimization is as efficient as smooth nonconvex optimization* (ICML 2023)
2. Koloskova et al. *Revisiting gradient clipping: Stochastic bias and tight convergence guarantees* (ICML 2023)
3. Li et al. *Convex and non-convex optimization under generalized smoothness* (NeurIPS 2024)
4. Malitsky & Mishchenko. *Adaptive gradient descent without descent* (ICML 2020)
5. Takezawa et al. *Parameter-free Clipped Gradient Descent Meets Polyak* (NeurIPS 2024)
6. Zhang et al. *Improved analysis of clipping algorithms for non- convex optimization* (NeurIPS 2020)
7. Zhang et al. *Why gradient clipping accelerates training: A theoretical justification for adaptivity* (ICLR 2020)

Gradient Descent with Polyak Stepsizes

Algorithm: $x^{k+1} = x^k - \frac{f(x^k)-f(x^*)}{\|\nabla f(x^k)\|^2}\nabla f(x^k)$

Rate: $f(x^N) - f(x^*) \le \varepsilon$ after $N = \mathcal{O}\left(\max\left\{\frac{L_0R_0^2}{\varepsilon}, L_1^2R_0^2\right\}\right)$

Gradient Descent with Smoothed Clipping

Algorithm: $x^{k+1} = x^k - \frac{\eta}{L_0 + L_1\|\nabla f(x^k)\|}\nabla f(x^k)$

Rate: $f(x^N) - f(x^*) \le \varepsilon$ after $N = \mathcal{O}\left(\max\left\{\frac{L_0R_0^2}{\varepsilon}, L_1^2R_0^2\right\}\right)$

- ✓ No exponential factors of $R_0 = \|x^0 - x^*\|$
- ✓ No additional assumptions
- **Proof sketch**
  - Using the convexity, $\frac{\nu\|\nabla f(x)\|^2}{2(L_0+L_1\|\nabla f(x)\|)} \le f(x) - f(x^*)$, and $\eta \le \frac{\nu}{2}$, we get

$$\|x^{k+1} - x^*\|^2 \le \|x^k - x^*\|^2 - \eta\frac{f(x^k) - f(x^*)}{L_0 + L_1\|\nabla f(x^k)\|}$$

  - Then, we consider two possible cases:

$$\|x^{k+1} - x^*\|^2 \le \begin{cases} \|x^k - x^*\|^2 - \frac{\nu\eta}{8L_1^2}, & \text{if } \|\nabla f(x^k)\| \ge \frac{L_0}{L_1} \\ \|x^k - x^*\|^2 - \frac{\eta}{2L_0}\left(f(x^k) - f(x^*)\right), & \text{if } \|\nabla f(x^k)\| < \frac{L_0}{L_1} \end{cases}$$

  - The first case occurs no more than $\mathcal{O}(L_1^2R_0^2)$ times; the second case is standard

Adaptive Gradient Descent (Malitsky & Mishchenko, (2020) [4])

Algorithm: $x^{k+1} = x^k - \lambda_k\nabla f(x^k)$, where

$$\lambda_k = \min\left\{\sqrt{1 + \frac{\lambda_{k-1}}{\lambda_{k-2}}}\lambda_{k-1}, \frac{\|x^k - x^{k-1}\|}{4\|\nabla f(x^k) - \nabla f(x^{k-1})\|}\right\}$$

Rate: $f(x^N) - f(x^*) \le \varepsilon$ after $N = \mathcal{O}\left(\max\left\{\frac{L_0D^2}{\varepsilon}, m^2(L_1^2D^2 + L_1^4D^4)\right\}\right)$

- $m := 1 + \log_{\sqrt{2}}\left\lceil\frac{(1+L_1D\exp(2L_1D))}{2}\right\rceil$
- $D^2 := \|x^1 - x^*\|^2 + \frac{3}{4}\|x^1 - x^0\|^2 + 2\lambda_1\theta_1(f(x^0) - f(x^*))$
- No explicit exponential factors of R_0
- In the paper, we also have: an accelerated method ((L_0, L_1)-STM) and stochastic extensions of the first two methods

Experiments

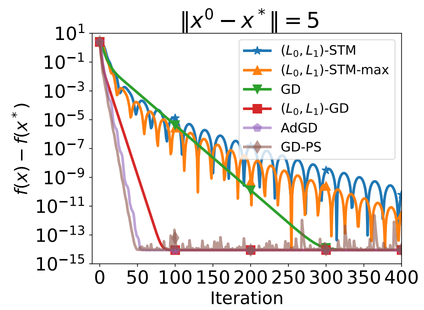

Figure 2: Logistic regression

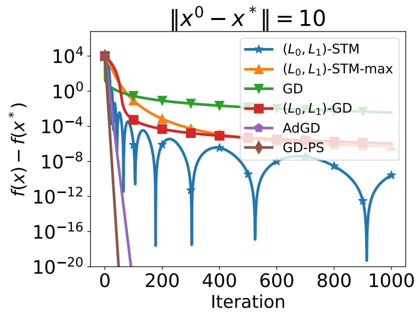

Figure 3: $f(x) = x^4$