High-Probability Convergence for Composite and Distributed Stochastic Minimization and Variational Inequalities with Heavy-Tailed Noise

Eduard Gorbunov [1] Abdurakhmon Sadiev [2] Marina Danilova [3] Samuel Horváth [1] Gauthier Gidel [4] Pavel Dvurechensky [5] Alexander Gasnikov [6,7,3,8] Peter Richtárik [2]

[1]MBZUAI [2]KAUST [3]MIPT [4]UdeM, CIFAR AI Chair [5] WIAS [6] Innopolis University [7] ISP RAS [8] Skoltech

1. Composite Stochastic Optimization

Composite minimization problem

$$\min_{x\in\mathbb{R}^d}\{\Phi(x) := f(x) + \Psi(x)\}$$

with stochastic first-order oracle:

$\nabla f_\xi(x)$ – an estimate of $\nabla f(x)$

- $f : \mathbb{R}^d \to \mathbb{R}$ – convex smooth function
- $\Psi : \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$ – proper, closed, convex function (composite/regularization term)

Examples:

- Regularized expectation minimization

$$\min_{x\in\mathbb{R}^d}\left\{\Phi(x) = \underbrace{\mathbb{E}_{\xi\sim\mathcal{D}}[f_\xi(x)]}_{f(x)} + \underbrace{\lambda_1\|x\|_1 + \lambda_2\|x\|_2^2}_{\Psi(x)}\right\}$$

- Constrained empirical risk minimization

$$\min_{x\in\mathbb{R}^d}\left\{\Phi(x) = \underbrace{\frac{1}{m}\sum_{i=1}^m f_{\xi_i}(x)}_{f(x)} + \Psi(x)\right\}, \quad \Psi(x) = \begin{cases}0, & \text{if } x\in\mathcal{X}\\ +\infty, & \text{if } x\notin\mathcal{X}\end{cases}$$

Heavy-tailed noise:

$$\mathbb{E}\|\nabla f_\xi(x) - \nabla f(x)\|^\alpha \le \sigma^\alpha, \quad 1 < \alpha \le 2$$

- Such noise appears in various ML problems, including training of LLMs [1] and GANs [2]

2. High-Probability Convergence

In-expectation guarantees:

$$\mathbb{E}[f(x) - f(x^*)] \le \varepsilon \qquad (1)$$

High-probability guarantees:

$$\mathbb{P}\{f(x) - f(x^*) \le \varepsilon\} \ge 1 - \beta \qquad (2)$$

✓ If for method $\mathcal{M}$ we know that (1) is satisfied for $x = x^{N(\varepsilon)}$ after $N(\varepsilon)$ iterations, then for the same method we can guarantee (2) after $N(\varepsilon\beta)$ iterations due to the Markov's inequality:

$$\mathbb{P}\{f(x^{N(\varepsilon\beta)}) - f(x^*) > \varepsilon\} < \frac{\mathbb{E}[f(x^{N(\varepsilon\beta)}) - f(x^*)]}{\varepsilon} \overset{(1)}{\le} \beta$$

✗ Typically $N(\varepsilon)$ has inverse power dependence on ε, e.g., $N(\varepsilon) \sim {}^1/_{\varepsilon^2}$ for SGD in the convex case $\longrightarrow$ this approach gives inverse power-dependence on β in high-probability complexity bounds

✓ High-probability guarantees are more accurate

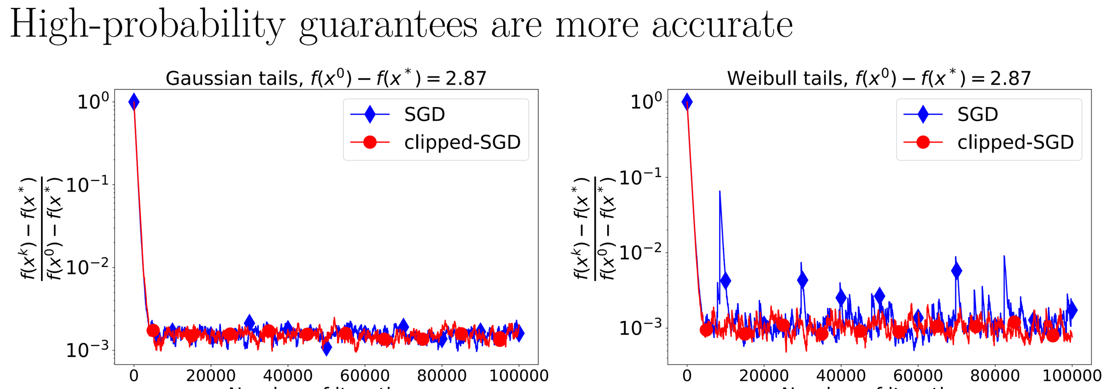

Figure: Typical trajectories of SGD and clipped-SGD applied to solve $\min_{x\in\mathbb{R}}\{f(x) := \|x\|^2/2\}$ with $\nabla f_\xi(x) = x + \xi$ and ξ having Gaussian or Weibull tails with the same variance. Plots are taken from [3].

✓ Gradient clipping improves high-probability convergence in theory (logarithmic dependence on β) and practice [2,3,4,5]

Resolved open question: *how to generalize the existing results to composite/distributed problems?*

Main contributions

- **Methods with clipping of gradient differences for distributed composite minimization**
  - **Key idea**: clip the difference between the stochastic gradients and the shifts that are updated on the fly
  - ✓ The first results showing linear speed-up under bounded α-th moment assumption
  - ✓ The first accelerated high-probability convergence rates and tight high-probability convergence rates for the non-accelerated method in the quasi-strongly convex case
- **Tight convergence rates**
  - ✓ In the known special cases ($\Psi \equiv 0$ and/or $n = 1$), the derived complexity bounds either recover or outperform previously known ones
  - ✓ In certain regimes, the results have optimal (up to logarithms) dependencies on ε
- **Generalization to the case of distributed composite variational inequalities**

3. Failure of Naïve Approach

Standard method for composite optimization:

$$x^{k+1} = \mathrm{prox}_{\gamma\Psi}\left(x^k - \gamma\nabla f(x^k)\right) \qquad \text{(Prox-GD)}$$

- Proximal operator: $\mathrm{prox}_{\gamma\Psi}(x) := \arg\min_{y\in\mathbb{R}^d}\left\{\gamma\Psi(y) + \frac{1}{2}\|y - x\|^2\right\}$

How to incorporate gradient clipping in Prox-GD?

Naïve approach:

$$x^{k+1} = \mathrm{prox}_{\gamma\Psi}\left(x^k - \gamma\,\texttt{clip}(\nabla f(x^k), \lambda_k)\right) \qquad \text{(Prox-clipped-GD)}$$

- Clipping operator: $\texttt{clip}(x,\lambda) := \begin{cases}\min\left\{1, \frac{\lambda}{\|x\|}\right\}x, & \text{if } x \ne 0\\ 0, & \text{if } x = 0\end{cases}$

✗ x^* is not a fixed point: if $\|\nabla f(x^*)\| > \lambda_k$ for all $k \ge k_0$, then

$$x^* \ne \mathrm{prox}_{\gamma\Psi}\left(x^* - \gamma\,\texttt{clip}(\nabla f(x^*), \lambda_k)\right)$$

‼ Decreasing stepsizes are needed for acceleration and tight convergence rates in (quasi-)strongly convex case [4,5]

4. Non-Implementable Solution

Clip the difference $\longrightarrow$ Prox-clipped-SGD-star

$$x^{k+1} = \mathrm{prox}_{\gamma\Psi}\left(x^k - \gamma\left(\nabla f(x^*) + \texttt{clip}(\nabla f_{\xi^k}(x^k) - \nabla f(x^*), \lambda_k)\right)\right)$$

✓ x^* is a fixed point (in the case of deterministic gradients)

✓ Provable high-probability convergence under heavy-tailed noise

✗ Non-implementable method: $\nabla f(x^*)$ is unknown

5. Clipping of Gradient Differences

Approximate $\nabla f(x^*)$ with a learnable shift:

$$x^{k+1} = \mathrm{prox}_{\gamma\Psi}\left(x^k - \gamma\widetilde{g}^k\right), \qquad \text{(Prox-clipped-SGD-shift)}$$

$$\widetilde{g}^k = h^k + \hat{\Delta}^k, \quad h^{k+1} = h^k + \nu\hat{\Delta}^k,$$

$$\hat{\Delta}^k = \texttt{clip}\left(\nabla f_{\xi^k}(x^k) - h^k, \lambda_k\right)$$

- $\nu > 0$ – stepsize for learning the shift

Distributed learning

$$f(x) := \frac{1}{n}\sum_{i=1}^n f_i(x), \quad f_i(x) := \mathbb{E}_{\xi_i\sim\mathcal{D}_i}[f_{\xi_i}(x)]$$

- n workers/clients are connected with a parameter-server
- $f_i(x)$ – loss on the data available on worker i

Distributed version – DProx-clipped-SGD-shift

$$x^{k+1} = \mathrm{prox}_{\gamma\Psi}\left(x^k - \gamma\widetilde{g}^k\right)$$

$$\widetilde{g}^k = \frac{1}{n}\sum_{i=1}^n \widetilde{g}_i^k, \quad \widetilde{g}_i^k = h_i^k + \hat{\Delta}_i^k, \quad h_i^{k+1} = h_i^k + \nu\hat{\Delta}_i^k$$

$$\hat{\Delta}_i^k = \texttt{clip}\left(\nabla f_{\xi_i^k}(x^k) - h_i^k, \lambda_k\right)$$

- Each worker updates its own shift h_i^k
- Even with $\Psi \equiv 0$ shifts are needed: otherwise we have

$$x^* \ne x^* - \frac{\gamma}{n}\sum_{i=1}^n \texttt{clip}\left(\nabla f_i(x^*), \lambda_k\right) \text{ in general}$$

- It is sufficient to store $h^k := \frac{1}{n}\sum_{i=1}^n h_i^k$ on the server

Table: Summary of known and new high-probability complexity results for solving (non-) composite (non-) distributed smooth optimization problem. Complexity is the number of stochastic oracle calls (per worker) needed for a method to guarantee that $\mathbb{P}\{\text{Metric} \le \varepsilon\} \ge 1 - \beta$ for some $\varepsilon > 0$, $\beta \in (0,1]$ and "Metric" is taken from the corresponding column. Numerical and logarithmic factors are omitted for simplicity. Notation: R = any upper bound on $\|x^0 - x^*\|$; ζ_* = any upper bound on $\sqrt{\frac{1}{n}\sum_{i=1}^n\|\nabla f_i(x^*)\|^2}$; $\widehat{R}^2 = R\left(3R + L^{-1}(2\eta\sigma + \|\nabla f(x^0)\|)\right)$ for some $\eta > 0$ (one can show that $\widehat{R}^2 = \Theta(R^2 + {}^{R\zeta_*}/_L)$ when $n = 1$).

Function	Method	Reference	Metric	Complexity	Composite?	Distributed?
Convex	Clipped-SMD(1)	[2]	$\Phi(\overline{x}^K) - \Phi(x^*)$	$\max\left\{\frac{L\widehat{R}^2}{\varepsilon}, \left(\frac{\sigma R}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right\}$	✓	✗
	Clipped-ASMD	[2]	$\Phi(y^K) - \Phi(x^*)$	$\max\left\{\sqrt{\frac{LR^2}{\varepsilon}}, \left(\frac{\sigma R}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right\}$	✓✗(2)	✗
	DProx-clipped-SGD-shift	This paper	$\Phi(\overline{x}^K) - \Phi(x^*)$	$\max\left\{\frac{LR^2}{\varepsilon}, \frac{R\zeta_*}{\sqrt{n}\varepsilon}, \frac{1}{n}\left(\frac{\sigma R}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right\}$	✓	✓
	DProx-clipped-SSTM-shift	This paper	$\Phi(y^K) - \Phi(x^*)$	$\max\left\{\sqrt{\frac{LR^2}{\varepsilon}}, \sqrt{\frac{R\zeta_*}{\sqrt{n}\varepsilon}}, \frac{1}{n}\left(\frac{\sigma R}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right\}$	✓	✓
Strongly convex	clipped-SGD	[1]	$\|x^K - x^*\|^2$	$\max\left\{\frac{L}{\mu}, \left(\frac{\sigma^2}{\mu^2\varepsilon}\right)^{\frac{\alpha}{2(\alpha-1)}}\right\}$	✗	✗
	DProx-clipped-SGD-shift	This paper	$\|x^K - x^*\|^2$	$\max\left\{\frac{L}{\mu}, \frac{1}{n}\left(\frac{\sigma^2}{\mu^2\varepsilon}\right)^{\frac{\alpha}{2(\alpha-1)}}\right\}$	✓	✓

(1) The authors additionally assume that for a chosen point $\widehat{x}$ from the domain and for $\eta > 0$ one can compute an estimate $\widehat{g}$ such that $\mathbb{P}\{\|\widehat{g} - \nabla f(\hat{x})\| > \eta\sigma\} \le \epsilon$. Such an estimate can be found using geometric median of $\mathcal{O}(\ln\epsilon^{-1})$ samples [6].

(2) The authors assume that $\nabla f(x^*) = 0$, which is not true for general composite optimization.

6. Convergence Results

Assumptions

For all $i = 1,\ldots,n$ and $x, y \in \mathbb{R}^d$ we have

A1. $\mathbb{E}\|\nabla f_{\xi}(x) - \nabla f_i(x)\|^\alpha \le \sigma^\alpha$ for some $\alpha \in (1,2]$

A2. Smoothness: $\|\nabla f_i(x) - \nabla f_i(y)\| \le L\|x - y\|$

A3. Strong convexity: $f(y) \ge f(x) + \langle\nabla f(x), y - x\rangle + \frac{\mu}{2}\|y - x\|^2$

Convergence of DProx-clipped-SGD-shift

Let the above assumptions hold with $\mu = 0$. Then, the iterates produced by DProx-clipped-SGD-shift after K iterations with

$$\gamma = \Theta\left(\min\left\{\frac{1}{LA}, \frac{R\sqrt{n}}{A\zeta_*}, \frac{Rn^{\frac{\alpha-1}{\alpha}}}{\sigma K^{\frac{1}{\alpha}}A^{\frac{\alpha-1}{\alpha}}}\right\}\right),$$

$$\lambda_k \equiv \lambda = \Theta\left(\frac{nR}{\gamma A}\right), \quad A = \ln\frac{48nK}{\beta}, \quad \zeta_* \ge \sqrt{\frac{1}{n}\sum_{i=1}^n\|\nabla f_i(x^*)\|^2}$$

with probability at least $1 - \beta$ satisfy

$$\Phi(\overline{x}^K) - \Phi(x^*) = \mathcal{O}\left(\max\left\{\frac{LR^2A}{K}, \frac{R\zeta_*A}{\sqrt{n}K}, \frac{\sigma RA^{\frac{\alpha-1}{\alpha}}}{(nK)^{\frac{\alpha-1}{\alpha}}}\right\}\right),$$

where $\overline{x}^K = \frac{1}{K+1}\sum_{k=0}^K x^k$.

✓ Logarithmic dependence on confidence level β

✓ Linear speed-up in the complexity (see the Table)

- $\nu = 0$ when $\mu = 0$ and $\nu = \Theta({}^1/_A)$ when $\mu > 0$

7. Acceleration

DProx-clipped-SSTM-shift: $x^0 = y^0 = z^0$, $A_0 = \alpha_0 = 0$, $\alpha_{k+1} = \frac{k+2}{2aL}$, $A_{k+1} = A_k + \alpha_{k+1}$ and

$$x^{k+1} = \frac{A_k y^k + \alpha_{k+1}z^k}{A_{k+1}}, \quad z^{k+1} = \mathrm{prox}_{\alpha_{k+1}\Psi}\left(z^k - \alpha_{k+1}\widetilde{g}(x^{k+1})\right),$$

$$y^{k+1} = \frac{A_k y^k + \alpha_{k+1}z^{k+1}}{A_{k+1}},$$

$$\widetilde{g}(x^{k+1}) = \frac{1}{n}\sum_{i=1}^n\widetilde{g}_i(x^{k+1}), \quad \widetilde{g}_i(x^{k+1}) = h_i^k + \hat{\Delta}_i^k,$$

$$h_i^{k+1} = h_i^k + \nu_k\hat{\Delta}_i^k, \quad \hat{\Delta}_i^k = \texttt{clip}\left(\nabla f_{\xi_i^k}(x^{k+1}) - h_i^k, \lambda_k\right)$$

Convergence of DProx-clipped-SSTM-shift

Let the above assumptions hold with $\mu = 0$. Then, the iterates produced by DProx-clipped-SSTM-shift after K iterations with

$$\nu_k = \begin{cases}\frac{2k+5}{(k+3)^2}, & \text{if } k > K_0,\\ \Theta\left(\frac{(k+2)^2}{A^2(K_0+2)^2}\right), & \text{if } k \le K_0,\end{cases} \quad \lambda_k = \Theta\left(\frac{nR}{\alpha_{k+1}A}\right),$$

$$K_0 = \Theta(A^2), \quad a = \Theta\left(\max\left\{2, \frac{A^4}{n}, \frac{A^3\zeta_*}{L\sqrt{n}R}, \frac{\sigma K^{(\alpha+1)/\alpha}A^{(\alpha-1)/\alpha}}{LRn^{\alpha-1/\alpha}}\right\}\right),$$

with probability at least $1 - \beta$ satisfy

$$\Phi(y^K) - \Phi(x^*) = \mathcal{O}\left(\max\left\{\frac{LR^2(1 + {}^{A^4}/_n)}{K^2}, \frac{R\zeta_*A^3}{\sqrt{n}K^2}, \frac{\sigma RA^{\frac{\alpha-1}{\alpha}}}{(nK)^{\frac{\alpha-1}{\alpha}}}\right\}\right).$$

References

[1] J. Zhang et al. "Why are adaptive methods good for attention models?." NeurIPS 2020.

[2] E. Gorbunov et al. "Clipped stochastic methods for variational inequalities with heavy-tailed noise." NeurIPS 2022.

[3] E. Gorbunov et al. "Stochastic optimization with heavy-tailed noise via accelerated gradient clipping." NeurIPS 2020.

[4] A. Sadiev et al. "High-probability bounds for stochastic optimization and variational inequalities: the case of unbounded variance." ICML 2023.

[5] T. D. Nguen et al. "Improved convergence in high probability of clipped gradient methods with heavy tails." NeurIPS 2023.

[6] S. Minsker. "Geometric median and robust estimation in banach spaces." Bernoulli 2015.