High-Probability Bounds for Stochastic Optimization and Variational Inequalities: the Case of Unbounded Variance

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

[1]King Abdullah University of Science and Technology [2]Moscow Institute of Physics and Technology [3]Mohamed bin Zayed University of Artificial Intelligence [4]Mila, Université de Montréal [5]Canada CIFAR AI Chair
[6]Weierstrass Institute for Applied Analysis and Stochastics [7]HSE University [8] Institute for Information Transmission Problems RAS

1.Preliminaries

Main Problems

- **Minimization problem:**

$$\min_{x\in\mathbb{R}^d}\left\{f(x)=\mathbb{E}_{\xi\sim\mathcal{D}}[f_\xi(x)]\right\}, \qquad (1)$$

where ξ is a random variable with distribution $\mathcal{D}$.

- **Variational inequality problem:**

$$\text{find } x^*\in\mathbb{R}^d \text{ such that } F(x^*)=0, \qquad (2)$$

where $F(x)=\mathbb{E}_{\xi\sim\mathcal{D}}[F_\xi(x)]$.

Bounded α-Moment Assumption

We assume that there exist some set $Q\subseteq\mathbb{R}^d$ and values $\sigma\geq 0$, $\alpha\in(1,2]$ such that for all $x\in Q$

A1 for problem (1) $\mathbb{E}_{\xi\sim\mathcal{D}}[\nabla f_\xi(x)]=\nabla f(x)$ and

$$\mathbb{E}_{\xi\sim\mathcal{D}}[\|\nabla f_\xi(x)-\nabla f(x)\|^\alpha]\leq\sigma^\alpha, \qquad (3)$$

A2 for problem (2) $\mathbb{E}_{\xi\sim\mathcal{D}}[F_\xi(x)]=F(x)$ and

$$\mathbb{E}_{\xi\sim\mathcal{D}}[\|F_\xi(x)-F(x)\|^\alpha]\leq\sigma^\alpha. \qquad (4)$$

Assumptions for Minimization Problem (1)

A3, **A4**: Smoothness and lower-boundedness: $\forall x,y\in Q$ we have $\|\nabla f(x)-\nabla f(y)\|\leq L\|x-y\|$ and $f_*=\inf_{x\in Q}f(x)>-\infty$

A5 Polyak-Łojasiewicz (PŁ) condition: $\forall x\in Q$ and $x^*=\arg\min_{x\in\mathbb{R}^d}f(x)$ we have $\|\nabla f(x)\|^2\geq 2\mu\left(f(x)-f(x^*)\right)$.

A6 μ-quasi-strong convexity: $\forall x\in Q$ and $x^*=\arg\min_{x\in\mathbb{R}^d}f(x)$ we have $f(x^*)\geq f(x)+\langle\nabla f(x),x^*-x\rangle+\frac{\mu}{2}\|x-x^*\|^2$.

A7 μ-strongly convexity: $\forall x,y\in Q$ we have $f(y)\geq f(x)+\langle\nabla f(x),y-x\rangle+\frac{\mu}{2}\|y-x\|^2$.
When $\mu=0$ function f is called convex.

Assumptions for Variational Inequality Problem (2)

A8 Lipschitzness: $\forall x,y\in Q$ we have $\|F(x)-F(y)\|\leq L\|x-y\|$.

A9 Monotonicity: $\forall x,y\in Q$ we have $\langle F(x)-F(y),x-y\rangle\geq 0$.

A10 μ-quasi-strong monotonicity: $\forall x\in Q$ and x^* such that $F(x^*)=0$ we have $\langle F(x),x-x^*\rangle\geq\mu\|x-x^*\|^2$.

A11 Star-cocoercivity: $\forall x\in Q$ x^* such that $F(x^*)=0$ we have $\|F(x)\|^2\leq\ell\langle F(x),x-x^*\rangle$.

2. In-Expectation vs High-Probability

In-expectation guarantees: $\mathbb{E}\left[\|x-x^*\|^2\right]\leq\varepsilon$, $\mathbb{E}\left[f(x)-f(x^*)\right]\leq\varepsilon$, $\mathbb{E}\left[\|\nabla f(x^*)\|^2\right]\leq\varepsilon$

✗ Typically, depend only on some moments of stochastic gradient, e.g., variance

High-probability guarantees: $\mathbb{P}\left\{\|x-x^*\|^2\leq\varepsilon\right\}\geq 1-\beta$, $\mathbb{P}\left\{f(x)-f(x^*)\leq\varepsilon\right\}\geq 1-\beta$, $\mathbb{P}\left\{\|\nabla f(x^*)\|^2\leq\varepsilon\right\}\geq 1-\beta$

✓ Sensitive to the distribution of the stochastic gradient noise

3. Our Contribution

◇ **New high-probability results under Assump. A1, A2**
- Smooth (quasi-strongly) convex minimization
- Monotone/quasi-strongly monotone VIP

◇ **Weaker assumptions in the non-convex case**
- We do not assume boundedness of the gradient
- Extension to the functions satisfying PŁ-condition **A5**

◇ **Failure of SGD**
- We construct an example of a strongly convex smooth problem and stochastic oracle with bounded variance such that to achieve $\mathbb{P}\{\|x^k-x^*\|^2>\varepsilon\}\leq\beta$ SGD requires $\Omega\left(\sigma^2/\mu\sqrt{\varepsilon\beta}\right)$ iterations

4. Failure of SGD

✗ SGD $x^{k+1}=x^k-\gamma\nabla f_{\xi^k}(x^k)$ can diverge in expectation, when Assumption **A1** is satisfied with $\alpha<2$.

✗ There are no high-probability convergence results for SGD having logarithmic dependence on $1/\beta$.

Theorem 1

For any $\varepsilon>0$ and sufficiently small $\beta\in(0,1)$ there exist problem (1) such that Assumptions **A1**, **A3**, and **A7** hold with $Q=\mathbb{R}^d$, $\alpha=2$, $0<\mu\leq L$ and for the iterates produced by SGD with any stepsize $\gamma>0$

$$\mathbb{P}\left\{\|x^k-x^*\|^2\geq\varepsilon\right\}\leq\beta \implies k=\Omega\left(\frac{\sigma}{\mu\sqrt{\varepsilon\beta}}\right).$$

- This partially justifies the need of applying some non-linearity to the stochastic gradient (e.g., clipping).

5. Gradient Clipping

The clipping operator is defined as

$$\texttt{clip}(x,\lambda)=\begin{cases}\min\left\{1,\frac{\lambda}{\|x\|}\right\}x, & \text{if } x\neq 0,\\ 0, & \text{otherwise.}\end{cases} \qquad (5)$$

- Clipping creates bias: $\mathbb{E}[\texttt{clip}(\nabla f_\xi(x),\lambda)]\neq\nabla f(x)$ in general

Lemma 1

Let X be a random vector in $\mathbb{R}^d$ and $\tilde{X}=\texttt{clip}(X,\lambda)$. Then, $\left\|\tilde{X}-\mathbb{E}[\tilde{X}]\right\|\leq 2\lambda$. Moreover, if for some $\sigma\geq 0$ and $\alpha\in[1,2)$ $\mathbb{E}[X]=x\in\mathbb{R}^d$, $\mathbb{E}[\|X-x\|^\alpha]\leq\sigma^\alpha$ and $\|x\|\leq\lambda/2$, then

$$\left\|\mathbb{E}[\tilde{X}]-x\right\|\leq\frac{2^\alpha\sigma^\alpha}{\lambda^{\alpha-1}}, \qquad (6)$$

$$\mathbb{E}\left[\left\|\tilde{X}-\mathbb{E}[\tilde{X}]\right\|^2\right]\leq 18\lambda^{2-\alpha}\sigma^\alpha. \qquad (7)$$

- clipped-SGD: $x^{k+1}=x^k-\gamma\cdot\texttt{clip}\left(\nabla f_{\xi^k}(x^k),\lambda_k\right)$
- In our proofs, we separate "stochastic" and "deterministic" parts
- In the analysis of clipped-SGD for convex problems, we derive

$$\gamma(k+1)\left(f(\overline{x}^k)-f(x^*)\right)\lessapprox R_0^2-R_{k+1}^2+\gamma\sum_{t=0}^k\langle\eta_t,\theta_t\rangle+\gamma^2\sum_{t=0}^k\|\theta_t\|^2,$$

where $R_t=\|x^t-x^*\|$, $\overline{x}^k=\frac{1}{k+1}\sum_{t=0}^k x^t$, $\eta_t=x^t-x^*-\gamma\nabla f(x^t)$, $\theta_t=\texttt{clip}(\nabla f_{\xi^t}(x^t),\lambda_t)-\nabla f(x_t)$

- We upper-bound the sums with θ^t using Bernstein's inequality for martingale differences and do it inductively (to ensure that R_t is bounded with high probability)

6. Results for clipped-SGD

Theorem 2

Let $k\geq 0$ and $\beta\in(0,1]$ are such that $A=\ln\frac{4(K+1)}{\beta}\geq 1$.

Case 1. Let Assumptions **A1**, **A3**, **A4** hold for $Q=\{x\in\mathbb{R}^d\mid\exists y\in\mathbb{R}^d: f(y)\leq f_*+2\Delta \text{ and } \|x-y\|\leq\sqrt{\Delta}/20\sqrt{L}\}$, $\Delta\geq f(x^0)-f_*$ and $0<\gamma\leq\mathcal{O}\left(\min\{1/LA,\sqrt{\Delta}/\sigma\sqrt{L}K^{1/\alpha}A^{(\alpha-1)/\alpha}\}\right)$, $\lambda_k=\lambda=\Theta(\sqrt{\Delta}/\sqrt{L}\gamma A)$.

Case 2. Let Assumptions **A1**, **A3**, **A5** hold for $Q=\{x\in\mathbb{R}^d\mid\exists y\in\mathbb{R}^d: f(y)\leq f_*+2\Delta \text{ and } \|x-y\|\leq\sqrt{\Delta}/20\sqrt{L}\}$, $\Delta\geq f(x^0)-f_*$ and $0<\gamma=\mathcal{O}\left(\min\{1/LA,\ln(B_K)/\mu(K+1)\}\right)$, $B_K=\Theta\left(\max\{2,(K+1)^{2(\alpha-1)/\alpha}\mu^2\Delta/L\sigma^2A^{2(\alpha-1)/\alpha}\ln^2(B_K)\}\right)$, $\lambda_k=\Theta(\exp(-\gamma\mu(1+k/2))\sqrt{\Delta}/\sqrt{L}\gamma A)$.

Case 3. Let Assumptions **A1**, **A3**, **A7** with $\mu=0$ hold for $Q=B_{3R}(x^*)$, $R\geq\|x^0-x^*\|$ and $0<\gamma\leq\mathcal{O}(\min\{1/LA,R/\sigma K^{1/\alpha}A^{(\alpha-1)/\alpha}\})$, $\lambda_k=\lambda=\Theta(R/\gamma A)$.

Case 4. Let Assumptions **A1**, **A3**, **A6** with $\mu>0$ hold for $Q=B_{3R}(x^*)$, $R\geq\|x^0-x^*\|$ and $0<\gamma=\mathcal{O}\left(\min\{1/LA,\ln(B_K)/\mu(K+1)\}\right)$, $B_K=\Theta\left(\max\{2,(K+1)^{2(\alpha-1)/\alpha}\mu^2R^2/\sigma^2A^{2(\alpha-1)/\alpha}\ln^2(B_K)\}\right)$, $\lambda_k=\Theta(\exp(-\gamma\mu(1+k/2))R/\gamma A)$.

Then to guarantee $\frac{1}{K+1}\sum_{k=0}^k\|\nabla f(x^k)\|^2\leq\varepsilon$ in **Case 1**, $f(x^K)-f(x^*)\leq\varepsilon$ in **Case 2**, $f(\bar{x}^K)-f(x^*)\leq\varepsilon$ in **Case 3** with $\bar{x}^K=\frac{1}{K+1}\sum_{k=0}^K x^k$, $\|x^K-x^*\|^2\leq\varepsilon$ in **Case 4** with probability $\geq 1-\beta$ clipped-SGD requires

$$\textbf{\textit{Case 1}}:\quad\widetilde{\mathcal{O}}\left(\max\left\{\frac{L\Delta}{\varepsilon},\left(\frac{\sqrt{L\Delta}\sigma}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right\}\right)$$

$$\textbf{\textit{Case 2}}:\quad\widetilde{\mathcal{O}}\left(\max\left\{\frac{L}{\mu},\left(\frac{L\sigma^2}{\mu^2\varepsilon}\right)^{\frac{\alpha}{2(\alpha-1)}}\right\}\right)$$

$$\textbf{\textit{Case 3}}:\quad\widetilde{\mathcal{O}}\left(\max\left\{\frac{LR^2}{\varepsilon},\left(\frac{\sigma R}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right\}\right)$$

$$\textbf{\textit{Case 4}}:\quad\widetilde{\mathcal{O}}\left(\max\left\{\frac{L}{\mu},\left(\frac{\sigma^2}{\mu^2\varepsilon}\right)^{\frac{\alpha}{2(\alpha-1)}}\right\}\right)$$

oracle calls.

- For $\alpha=2$ the derived complexity bounds match the best-known ones for clipped-SGD
- The second term under the maximum in (8) (quasi-strongly convex functions) is optimal up to logarithmic factors

7. Results for clipped-SSTM

- Clipped Stochastic Similar Triangles Method:

$$x^{k+1}=\frac{A_ky^k+\alpha_{k+1}z^k}{A_{k+1}},$$
$$z^{k+1}=z^k-\alpha_{k+1}\cdot\texttt{clip}\left(\nabla f_{\xi^k}(x^{k+1}),\lambda_k\right),$$
$$y^{k+1}=\frac{A_ky^k+\alpha_{k+1}z^{k+1}}{A_{k+1}},$$

where $A_0=\alpha_0=0$, $\alpha_{k+1}=\frac{k+2}{2aL}$, $A_{k+1}=A_k+\alpha_{k+1}$, and ξ^k is sampled from $\mathcal{D}_k$ independently from previous steps.

Theorem 3

Let Assumptions **A1**, **A3**, **A7** with $\mu=0$ hold for $Q=B_{3R}(x^*)$, $R\geq\|x^0-x^*\|^2$ and $a=\Theta(\max\{A^2,\sigma K^{(\alpha+1)/\alpha}A^{(\alpha-1)/\alpha}/LR\})$, $\lambda_k=\Theta(R/(\alpha_{k+1}A))$, where $\beta\in(0,1]$ are such that $A=\ln\frac{4K}{\beta}\geq 1$. Then to guarantee $f(y^K)-f(x^*)\leq\varepsilon$ with probability $\geq 1-\beta$ clipped-SSTM requires

$$\widetilde{\mathcal{O}}\left(\max\left\{\sqrt{\frac{LR^2}{\varepsilon}},\left(\frac{\sigma R}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right\}\right)\quad\text{oracle calls.}$$

Moreover, with probability $\geq 1-\beta$ the iterates of clipped-SSTM stay in the ball $B_{2R}(x^*)$: $\{x^k\}_{k=0}^{K+1},\{y^k\}_{k=0}^K,\{z^k\}_{k=0}^K\subseteq B_{2R}(x^*)$.

- For $\alpha=2$ the derived complexity bounds match the best-known ones for clipped-SSTM
- For strongly convex problems, we have a restarted version (R-clipped-SSTM)

8. Comparison with Prior Work

◇ Minimization Problem:

Setup	Method	Complexity	α
A7, ($\mu=0$)	RSMD [1]	$\max\left\{\frac{LD^2}{\varepsilon},\frac{\sigma^2D^2}{\varepsilon}\right\}$	2
	clipped-SGD [2]	$\max\left\{\frac{LR^2}{\varepsilon},\frac{\sigma^2R^2}{\varepsilon}\right\}$	2
	clipped-SSTM [2]	$\max\left\{\sqrt{\frac{LR^2}{\varepsilon}},\frac{\sigma^2R^2}{\varepsilon}\right\}$	2
	clipped-SGD	$\max\left\{\frac{LR^2}{\varepsilon},\left(\frac{\sigma R}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right\}$	$(1,2]$
	clipped-SSTM	$\max\left\{\sqrt{\frac{LR^2}{\varepsilon}},\left(\frac{\sigma R}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right\}$	$(1,2]$
A7, ($\mu>0$)	restarted-RSMD [1]	$\max\left\{\frac{L}{\mu},\frac{\sigma^2}{\mu\varepsilon}\right\}$	2
	proxBoost [4]	$\max\left\{\frac{L}{\mu},\frac{\sigma^2}{\mu\varepsilon}\right\}$	2
	R-clipped-SGD [2]	$\max\left\{\frac{L}{\mu},\frac{\sigma^2}{\mu\varepsilon}\right\}$	2
	R-clipped-SSTM [2]	$\max\left\{\sqrt{\frac{L}{\mu}},\frac{\sigma^2}{\mu\varepsilon}\right\}$	2
	R-clipped-SSTM	$\max\left\{\sqrt{\frac{L}{\mu}},\left(\frac{\sigma^2}{\mu\varepsilon}\right)^{\frac{\alpha}{2(\alpha-1)}}\right\}$	$(1,2]$
A6 ($\mu>0$)	clipped-SGD	$\max\left\{\frac{L}{\mu},\left(\frac{\sigma^2}{\mu^2\varepsilon}\right)^{\frac{\alpha}{2(\alpha-1)}}\right\}$	$(1,2]$
A4	MSGD [L5]	$\max\left\{\frac{L^2\Delta^2}{\varepsilon},\frac{\sigma^4}{\varepsilon^2}\right\}$	✗
	clipped-NMSGD [6]	$\left(\frac{G^2}{\varepsilon}\right)^{\frac{3\alpha-2}{2\alpha-2}}$	$(1,2]$
	clipped-SGD	$\max\left\{\frac{L\Delta}{\varepsilon},\left(\frac{\sqrt{L\Delta}\sigma}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right\}$	$(1,2]$
A5	clipped-SGD	$\max\left\{\frac{L}{\mu},\left(\frac{L\sigma^2}{\mu^2\varepsilon}\right)^{\frac{\alpha}{2(\alpha-1)}}\right\}$	$(1,2]$

- Column "Setup" indicates the assumptions made in addition to Assumptions **A1**, **A3**

◇ Variational Inequality Problem:

Setup	Method	Complexity	α
A8, **A9**	Mirror-Prox [7]	$\max\left\{\frac{LD^2}{\varepsilon},\frac{\sigma^2D^2}{\varepsilon}\right\}$	✗
	clipped-SEG [3]	$\max\left\{\frac{LR^2}{\varepsilon},\frac{\sigma^2R^2}{\varepsilon}\right\}$	2
	clipped-SEG	$\max\left\{\frac{LR^2}{\varepsilon},\left(\frac{\sigma R}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right\}$	$(1,2]$
A8, **A10**	clipped-SEG [3]	$\max\left\{\frac{L}{\mu},\frac{\sigma^2}{\mu\varepsilon}\right\}$	2
	clipped-SEG	$\max\left\{\frac{L}{\mu},\left(\frac{\sigma^2}{\mu^2\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right\}$	$(1,2]$
A9, **A11**	clipped-SGDA [3]	$\max\left\{\frac{\ell R^2}{\varepsilon},\frac{\sigma^2R^2}{\varepsilon^2}\right\}$	2
	clipped-SGDA	$\max\left\{\frac{\ell R^2}{\varepsilon},\left(\frac{\sigma R}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right\}$	$(1,2]$
A11	clipped-SGDA [3]	$\max\left\{\frac{\ell^2R^2}{\varepsilon},\frac{\ell^2\sigma^2R^2}{\varepsilon^2}\right\}$	2
	clipped-SGDA	$\max\left\{\frac{\ell^2R^2}{\varepsilon},\left(\frac{\ell\sigma R}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right\}$	$(1,2]$
A10, **A11**	clipped-SGDA [3]	$\max\left\{\frac{\ell}{\mu},\frac{\sigma^2}{\mu^2\varepsilon}\right\}$	2
	clipped-SGDA	$\max\left\{\frac{\ell}{\mu},\left(\frac{\sigma^2}{\mu^2\varepsilon}\right)^{\frac{\alpha}{2(\alpha-1)}}\right\}$	$(1,2]$

References

[1] Nazin, A. V., Nemirovsky, A., Tsybakov, A. B., and Juditsky, A. Algorithms of robust stochastic optimization based on mirror descent method. Automation and Remote Control 2019.

[2] Gorbunov, E., Danilova, M., and Gasnikov, A. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. NeurIPS 2020.

[3] Gorbunov, E., Danilova, M., Dobre, D., Dvurechensky, P., Gasnikov, A., and Gidel, G. Clipped stochastic methods for variational inequalities with heavy-tailed noise. NeurIPS 2022.

[4] Davis, D., Drusvyatskiy, D., Xiao, L., and Zhang, J. From low probability to high confidence in stochastic convex optimization. JMLR 2021

[5] Li, X. and Orabona, F. A high probability analysis of adaptive SGD with momentum. arXiv preprint arXiv:2007.14294, 2020.

[6] Cutkosky, A. and Mehta, H. High-probability bounds for non-convex stochastic optimization with heavy tails. NeurIPS 2021.

[7] Juditsky, A., Nemirovski, A., and Tauvel, C. Solving variational inequalities with stochastic mirror-prox algorithm. Stochastic Systems, 2011