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). *[illegible]*. NeurIPS 2020
- Gorbunov, E., Danilova, M., Shibaev, I., Dvurechensky, P., & Gasnikov, A. (2021). *[illegible]*. arXiv:2106.05958
- Gorbunov, E., Danilova, M., Dobre, D., Dvurechenskii, P., Gasnikov, A., & Gidel, G. (2022). *[illegible]*. NeurIPS 2022.
- Sadiev, A., Danilova, M., Gorbunov, E., Horváth, S., Gidel, G., Dvurechensky, P., Gasnikov, A., & Richtárik, P. (2023). *[illegible]*. ICML 2023.

Clipping and Heavy-Tailed Noise

Stochastic Gradient Descent (*RFC*)

$$\theta_{j+1} = \theta_j \qquad r\tilde{\imath}(\theta_j;\ \ _j) \tag{1}$$

- $\tilde{\imath}$ – the function to be minimized
- $r\tilde{\imath}(\theta_j;\ \ _j)$ – stochastic gradient, i.e., $\cdot\hbar\mathcal{E}\hat{\mathbb{0}}hu'\bar{a}\check{a}$ estimate of $r\tilde{\imath}(\theta_j)$: $\mathrm{E}_{\ j}[\,r\tilde{\imath}(\theta_j;\ \ _j)] = r\tilde{\imath}(\theta_j)$

Clipped Stochastic Gradient Descent (*bkhoodc*, *RFC*)

$$’^{j+1} = ’^{j} \quad \textit{bkho} \quad r\,ì(’^{j};\ ^{j}); \qquad (2)$$

- $\textit{bkho}(’;\ ) = \min f1;\ =k’k g’$
- $\textit{bkho}(r\,ì(’^{j};\ ^{j});\ )$ – ɛ̃ħ’āǎ estimate of $r\,ì(’^{j})$:
  $E_{j}[\textit{bkho}(r\,ì(’^{j};\ ^{j});\ )] \neq r\,ì(’^{j})$

Origin of Clipping

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

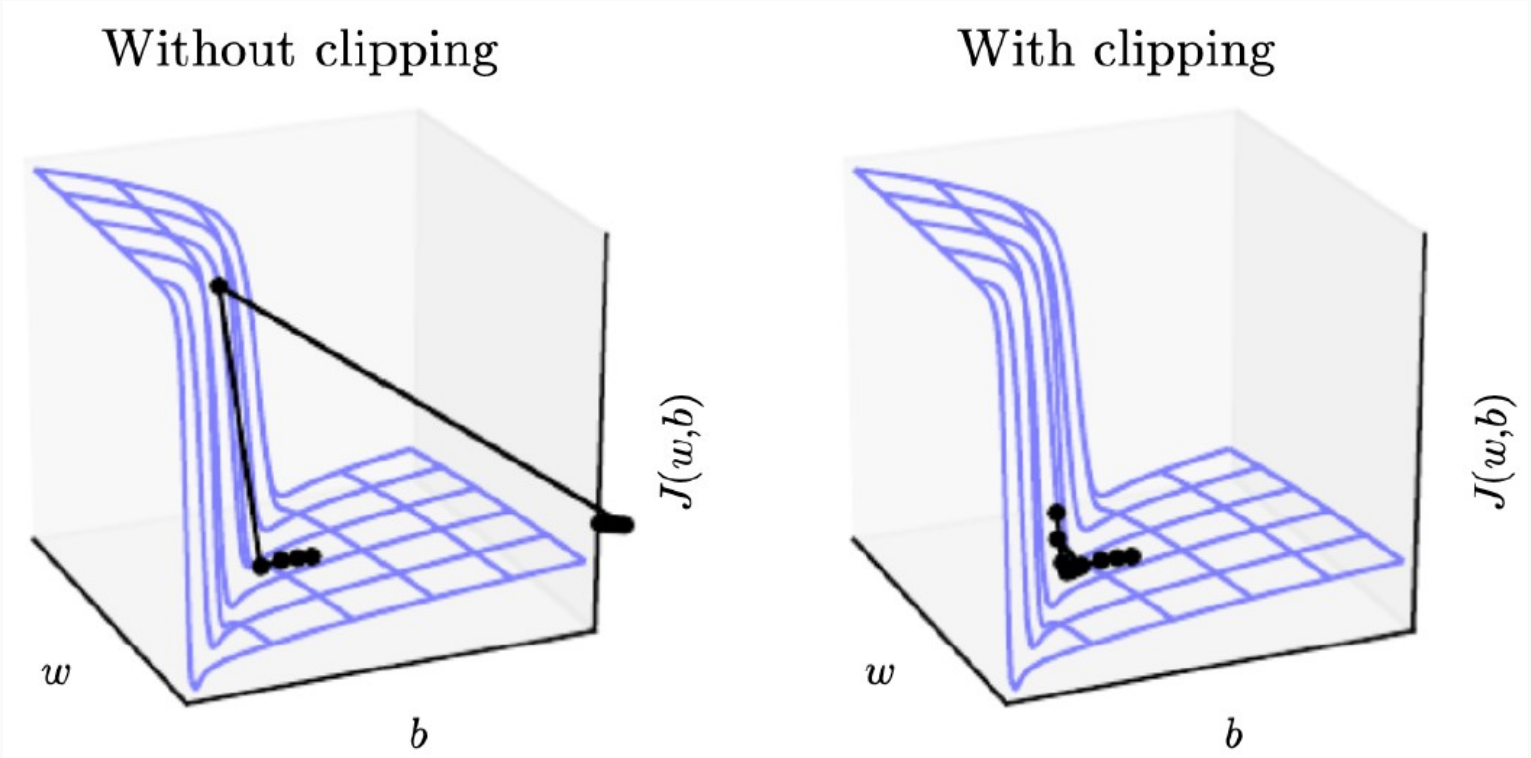

Figure 1: from (Goodfellow et al., 2016)

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 *@c `l* + clipping
- Mosbach et al. (2020) fine-tune BERT using *@c `lV* + 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 *@c `l* + clipping
- Mosbach et al. (2020) fine-tune BERT using *@c `lV* + 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 $k\,r\,\grave{\imath}('\,;\ \ )\qquad r\,\grave{\imath}('\,)k$ 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

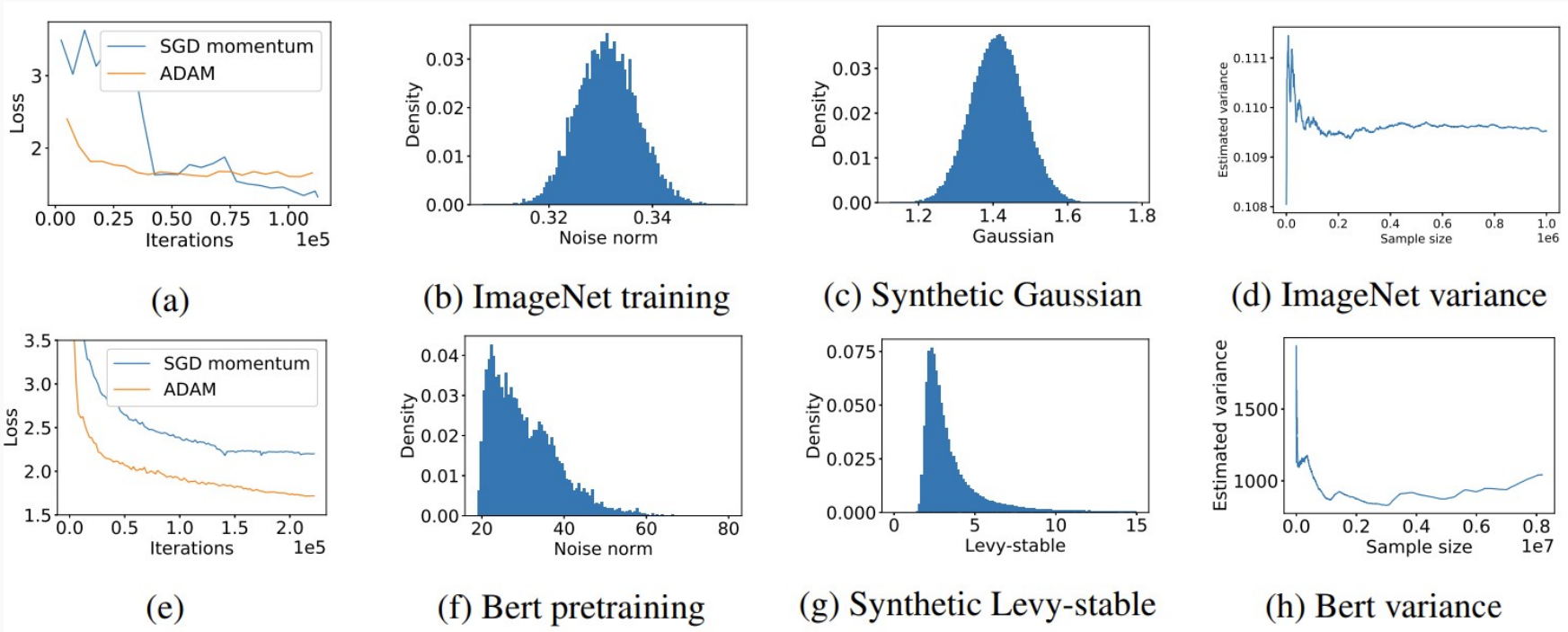

Figure 2: from (Zhang et al., 2020)

We see that ***@C@L*** is much better than ***RFC*** when the noise in the stochastic gradient is heavy-tailed

@c`l and bkhoodc, RFC

- bkhoodc, RFC:

$$, _{j+1} = , _{j} \qquad bkho \quad r\ddot{i}(' _{j}; \ _{j}); \ _{j}$$

- @c`l:

$$g_{j} = \ _{1} g_{j \ 1} + (1 \qquad _{1}) \, r\ddot{i}(' _{j}; \ _{j});$$

$$\dot{}_{j} = \ _{2} \dot{}_{j \ 1} + (1 \qquad _{2})(r\ddot{i}(' _{j}; \ _{j}))^{2};$$

$$, _{j+1} = , _{j} \qquad \frac{p}{\dot{}_{j} + } g^{j}$$

  - When $\beta_1 = 0$ **@c`l** (**QLRoqno**) can be seen as **bkhoodc, RFC** with "adaptive" λ_j

Definition of Heavy-Tailed Noise in Stochastic Gradients

- Random vector t has light tails if

$$\mathrm{P}\,fkt \quad \mathrm{E}[t]k \quad \varepsilon g \quad 2\exp \quad \frac{\varepsilon^2}{2^{\ 2}} \qquad \forall \varepsilon > 0: \tag{3}$$

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

$$\mathrm{E} \quad \exp \quad \frac{kt \quad \mathrm{E}[t]k^2}{2} \qquad \exp(1): \tag{4}$$

Definition of Heavy-Tailed Noise in Stochastic Gradients

- Random vector t has light tails if

$$\mathrm{P}fkt \quad \mathrm{E}[t]k \quad \varepsilon g \quad 2\exp \quad \frac{\varepsilon^2}{2^{\ 2}} \qquad \forall \varepsilon > 0: \tag{3}$$

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

$$\mathrm{E} \quad \exp \quad \frac{kt \quad \mathrm{E}[t]k^2}{2} \qquad \exp(1): \tag{4}$$

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

$$\mathrm{E}\,[kt \quad \mathrm{E}[t]k\ ] \tag{5}$$

In-Expectation Guarantees vs High-Probability Convergence

Problem and Assumptions

$$\min_{\,2\,\mathrm{R}^{\eta}} f\ddot{\imath}(\,) = \mathrm{E}\ \ [\ddot{\imath}(\,;\ \ )]g \tag{6}$$

- $\ddot{\imath} : \mathrm{R}^{\eta}\ \ !\ \ \mathrm{R}^{\eta}$ is convex and $\acute{I}$-smooth, i.e., $8\,;\ \ 2\ \mathrm{R}^{\eta}$

$$\ddot{\imath}(\,)\quad \ddot{\imath}(\,) + h\,r\,\ddot{\imath}(\,);\,\quad \ i; \tag{7}$$

$$k\,r\,\ddot{\imath}(\,)\quad r\,\ddot{\imath}(\,)k\quad \acute{I}k\,\quad \ k: \tag{8}$$

Problem and Assumptions

$$\min_{\text{'} \; 2 \, \mathrm{R}^{\text{ɧ}}} f\text{ì}(\text{'}) = \mathrm{E} \; [\text{ì}(\text{'}; \;\;)]g \tag{6}$$

- $\text{ì} : \mathrm{R}^{\text{ɧ}} \;\; ! \;\; \mathrm{R}^{\text{ɧ}}$ is convex and $\acute{I}$-smooth, i.e., $8\text{'}; \text{`} \;\; 2 \;\; \mathrm{R}^{\text{ɧ}}$

$$\text{ì}(\text{'}) \quad \text{ì}(\text{`}) + h \, r \, \text{ì}(\text{`}); \text{'} \quad \text{`} \; i; \tag{7}$$

$$k \, r \, \text{ì}(\text{'}) \quad r \, \text{ì}(\text{`})k \quad \acute{I}k\text{'} \quad \text{`} \; k: \tag{8}$$

- Stochastic gradient $r \, \text{ì}(\text{'}; \;\;)$ with bounded central $\;\;$-th moment ($\;\; 2 \; (1; 2]$) is available, i.e., $8\text{'} \;\; 2 \;\; \mathrm{R}^{\text{ɧ}}$

$$\mathrm{E} \;\; [\, r \, \text{ì}(\text{'}; \;\;)] = \;\; r \, \text{ì}(\text{'}); \quad \mathrm{E} \;\; [k \, r \, \text{ì}(\text{'}; \;\;) \quad r \, \text{ì}(\text{'})k \;\;] \quad : \tag{9}$$

RFC Does Not Converge When < 2

- In-expectation guarantees: $\mathrm{E}[k' \quad ' \quad k^2] \quad ''$, $\mathrm{E}[\grave{\imath}(' \,) \quad \grave{\imath}(' \quad)] \quad ''$, $\mathrm{E}[k \, r \, \grave{\imath}(' \,) k^2] \quad ''$

RFC Does Not Converge When < 2

- In-expectation guarantees: $\mathrm{E}[k' \quad ' \quad k^2] \quad '', \mathrm{E}[\grave{\imath}(') \quad \grave{\imath}(' \quad)] \quad '',$ $\mathrm{E}[k\, r\, \grave{\imath}(')k^2] \quad ''$
- Consider the example from (Zhang et al., 2020): $\grave{\imath}(') = \frac{1}{2}k' \, k^2$ and $r\, \grave{\imath}(' ; \ ) = ' + \ $, where $\mathrm{E}[\ ] = 0$ and $\mathrm{E}k \ \ k$ but $\mathrm{E}k \ \ k^2 = 1$ (e.g., can Levý -stable distribution)

RFC Does Not Converge When $\alpha < 2$

- In-expectation guarantees: $\mathbb{E}[k' - ' k^2] \leq \varepsilon$, $\mathbb{E}[\hat{\imath}(') - \hat{\imath}(')] \leq \varepsilon$, $\mathbb{E}[k r\hat{\imath}(')k^2] \leq \varepsilon$
- Consider the example from (Zhang et al., 2020): $\hat{\imath}(') = \frac{1}{2}k' k^2$ and $r\hat{\imath}(';) = ' +$, where $\mathbb{E}[\ ] = 0$ and $\mathbb{E}k \ k$ but $\mathbb{E}k \ k^2 = 1$ (e.g., can Levý -stable distribution)
- Then, after one step of ***RFC*** we have

$$\begin{aligned}
\mathbb{E}k'^{1} - ' k^2 &= \mathbb{E}k'^{0} - ' r\hat{\imath}('^{0}; {}^{0})k^2 \\
&= k'^{0} - ' k^2 \underbrace{2\,\mathbb{E}\,'^{0} - ' ; r\hat{\imath}('^{0}; {}^{0})}_{\text{infinite}} \\
&\quad + {}^{2}\underbrace{\mathbb{E}k r\hat{\imath}('^{0}; {}^{0})k^2}_{=1} \\
&= 1
\end{aligned}$$

The method does not converge in expectation (in $\acute{I}_2$) when $\alpha < 2$!

What about the case when $\alpha = 2$ (bounded variance)?

In-Expectation Guarantees and Trajectories of the Method

Consider ***RFC*** with constant stepsize

$$\theta_{j+1} = \theta_j \qquad r\,\widehat{\nabla} f(\theta_j;\ \xi_j)$$

applied to a toy stochastic quadratic problem:

$$\min_{\theta \in \mathbb{R}^d} f(\theta) = \mathrm{E}\,[f(\theta;\ \xi)]; \quad \widehat{\nabla} f(\theta;\ \xi) = \frac{1}{2} k\,\theta\,\|k\|^2 + h\ ;\ \theta\ i;$$

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

In-Expectation Guarantees and Trajectories of the Method

Consider ***RFC*** with constant stepsize

$$, _{j+1} = , _{j} \qquad r \ddot{\imath}(, _{j}; \ _{j})$$

applied to a toy stochastic quadratic problem:

$$\min_{, \ 2 \ \mathrm{R}^{\eth}} f\ddot{\imath}(,) = \mathrm{E}\ [\ddot{\imath}(, ; \ )]g; \quad \ddot{\imath}(, ; \ ) = \frac{1}{2} k, \ k^2 + h \ ; , \ i;$$

where $\mathrm{E}[\ ] = 0$ and $\mathrm{E}[k \ \ k^2] = \ ^2$. We consider three scenarios:

- has Gaussian distribution
- has Weibull distribution (non-sub-Gaussian)
- 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 ***RFC*** (Ghadimi and Lan, 2013) says

$$\mathrm{E}\left[\ddot{\imath}(\cdot^{j}) \quad \ddot{\imath}(\cdot \quad)\right] \quad (1 \quad)^{j} \quad \ddot{\imath}(\cdot^{0}) \quad \ddot{\imath}(\cdot \quad) \quad + \frac{^{2}}{2}: \tag{10}$$

In-Expectation Guarantees and Trajectories of the Method

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

$$\mathrm{E}\left[\tilde{\imath}(\cdot^{j}) \quad \tilde{\imath}(\cdot \quad)\right] \quad (1 \quad)^{j} \quad \tilde{\imath}(\cdot^{0}) \quad \tilde{\imath}(\cdot \quad) \quad + \frac{^{2}}{2}: \qquad (10)$$

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: $\mathrm{E}[k' \quad ' \quad k^2] \quad ''$, $\mathrm{E}[\grave{\imath}(' \quad) \quad \grave{\imath}(' \quad)] \quad ''$, $\mathrm{E}[k\, r\, \grave{\imath}(' \,)k^2] \quad ''$
  - Typically, depend only on some moments of stochastic gradient, e.g., variance

In-Expectation Guarantees vs High-Probability Guarantees

- In-expectation guarantees: $\mathrm{E}[k' \quad ' \quad k^2] \quad ''$, $\mathrm{E}[\grave{\imath}(') \quad \grave{\imath}(')] \quad ''$, $\mathrm{E}[k\, r\, \grave{\imath}(')k^2] \quad ''$
  - Typically, depend only on some moments of stochastic gradient, e.g., variance
- High-probability guarantees: $\mathrm{P}\{k' \quad ' \quad k^2 \quad ''g \quad 1 \quad ,$ $\mathrm{P}\{\grave{\imath}(') \quad \grave{\imath}(') \quad ''g \quad 1 \quad , \mathrm{P}\{k\, r\, \grave{\imath}(')k^2 \quad ''g \quad 1$
  - Sensitive to the distribution of the stochastic gradient noise

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

Natural idea: apply Markov's inequality:

$$\mathrm{P}\left\{\tilde{\imath}(\hat{\imath}) \quad \tilde{\imath}(\quad) > \varepsilon\right\} < \frac{\mathrm{E}\left[\tilde{\imath}(\hat{\imath}) \quad \tilde{\imath}(\quad)\right]}{\varepsilon}:$$

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

Natural idea: apply Markov's inequality:

$$\mathrm{P}\,\{\tilde{\imath}(\nwarrow) \quad \tilde{\imath}(\,') > \varepsilon\} < \frac{\mathrm{E}\,[\tilde{\imath}(\nwarrow) \quad \tilde{\imath}(\,')]}{\varepsilon}.$$

Taking $\bar{a}\hbar\omega\cdot\hat{O}\not{q}\ '^{-}\bar{a}\pounds'$ of ***RFC***, we can guarantee $\mathrm{E}\,[\tilde{\imath}(\nwarrow) \quad \tilde{\imath}(\,')] \quad \varepsilon$ that implies $\mathrm{P}\,\{\tilde{\imath}(\nwarrow) \quad \tilde{\imath}(\,') > \varepsilon\}$ or, equivalently, $\mathrm{P}\,\{\tilde{\imath}(\nwarrow) \quad \tilde{\imath}(\,') \quad \varepsilon\} \quad 1 \quad .$

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

Natural idea: apply Markov's inequality:

$$\mathrm{P}\,\{\tilde{\imath}(\hat{\imath}) \quad \tilde{\imath}(\quad) > \varepsilon\} < \frac{\mathrm{E}\,[\tilde{\imath}(\hat{\imath}) \quad \tilde{\imath}(\quad)]}{\varepsilon}:$$

Taking $\bar{a}\hbar\omega\cdot\hat{O}\not{q}$ '$\bar{a}\pounds$' of ***RFC***, we can guarantee $\mathrm{E}\,[\tilde{\imath}(\hat{\imath}) \quad \tilde{\imath}(\quad)] \quad \varepsilon$ that implies $\mathrm{P}\,\{\tilde{\imath}(\hat{\imath}) \quad \tilde{\imath}(\quad) > \varepsilon\}$ or, equivalently, $\mathrm{P}\,\{\tilde{\imath}(\hat{\imath}) \quad \tilde{\imath}(\quad) \quad \varepsilon\} \quad 1 \quad .$

Bad news: to ensure $\mathrm{E}\,[\tilde{\imath}(\hat{\imath}) \quad \tilde{\imath}(\quad)] \quad \varepsilon$ ***RFC*** needs

$$O \quad \max \quad \frac{\acute{I}\hat{G}_0^2}{\varepsilon}; \frac{^2\hat{G}_0^2}{\varepsilon^2 \quad ^2} \qquad \text{oracle calls}$$

Negative-power dependence on :(

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

Natural idea: apply Markov's inequality:

$$\mathrm{P}\,\{\tilde{\imath}(\hat{\imath}) \quad \tilde{\imath}(\quad) > \varepsilon\} < \frac{\mathrm{E}\,[\tilde{\imath}(\hat{\imath}) \quad \tilde{\imath}(\quad)]}{\varepsilon}:$$

Taking $\bar{a}\hbar\omega\cdot\hat{O}\not{q}$ '$\bar{a}\pounds$' of ***RFC***, we can guarantee $\mathrm{E}\,[\tilde{\imath}(\hat{\imath}) \quad \tilde{\imath}(\quad)] \quad \varepsilon$ that implies $\mathrm{P}\,\{\tilde{\imath}(\hat{\imath}) \quad \tilde{\imath}(\quad) > \varepsilon\}$ or, equivalently, $\mathrm{P}\,\{\tilde{\imath}(\hat{\imath}) \quad \tilde{\imath}(\quad) \quad \varepsilon\} \quad 1 \quad .$

Bad news: to ensure $\mathrm{E}\,[\tilde{\imath}(\hat{\imath}) \quad \tilde{\imath}(\quad)] \quad \varepsilon$ ***RFC*** needs

$$O \quad \max \quad \frac{\acute{I}\hat{G}_0^2}{\varepsilon}; \frac{{}^2\hat{G}_0^2}{\varepsilon^2 \quad {}^2} \qquad \text{oracle calls}$$

Negative-power dependence on :(

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

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

Failure of *RFC*

For any $\varepsilon > 0$, $\alpha \in (0; 1)$, and ***RFC*** 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 $\alpha = 2$, $0 < \sigma < L$ such that for the iterates produced by ***RFC*** with any stepsize $0 < \gamma \leq \frac{1}{L}$

$$\mathrm{P}\left\{\|x^K - x^*\|^2 \geq \varepsilon\right\} = 1 \quad K = \Omega\left(\frac{\sigma^2}{\mu^2 \varepsilon}\right). \tag{11}$$

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

Main Results

Key Challenge in the Analysis of *bkhoodc, RFC*

$$\theta_{j+1} = \theta_j \qquad \underbrace{bkho \quad r\tilde{\imath}(\theta_j; \quad ^j)}_{\widetilde{r}\,\tilde{\imath}(\theta_j;\xi^j)};$$

- Key challenge: $\mathrm{E}\left[\widetilde{r}\,\tilde{\imath}(\theta_j; \quad ^j)\, j\, \theta_j\right] \neq r\tilde{\imath}(\theta_j)$

Analysis of *bkhoodc, RFC*: Key Idea

- We start the proof classically:

$$\begin{aligned} k' ^{j+1} \quad ' \quad k^2 \quad &= \quad k' ^{j} \quad ' \quad k^2 \quad 2 \; h' ^{j} \quad ' \quad ; \; \text{Re}\, i(' ^{j}; \; ^{j}) i \\ &\quad + \; ^2 k \, \text{Re}\, i(' ^{j}; \; ^{j}) k^2 \\ &\quad \vdots \vdots \vdots \end{aligned}$$

Analysis of *bkhoodc*, *RFC*: Key Idea

- We start the proof classically:

$$\|x^{k+1} - x^*\|^2 = \|x^k - x^*\|^2 - 2\gamma\langle x^k - x^*, \nabla f(x^k; \xi^k)\rangle + \gamma^2\|\nabla f(x^k; \xi^k)\|^2 \ldots$$

- Using convexity and smoothness of f and simple rearrangements, we eventually get for $\delta_k = f(x^k) - f(x^*)$, $\hat{G}_k = \|x^k - x^*\|$, $\omega_k = \nabla f(x^k; \xi^k) - \nabla f(x^k)$

$$\frac{2\gamma(1 - 2L\gamma)}{K}\sum_{k=0}^{K-1}\delta_k \leq \frac{1}{K}\left(\hat{G}_0^2 - \hat{G}_K^2\right) + \frac{2\gamma}{K}\sum_{k=0}^{K-1}\langle x^* - x^k, \omega_k\rangle + \frac{2\gamma^2}{K}\sum_{k=0}^{K-1}\|\omega_k\|^2$$

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 $f t_i g_{i=1}$ form a martingale difference sequence, i.e. $\mathrm{E}[t_i \mid t_{i-1}, \ldots, t_1] = 0$ for all $i \geq 1$. Assume that conditional variances $\sigma_i^2 \stackrel{\text{def}}{=} \mathrm{E}\left[t_i^2 \mid t_{i-1}, \ldots, t_1\right]$ exist and are bounded and assume also that there exists deterministic constant $L > 0$ such that $|t_i| \leq L$ 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 $f t_i g_{i=1}$ form a martingale difference sequence, i.e. $\mathrm{E}[t_i \mid t_{i-1}; \ldots; t_1] = 0$ for all $i \geq 1$. Assume that conditional variances $\sigma_i^2 \stackrel{\text{def}}{=} \mathrm{E}\left[t_i^2 \mid t_{i-1}; \ldots; t_1\right]$ exist and are bounded and assume also that there exists deterministic constant $A > 0$ such that $|t_i| \leq A$ almost surely for all $i \geq 1$. Then for all $\varepsilon > 0$, $v > 0$ and $n \geq 1$

$$\mathrm{P}\left(\sum_{i=1}^{n} t_i > \varepsilon \text{ and } \sum_{i=1}^{n} \sigma_i^2 \leq v\right) \leq \exp\left(\frac{\varepsilon^2}{2v + 2A\varepsilon/3}\right).$$

Bernstein Inequality for Martingale Differences

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

Let the sequence of random variables $f t_i g_{i=1}$ form a martingale difference sequence, i.e. $\mathrm{E}[t_i \mid t_{i-1}; \ldots; t_1] = 0$ for all $i \geq 1$. Assume that conditional variances $\sigma_i^2 \stackrel{\text{def}}{=} \mathrm{E}\left[t_i^2 \mid t_{i-1}; \ldots; t_1\right]$ exist and are bounded and assume also that there exists deterministic constant $A > 0$ such that $|t_i| \leq A$ almost surely for all $i \geq 1$. Then for all $\varepsilon > 0$, $v > 0$ and $m \geq 1$

$$\mathrm{P}\left(\sum_{i=1}^{m} t_i > \varepsilon \text{ and } \sum_{i=1}^{m} \sigma_i^2 \leq v\right) \leq \exp\left(\frac{\varepsilon^2}{2v + 2A\varepsilon/3}\right).$$

To bound $\frac{2}{m}\sum_{j=0}^{m-1} h(X_j, Y_j) - \frac{2}{m(m-1)}\sum_{j=0}^{m-1} k(X_j, X_j)$ we need to

- upper bound bias, variance, and distortion of $\hat{\theta}_j$
- have high-prob. upper bounds for $k(X_j, Y_j)$, $k(X_j, X_j)$ and $k(X_j, X_j)$

Magnitude, Bias, Variance, Distortion

Lemma 2

Let t be a random vector in $\mathrm{R}^{\check{a}}$ and $\mathfrak{e} = \mathit{bkho}(t;\ )$. Then, $k\mathfrak{e}\quad \mathrm{E}[\mathfrak{e}]k\quad 2$. Moreover, if for some $\quad 0$ and $\quad 2\ (1;2]$ we have $\mathrm{E}[t] = '\ 2\ \mathrm{R}^{\check{a}}$, $\mathrm{E}[kt\quad 'k\ ]\quad$, and $k'k\quad =_2$, then

$$\mathrm{E}[\mathfrak{e}]\quad ,\quad \frac{2}{1}; \tag{12}$$

$$\mathrm{E}\quad \mathfrak{e}\quad ,\ ^2\quad 18\ ^2\quad ; \tag{13}$$

$$\mathrm{E}\quad \mathfrak{e}\quad \mathrm{E}[\mathfrak{e}]\ ^2\quad 18\ ^2\quad : \tag{14}$$

Bound on the Distance to the Solution

Inequality

$$\frac{2\ (1\quad 2\ \acute{I})}{\acute{Y}} \sum_{j=0}^{\acute{Y}\mathrm{X}^{1}} \quad j \qquad \frac{1}{\acute{Y}}\ \hat{G}_{0}^{2} \quad \hat{G}_{\acute{Y}}^{2}$$

$$+\frac{2}{\acute{Y}} \sum_{j=0}^{\acute{Y}\mathrm{X}^{1}} h' \qquad ,\ ^{j};\ _{j}i + \frac{2}{\acute{Y}}^{\ 2} \sum_{j=0}^{\acute{Y}\mathrm{X}^{1}} k\ _{j}k^{2}$$

implies

$$\hat{G}_{\acute{Y}}^{2} \qquad \hat{G}_{0}^{2} + 2 \sum_{j=0}^{\acute{Y}\mathrm{X}^{1}} h' \qquad ,\ ^{j};\ _{j}i + 2\ ^{2} \sum_{j=0}^{\acute{Y}\mathrm{X}^{1}} k\ _{j}k^{2}:$$

Bound on the Distance to the Solution

Inequality

$$\frac{2\ (1\quad 2\ \acute{I})}{\acute{Y}} \overset{\acute{Y}}{\underset{j=0}{\mathrm{X}}}{}^{1} \quad {}_{j} \qquad \frac{1}{\acute{Y}}\ \hat{G}_{0}^{2} \quad \hat{G}_{\acute{Y}}^{2}$$

$$+\frac{2}{\acute{Y}} \overset{\acute{Y}}{\underset{j=0}{\mathrm{X}}}{}^{1} h' \qquad ,{}^{j};\ {}_{j}i + \frac{2}{\acute{Y}}{}^{2} \overset{\acute{Y}}{\underset{j=0}{\mathrm{X}}}{}^{1} k\ {}_{j}k^{2}$$

implies

$$\hat{G}_{\acute{Y}}^{2} \qquad \hat{G}_{0}^{2} + 2 \overset{\acute{Y}}{\underset{j=0}{\mathrm{X}}}{}^{1} h' \qquad ,{}^{j};\ {}_{j}i + 2\ {}^{2} \overset{\acute{Y}}{\underset{j=0}{\mathrm{X}}}{}^{1} k\ {}_{j}k^{2}:$$

Key idea: prove $\hat{G}_{\acute{Y}} \quad G\hat{G}_{0}$ with high probability for some numerical constant G using the induction!

High-Probability Convergence of *bkhoodc, RFC*

Theorem 1

Let ì be convex and $\acute{I}$-smooth on
$@_{7\hat{G}_0}(\text{’}\ \ ) = f\text{’}\ \ 2\ \mathrm{R}^{\eta}\ j\ k\text{’}\ \ \ \text{’}\ \ k\ \ \ 7\hat{G}_0 g$ and (9) holds on $@_{7\hat{G}_0}(\text{’}\ \ )$.

High-Probability Convergence of *bkhoodc, RFC*

Theorem 1

Let $\ddot{\imath}$ be convex and $\acute{I}$-smooth on $@_{7\hat{G}_0}(\text{’}\ \ ) = f\text{’}\ \ 2\ \mathrm{R}^{\eta}\ j\ k\text{’}\ \ \text{’}\ \ k\ \ \ 7\hat{G}_0 g$ and (9) holds on $@_{7\hat{G}_0}(\text{’}\ \ )$. Then, for all $\ \ 2\ (0;1)$, $\text{"}\ \ \ 0$ such that $\ln(^{\acute{I}\hat{G}_0^2=\text{"}}\ )\ \ \ 2$ there exists a choice of such that ***bkhoodc, RFC*** with clipping level $\ \ ^{1=}$ and batchsize $g_{\ j}\ = 1$ finds $\text{’}\ ^{\acute{Y}}$ satisfying $\ddot{\imath}(\text{’}\ ^{\acute{Y}})\ \ \ \ddot{\imath}(\text{’}\ \ )\ \ \ \text{"}$ with probability at least 1 using

High-Probability Convergence of *bkhoodc, RFC*

Theorem 1

Let $\ddot{\imath}$ be convex and $\acute{I}$-smooth on $@_{7\hat{G}_0}(\text{’}\ \ ) = f\text{’}\ \ 2\ \mathrm{R}^{\eta}\ j\ k\text{’}\ \ \text{’}\ \ k\ \ \ 7\hat{G}_0 g$ and (9) holds on $@_{7\hat{G}_0}(\text{’}\ \ )$. Then, for all $\ \ 2\ (0;1)$, $"\ \ 0$ such that $\ln(^{\acute{I}\hat{G}_0^2="}\ )\ \ \ 2$ there exists a choice of such that ***bkhoodc, RFC*** with clipping level $\ \ ^{1=}$ and batchsize $g_j = 1$ finds $\text{’}^{\acute{Y}}$ satisfying $\ddot{\imath}(\text{’}^{\acute{Y}})\ \ \ddot{\imath}(\text{’}\ \ )\ \ "$ with probability at least 1 using

$$O\ \ \max\ \ \frac{\acute{I}\hat{G}^2}{"};\ \ \frac{\hat{G}}{"}^{\overline{\ \ 1}}\ \ln\ \ \frac{1}{-}\ \ \frac{\hat{G}}{"}^{\overline{\ \ 1}}\ \ !\ )\ !$$

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]:

- ***ADQS*** (0:6*Ù* parameters) fine-tuning on ***BnK@*** dataset. We use pretrained ***ADQS*** 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.
- ***QdrMds,07*** (11:7*Ù* parameters) training on ***Hl`fdMds,0//*** (first 100 classes of ***Hl`fdMds***). It has 134395 images.

[1]The code is available at ***gssor9..fhsgta-bnl.BkhoodcRsnbg`rshbLdsgncr.bkhoodc,RRSL***

Numerical Experiments: Noise Distribution

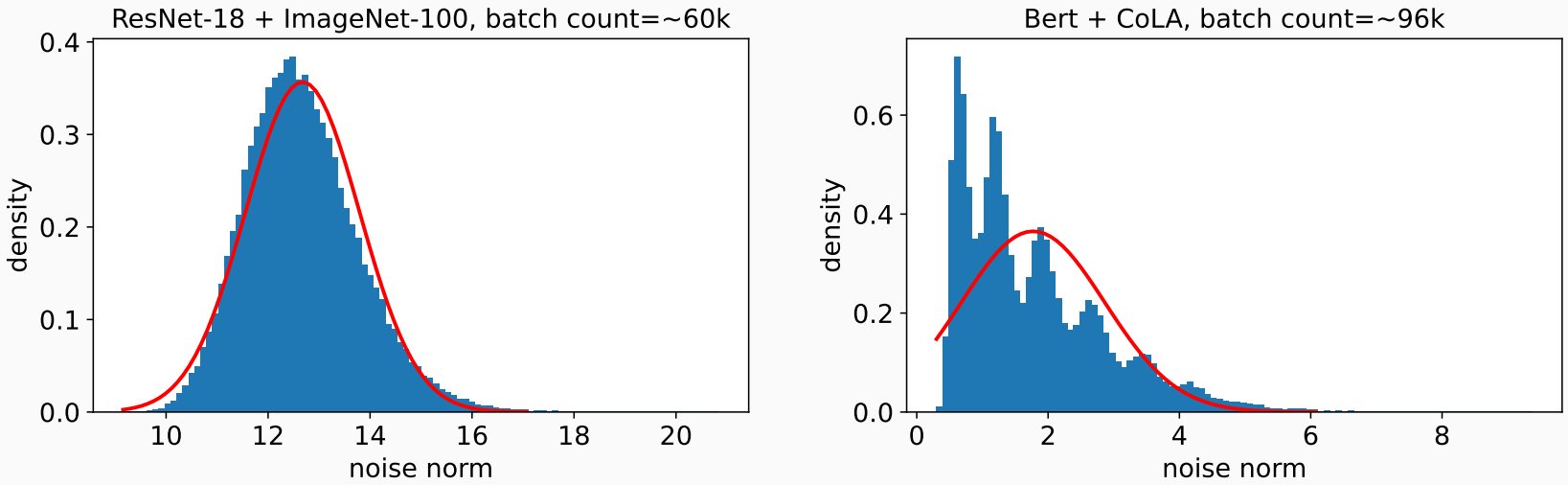

Figure 4: Noise distribution of the stochastic gradients for *QdrMds,07* on *Hl`fdMds,0//* and *ADQS* fine-tuning on the *BnK@* dataset before the training. Red lines: probability density functions of normal distributions with means and variances empirically estimated by the samples. Batch count is the total number of samples used to build a histogram.

Numerical Results, Image Classification

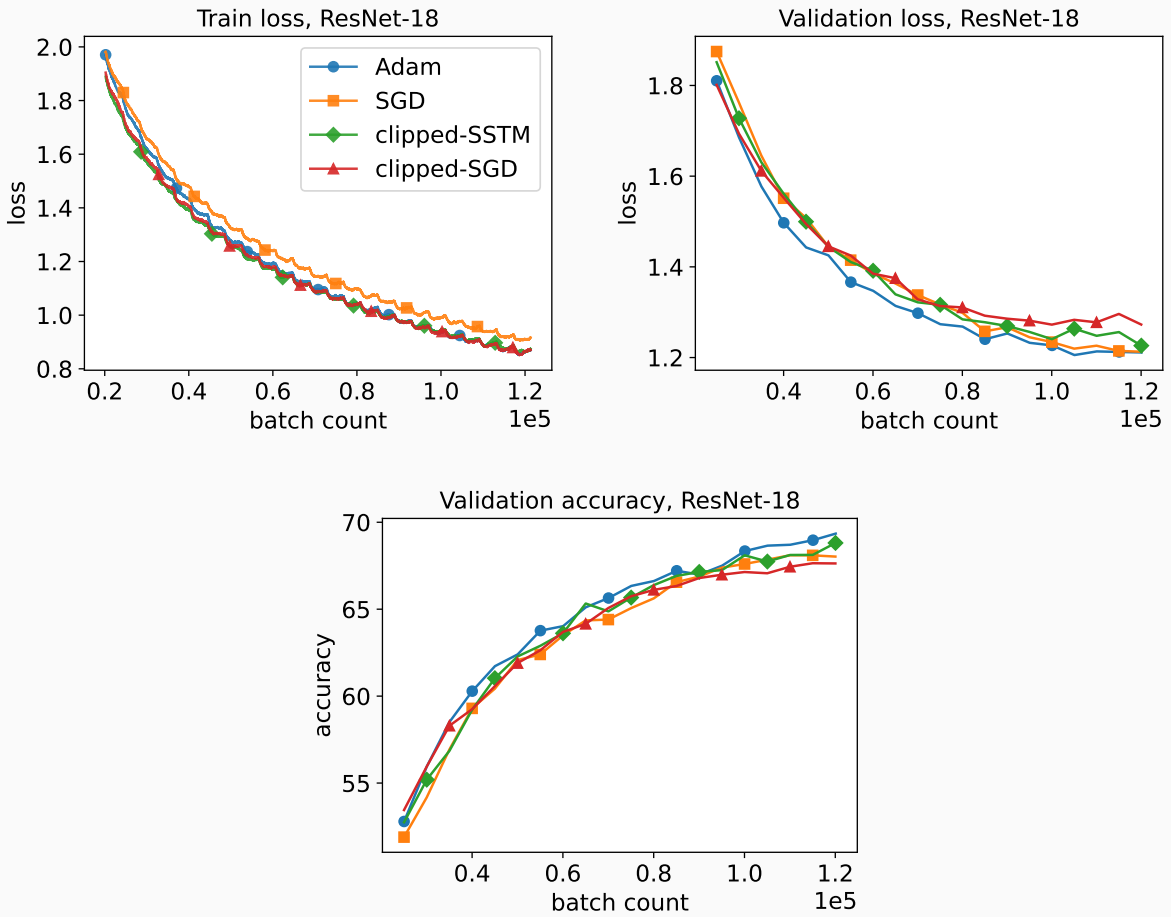

Figure 5: Train and validation loss + accuracy for different optimizers on *QdrMds,07 + Hl`fdMds,0//* problem. Here, "batch count" denotes the total number of used stochastic gradients. The noise distribution is almost Gaussian, even vanilla ***RFC*** performs well.

Numerical Results, Text Classification

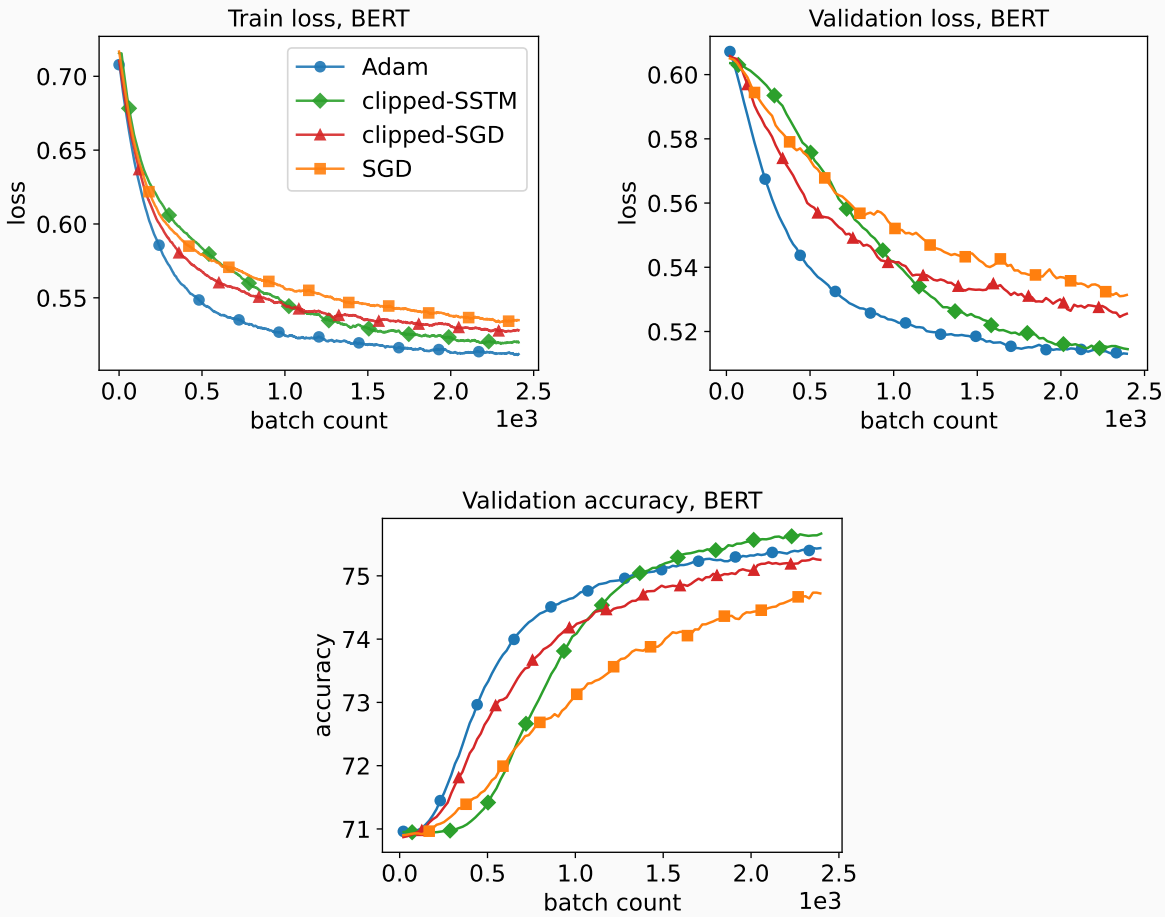

Figure 6: Train and validation loss + accuracy for different optimizers on *ADQS + BnK@* problem. The noise distribution is heavy-tailed, the methods with clipping outperform *RFC* by a large margin.

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 @c `l 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: *gssor9..vvv-`q`amdvr-bnl.mncd.0613000.`lo*

References

Bennett, G. (1962). Probability inequalities for the sum of independent random variables. *[illegible]*, 57(297):33–45.

Dzhaparidze, K. and Van Zanten, J. (2001). On bernstein-type inequalities for martingales. *[illegible]*, 93(1):109–117.

Freedman, D. A. et al. (1975). On tail probabilities for martingales. *[illegible]*, 3(1):100–118.

Ghadimi, S. and Lan, G. (2013). Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *[illegible]*, 23(4):2341–2368.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *[illegible]*. MIT Press. `gsso9..vvv-cddokd`qmhmfannj-nqf`.

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

Gorbunov, E., Danilova, M., and Gasnikov, A. (2020). Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. *[illegible]*, 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. *[illegible]*.

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. [illegible].

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? [illegible], 33.