Methods with Clipping for Stochastic Optimization and Variational Inequalities with Heavy-Tailed Noise

Eduard Gorbunov
Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE

All-Russian Optimization Seminar

September 9, 2022
Outline

1. Clipping and Heavy-Tailed Noise
2. Minimization Problems
3. Variational Inequalities
The Talk is Based on Three Papers

Stochastic Gradient Descent (SGD)

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

- \( f \) – 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)

\[ x^{k+1} = x^k - \gamma \cdot \text{clip}(\nabla f(x^k, \xi^k), \lambda) \quad (2) \]

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

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

![Without clipping](image1.png) ![With clipping](image2.png)

**Figure:** 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

Seems that gradient clipping is an important component in training these models. But why?
Let us look at the distribution of $\|\nabla f(x, \xi) - \nabla f(x)\|$ in two settings:

- Standard vision task: training ResNet50 on ImageNet dataset
- Standard NLP task: training BERT on Wikipedia+Books dataset
Heavy-Tailed Noise in Stochastic Gradients

Figure: from [Zhang et al., 2020]
Definition of Heavy-Tailed Noise in Stochastic Gradients

• Random vector $X$ has light tails if

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

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

$$
\mathbb{E}\left[\exp\left(\frac{\|X - \mathbb{E}[X]\|^2}{\sigma^2}\right)\right] \leq \exp(1). \quad (4)
$$
Definition of Heavy-Tailed Noise in Stochastic Gradients

- Random vector $X$ has light tails if

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

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

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

- Otherwise we say that $X$ has heavy tails. However, in this talk, we will assume that it has bounded variance:

$$
\mathbb{E}\left[\|X - \mathbb{E}[X]\|^2\right] \leq \sigma^2 \quad (5)
$$
Problem and Assumptions

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

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

\[
\begin{align*}
\n f(x) & \geq f(y) + \langle \nabla f(y), x - y \rangle, \\
\| \nabla f(x) - \nabla f(y) \| & \leq L \| x - y \|. 
\end{align*}
\] (7)

\[
\begin{align*}
\| \nabla f(x) - \nabla f(y) \| & \leq L \| x - y \|. 
\end{align*}
\] (8)
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., \( \forall x, y \in \mathbb{R}^n \)
  \[
  f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle, \tag{7}
  \]
  \[
  \| \nabla f(x) - \nabla f(y) \| \leq L \| x - y \|. \tag{8}
  \]

- Stochastic gradient \( \nabla f(x, \xi) \) with bounded variance is available, i.e., \( \forall x \in \mathbb{R}^n \)
  \[
  \mathbb{E}_\xi [\nabla f(x, \xi)] = \nabla f(x), \quad \mathbb{E}_\xi [\| \nabla f(x, \xi) - \nabla f(x) \|^2] \leq \sigma^2. \tag{9}
  \]
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
In-Expectation Guarantees are Less Sensitive to Distribution

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:

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

where $\mathbb{E}[\xi] = 0$ and $\mathbb{E}[\|\xi\|^2] = \sigma^2.$
In-Expectation Guarantees are Less Sensitive to Distribution

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:

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

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

- \( \xi \) has Gaussian distribution
- \( \xi \) has Weibull distribution (non-sub-Gaussian)
- \( \xi \) has Burr Type XII distribution (non-sub-Gaussian)
In-Expectation Guarantees are Less Sensitive to Distribution

For all of three cases, state-of-the-art theory on SGD [Ghadimi and Lan, 2013] says

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

(10)
In-Expectation Guarantees are Less Sensitive to Distribution

For all of three cases, state-of-the-art theory on SGD [Ghadimi and Lan, 2013] says

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

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

**Figure:** from [Gorbunov et al., 2020]
High-Probability Results under Light-Tails Assumption

Light-tails assumption (classical one):

$$\mathbb{E} \left[ \exp \left( \frac{\| \nabla f(x, \xi) - \nabla f(x) \|^2}{\sigma^2} \right) \right] \leq \exp(1).$$ (11)
High-Probability Results under Light-Tails Assumption

Light-tails assumption (classical one):

\[
E \left[ \exp \left( \frac{\| \nabla f(x, \xi) - \nabla f(x) \|^2}{\sigma^2} \right) \right] \leq \exp(1). \tag{11}
\]

Under this assumption (+ convexity and L-smoothness of f)

- Devolder et al. [2011] proved that SGD finds \( \hat{x} \) such that \( f(\hat{x}) - f(x^*) \leq \epsilon \) with probability at least \( 1 - \beta \) using

\[
O \left( \max \left\{ \frac{LR_0^2}{\epsilon}, \frac{\sigma^2 R_0^2}{\epsilon^2} \ln^2 \left( \frac{1}{\beta} \right) \right\} \right) \text{ oracle calls}
\]
High-Probability Results under Light-Tails Assumption

Light-tails assumption (classical one):

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

Under this assumption (+ convexity and \(L\)-smoothness of \(f\))

- Devolder et al. [2011] proved that SGD finds \(\hat{x}\) such that \(f(\hat{x}) - f(x^*) \leq \varepsilon\) with probability at least \(1 - \beta\) using

  \[O \left( \max \left\{ \frac{LR_0^2}{\varepsilon}, \frac{\sigma^2 R_0^2}{\varepsilon^2} \ln^2 \left( \frac{1}{\beta} \right) \right\} \right) \text{ oracle calls}\]

- Ghadimi and Lan [2012] proved that AC-SA (an accelerated version of SGD) finds \(\hat{x}\) such that \(f(\hat{x}) - f(x^*) \leq \varepsilon\) with probability at least \(1 - \beta\) using

  \[O \left( \max \left\{ \sqrt{\frac{LR_0^2}{\varepsilon}}, \frac{\sigma^2 R_0^2}{\varepsilon^2} \ln^2 \left( \frac{1}{\beta} \right) \right\} \right) \text{ oracle calls}\]
High-Probability Results under Heavy-Tails Assumption

- Nazin et al. [2019] proposed Robust Stochastic Mirror Descent (RSMD), which reminds clipped-SGD, and proved the following complexity bound:

\[
\mathcal{O} \left( \max \left\{ \frac{LD^2}{\varepsilon}, \frac{\sigma^2 D^2}{\varepsilon^2} \right\} \ln \left( \frac{1}{\beta} \right) \right)
\]

✓ The first work in the area (in my opinion, it is breakthrough)
✗ \(D\) – diameter of the domain; the proof relies on \(D < +\infty\)
✗ No acceleration
Davis et al. [2021] proposed proxBoost based on robust distance estimation and Proximal Point method. They proved the following complexity bound (in the strongly convex case):

$$\mathcal{O} \left( \max \left\{ \sqrt{\frac{L}{\mu}} \ln \left( \frac{LR_0^2 \ln \frac{L}{\mu}}{\varepsilon} \right), \frac{\sigma^2 \ln \frac{L}{\mu}}{\mu \varepsilon} \right\} \ln \left( \frac{L}{\mu} \right) \ln \left( \frac{\ln \frac{L}{\mu}}{\beta} \right) \right)$$

- Accelerated results
- Valid for any convex closed domain (bounded/unbounded)
- Requires to solve an auxiliary problem at each iteration
- Extra logarithm of the condition number
Key Challenge in the Analysis of clipped-SGD

\[ x^{k+1} = x^k - \gamma \cdot \text{clip} \left( \nabla f(x^k, \xi^k), \lambda \right) \]

\[ \nabla f(x^k, \xi^k) = \frac{1}{m_k} \sum_{i=1}^{m_k} \nabla f(x^k, \xi^k_i), \text{ where } \xi^k_1, \ldots, \xi^k_{m_k} \text{ are i.i.d. samples} \]
Key Challenge in the Analysis of clipped-SGD

\[ x^{k+1} = x^k - \gamma \cdot \text{clip} \left( \nabla f(x^k, \xi^k), \lambda \right) \]

\[ \tilde{\nabla} f(x^k, \xi^k) \]

- \[ \nabla f(x^k, \xi^k) = \frac{1}{m_k} \sum_{i=1}^{m_k} \nabla f(x^k, \xi^i_k) \], where \( \xi^1_k, \ldots, \xi^{m_k}_k \) are i.i.d. samples

- Key challenge: \( \mathbb{E} \left[ \tilde{\nabla} f(x^k, \xi^k) \mid x^k \right] \neq \nabla f(x^k) \)
Analysis of clipped-SGD: Key Idea

• We start the proof classically:

\[
\|x^{k+1} - x^*\|^2 = \|x^k - x^*\|^2 - 2\gamma \langle x^k - x^*, \nabla f(x^k, \xi^k) \rangle + \gamma^2 \|\nabla f(x^k, \xi^k)\|^2 \\
\leq \ldots
\]
Analysis of clipped-SGD: Key Idea

• We start the proof classically:

\[
\|x^{k+1} - x^*\|^2 = \|x^k - x^*\|^2 - 2\gamma \langle x^k - x^*, \tilde{\nabla} f(x^k, \xi^k) \rangle + \gamma^2 \|\tilde{\nabla} f(x^k, \xi^k)\|^2 \\
\leq \ldots
\]

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

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

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. \( \mathbb{E}[X_i | X_{i-1}, \ldots, X_1] = 0 \) for all \( i \geq 1 \). Assume that conditional variances \( \sigma_i^2 \overset{\text{def}}{=} \mathbb{E}[X_i^2 | X_{i-1}, \ldots, X_1] \) 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 \( \sigma_i^2 \overset{\text{def}}{=} \mathbb{E}[X_i^2 \mid X_{i-1}, \ldots, X_1] \) 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 \( N \geq 1 \)

\[
\mathbb{P} \left\{ \left| \sum_{i=1}^{N} X_i \right| > b \text{ and } \sum_{i=1}^{N} \sigma_i^2 \leq G \right\} \leq 2 \exp \left( -\frac{b^2}{2G + 2cb/3} \right).
\]
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 \( \sigma_i^2 \equiv \mathbb{E} [X_i^2 \mid X_{i-1}, \ldots, X_1] \) 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 \( N \geq 1 \)

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

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 \( \theta_k \)
- have upper bounds for \( \|x^k - x^*\| \) and \( \|\theta_k\| \) that hold with large probability
Lemma 2

Let $X$ be a random vector in $\mathbb{R}^n$ and $\tilde{X} = \text{clip}(X, \lambda)$. Then,

$$\|\tilde{X} - \mathbb{E}[\tilde{X}]\| \leq 2\lambda. \quad (12)$$

Moreover, if for some $\sigma \geq 0$ we have $\mathbb{E}[X] = x \in \mathbb{R}^n$, $\mathbb{E}[\|X - x\|^2] \leq \sigma^2$, and $x \leq \lambda/2$, then

$$\|\mathbb{E}[\tilde{X}] - x\| \leq \frac{4\sigma^2}{\lambda},$$

$$\mathbb{E}\left[\|\tilde{X} - x\|^2\right] \leq 18\sigma^2,$$  \quad (14)

$$\mathbb{E}\left[\|\tilde{X} - \mathbb{E}[\tilde{X}]\|^2\right] \leq 18\sigma^2.$$  \quad (15)
Bound on the Distance to the Solution

Inequality

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

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

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

implies

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

Key idea: prove \( R_N \leq CR_0 \) with high probability for some numerical constant \( C \) using the induction!
High-Probability Convergence of clipped-SGD

It is sufficient to make all assumptions on a ball around the solution!
High-Probability Convergence of clipped-SGD

It is sufficient to make all assumptions on a ball around the solution!

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

It is sufficient to make all assumptions on a ball around the solution!

**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 $\ln \left( LR_0^2 / \varepsilon \beta \right) \geq 2$ there exists a choice of $\gamma$ 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

It is sufficient to make all assumptions on a ball around the solution!

**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 $\gamma$ 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

$$O \left( \max \left\{ \frac{LR_0^2}{\varepsilon}, \frac{\sigma^2R_0^2}{\varepsilon^2} \ln \left( \frac{LR_0^2}{\varepsilon \beta} + \frac{\sigma^2R_0^2}{\varepsilon^2 \beta} \right) \right\} \right)$$ iterations/oracle calls.
High-Probability Convergence of clipped-SGD

It is sufficient to make all assumptions on a ball around the solution!

**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 $\gamma$ 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

$$O \left( \max \left\{ \frac{LR_0^2}{\varepsilon}, \frac{\sigma^2 R_0^2}{\varepsilon^2} \ln \left( \frac{LR_0^2}{\varepsilon \beta} + \frac{\sigma^2 R_0^2}{\varepsilon^2 \beta} \right) \right\} \right)$$ iterations/oracle calls.

- Same result (up to the difference in logarithmic factors) as for SGD in the light-tailed case
High-Probability Convergence of \textit{clipped-SGD}

It is sufficient to make all assumptions on a ball around the solution!

\begin{theorem}
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 $\gamma$ such that \textit{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

$$O \left( \max \left\{ \frac{LR_0^2}{\varepsilon}, \frac{\sigma^2 R_0^2}{\varepsilon^2} \ln \left( \frac{LR_0^2}{\varepsilon \beta} + \frac{\sigma^2 R_0^2}{\varepsilon^2 \beta} \right) \right\} \right) \text{ iterations/oracle calls.}$$

\end{theorem}

- Same result (up to the difference in logarithmic factors) as for \textit{SGD} in the light-tailed case
- Same result (up to the difference in logarithmic factors) as for \textit{RSMD} in the heavy-tailed case, but for unconstrained case
Accelerated clipped-SGD: clipped-SSTM

- Stochastic Similar Triangles Method was proposed by Gasnikov and Nesterov [2016]
Accelerated clipped-SGD: clipped-SSTM

- Stochastic Similar Triangles Method was proposed by Gasnikov and Nesterov [2016]
- We combine it with a gradient clipping:

\[
\begin{align*}
\alpha_{k+1} &= \frac{k + 2}{2aL}, \\
A_{k+1} &= A_k + \alpha_{k+1}, \\
\lambda_{k+1} &= \frac{B}{\alpha_{k+1}} \\
\end{align*}
\]

\[
\begin{align*}
x^{k+1} &= \frac{A_k y^k + \alpha_{k+1} z^k}{A_{k+1}} \\
z^{k+1} &= z^k - \alpha_{k+1} \left\{ \nabla f(x^{k+1}, \xi^k) \text{ clip}(\nabla f(x^{k+1}, \xi^k), \lambda_{k+1}) \right\} \\
y^{k+1} &= \frac{A y^k + \alpha_{k+1} z^{k+1}}{A_{k+1}}
\end{align*}
\]
Accelerated clipped-SGD: clipped-SSTM

- Stochastic Similar Triangles Method was proposed by Gasnikov and Nesterov [2016]
- We combine it with a gradient clipping:

\[ \alpha_{k+1} = \frac{k + 2}{2aL}, \quad A_{k+1} = A_k + \alpha_{k+1}, \quad \lambda_{k+1} = \frac{B}{\alpha_{k+1}} \]

\[ x^{k+1} = \frac{A_k y^k + \alpha_{k+1} z^k}{A_{k+1}} \]

\[ z^{k+1} = z^k - \alpha_{k+1} \left( \nabla f(x^{k+1}, \xi^k) \right) \]

\[ y^{k+1} = \frac{A y^k + \alpha_{k+1} z^{k+1}}{A_{k+1}} \]

- Why factor \( a \) is needed?
- Why \( \lambda_{k+1} \) is chosen this way?
clipped-SSTM: Intuition Behind the Proof

- The key idea is the same: prove that $R_N \leq CR_0$ with high probability using the induction
clipped-SSTM: Intuition Behind the Proof

- The key idea is the same: prove that $R_N \leq CR_0$ with high probability using the induction
- The method is accelerated – it is more sensitive to the quality of estimate $\tilde{\nabla}f(x^{k+1}, \xi^k)$
clipped-SSTM: Intuition Behind the Proof

• The key idea is the same: prove that $R_N \leq CR_0$ with high probability using the induction.

• The method is accelerated – it is more sensitive to the quality of estimate $\tilde{\nabla}f(x^{k+1},\xi^k)$.
  - For deterministic SSTM (i.e., STM) one can prove $\|\nabla f(x^{k+1})\| = O(1/\alpha_{k+1})$.
  - This hints to choose $\lambda_{k+1} \sim 1/\alpha_{k+1}$ (in the hope that $\|\nabla f(x^{k+1})\| = O(1/\alpha_{k+1})$ in the stochastic case with high probability).
clipped-SSTM: Intuition Behind the Proof

- The key idea is the same: prove that $R_N \leq CR_0$ with high probability using the induction
- The method is accelerated – it is more sensitive to the quality of estimate $\nabla f(x^{k+1}, \xi^k)$
  - For deterministic SSTM (i.e., STM) one can prove $\|\nabla f(x^{k+1})\| = O(1/\alpha_{k+1})$
  - This hints to choose $\lambda_{k+1} \sim 1/\alpha_{k+1}$ (in the hope that $\|\nabla f(x^{k+1})\| = O(1/\alpha_{k+1})$
    in the stochastic case with high probability)
  - Parameter $a$ allows to choose smaller stepsizes and, as the result, batchsizes $m_k = 1$
High-Probability Convergence of clipped-SSTM

It is sufficient to make all assumptions on a ball around the solution!
High-Probability Convergence of clipped-SSTM

It is sufficient to make all assumptions on a ball around the solution!

**Theorem 2**

Let $f$ be convex and $L$-smooth on $B_{3R_0}(x^*)$ and (9) holds on $B_{3R_0}(x^*)$. 
High-Probability Convergence of clipped-SSTM

It is sufficient to make all assumptions on a ball around the solution!

**Theorem 2**

Let $f$ be convex and $L$-smooth on $B_{3R_0}(x^*)$ and (9) holds on $B_{3R_0}(x^*)$. Then, for all $\beta \in (0, 1)$, $\varepsilon \geq 0$ such that $\ln(\sqrt{L}R_0/\sqrt{\varepsilon \beta}) \geq 2$ there exists a choice of $a$ such that clipped-SSTM with clipping level $\lambda \sim 1/\alpha_{k+1}$ and batchsize $m_k = 1$ finds $y^N$ satisfying $f(y^N) - f(x^*) \leq \varepsilon$ with probability at least $1 - \beta$ using
High-Probability Convergence of clipped-SSTM

It is sufficient to make all assumptions on a ball around the solution!

Theorem 2

Let $f$ be convex and $L$-smooth on $B_{3R_0}(x^*)$ and (9) holds on $B_{3R_0}(x^*)$. Then, for all $\beta \in (0, 1)$, $\varepsilon \geq 0$ such that $\ln(\sqrt{LR_0}/\sqrt{\varepsilon \beta}) \geq 2$ there exists a choice of $\alpha_{k+1}$ and batchsize $m_k = 1$ finds $y^N$ satisfying $f(y^N) - f(x^*) \leq \varepsilon$ with probability at least $1 - \beta$ using

$$O\left(\max\left\{\frac{\sqrt{LR_0^2}}{\varepsilon} \ln \left(\frac{\sqrt{LR_0^2}}{\varepsilon \beta^2}\right), \frac{\sigma^2 R_0^2}{\varepsilon^2} \ln \left(\frac{\sigma^2 R_0^2}{\varepsilon^2 \beta}\right)\right\}\right)$$

iterations/oracle calls.
High-Probability Convergence of clipped-SSTM

It is sufficient to make all assumptions on a ball around the solution!

**Theorem 2**

Let $f$ be convex and $L$-smooth on $B_{3R_0}(x^*)$ and (9) holds on $B_{3R_0}(x^*)$. Then, for all $\beta \in (0, 1)$, $\varepsilon \geq 0$ such that $\ln(\sqrt{LR_0}/\sqrt{\varepsilon \beta}) \geq 2$ there exists a choice of $\alpha$ such that clipped-SSTM with clipping level $\lambda \sim 1/\alpha_{k+1}$ and batchsize $m_k = 1$ finds $y^N$ satisfying $f(y^N) - f(x^*) \leq \varepsilon$ with probability at least $1 - \beta$ using

$$O \left( \max \left\{ \sqrt{\frac{LR_0^2}{\varepsilon}} \ln \left( \sqrt{\frac{LR_0^2}{\varepsilon \beta^2}} \right), \frac{\sigma^2 R_0^2}{\varepsilon^2} \ln \left( \frac{\sigma^2 R_0^2}{\varepsilon^2 \beta} \right) \right\} \right)$$ iterations/oracle calls.

- Same result (up to the difference in logarithmic factors) as for AC-SA in the light-tailed case
It is sufficient to make all assumptions on a ball around the solution!

**Theorem 2**

Let $f$ be convex and $L$-smooth on $B_{3R_0}(x^*)$ and (9) holds on $B_{3R_0}(x^*)$. Then, for all $\beta \in (0, 1)$, $\varepsilon \geq 0$ such that $\ln(\sqrt{LR_0}/\sqrt{\varepsilon \beta}) \geq 2$ there exists a choice of $\alpha$ such that clipped-SSTM with clipping level $\lambda \sim 1/\alpha_{k+1}$ and batchsize $m_k = 1$ finds $y^N$ satisfying $f(y^N) - f(x^*) \leq \varepsilon$ with probability at least $1 - \beta$ using

$$O \left( \max \left\{ \frac{LR_0^2}{\varepsilon} \ln \left( \frac{LR_0^2}{\varepsilon \beta^2} \right), \frac{\sigma^2 R_0^2}{\varepsilon^2} \ln \left( \frac{\sigma^2 R_0^2}{\varepsilon^2 \beta} \right) \right\} \right)$$ iterations/oracle calls.

- Same result (up to the difference in logarithmic factors) as for AC-SA in the light-tailed case
- Better result than for clipped-SGD
Theoretical Extensions

In [Gorbunov et al., 2020, 2021] we also have

- Results for the strongly convex objectives
- Results for the functions with Hölder continuous gradient
Numerical Experiments: Setup

We tested the performance of the methods on the following problems:\footnote{The code is available at \url{https://github.com/ClippedStochasticMethods/clipped-SSTM}}:

- **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.
Figure: 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.
Evolution of the Noise Distribution, Image Classification

Figure: Evolution of the noise distribution for ResNet-18 + ImageNet-100 task.
Figure: Evolution of the noise distribution for BERT + CoLA task.
Evolution of the Noise Distribution, Text Classification

**Figure**: Evolution of the noise distribution for BERT + CoLA task, from iteration 0 (before the training) to iteration 500.
Numerical Results, Image Classification

**Figure**: 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, i.e., gradient clipping is not required.
**Figure:** 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.
Variational Inequality Problem

\[
\text{find } x^* \in Q \subseteq \mathbb{R}^n \text{ such that } \langle F(x^*), x - x^* \rangle \geq 0, \forall x \in Q \quad (\text{VIP-C})
\]
Variational Inequality Problem

find $x^* \in Q \subseteq \mathbb{R}^n$ such that $\langle F(x^*), x - x^* \rangle \geq 0$, $\forall x \in Q$ (VIP-C)

- $F : Q \rightarrow \mathbb{R}^n$ is $L$-Lipschitz operator: $\forall x, y \in Q$

$$\|F(x) - F(y)\| \leq L\|x - y\|$$ (16)
Variational Inequality Problem

find \( x^* \in Q \subseteq \mathbb{R}^n \) such that \( \langle F(x^*), x - x^* \rangle \geq 0, \forall x \in Q \) \hfill (VIP-C)

- \( F : Q \rightarrow \mathbb{R}^n \) is \( L \)-Lipschitz operator: \( \forall x, y \in Q \)

\[
\| F(x) - F(y) \| \leq L \| x - y \| \quad (16)
\]

- \( F \) is monotone: \( \forall x, y \in Q \)

\[
\langle F(x) - F(y), x - y \rangle \geq 0 \quad (17)
\]
Variational Inequality Problem: Examples

- Min-max problems:

\[
\min_{u \in U} \max_{v \in V} f(u, v) \tag{18}
\]
Min-max problems:

$$\min_{u \in U} \max_{v \in V} f(u, v)$$  \hspace{1cm} (18)$$

If $f$ is convex-concave, then (18) is equivalent to finding $(u^*, v^*) \in U \times V$ such that $\forall (u, v) \in U \times V$

$$\langle \nabla_u f(u^*, v^*), u - u^* \rangle \geq 0, \quad -\langle \nabla_v f(u^*, v^*), v - v^* \rangle \geq 0,$$

These problems appear in various applications such as robust optimization [Ben-Tal et al., 2009] and control [Hast et al., 2013], adversarial training [Goodfellow et al., 2015, Madry et al., 2018] and generative adversarial networks (GANs) [Goodfellow et al., 2014].
Variational Inequality Problem: Examples

- Min-max problems:

\[
\min_{u \in U} \max_{v \in V} f(u, v) \tag{18}
\]

If \( f \) is convex-concave, then \( (18) \) is equivalent to finding \((u^*, v^*) \in U \times V\) such that \( \forall (u, v) \in U \times V \)

\[
\langle \nabla_u f(u^*, v^*), u - u^* \rangle \geq 0, \quad -\langle \nabla_v f(u^*, v^*), v - v^* \rangle \geq 0,
\]

which is equivalent to (VIP-C) with \( Q = U \times V, x = (u^T, v^T)^T \), and

\[
F(x) = \begin{pmatrix} \nabla_u f(u, v) \\ -\nabla_v f(u, v) \end{pmatrix}
\]

These problems appear in various applications such as robust optimization [Ben-Tal et al., 2009] and control [Hast et al., 2013], adversarial training [Goodfellow et al., 2015, Madry et al., 2018] and generative adversarial networks (GANs) [Goodfellow et al., 2014].
Variational Inequality Problem: Examples

- Min-max problems:

\[
\min_{u \in U} \max_{v \in V} f(u, v) \tag{18}
\]

If \( f \) is convex-concave, then (18) is equivalent to finding \((u^*, v^*) \in U \times V\) such that \( \forall (u, v) \in U \times V \)

\[
\langle \nabla_u f(u^*, v^*), u - u^* \rangle \geq 0, \quad -\langle \nabla_v f(u^*, v^*), v - v^* \rangle \geq 0,
\]

which is equivalent to (VIP-C) with \( Q = U \times V \), \( x = (u^T, v^T)^T \), and

\[
F(x) = \begin{pmatrix} \nabla_u f(u, v) \\ -\nabla_v f(u, v) \end{pmatrix}
\]

These problems appear in various applications such as robust optimization [Ben-Tal et al., 2009] and control [Hast et al., 2013], adversarial training [Goodfellow et al., 2015, Madry et al., 2018] and generative adversarial networks (GANs) [Goodfellow et al., 2014].
Variational Inequality Problem: Examples

• Minimization problems:

\[ \min_{x \in Q} f(x) \tag{19} \]
Variational Inequality Problem: Examples

• Minimization problems:

\[
\min_{x \in Q} f(x) \tag{19}
\]

If \( f \) is convex, then (19) is equivalent to finding a stationary point of \( f \), i.e., it is equivalent to (VIP-C) with

\[
F(x) = \nabla f(x)
\]
Variational Inequality Problem: Unconstrained Case

When $Q = \mathbb{R}^n$ (VIP-C) can be rewritten as

$$\text{find } x^* \in \mathbb{R}^n \text{ such that } F(x^*) = 0$$  \hspace{1cm} (VIP)

In this talk, we focus on (40) rather than (VIP-C)
Gradient Descent-Ascent (GDA) and Extragradient (EG)

- GDA [Krasnosel’skii, 1955, Mann, 1953]:
  \[ x^{k+1} = x^k - \gamma F(x^k) \]

  ✓ Very simple
  ✗ Does not converge for some simple problems (like bilinear games)
Gradient Descent-Ascent (GDA) and Extragradient (EG)

- **GDA** [Krasnosel’skii, 1955, Mann, 1953]:
  \[ x^{k+1} = x^k - \gamma F(x^k) \]
  - ✓ Very simple
  - ✗ Does not converge for some simple problems (like bilinear games)

- **EG** [Korpelevich, 1976]
  \[ x^{k+1} = x^k - \gamma F (x^k - \gamma F(x^k)) \]
  - ✓ Converges for any monotone and \( L \)-Lipschitz operator
  - ✗ Requires two oracle calls per step (although this can be easily fixed)
  - ✗ Converges worse than Alternating GDA for some popular tasks (GANs)
Stochastic VIP

We consider with

\[ F(x) = \mathbb{E}_\xi [F_\xi(x)] \]

• We have access to \( F_\xi \) such that for all \( x \in \mathbb{R}^n \)

\[ \mathbb{E}_\xi [\|F_\xi(x) - F(x)\|^2] \leq \sigma^2 \quad (20) \]
We consider with

\[ F(x) = \mathbb{E}_\xi[F_\xi(x)] \]

- We have access to \( F_\xi \) such that for all \( x \in \mathbb{R}^n \)

\[ \mathbb{E}_\xi [\|F_\xi(x) - F(x)\|^2] \leq \sigma^2 \tag{20} \]

- For GDA-based methods we assume \( \ell \)-star-cocoercivity: \( \forall x \in \mathbb{R}^n \)

\[ \ell \langle F(x), x - x^* \rangle \geq \|F(x)\|^2 \]
Stochastic VIP

We consider with

\[ F(x) = \mathbb{E}_\xi [F_\xi(x)] \]

- We have access to \( F_\xi \) such that for all \( x \in \mathbb{R}^n \)

\[ \mathbb{E}_\xi \left[ \| F_\xi(x) - F(x) \|^2 \right] \leq \sigma^2 \quad (20) \]

- For GDA-based methods we assume \( \ell \)-star-cocoercivity: \( \forall x \in \mathbb{R}^n \)

\[ \ell \langle F(x), x - x^* \rangle \geq \| F(x) \|^2 \]

- For EG-based methods we assume monotonicity and \( L \)-Lipschitzness: \( \forall x, y \in \mathbb{R}^n \)

\[ \langle F(x) - F(y), x - y \rangle \geq 0, \]
\[ \| F(x) - F(y) \| \leq L \| x - y \| \]
Stochastic GDA (SGDA) and Stochastic EG (SEG)

- **SGDA:**
  \[ x^{k+1} = x^k - \gamma F_{\xi_k}(x^k) \]

- **SEG:**
  \[ x^{k+1} = x^k - \gamma_2 F_{\xi_2} \left( x^k - \gamma_1 F_{\xi_1}(x^k) \right) \]
Stochastic GDA (SGDA) and Stochastic EG (SEG)

- **SGDA:**
  \[ x^{k+1} = x^k - \gamma F_{\xi_k}(x^k) \]

- **SEG:**
  \[ x^{k+1} = x^k - \gamma_2 F_{\xi_2} \left( x^k - \gamma_1 F_{\xi_1}(x^k) \right) \]

- \( \xi_1^k, \xi_2^k \) are i.i.d. samples
- \( \gamma_2 \leq \gamma_1 \)
Prior Work on High-Probability Convergence

For the case of bounded domain (with diameter $D$) and under light-tails assumption

$$
\mathbb{E} \left[ \exp \left( \frac{\|F_\xi(x) - F(x)\|^2}{\sigma^2} \right) \right] \leq \exp(1),
$$

(21)

Juditsky et al. [2011] proved that projected version of SEG (Mirror-Prox) finds $\hat{x}$ such that $^2\text{Gap}_D(\hat{x}) \leq \varepsilon$ with probability at least $1 - \beta$ using

$$
\mathcal{O} \left( \max \left\{ \frac{LD^2}{\varepsilon}, \frac{\sigma^2D^2}{\varepsilon^2} \ln^2 \left( \frac{1}{\beta} \right) \right\} \right) \text{ oracle calls}
$$
clipped-SGDA and clipped-SEG

- **SGDA:**
  \[ x^{k+1} = x^k - \gamma \cdot \text{clip} (F_{\xi_k} (x^k), \lambda_k) \]

- **SEG:**
  \[ x^{k+1} = x^k - \gamma_2 \cdot \text{clip} (F_{\xi_2} (\tilde{x}^k), \lambda_2,k), \quad \tilde{x}^k = x^k - \gamma_1 \cdot \text{clip} (F_{\xi_1} (x^k), \lambda_1,k) \]

- \( \xi_1^k, \xi_2^k \) are i.i.d. samples
- \( \gamma_2 \leq \gamma_1 \)
clipped-SGDA and clipped-SEG

- **SGDA:**
  \[ x^{k+1} = x^k - \gamma \cdot \text{clip}(F_{\xi_{k}}(x^k), \lambda_k) \]

- **SEG:**
  \[ x^{k+1} = x^k - \gamma_2 \cdot \text{clip}(F_{\xi_{k}^2}(\tilde{x}^k), \lambda_{2,k}) \]
  \[ \tilde{x}^k = x^k - \gamma_1 \cdot \text{clip}(F_{\xi_{k}^1}(x^k), \lambda_{1,k}) \]

  - \( \xi_{k}^1, \xi_{k}^2 \) are i.i.d. samples
  - \( \gamma_2 \leq \gamma_1 \)

  The key idea behind the proof is exactly the same as in minimization!
High-Probability Convergence of clipped-SEG

It is sufficient to make all assumptions on a ball around the solution!

**Theorem 3**

Let $F$ be monotone and $L$-Lipschitz on $B_{4R}(x^*)$ and (20) holds on $B_{4R}(x^*)$, $R \geq R_0$. 

• Same result (up to the difference in logarithmic factors) as for Mirror-Prox in the light-tailed case
• Derived for unconstrained case
High-Probability Convergence of clipped-SEG

It is sufficient to make all assumptions on a ball around the solution!

Theorem 3

Let $F$ be monotone and $L$-Lipschitz on $B_{4R}(x^*)$ and (20) holds on $B_{4R}(x^*)$, $R \geq R_0$. Then, for all $\beta \in (0,1)$, $\varepsilon \geq 0$ such that $\ln(\frac{6LR_0^2}{\varepsilon \beta}) \geq 1$ there exists a choice of $\gamma_1 = \gamma_2 = \gamma$ such that clipped-SEG with clipping level $\lambda \sim \frac{1}{\gamma}$ finds $\hat{x}$ satisfying $\text{Gap}_R(\hat{x}) \leq \varepsilon$ with probability at least $1 - \beta$ using...
High-Probability Convergence of clipped-SEG

It is sufficient to make all assumptions on a ball around the solution!

**Theorem 3**

Let $F$ be monotone and $L$-Lipschitz on $B_{4R}(x^*)$ and (20) holds on $B_{4R}(x^*)$, $R \geq R_0$. Then, for all $\beta \in (0, 1)$, $\varepsilon \geq 0$ such that $\ln(\frac{6LR_0^2}{\varepsilon \beta}) \geq 1$ there exists a choice of $\gamma_1 = \gamma_2 = \gamma$ such that clipped-SEG with clipping level $\lambda \sim \frac{1}{\gamma}$ finds $\hat{x}$ satisfying $\text{Gap}_R(\hat{x}) \leq \varepsilon$ with probability at least $1 - \beta$ using

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

iterations/oracle calls.
High-Probability Convergence of clipped-SEG

It is sufficient to make all assumptions on a ball around the solution!

**Theorem 3**

Let $F$ be monotone and $L$-Lipschitz on $B_{4R}(x^*)$ and (20) holds on $B_{4R}(x^*)$, $R \geq R_0$. Then, for all $\beta \in (0,1)$, $\epsilon \geq 0$ such that $\ln(\frac{6L^2R_0^2}{\epsilon \beta}) \geq 1$ there exists a choice of $\gamma_1 = \gamma_2 = \gamma$ such that clipped-SEG with clipping level $\lambda \sim \frac{1}{\gamma}$ finds $\hat{x}$ satisfying $\text{Gap}_R(\hat{x}) \leq \epsilon$ with probability at least $1 - \beta$ using

$$O \left( \max \left\{ \frac{LR^2}{\epsilon} \ln \left( \frac{LR^2}{\epsilon \beta} \right), \frac{\sigma^2 R^2}{\epsilon^2} \ln \left( \frac{\sigma^2 R^2}{\epsilon^2 \beta} \right) \right\} \right)^\gamma$$ iterations/oracle calls.

• Same result (up to the difference in logarithmic factors) as for Mirror-Prox in the light-tailed case
• Derived for unconstrained case
High-Probability Convergence of clipped-SGDA

It is sufficient to make all assumptions on a ball around the solution!

Theorem 4

Let $F$ be $\ell$-star-cocoercive on $B_{2R}(x^*)$ and (20) holds on $B_{2R}(x^*)$, $R \geq R_0$. 
High-Probability Convergence of clipped-SGDA

It is sufficient to make all assumptions on a ball around the solution!

**Theorem 4**

Let $F$ be $\ell$-star-cocoercive on $B_{2R}(x^*)$ and (20) holds on $B_{2R}(x^*)$, $R \geq R_0$. Then, for all $\beta \in (0,1)$, $\varepsilon \geq 0$ such that $\ln(6LR_0^2/\varepsilon \beta) \geq 1$ there exists a choice of $\gamma$ such that clipped-SGDA with clipping level $\lambda \sim 1/\gamma$ finds $\hat{x}$ satisfying

$$\frac{1}{K+1} \sum_{k=0}^{K} \|F(x^k)\|^2 \leq \varepsilon \quad \text{with probability at least } 1 - \beta$$

using
High-Probability Convergence of clipped-SGDA

It is sufficient to make all assumptions on a ball around the solution!

**Theorem 4**

Let $F$ be $\ell$-star-cocoercive on $B_{2R}(x^*)$ and (20) holds on $B_{2R}(x^*)$, $R \geq R_0$. Then, for all $\beta \in (0, 1)$, $\epsilon \geq 0$ such that $\ln(6LR_0^2/\epsilon \beta) \geq 1$ there exists a choice of $\gamma$ such that clipped-SGDA with clipping level $\lambda \sim 1/\gamma$ finds $\hat{x}$ satisfying

$$\frac{1}{K+1} \sum_{k=0}^{K} \|F(x^k)\|^2 \leq \epsilon$$

with probability at least $1 - \beta$ using

$$O \left( \max \left\{ \frac{\ell^2 R^2}{\epsilon} \ln \left( \frac{\ell^2 R^2}{\epsilon \beta} \right), \frac{\ell^2 \sigma^2 R^2}{\epsilon^2} \ln \left( \frac{\ell^2 \sigma^2 R^2}{\epsilon^2 \beta} \right) \right\} \right)$$

iterations/oracle calls.
High-Probability Convergence of clipped-SGDA

It is sufficient to make all assumptions on a ball around the solution!

**Theorem 4**

Let $F$ be $\ell$-star-cocoercive on $B_{2R}(x^*)$ and (20) holds on $B_{2R}(x^*)$, $R \geq R_0$. Then, for all $\beta \in (0, 1)$, $\varepsilon \geq 0$ such that $\ln(\frac{6LR_0^2}{\varepsilon\beta}) \geq 1$ there exists a choice of $\gamma$ such that clipped-SGDA with clipping level $\lambda \sim \frac{1}{\gamma}$ finds $\hat{x}$ satisfying

$$\frac{1}{K+1} \sum_{k=0}^{K} \|F(x^k)\|^2 \leq \varepsilon \text{ with probability at least } 1 - \beta$$

using

$$O\left(\max \left\{ \frac{\ell^2 R^2}{\varepsilon} \ln \left( \frac{\ell^2 R^2}{\varepsilon\beta} \right), \frac{\ell^2 \sigma^2 R^2}{\varepsilon^2} \ln \left( \frac{\ell^2 \sigma^2 R^2}{\varepsilon^2\beta} \right) \right\} \right) \text{ iterations/oracle calls.}$$

- The first high-probability complexity result for SGDA-based methods
Theoretical Extensions

In [Gorbunov et al., 2022] we also have

- extensions to the quasi-strongly monotone and star-negative comonotone problems for clipped-SEG
- extensions to the (quasi-strongly) monotone + star-cocoercive problems for clipped-SGDA
Numerical Experiments

In the experiments in training GANs, we tested the following methods:

- clipped-SGDA with alternating updates
- Coord-clipped-SGDA – clipped-SGDA with coordinate-wise clipping and alternating updates
- clipped-SEG
- Coord-clipped-SEG
WGAN-GP on CIFAR10 Has Heavy-Tailed Gradients

- $\rho_{mR}$: relative fraction of mass after $Q_3 + 1.5 \cdot (Q_3 - Q_1)$
  - For normal distribution there is $\approx 0.35\%$ of the mass
  - In this plot: $\approx 12$ times more
- $\rho_{meR}$: relative fraction of mass after $Q_3 + 3 \cdot (Q_3 - Q_1)$
  - For normal distribution there is $\approx 10^{-4}\%$ of the mass
  - In this plot: $\approx 4603$ times more
WGAN-GP on CIFAR10 Has Heavy-Tailed Gradients

Eduard Gorbunov

Clipping, Heavy Tails, High Prob. Analysis

September 9, 2022 49 / 59
Clipping Helps for WGAN-GP on CIFAR10

(a) SGDA (67.4)  
(b) clipped-SGDA (19.7)  
(c) clipped-SEG (25.3)

The graph shows FID (Fréchet Inception Distance) for different optimization methods with and without clipping.

- With Clipping
- Without Clipping

Diverged: False, True

Clipping and Heavy-Tailed Noise  
Minimization Problems  
Variational Inequalities  
References
StyleGAN2 on FFHQ Has Heavy-Tailed Gradients

(a) Initialization

(b) clipped-SGDA
Clipping Helps for StyleGAN2 on FFHQ

(c) SGDA

(d) clipped-SGDA
Clipping Helps for StyleGAN2 on FFHQ

• Still not matching Adam (on this GAN)
• StyleGan2 is full of trick and heuristics
• Has been tuned for Adam!
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


References IV


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

robust stochastic optimization based on mirror descent method. *Automation

R. Pascanu, T. Mikolov, and Y. Bengio. On the difficulty of training recurrent

sequence tagging with bidirectional language models. *arXiv preprint

Why are adaptive methods good for attention models? *Advances in Neural