Stochastic Optimization with Heavy-Tailed Noise via Accelerated Gradient Clipping

1. The Problem

Problem: expectation minimization

$$\min_{x \in \mathbb{R}^n} \left\{ f(x) = \mathbb{E}[f(x, \xi)] \right\}$$

Assumptions: convexity and smoothness

$$f(x) - f(y) \geq \langle \nabla f(y), x - y \rangle$$

$$\| \nabla f(x) - \nabla f(y) \| \leq L \| x - y \|$$

One of the most popular methods to solve such problems is SGD:

$$x^{k+1} = x^k - \gamma \nabla f(x^k, \xi^k)$$

There is a lot of literature on the convergence in expectation. However, we focus on the convergence with high probability.

2. Motivational Example

Consider the following instance of the problem described above:

$$f(x, \xi) = \frac{1}{2} \| x - \xi \|^2 + \langle \xi, x \rangle$$

$$\mathbb{E}[\xi] = 0$$

$$\mathbb{E} \left[ \| \xi \|^2 \right] = \sigma^2$$

The state-of-the-art analysis of SGD for this problem gives:

$$\mathbb{E} \left[ f(x^k) - f(x^*) \right] \leq (1 - \gamma)^k \left( f(x^0) - f(x^*) \right) + \frac{2\sigma^2}{\gamma}$$

This bound cannot explain the following phenomenon

3. Light and Heavy Tails

**Light tails:**

$$\mathbb{E} \left[ \| \nabla f(x, \xi) - \nabla f(x) \|^2 \right] \leq \sigma^2$$

**Heavy tails:**

$$\mathbb{E} \left[ \| \nabla f(x, \xi) - \nabla f(x) \|^2 \right] \leq \sigma^2$$

- **Light-tailed case is well-understood:** there exist results for SGD [1] and accelerated SGD (AC-SA) [2,3] that coincide with corresponding convergence bounds in expectation.
- **Heavy-tailed case is partially studied:** in convex case there exist non-accelerated result matching the complexity of Light-tailed SGD [4].

4. Our Contributions

- The first accelerated stochastic method converging with the same rate as AC-SA but without light-tailed assumption — **Clipped Stochastic Similar Triangles Method (clipped-SSTM)**
- The first high-probability complexity guarantees for clipped-SGD in convex and strongly convex cases

5. Accelerated SGD with Clipping

**Clipped Stochastic Similar Triangles Method (clipped-SSTM)**

**Initialization:**

- starting point $x^0$
- number of iterations $N$
- batchesizes $\{m_i\}$
- stepsize parameter $\alpha$
- clipping parameter $\beta$

**Step k:**

1. $\alpha_{k+1} = \frac{k + 2}{2 \alpha L} \left( A_{k+1} \right)_{k+1} = A_k + \alpha_{k+1}, \lambda_{k+1} = \alpha_{k+1}$

2. $x^{k+1} = A_{k} y^k + \alpha_{k+1} z^k$

**Batching:**

- draw fresh i.i.d. samples $\xi^0, \ldots, \xi^m$
- and compute $\nabla f(x^{k+1}, \xi^k) = \frac{1}{m_k} \sum_{i=0}^{m_k} \nabla f(x^{k+1}, \xi_i)$

**Clipping:**

- $z^{k+1} = z^k - \alpha_{k+1} \nabla f(x^{k+1}, \xi^k)$

**Output:**

$$y^N$$

6. Comparison of Complexities

**Complexity** = number of stochastic first-order oracle calls needed by the method to find such point $x$ that

$$\mathbb{P} \left( \left\{ f(x) - f(x^*) \right\} < \varepsilon \right) < \beta$$

The red color is used to indicate the restrictions we eliminated in our analysis.

- $\varepsilon$ — diameter of the domain (if it is bounded), $R_0 = \left\| x^0 - x^* \right\|$

7. Numerical Experiments

We conducted several numerical experiments on logistic regression problem

References

[1] Devolder, Olivier. Stochastic first order methods in smooth convex optimization. No. UCL-Université Catholique de Louvain. CORE, 2011.

[2] Ghadimi, Saeed, and Guanghui Lan. “Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization I: A generic algorithmic framework.” SIAM Journal on Optimization 22, no. 4 (2012): 1469-1492.

[3] Lan, Guanghui. “An optimal method for stochastic composite optimization.” Mathematical Programming 133, no. 1-2 (2012): 365-397.

[4] Nazin, Alexander V., Arkadi S. Nemirovsky, Alexandre B. Tsybakov, and Anatoli B. Juditsky. “Algorithms of robust stochastic optimization based on mirror descent method.” Automation and Remote Control 80, no. 9 (2019): 1607-1627.

[5] Davis, Damek, Dmitriy Drusvyatskiy, Lin Xiao, and Junyu Zhang. “From low probability to high confidence in stochastic convex optimization.” arXiv preprint arXiv:1907.13307 (2019).