Clipped Methods for Stochastic Optimization with Heavy-Tailed Noise

TES Conference on Mathematical Optimization for Machine Learning, Berlin

Eduard Gorbunov

September 15, 2023

Mohamed bin Zayed University of Artificial Intelligence

Outline

1. Clipping and Heavy-Tailed Noise
2. In-Expectation Guarantees vs High-Probability Convergence
3. Main Results

The Talk is Based on Four Papers

- Gorbunov, E., Danilova, M., & Gasnikov, A. (2020). *Stochastic optimization with heavy-tailed noise via accelerated gradient clipping.* NeurIPS 2020
- Gorbunov, E., Danilova, M., Shibaev, I., Dvurechensky, P., & Gasnikov, A. (2021). *Near-optimal high probability complexity bounds for non-smooth stochastic optimization with heavy-tailed noise.* arXiv:2106.05958
- Gorbunov, E., Danilova, M., Dobre, D., Dvurechenskii, P., Gasnikov, A., & Gidel, G. (2022). *Clipped stochastic methods for variational inequalities with heavy-tailed noise.* NeurIPS 2022.
- Sadiev, A., Danilova, M., Gorbunov, E., Horváth, S., Gidel, G., Dvurechensky, P., Gasnikov, A., & Richtárik, P. (2023). *High-probability bounds for stochastic optimization and variational inequalities: the case of unbounded variance.* ICML 2023.

Clipping and Heavy-Tailed Noise

Stochastic Gradient Descent (*SGD*)

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

- f – the function to be minimized
- $\nabla f(x^k, \xi^k)$ – stochastic gradient, i.e., *unbiased* estimate of $\nabla f(x^k)$: $\mathbb{E}_{\xi^k}[\nabla f(x^k, \xi^k)] = \nabla f(x^k)$

Clipped Stochastic Gradient Descent (*clipped-SGD*)

$$x^{k+1} = x^k - \gamma \cdot \texttt{clip}\left(\nabla f(x^k, \xi^k), \lambda\right) \qquad (2)$$

- $\texttt{clip}(x, \lambda) = \min\{1, \lambda/\|x\|\}x$
- $\texttt{clip}(\nabla f(x^k, \xi^k), \lambda)$ – *biased* estimate of $\nabla f(x^k)$: $\mathbb{E}_{\xi^k}[\texttt{clip}(\nabla f(x^k, \xi^k), \lambda)] \neq \nabla f(x^k)$

Origin of Clipping

- Gradient clipping was proposed in (Pascanu et al., 2013). Originally it was used to handle exploding and vanishing gradients in RNNs.

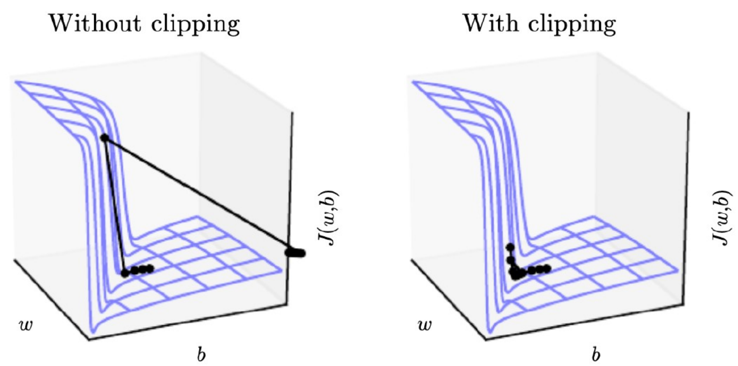

Figure 1: from (Goodfellow et al., 2016)

Few Years Later in NLP...

- Merity et al. (2017) use gradient clipping for LSTM
- Peters et al. (2017) trained their deep bidirectional language model with *Adam* + clipping
- Mosbach et al. (2020) fine-tune BERT using *AdamW* + clipping

Few Years Later in NLP...

- Merity et al. (2017) use gradient clipping for LSTM
- Peters et al. (2017) trained their deep bidirectional language model with *Adam* + clipping
- Mosbach et al. (2020) fine-tune BERT using *AdamW* + clipping

It seems that gradient clipping is an important component in training these models. Why?

Heavy-Tailed Noise in Stochastic Gradients

Let us look at the distribution of $\|\nabla f(x,\xi) - \nabla f(x)\|$ in two settings:

- Standard CV task: training ResNet50 on ImageNet dataset
- Standard NLP task: training BERT on Wikipedia+Books dataset

Heavy-Tailed Noise in Stochastic Gradients

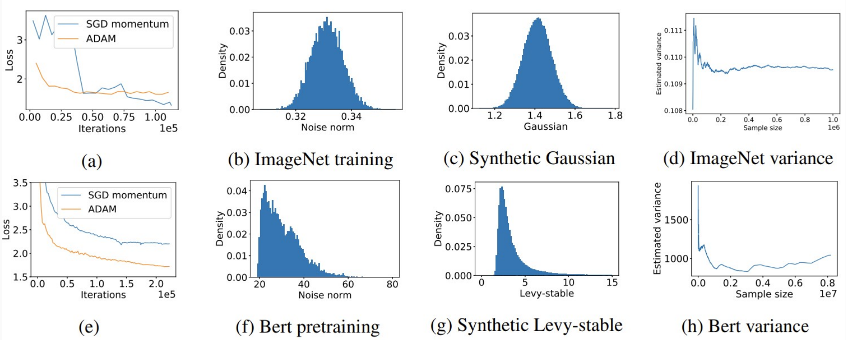

Figure 2: from (Zhang et al., 2020)

We see that *ADAM* is much better than *SGD* when the noise in the stochastic gradient is heavy-tailed

Adam and *clipped-SGD*

- *clipped-SGD*:

$$x^{k+1} = x^k - \gamma \cdot \texttt{clip}\left(\nabla f(x^k, \boldsymbol{\xi}^k), \lambda_k\right)$$

- *Adam*:

$$m_k = \beta_1 m_{k-1} + (1 - \beta_1)\nabla f(x^k, \boldsymbol{\xi}^k),$$
$$v_k = \beta_2 v_{k-1} + (1 - \beta_2)(\nabla f(x^k, \boldsymbol{\xi}^k))^2,$$
$$x^{k+1} = x^k - \frac{\gamma}{\sqrt{v^k + \delta}} m^k$$

  - When $\beta_1 = 0$ *Adam* (*RMSprop*) can be seen as *clipped-SGD* with "adaptive" λ_k

Definition of Heavy-Tailed Noise in Stochastic Gradients

- Random vector X has light tails if

$$\mathbb{P}\{\|X - \mathbb{E}[X]\| \geq b\} \leq 2\exp\left(-\frac{b^2}{2\sigma^2}\right) \quad \forall b > 0. \tag{3}$$

  The above condition is equivalent (up to the numerical factor in σ) to

$$\mathbb{E}\left[\exp\left(\frac{\|X - \mathbb{E}[X]\|^2}{\sigma^2}\right)\right] \leq \exp(1). \tag{4}$$

Definition of Heavy-Tailed Noise in Stochastic Gradients

- Random vector X has light tails if

$$\mathbb{P}\{\|X - \mathbb{E}[X]\| \geq b\} \leq 2\exp\left(-\frac{b^2}{2\sigma^2}\right) \quad \forall b > 0. \tag{3}$$

  The above condition is equivalent (up to the numerical factor in σ) to

$$\mathbb{E}\left[\exp\left(\frac{\|X - \mathbb{E}[X]\|^2}{\sigma^2}\right)\right] \leq \exp(1). \tag{4}$$

- Otherwise we say that X has heavy tails. However, in this talk, we will assume that it has bounded central α-th moment for some $\alpha \in (1, 2]$:

$$\mathbb{E}\left[\|X - \mathbb{E}[X]\|^\alpha\right] \leq \sigma^\alpha \tag{5}$$

In-Expectation Guarantees vs High-Probability Convergence

Problem and Assumptions

$$\min_{x \in \mathbb{R}^n} \{f(x) = \mathbb{E}_\xi [f(x, \xi)]\} \tag{6}$$

- $f : \mathbb{R}^n \to \mathbb{R}^n$ is convex and L-smooth, i.e., $\forall x, y \in \mathbb{R}^n$

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle, \tag{7}$$

$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|. \tag{8}$$

Problem and Assumptions

$$\min_{x\in\mathbb{R}^n} \{f(x) = \mathbb{E}_\xi\left[f(x,\xi)\right]\} \quad (6)$$

- $f : \mathbb{R}^n \to \mathbb{R}^n$ is convex and L-smooth, i.e., $\forall x, y \in \mathbb{R}^n$

$$f(x) \geq f(y) + \langle \nabla f(y), x - y\rangle, \quad (7)$$

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|. \quad (8)$$

- Stochastic gradient $\nabla f(x,\xi)$ with bounded central α-th moment ($\alpha \in (1,2]$) is available, i.e., $\forall x \in \mathbb{R}^n$

$$\mathbb{E}_\xi\left[\nabla f(x,\xi)\right] = \nabla f(x), \quad \mathbb{E}_\xi\left[\|\nabla f(x,\xi) - \nabla f(x)\|^\alpha\right] \leq \sigma^\alpha. \quad (9)$$

SGD Does Not Converge When $\alpha < 2$

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

SGD Does Not Converge When $\alpha < 2$

- In-expectation guarantees: $\mathbb{E}[\|x - x^*\|^2] \le \varepsilon$, $\mathbb{E}[f(x) - f(x^*)] \le \varepsilon$, $\mathbb{E}[\|\nabla f(x)\|^2] \le \varepsilon$
- Consider the example from (Zhang et al., 2020): $f(x) = \frac{1}{2}\|x\|^2$ and $\nabla f(x, \xi) = x + \xi$, where $\mathbb{E}[\xi] = 0$ and $\mathbb{E}\|\xi\|^\alpha \le \sigma^\alpha$ but $\mathbb{E}\|\xi\|^2 = \infty$ (e.g., ξ can Levý α-stable distribution)

SGD Does Not Converge When $\alpha < 2$

- In-expectation guarantees: $\mathbb{E}[\|x - x^*\|^2] \leq \varepsilon$, $\mathbb{E}[f(x) - f(x^*)] \leq \varepsilon$, $\mathbb{E}[\|\nabla f(x)\|^2] \leq \varepsilon$
- Consider the example from (Zhang et al., 2020): $f(x) = \frac{1}{2}\|x\|^2$ and $\nabla f(x,\xi) = x + \xi$, where $\mathbb{E}[\xi] = 0$ and $\mathbb{E}\|\xi\|^\alpha \leq \sigma^\alpha$ but $\mathbb{E}\|\xi\|^2 = \infty$ (e.g., ξ can Levý α-stable distribution)
- Then, after one step of ***SGD*** we have

$$
\begin{aligned}
\mathbb{E}\|x^1 - x^*\|^2 &= \mathbb{E}\|x^0 - x^* - \gamma\nabla f(x^0,\xi^0)\|^2 \\
&= \underbrace{\|x^0 - x^*\|^2 - 2\gamma\mathbb{E}\left[x^0 - x^*, \nabla f(x^0,\xi^0)\right]}_{\text{infinite}} \\
&\quad + \gamma^2 \underbrace{\mathbb{E}\|\nabla f(x^0,\xi^0)\|^2}_{=\infty} \\
&= \infty
\end{aligned}
$$

The method does not converge in expectation (in L_2) when $\alpha < 2$!
What about the case when $\alpha = 2$ (bounded variance)?

In-Expectation Guarantees and Trajectories of the Method

Consider *SGD* with constant stepsize

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

applied to a toy stochastic quadratic problem:

$$\min_{x \in \mathbb{R}^n} \left\{ f(x) = \mathbb{E}_\xi[f(x, \xi)] \right\}, \quad f(x, \xi) = \frac{1}{2}\|x\|^2 + \langle \xi, x \rangle,$$

where $\mathbb{E}[\xi] = 0$ and $\mathbb{E}[\|\xi\|^2] = \sigma^2$.

In-Expectation Guarantees and Trajectories of the Method

Consider *SGD* with constant stepsize

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

applied to a toy stochastic quadratic problem:

$$\min_{x \in \mathbb{R}^n} \left\{ f(x) = \mathbb{E}_\xi[f(x,\xi)] \right\}, \quad f(x,\xi) = \frac{1}{2}\|x\|^2 + \langle \xi, x \rangle,$$

where $\mathbb{E}[\xi] = 0$ and $\mathbb{E}[\|\xi\|^2] = \sigma^2$. We consider three scenarios:

- ξ has Gaussian distribution
- ξ has Weibull distribution (non-sub-Gaussian)
- ξ has Burr Type XII distribution (non-sub-Gaussian)

In-Expectation Guarantees and Trajectories of the Method

For all of three cases, state-of-the-art theory on ***SGD*** (Ghadimi and Lan, 2013) says

$$\mathbb{E}\left[f(x^k) - f(x^*)\right] \leq (1-\gamma)^k \left(f(x^0) - f(x^*)\right) + \frac{\gamma\sigma^2}{2}. \tag{10}$$

In-Expectation Guarantees and Trajectories of the Method

For all of three cases, state-of-the-art theory on ***SGD*** (Ghadimi and Lan, 2013) says

$$\mathbb{E}\left[f(x^k) - f(x^*)\right] \leq (1-\gamma)^k \left(f(x^0) - f(x^*)\right) + \frac{\gamma\sigma^2}{2}. \qquad (10)$$

However, the behavior in practice does depend on the distribution:

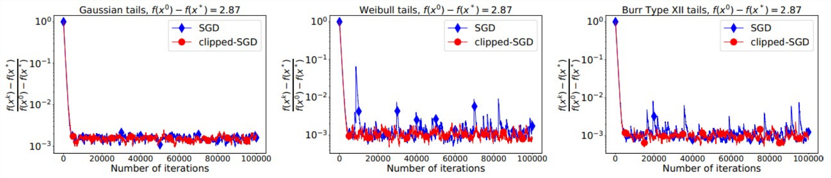

Figure 3: from (Gorbunov et al., 2020)

In-Expectation Guarantees vs High-Probability Guarantees

- In-expectation guarantees: $\mathbb{E}[\|x - x^*\|^2] \leq \varepsilon$, $\mathbb{E}[f(x) - f(x^*)] \leq \varepsilon$, $\mathbb{E}[\|\nabla f(x)\|^2] \leq \varepsilon$
    - Typically, depend only on some moments of stochastic gradient, e.g., variance

In-Expectation Guarantees vs High-Probability Guarantees

- In-expectation guarantees: $\mathbb{E}[\|x - x^*\|^2] \leq \varepsilon$, $\mathbb{E}[f(x) - f(x^*)] \leq \varepsilon$, $\mathbb{E}[\|\nabla f(x)\|^2] \leq \varepsilon$
    - Typically, depend only on some moments of stochastic gradient, e.g., variance
- High-probability guarantees: $\mathbb{P}\{\|x - x^*\|^2 \leq \varepsilon\} \geq 1 - \beta$, $\mathbb{P}\{f(x) - f(x^*) \leq \varepsilon\} \geq 1 - \beta$, $\mathbb{P}\{\|\nabla f(x)\|^2 \leq \varepsilon\} \geq 1 - \beta$
    - Sensitive to the distribution of the stochastic gradient noise

High-Probability Convergence of *SGD* under Bounded Variance Assumption

Natural idea: apply Markov's inequality:

$$\mathbb{P}\left\{f(\hat{x}) - f(x^*) > \varepsilon\right\} < \frac{\mathbb{E}\left[f(\hat{x}) - f(x^*)\right]}{\varepsilon}.$$

High-Probability Convergence of *SGD* under Bounded Variance Assumption

Natural idea: apply Markov's inequality:

$$\mathbb{P}\left\{f(\hat{x}) - f(x^*) > \varepsilon\right\} < \frac{\mathbb{E}\left[f(\hat{x}) - f(x^*)\right]}{\varepsilon}.$$

Taking *enough steps* of ***SGD***, we can guarantee $\mathbb{E}\left[f(\hat{x}) - f(x^*)\right] \leq \varepsilon\beta$ that implies $\mathbb{P}\left\{f(\hat{x}) - f(x^*) > \varepsilon\right\} \leq \beta$ or, equivalently, $\mathbb{P}\left\{f(\hat{x}) - f(x^*) \leq \varepsilon\right\} \geq 1 - \beta$.

High-Probability Convergence of *SGD* under Bounded Variance Assumption

Natural idea: apply Markov's inequality:

$$\mathbb{P}\left\{f(\hat{x}) - f(x^*) > \varepsilon\right\} < \frac{\mathbb{E}\left[f(\hat{x}) - f(x^*)\right]}{\varepsilon}.$$

Taking *enough steps* of ***SGD***, we can guarantee $\mathbb{E}\left[f(\hat{x}) - f(x^*)\right] \leq \varepsilon\beta$ that implies $\mathbb{P}\left\{f(\hat{x}) - f(x^*) > \varepsilon\right\} \leq \beta$ or, equivalently, $\mathbb{P}\left\{f(\hat{x}) - f(x^*) \leq \varepsilon\right\} \geq 1 - \beta$.

Bad news: to ensure $\mathbb{E}\left[f(\hat{x}) - f(x^*)\right] \leq \varepsilon\beta$ ***SGD*** needs

$$\mathcal{O}\left(\max\left\{\frac{LR_0^2}{\varepsilon\beta}, \frac{\sigma^2 R_0^2}{\varepsilon^2\beta^2}\right\}\right) \quad \text{oracle calls}$$

Negative-power dependence on β :(

High-Probability Convergence of *SGD* under Bounded Variance Assumption

Natural idea: apply Markov's inequality:

$$\mathbb{P}\left\{f(\hat{x}) - f(x^*) > \varepsilon\right\} < \frac{\mathbb{E}\left[f(\hat{x}) - f(x^*)\right]}{\varepsilon}.$$

Taking *enough steps* of ***SGD***, we can guarantee $\mathbb{E}\left[f(\hat{x}) - f(x^*)\right] \leq \varepsilon\beta$ that implies $\mathbb{P}\left\{f(\hat{x}) - f(x^*) > \varepsilon\right\} \leq \beta$ or, equivalently, $\mathbb{P}\left\{f(\hat{x}) - f(x^*) \leq \varepsilon\right\} \geq 1 - \beta$.

Bad news: to ensure $\mathbb{E}\left[f(\hat{x}) - f(x^*)\right] \leq \varepsilon\beta$ ***SGD*** needs

$$\mathcal{O}\left(\max\left\{\frac{LR_0^2}{\varepsilon\beta}, \frac{\sigma^2 R_0^2}{\varepsilon^2\beta^2}\right\}\right) \quad \text{oracle calls}$$

Negative-power dependence on β :(

Natural question: can we analyze high-probability convergence of ***SGD*** better?

High-Probability Convergence of *SGD* under Bounded Variance Assumption

Failure of *SGD*

For any $\varepsilon > 0$, $\beta \in (0,1)$, and ***SGD*** parameterized by the number of steps K and stepsize γ, there exists μ-strongly convex L-smooth problem and stochastic oracle with noise having bounded α-th moment with $\alpha = 2$, $0 < \mu \leq L$ such that for the iterates produced by ***SGD*** with any stepsize $0 < \gamma \leq 1/\mu$

$$\mathbb{P}\left\{\|x^K - x^*\|^2 \geq \varepsilon\right\} \leq \beta \implies K = \Omega\left(\frac{\sigma}{\mu\sqrt{\beta\varepsilon}}\right). \tag{11}$$

This illustrates the necessity of modifying the method, e.g., one can use gradient clipping

Main Results

Key Challenge in the Analysis of `clipped-SGD`

$$x^{k+1} = x^k - \gamma \cdot \underbrace{\texttt{clip}\left(\nabla f(x^k, \boldsymbol{\xi}^k), \lambda\right)}_{\widetilde{\nabla} f(x^k, \boldsymbol{\xi}^k)}$$

- Key challenge: $\mathbb{E}\left[\widetilde{\nabla} f(x^k, \boldsymbol{\xi}^k) \mid x^k\right] \neq \nabla f(x^k)$

Analysis of *clipped-SGD*: Key Idea

- We start the proof classically:

$$\begin{aligned}\|x^{k+1} - x^*\|^2 &= \|x^k - x^*\|^2 - 2\gamma\langle x^k - x^*, \widetilde{\nabla} f(x^k, \boldsymbol{\xi}^k)\rangle \\ &\quad + \gamma^2\|\widetilde{\nabla} f(x^k, \boldsymbol{\xi}^k)\|^2 \\ &\leq \ldots\end{aligned}$$

Analysis of `clipped-SGD`: Key Idea

- We start the proof classically:

$$
\begin{aligned}
\|x^{k+1} - x^*\|^2 &= \|x^k - x^*\|^2 - 2\gamma\langle x^k - x^*, \widetilde{\nabla} f(x^k, \boldsymbol{\xi}^k)\rangle \\
&\quad + \gamma^2 \|\widetilde{\nabla} f(x^k, \boldsymbol{\xi}^k)\|^2 \\
&\leq \ldots
\end{aligned}
$$

- Using convexity and smoothness of f and simple rearrangements, we eventually get for $\Delta_k = f(x^k) - f(x^*)$, $R_k = \|x^k - x^*\|$, $\theta_k = \widetilde{\nabla} f(x^k, \boldsymbol{\xi}^k) - \nabla f(x^k)$

$$
\begin{aligned}
\frac{2\gamma(1 - 2\gamma L)}{N}\sum_{k=0}^{N-1}\Delta_k &\leq \frac{1}{N}\left(R_0^2 - R_N^2\right) \\
&\quad + \frac{2\gamma}{N}\sum_{k=0}^{N-1}\langle x^* - x^k, \theta_k\rangle + \frac{2\gamma^2}{N}\sum_{k=0}^{N-1}\|\theta_k\|^2
\end{aligned}
$$

How to upper bound the sums in red?

Bernstein Inequality for Martingale Differences

Lemma 1 (Bennett, 1962; Dzhaparidze and Van Zanten, 2001; Freedman et al., 1975)

Let the sequence of random variables $\{X_i\}_{i \geq 1}$ form a martingale difference sequence, i.e. $\mathbb{E}[X_i \mid X_{i-1}, \ldots, X_1] = 0$ for all $i \geq 1$. Assume that conditional variances $\sigma_i^2 \stackrel{\text{def}}{=} \mathbb{E}\left[X_i^2 \mid X_{i-1}, \ldots, X_1\right]$ exist and are bounded and assume also that there exists deterministic constant $c > 0$ such that $|X_i| \leq c$ almost surely for all $i \geq 1$.

Bernstein Inequality for Martingale Differences

Lemma 1 (Bennett, 1962; Dzhaparidze and Van Zanten, 2001; Freedman et al., 1975)

Let the sequence of random variables $\{X_i\}_{i\geq 1}$ form a martingale difference sequence, i.e. $\mathbb{E}[X_i \mid X_{i-1}, \ldots, X_1] = 0$ for all $i \geq 1$. Assume that conditional variances $\sigma_i^2 \overset{\text{def}}{=} \mathbb{E}\left[X_i^2 \mid X_{i-1}, \ldots, X_1\right]$ exist and are bounded and assume also that there exists deterministic constant $c > 0$ such that $|X_i| \leq c$ almost surely for all $i \geq 1$. Then for all $b > 0$, $G > 0$ and $N \geq 1$

$$\mathbb{P}\left\{\left|\sum_{i=1}^{N} X_i\right| > b \text{ and } \sum_{i=1}^{N} \sigma_i^2 \leq G\right\} \leq 2\exp\left(-\frac{b^2}{2G + {}^{2cb}/_{3}}\right).$$

Bernstein Inequality for Martingale Differences

Lemma 1 (Bennett, 1962; Dzhaparidze and Van Zanten, 2001; Freedman et al., 1975)

Let the sequence of random variables $\{X_i\}_{i\geq 1}$ form a martingale difference sequence, i.e. $\mathbb{E}\left[X_i \mid X_{i-1},\ldots,X_1\right] = 0$ for all $i \geq 1$. Assume that conditional variances $\sigma_i^2 \overset{\text{def}}{=} \mathbb{E}\left[X_i^2 \mid X_{i-1},\ldots,X_1\right]$ exist and are bounded and assume also that there exists deterministic constant $c > 0$ such that $|X_i| \leq c$ almost surely for all $i \geq 1$. Then for all $b > 0$, $G > 0$ and $N \geq 1$

$$\mathbb{P}\left\{\left|\sum_{i=1}^{N} X_i\right| > b \text{ and } \sum_{i=1}^{N}\sigma_i^2 \leq G\right\} \leq 2\exp\left(-\frac{b^2}{2G + {}^{2cb}/_{3}}\right).$$

To bound $\frac{2\gamma}{N}\sum_{k=0}^{N-1}\langle x^* - x^k, \theta_k\rangle + \frac{2\gamma^2}{N}\sum_{k=0}^{N-1}\|\theta_k\|^2$ we need to

- upper bound bias, variance, and distortion of θ_k
- have high-prob. upper bounds for $\|x^k - x^*\|$ and $\|\theta_k\|$

Lemma 2

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

$$\left\|\mathbb{E}[\widetilde{X}] - x\right\| \leq \frac{2^\alpha \sigma^\alpha}{\lambda^{\alpha-1}}, \tag{12}$$

$$\mathbb{E}\left[\left\|\widetilde{X} - x\right\|^2\right] \leq 18\lambda^{2-\alpha}\sigma^\alpha, \tag{13}$$

$$\mathbb{E}\left[\left\|\widetilde{X} - \mathbb{E}[\widetilde{X}]\right\|^2\right] \leq 18\lambda^{2-\alpha}\sigma^\alpha. \tag{14}$$

Bound on the Distance to the Solution

Inequality

$$\frac{2\gamma(1-2\gamma L)}{N}\sum_{k=0}^{N-1}\Delta_k \leq \frac{1}{N}\left(R_0^2 - R_N^2\right) + \frac{2\gamma}{N}\sum_{k=0}^{N-1}\langle x^* - x^k, \theta_k\rangle + \frac{2\gamma^2}{N}\sum_{k=0}^{N-1}\|\theta_k\|^2$$

implies

$$R_N^2 \leq R_0^2 + 2\gamma\sum_{k=0}^{N-1}\langle x^* - x^k, \theta_k\rangle + 2\gamma^2\sum_{k=0}^{N-1}\|\theta_k\|^2.$$

Bound on the Distance to the Solution

Inequality

$$\frac{2\gamma(1-2\gamma L)}{N}\sum_{k=0}^{N-1}\Delta_k \quad\leq\quad \frac{1}{N}\left(R_0^2 - R_N^2\right)$$
$$+\frac{2\gamma}{N}\sum_{k=0}^{N-1}\langle x^* - x^k, \theta_k\rangle + \frac{2\gamma^2}{N}\sum_{k=0}^{N-1}\|\theta_k\|^2$$

implies

$$R_N^2 \quad\leq\quad R_0^2 + 2\gamma\sum_{k=0}^{N-1}\langle x^* - x^k, \theta_k\rangle + 2\gamma^2\sum_{k=0}^{N-1}\|\theta_k\|^2.$$

Key idea: prove $R_N \leq CR_0$ with high probability for some numerical constant C using the induction!

Theorem 1

Let f be convex and L-smooth on $B_{7R_0}(x^*) = \{x \in \mathbb{R}^n \mid \|x - x^*\| \leq 7R_0\}$ and (9) holds on $B_{7R_0}(x^*)$.

High-Probability Convergence of *clipped-SGD*

Theorem 1

Let f be convex and L-smooth on $B_{7R_0}(x^*) = \{x \in \mathbb{R}^n \mid \|x - x^*\| \leq 7R_0\}$ and (9) holds on $B_{7R_0}(x^*)$. Then, for all $\beta \in (0, 1)$, $\varepsilon \geq 0$ such that $\ln(LR_0^2/\varepsilon\beta) \geq 2$ there exists a choice of γ such that ***clipped-SGD*** with clipping level $\lambda \sim 1/\gamma$ and batchsize $m_k = 1$ finds $\bar{x}^N$ satisfying $f(\bar{x}^N) - f(x^*) \leq \varepsilon$ with probability at least $1 - \beta$ using

High-Probability Convergence of `clipped-SGD`

Theorem 1

Let f be convex and L-smooth on $B_{7R_0}(x^*) = \{x \in \mathbb{R}^n \mid \|x - x^*\| \leq 7R_0\}$ and (9) holds on $B_{7R_0}(x^*)$. Then, for all $\beta \in (0,1)$, $\varepsilon \geq 0$ such that $\ln(\nicefrac{LR_0^2}{\varepsilon\beta}) \geq 2$ there exists a choice of γ such that **`clipped-SGD`** with clipping level $\lambda \sim \nicefrac{1}{\gamma}$ and batchsize $m_k = 1$ finds $\bar{x}^N$ satisfying $f(\bar{x}^N) - f(x^*) \leq \varepsilon$ with probability at least $1 - \beta$ using

$$\mathcal{O}\left(\max\left\{\frac{LR^2}{\varepsilon}, \left(\frac{\sigma R}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}} \ln\left(\frac{1}{\beta}\left(\frac{\sigma R}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right)\right\}\right)$$

iterations/oracle calls.

Theoretical Extensions

In (Gorbunov et al., 2020, 2021, 2022; Sadiev et al., 2023) we also have

- Accelerated method (Clipped Stochastic Similar Triangles Method)
- Results for the non-convex objectives
- Results for the strongly convex objectives
- Results for the functions with Hölder continuous gradient
- Results for the variational inequalities

Numerical Experiments: Setup

We tested the performance of the methods on the following problems[1]:

- *BERT* ($\approx 0.6M$ parameters) fine-tuning on `CoLA` dataset. We use pretrained *BERT* and freeze all layers except the last two linear ones. This dataset contains 8551 sentences, and the task is binary classification – to determine if sentence is grammatically correct.
- *ResNet-18* ($\approx 11.7M$ parameters) training on `ImageNet-100` (first 100 classes of `ImageNet`). It has 134395 images.

[1] The code is available at https://github.com/ClippedStochasticMethods/clipped-SSTM

Numerical Experiments: Noise Distribution

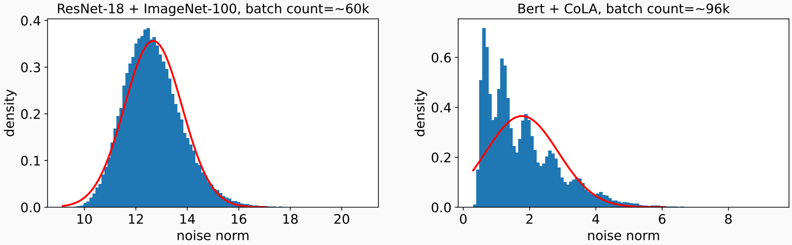

Figure 4: Noise distribution of the stochastic gradients for *ResNet-18* on *ImageNet-100* and *BERT* fine-tuning on the *CoLA* dataset before the training. Red lines: probability density functions of normal distributions with means and variances empirically estimated by the samples. Batch count is the total number of samples used to build a histogram.

Numerical Results, Image Classification

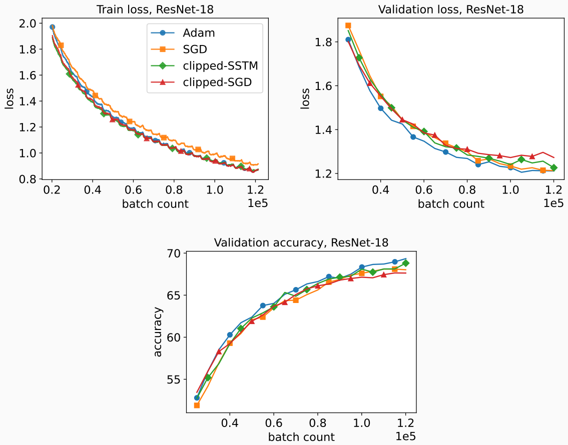

Figure 5: Train and validation loss + accuracy for different optimizers on *ResNet-18* + *ImageNet-100* problem. Here, “batch count” denotes the total number of used stochastic gradients. The noise distribution is almost Gaussian, even vanilla *SGD* performs well.

Numerical Results, Text Classification

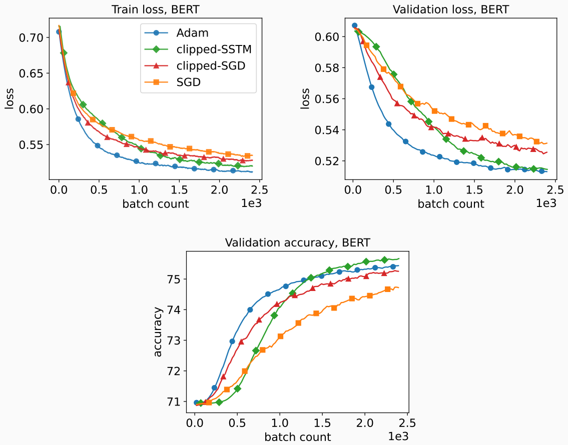

Figure 6: Train and validation loss + accuracy for different optimizers on *BERT* + *CoLA* problem. The noise distribution is heavy-tailed, the methods with clipping outperform *SGD* by a large margin.

Conclusion

- Some popular problems have heavy-tailed noise: in NLP it was observed before, for GANs we demonstrated empirically
- Clipping is a simple way to deal with heavy-tailed noise
- High-probability convergence results for methods with clipping are better than known high-probability convergence results for methods without it
- Partial explanation of the success of adaptive methods like *Adam* on GANs and NLP tasks

About MBZUAI

- Established in 2019, located in Masdar City (Abu Dhabi, UAE)
- First classes started in January 2021 (because of COVID-19)
- Three departments: NLP, CV, and ML
- Some numbers: ≈ 300 students, ≈ 50 faculties, 20th in CSRankings (AI, CV, ML, and NLP)

Figure 7: *https://www.arabnews.com/node/1724111/amp*

References

Bennett, G. (1962). Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45.

Dzhaparidze, K. and Van Zanten, J. (2001). On bernstein-type inequalities for martingales. *Stochastic processes and their applications*, 93(1):109–117.

Freedman, D. A. et al. (1975). On tail probabilities for martingales. *the Annals of Probability*, 3(1):100–118.

Ghadimi, S. and Lan, G. (2013). Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368.

References ii

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Gorbunov, E., Danilova, M., Dobre, D., Dvurechensky, P., Gasnikov, A., and Gidel, G. (2022). Clipped stochastic methods for variational inequalities with heavy-tailed noise. *arXiv preprint arXiv:2206.01095*.

Gorbunov, E., Danilova, M., and Gasnikov, A. (2020). Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. *Advances in Neural Information Processing Systems*, 33:15042–15053.

Gorbunov, E., Danilova, M., Shibaev, I., Dvurechensky, P., and Gasnikov, A. (2021). Near-optimal high probability complexity bounds for non-smooth stochastic optimization with heavy-tailed noise. *arXiv preprint arXiv:2106.05958*.

Merity, S., Keskar, N. S., and Socher, R. (2017). Regularizing and optimizing lstm language models. *arXiv preprint arXiv:1708.02182*.

Mosbach, M., Andriushchenko, M., and Klakow, D. (2020). On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. *arXiv preprint arXiv:2006.04884*.

Pascanu, R., Mikolov, T., and Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pages 1310–1318.

Peters, M. E., Ammar, W., Bhagavatula, C., and Power, R. (2017). Semi-supervised sequence tagging with bidirectional language models. *arXiv preprint arXiv:1705.00108*.

Sadiev, A., Danilova, M., Gorbunov, E., Horváth, S., Gidel, G., Dvurechensky, P., Gasnikov, A., and Richtárik, P. (2023). High-probability bounds for stochastic optimization and variational inequalities: the case of unbounded variance. *arXiv preprint arXiv:2302.00999*.

Zhang, J., Karimireddy, S. P., Veit, A., Kim, S., Reddi, S. J., Kumar, S., and Sra, S. (2020). Why are adaptive methods good for attention models? *Advances in Neural Information Processing Systems*, 33.