Stochastic Optimization with Heavy-Tailed Noise via Accelerated Gradient Clipping

Eduard Gorbunov[1,2] **Marina Danilova**[3,1] **Alexander Gasnikov**[1,2]

[1]**MIPT (Russia)** [2]**HSE (Russia)** [3]**ICS RAS (Russia)**

1. The Problem

Problem: expectation minimization

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

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

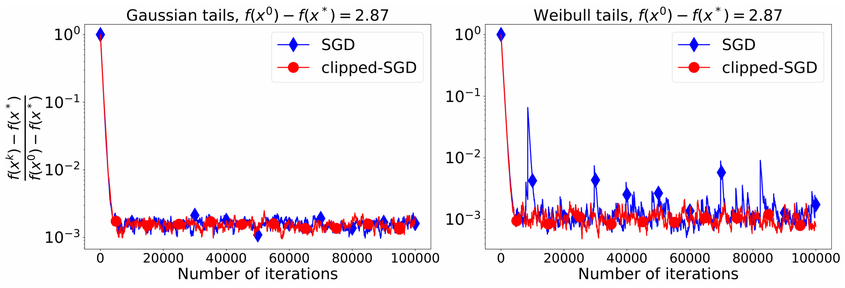
Consider the following instance of the problem described above:

$$f(x,\xi)=\frac{\|x\|_2^2}{2}+\langle\xi,x\rangle\qquad\mathbb{E}[\xi]=0\qquad\mathbb{E}\left[\|\xi\|^2\right]=\sigma^2$$

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

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

This bound cannot explain the following phenomenon



All parameters that the bound above depends on are the same for both runs of SGD. The difference is only in distributions of stochastic gradients and it is clear how oscillations of SGD depends on this.

- Convergence in expectation is not able to reflect this behavior of SGD while convergence with high probability is
- Clipping helps to reduce oscillations without sacrificing the rate

clipped-SGD: $x^{k+1}=x^k-\text{clip}(\nabla f(x^k,\xi^k),\lambda)$

$$\text{clip}(\nabla f(x^k,\xi^k),\lambda)=\min\left\{1,\frac{\lambda}{\|\nabla f(x^k,\xi^k)\|}\right\}\nabla f(x^k,\xi^k)$$

3. Light and Heavy Tails

$$\mathbb{E}_\xi[\nabla f(x,\xi)]=\nabla f(x)$$

Light tails: $\mathbb{E}_\xi\left[\exp\left(\frac{\|\nabla f(x,\xi)-\nabla f(x))\|^2}{\sigma^2}\right)\right]\leq\exp(1)$

Heavy tails: $\mathbb{E}_\xi\left[\|\nabla f(x,\xi)-\nabla f(x)\|_2^2\right]\leq\sigma^2$

- **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 generalization of clipped-SSTM to the **strongly convex case**
- **The first high-probability complexity gurantees 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, batchsizes $\{m_k\}_{k=1}^N$, stepsize parameter a, clipping parameter B

$$A_0=\alpha_0=0,y^0=z^0=x^0$$

Step k:

1. $\alpha_{k+1}=\frac{k+2}{2aL},A_{k+1}=A_k+\alpha_{k+1},\lambda_{k+1}=\frac{B}{\alpha_{k+1}}$
2. $x^{k+1}=\frac{A_ky^k+\alpha_{k+1}z^k}{A_{k+1}}$
3. **Batching:** draw fresh i.i.d. samples $\xi_1^k,\ldots,\xi_{m_k}^k$ and compute $\nabla f\left(x^{k+1},\boldsymbol{\xi}^k\right)=\frac{1}{m_k}\sum_{i=1}^{m_k}\nabla f\left(x^{k+1},\xi_i^k\right)$
4. **Clipping:** $\widetilde{\nabla}f\left(x^{k+1},\boldsymbol{\xi}^k\right)=\text{clip}\left(\nabla f\left(x^{k+1},\boldsymbol{\xi}^k\right),\lambda_{k+1}\right)$
5. $z^{k+1}=z^k-\alpha_{k+1}\widetilde{\nabla}f\left(x^{k+1},\boldsymbol{\xi}^k\right)$
6. $y^{k+1}=\frac{A_ky^k+\alpha_{k+1}z^{k+1}}{A_{k+1}}$

Output: y^N

6. Comparison of Complexities

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

$$\text{Prob}\left\{f(\hat{x})-f(x^*)>\varepsilon\right\}<\beta$$

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

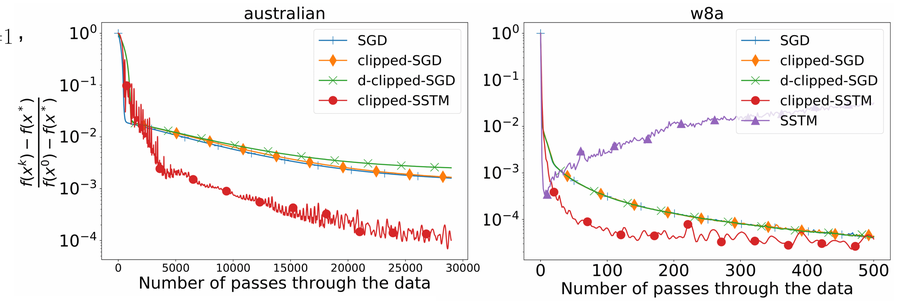
Θ — diameter of the domain (if it is bounded), $R_0=\|x^0-x^*\|$

Method	Complexity	Tails	Domain
SGD [1]	$O\left(\max\left\{\frac{LR_0^2}{\varepsilon},\frac{\sigma^2R_0^2}{\varepsilon^2}\ln^2(\beta^{-1})\right\}\right)$	light	bounded
AC-SA [2,3]	$O\left(\max\left\{\sqrt{\frac{LR_0^2}{\varepsilon}},\frac{\sigma^2R_0^2}{\varepsilon^2}\ln(\beta^{-1})\right\}\right)$	light	arbitrary
RSMD [4]	$O\left(\max\left\{\frac{L\Theta^2}{\varepsilon},\frac{\sigma^2\Theta^2}{\varepsilon^2}\right\}\ln(\beta^{-1})\right)$	heavy	bounded
clipped-SGD [This work]	$O\left(\max\left\{\frac{LR_0^2}{\varepsilon},\frac{\sigma^2R_0^2}{\varepsilon^2}\right\}\ln(\beta^{-1})\right)$	heavy	$\mathbb{R}^n$
clipped-SSTM [This work]	$O\left(\max\left\{\sqrt{\frac{LR_0^2}{\varepsilon}},\frac{\sigma^2R_0^2}{\varepsilon^2}\right\}\ln\frac{LR_0^2+\sigma R_0}{\varepsilon\beta}\right)$	heavy	$\mathbb{R}^n$

7. Numerical Experiments

We conducted several numerical experiments on logistic regression problem

$$\min_{x\in\mathbb{R}^n}\left\{f(x)=\frac{1}{r}\sum_{i=1}^r\log\left(1+\exp\left(-y_i\cdot(Ax)_i\right)\right)\right\}$$



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).