Algorithms for Stochastic Optimization with Heavy-Tailed Noise and Connections with the Training of Large Language Models

Oberseminar at LT Group, University of Hamburg

Eduard Gorbunov
June 6, 2023

Mohamed bin Zayed University of Artificial Intelligence
About me

- Postdoc at Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, UAE
- Got PhD at Moscow Institute of Physics and Technology in December 2021
- Research interests: stochastic optimization, variational inequalities, computer-aided proofs, federated learning
- Hobbies: wakesurf, gym, hiking, football
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: ≈ 200 students, ≈ 50 faculties, 19th in CSRankings (AI, CV, ML, and NLP)
- 1 hour to Dubai :)

Figure 1: https://www.arabnews.com/node/1724111/amp
1. Clipping and Heavy-Tailed Noise

2. In-Expectation Guarantees vs High-Probability Convergence

3. Main Results for Minimization Problems

4. Main Results for Variational Inequalities
The Talk is Based on Four Papers


Clipping and Heavy-Tailed Noise
Stochastic Gradient Descent (SGD)

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

- \( f \) – 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)

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

- \( clip(x, \lambda) = \min\{1, \lambda/\|x\|\}x \)
- \( clip(\nabla f(x^k, \xi^k), \lambda) \) - biased estimate of \( \nabla f(x^k) \):
  \[ \mathbb{E}_{\xi^k}[clip(\nabla f(x^k, \xi^k), \lambda)] \neq \nabla f(x^k) \]
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 2: 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?
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
Figure 3: from (Zhang et al., 2020)
• Random vector $X$ has light tails if

$$\mathbb{P}\{\|X - \mathbb{E}[X]\| \geq b\} \leq 2 \exp\left(-\frac{b^2}{2\sigma^2}\right) \quad \forall b > 0.$$  \hspace{1cm} (3)

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

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

- Random vector $X$ has light tails if

$$\mathbb{P}\{\|X - \mathbb{E}[X]\| \geq b\} \leq 2 \exp \left( -\frac{b^2}{2\sigma^2} \right) \quad \forall b > 0.$$  \hspace{1cm} (3)

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

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

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

$$\mathbb{E} [\|X - \mathbb{E}[X]\|^\alpha] \leq \sigma^\alpha$$ \hspace{1cm} (5)
In-Expectation Guarantees vs High-Probability Convergence
Problem and Assumptions

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

- \( f : \mathbb{R}^n \to \mathbb{R}^n \) is convex and \( L \)-smooth, i.e., \( \forall x, y \in \mathbb{R}^n \)
  \[
  f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle, \quad (7)
  \]
  \[
  \| \nabla f(x) - \nabla f(y) \| \leq L \| x - y \|. \quad (8)
  \]
Problem and Assumptions

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

\begin{itemize}
    \item \( f : \mathbb{R}^n \rightarrow \mathbb{R}^n \) is convex and \( L \)-smooth, i.e., \( \forall x, y \in \mathbb{R}^n \)
    \[
    f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle, \tag{7}
    \]
    \[
    \| \nabla f(x) - \nabla f(y) \| \leq L \| x - y \|. \tag{8}
    \]
    \item Stochastic gradient \( \nabla f(x, \xi) \) with bounded central \( \alpha \)-th moment \( (\alpha \in (1, 2]) \) is available, i.e., \( \forall x \in \mathbb{R}^n \)
    \[
    \mathbb{E}_\xi [\nabla f(x, \xi)] = \nabla f(x), \quad \mathbb{E}_\xi [\| \nabla f(x, \xi) - \nabla f(x) \|^\alpha] \leq \sigma^\alpha. \tag{9}
    \]
\end{itemize}
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., $\xi$ can be a Lévy $\alpha$-stable distribution).

Then, after one step of SGD we have

$\mathbb{E}\|x_1 - x^*\|^2 = \mathbb{E}\|x_0 - x^* - \gamma \nabla f(x_0, \xi_0)\|^2 = \|x_0 - x^*\|^2 - 2\gamma \mathbb{E}\|\xi_0\|^{\alpha} + \gamma^2 \mathbb{E}\|\nabla f(x_0, \xi_0)\|^2$.

The method does not converge in expectation ($L^2$) when $\alpha < 2$!
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., $\xi$ can Levý $\alpha$-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., $\xi$ can be a Lévy $\alpha$-stable distribution)
- Then, after one step of SGD we have

\[
\begin{align*}
\mathbb{E}\|x^1 - x^*\|^2 &= \mathbb{E}\|x^0 - x^* - \gamma \nabla f(x^0, \xi^0)\|^2 \\
&= \|x^0 - x^*\|^2 - 2\gamma \mathbb{E} [x^0 - x^*, \nabla f(x^0, \xi^0)] \\
&\quad + \gamma^2 \mathbb{E}\|\nabla f(x^0, \xi^0)\|^2 \\
&= \infty
\end{align*}
\]

The method does not converge in expectation (in $L_2$) when $\alpha < 2$!

What about the case when $\alpha = 2$ (bounded variance)?
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 \( \mathbb{E}[\|\xi\|^2] = \sigma^2 \).
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)
For all of three cases, state-of-the-art theory on SGD (Ghadimi and Lan, 2013) says

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

(10)
For all of three cases, state-of-the-art theory on SGD (Ghadimi and Lan, 2013) says

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

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

**Figure 4:** 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

- High-probability guarantees:
  \( \mathbb{P}\{\|x - x^*\|^2 \leq \varepsilon\} \geq 1 - \beta, \mathbb{P}\{f(x) - f(x^*) \leq \varepsilon\} \geq 1 - \beta, \mathbb{P}\{\|\nabla f(x)\|^2 \leq \varepsilon\} \geq 1 - \beta \)
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 Results under Light-Tails Assumption

Light-tails assumption (classical one):

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

Light-tails assumption (classical one):

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

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

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

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

Light-tails assumption (classical one):

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

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

- Devolder et al. (2011) proved that \(SGD\) finds \(\hat{x}\) such that \(f(\hat{x}) - f(x^*) \leq \varepsilon\) with probability at least \(1 - \beta\) using
  \[
  \mathcal{O} \left( \max \left\{ \frac{LR_0^2}{\varepsilon}, \frac{\sigma^2 R_0^2}{\varepsilon^2} \ln^2 \left( \frac{1}{\beta} \right) \right\} \right) \text{ oracle calls}
  \]

- 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
  \[
  \mathcal{O} \left( \max \left\{ \sqrt{\frac{LR_0^2}{\varepsilon}}, \frac{\sigma^2 R_0^2}{\varepsilon^2} \ln^2 \left( \frac{1}{\beta} \right) \right\} \right) \text{ oracle calls}
  \]
Natural idea: apply Markov’s inequality:

\[ \mathbb{P} \{ f(\hat{x}) - f(x^*) > \varepsilon \} < \frac{\mathbb{E} [f(\hat{x}) - f(x^*)]}{\varepsilon} . \]
High-Probability Convergence of SGD under Bounded Variance Assumption

**Natural idea:** apply Markov’s inequality:

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

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

**Bad news:** to ensure $\mathbb{E}[f(\hat{x}) - f(x^*)] \leq \varepsilon \beta$, SGD needs $O(\max\{L^2 R^2 / \varepsilon^2 \beta^2, \varepsilon \beta\})$ oracle calls.
Natural idea: apply Markov's inequality:

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

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

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

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

Negative-power dependence on $$\beta$$ :(
Natural idea: apply Markov's inequality:

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

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

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

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

Negative-power dependence on \( \beta \):(

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

For any $\varepsilon > 0$, $\beta \in (0, 1)$, and SGD parameterized by the number of steps $K$ and stepsize $\gamma$, there exists $\mu$-strongly convex $L$-smooth problem (19) and stochastic oracle with noise having bounded $\alpha$-th moment with $\alpha = 2$, $0 < \mu \leq L$ such that for the iterates produced by SGD with any stepsize $0 < \gamma \leq 1/\mu$

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

(12)

This illustrates the necessity of modifying the method, e.g., one can use gradient clipping.
Main Results for Minimization Problems
Key Challenge in the Analysis of clipped-SGD

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

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

- 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:

\[
\begin{align*}
\|x^{k+1} - x^*\|^2 &= \|x^k - x^*\|^2 - 2\gamma \langle x^k - x^*, \tilde{\nabla}f(x^k, \xi^k) \rangle \\
&\quad + \gamma^2 \|\tilde{\nabla}f(x^k, \xi^k)\|^2 \\
&\leq \ldots
\end{align*}
\]
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?
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 \( \sigma_i^2 \stackrel{\text{def}}{=} \mathbb{E}[X_i^2 \mid X_{i-1}, \ldots, X_1] \) exist and are bounded and assume also that there exists deterministic constant \( c > 0 \) such that \( |X_i| \leq c \) almost surely for all \( i \geq 1 \).
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 \( \sigma_i^2 \stackrel{\text{def}}{=} \mathbb{E}[X_i^2 \mid X_{i-1}, \ldots, X_1] \) 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 \( N \geq 1 \)

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

\[
\begin{align*}
\|\mathbb{E}[\tilde{X}] - x\| & \leq \frac{2^\alpha \sigma^\alpha}{\lambda^{\alpha-1}}, \\
\mathbb{E} \left[ \|\tilde{X} - x\|^2 \right] & \leq 18\lambda^{2-\alpha} \sigma^{\alpha}, \\
\mathbb{E} \left[ \|\tilde{X} - \mathbb{E}[\tilde{X}]\|^2 \right] & \leq 18\lambda^{2-\alpha} \sigma^{\alpha}.
\end{align*}
\]
**Bound on the Distance to the Solution**

Inequality

\[
\frac{2\gamma(1 - 2\gamma L)}{N} \sum_{k=0}^{N-1} \Delta_R \leq \frac{1}{N} \left( R_0^2 - R_N^2 \right)
\]

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

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

\[ \frac{2\gamma(1 - 2\gamma L)}{N} \sum_{k=0}^{N-1} \Delta_k \leq \frac{1}{N} \left( R_0^2 - R_N^2 \right) \]

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

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 \textit{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 $B_{7R_0}(x^*)$. 
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 $\ln\left(\frac{LR_0^2}{\varepsilon\beta}\right) \geq 2$ there exists a choice of $\gamma$ such that clipped-SGD with clipping level $\lambda \sim \frac{1}{\gamma}$ 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 $O\left(\max\left(\frac{LR_0^2}{\varepsilon}, \frac{1}{\alpha \varepsilon} \right) \right)$ 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 $\ln \left( LR_0^2 / \varepsilon \beta \right) \geq 2$ 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

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

iterations/oracle calls.
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:

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

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

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

\[
y_{k+1} = \frac{Ay^k + \alpha_{k+1} z^{k+1}}{A_{k+1}}
\]
Accelerated \textit{clipped-SGD}: \textit{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}{2\alpha L}, \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} \underbrace{\nabla f(x^{k+1}, \xi^k)}_{\text{clip}(\nabla f(x^{k+1}, \xi^k), \lambda_{k+1})} \\
y^{k+1} &= \frac{A y^k + \alpha_{k+1} z^{k+1}}{A_{k+1}}
\end{align*}

- Why factor $\alpha$ is needed?
- Why $\lambda_{k+1}$ is chosen this way?
• 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 $\tilde{\nabla}f(x^{k+1}, \xi^k)$.

Parameter $a$ allows to choose smaller stepsizes and, as the result, batchsizes $m_k = \frac{1}{27}$. 
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
    \[ \| \nabla f(x^{k+1}) \| = O\left(\frac{1}{\alpha_{k+1}}\right) \]
  - This hints to choose $\lambda_{k+1} \sim 1/\alpha_{k+1}$ (in the hope that
    \[ \| \nabla f(x^{k+1}) \| = O\left(\frac{1}{\alpha_{k+1}}\right) \] in the stochastic case with high probability).
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
\[ \| \nabla f(x^{k+1}) \| = O\left(\frac{1}{\alpha_{k+1}}\right) \]

This hints to choose $\lambda_{k+1} \sim \frac{1}{\alpha_{k+1}}$ (in the hope that
\[ \| \nabla f(x^{k+1}) \| = O\left(\frac{1}{\alpha_{k+1}}\right) \]
in the stochastic case with high probability).

Parameter $a$ allows to choose smaller stepsizes and, as the result, batchesizes $m_k = 1$. 

*clipped-SSTM*: Intuition Behind the Proof
High-Probability Convergence of $clipped$--$SSTM$

It is sufficient to make all assumptions on a ball around the solution!
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 $B_{3R_0}(x^*)$. 

---

Better result than for clipped-SGD.
High-Probability Convergence of \textit{clipped-SSTM}

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

\begin{table}[h]
\centering
\begin{tabular}{|l|}
\hline
\textbf{Theorem 2} \\
\hline
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 $\ln(\sqrt{L_0}/\sqrt{\varepsilon \beta}) \geq 2$ there exists a choice of $a$ such that \textit{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 \\
\hline
\end{tabular}
\end{table}
High-Probability Convergence of \textit{clipped-SSTM}

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

\textbf{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 $a$ such that \textit{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

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

iterations/oracle calls.
High-Probability Convergence of \textit{clipped-SSTM}

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

\textbf{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 $\beta \in (0, 1)$, $\epsilon \geq 0$ such that $\ln(\sqrt{L}R_0/\sqrt{\epsilon}\beta) \geq 2$ there exists a choice of $a$ such that \textit{clipped-SSTM} with clipping level $\lambda \sim 1/\alpha_{k+1}$ and batchsize $m_k = 1$ finds $y^N$ satisfying $f(y^N) - f(x^*) \leq \epsilon$ with probability at least $1 - \beta$ using

$$O \left( \sqrt{\frac{LR_0^2}{\epsilon}} \ln \frac{LR_0^2}{\epsilon\beta}, \left( \frac{\sigma R}{\epsilon} \right)^{\frac{\alpha}{\alpha - 1}} \ln \left( \frac{1}{\beta} \left( \frac{\sigma R}{\epsilon} \right)^{\frac{\alpha}{\alpha - 1}} \right) \right)$$

iterations/oracle calls.

- Better result than for \textit{clipped-SGD}
In (Gorbunov et al., 2021; Sadiev et al., 2023) we also have

- Results for the non-convex objectives
- 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:

- **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.

---

1The code is available at [https://github.com/ClippedStochasticMethods/clipped-SSTM](https://github.com/ClippedStochasticMethods/clipped-SSTM)
Numerical Experiments: Noise Distribution

Figure 5: 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.
Figure 6: Evolution of the noise distribution for ResNet-18 + ImageNet-100 task.
Figure 7: Evolution of the noise distribution for BERT + CoLA task.
Figure 8: Evolution of the noise distribution for BERT + CoLA task, from iteration 0 (before the training) to iteration 500.
Figure 9: 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.
**Figure 10:** 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.
Adam and clipped-SGD

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

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

- When $\beta_1 = 0$ Adam (RMSprop) can be seen as **clipped-SGD** with “adaptive” $\lambda_k$
Main Results for Variational Inequalities
find $x^* \in Q \subseteq \mathbb{R}^n$ such that \[ \langle F(x^*), x - x^* \rangle \geq 0, \quad \forall x \in Q \quad (\text{VIP-C}) \]
find $x^* \in Q \subseteq \mathbb{R}^n$ such that $\langle F(x^*), x - x^* \rangle \geq 0$, $\forall x \in Q$ (VIP-C)

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

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

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

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

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

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

\[ \langle F(x) - F(y), x - y \rangle \geq 0 \quad (17) \]
Min-max problems:

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

(18)

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

\[
\forall (u, v) \in U \times V \quad \langle \nabla_u f(u^*, v^*), u - u^* \rangle \geq 0, \\
\quad \langle \nabla_v f(u^*, v^*), v - v^* \rangle \geq 0,
\]

which is equivalent to \((\text{VIP-C})\) with \(Q = U \times V\), \(x = (u^\top, v^\top)^\top\), and \(F(x) = \nabla_u f(u, v) - \nabla_v f(u, v)\).

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 \(\forall (u, v) \in U \times V\)

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

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).
• Minimization problems:

\[
\min_{x \in Q} f(x)
\]  \hspace{1cm} (19)
• 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

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

(VIP)

In this talk, we focus on (43) 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)

- **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)
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)

  \[ x^{k+1} = x^k - \gamma F \left( x^k - \gamma F(x^k) \right) \]

  - ✔️ 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)
We consider with

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

- We have access to \( F_\xi \) such that for some \( \alpha \in (1, 2] \) and for all \( x \in \mathbb{R}^n \)

\[ \mathbb{E}_\xi [||F_\xi(x) - F(x)||^\alpha] \leq \sigma^\alpha \]  \hspace{1cm} (20)
We consider with

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

• We have access to \( F_\xi \) such that for some \( \alpha \in (1, 2] \) and for all \( x \in \mathbb{R}^n \)

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

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

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

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

- We have access to \( F_\xi \) such that for some \( \alpha \in (1, 2] \) and for all \( x \in \mathbb{R}^n \)

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

(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 \textit{GDA (SGDA)} and Stochastic \textit{EG (SEG)}

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

- \textit{SEG}:
  \[ x^{k+1} = x^k - \gamma_2 F_{\xi^k_2} \left( x^k - \gamma_1 F_{\xi^k_1}(x^k) \right) \]

\( \xi^k_1, \xi^k_2 \) are i.i.d. samples

\( \gamma_2 \leq \gamma_1 \)
Stochastic **GDA (SGDA)** and Stochastic **EG (SEG)**

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

- **SEG**:
  \[ x^{k+1} = x^k - \gamma_2 F_{\xi_{k_2}}(x^k) - \gamma_1 F_{\xi_{k_1}}(x^k) \]

- \( \xi_{k_1}, \xi_{k_2} \) 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

$$\mathbb{E} \left[ \exp \left( \frac{\|F_{\xi}(x) - F(x)\|^2}{\sigma^2} \right) \right] \leq \exp(1), \quad (21)$$

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

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

---

$^2 \text{Gap}_D(y) = \max_{x: \|x - x^*\| \leq D} \langle F(x), y - x \rangle$
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:**
  \[ x^{k+1} = x^k - \gamma \cdot \text{clip}\left(F_{\xi_k}(x^k), \lambda_k\right) \]

- **SEG:**
  \[ x^{k+1} = x^k - \gamma_2 \cdot \text{clip}\left(F_{\xi_2}(\tilde{x}^k), \lambda_{2,k}\right), \quad \tilde{x}^k = x^k - \gamma_1 \cdot \text{clip}\left(F_{\xi_1}(x^k), \lambda_{1,k}\right) \]

- \( \xi^k_1, \xi^k_2 \) are i.i.d. samples
- \( \gamma_2 \leq \gamma_1 \)

The key idea behind the proof is exactly the same as in minimization! For simplicity, we skip the convergence results in this part.
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

Generators and discriminators show heavy-tailed gradients at various steps.
Clipping Helps for WGAN-GP on CIFAR10

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

The diagram shows the FID values for different optimization methods with and without clipping. The chart indicates a significant improvement in the FID values when clipping is applied.
StyleGAN2 on FFHQ Has Heavy-Tailed Gradients

![Histograms of gradients for Generator and Discriminator](image)

- **Generator**
  - Initialization: $\mu = 3.3$, $\sigma = 0.21$, $\rho_{mR} = 5.486$, $\rho_{eR} = 634.9$
  - Clipped-SGD: $\mu = 7.93$, $\sigma = 1.79$, $\rho_{mR} = 6.934$, $\rho_{eR} = 2963$

- **Discriminator**
  - Initialization: $\mu = 0.17$, $\sigma = 0.03$, $\rho_{mR} = 8.991$, $\rho_{eR} = 2116$
  - Clipped-SGD: $\mu = 5.98$, $\sigma = 2.01$, $\rho_{mR} = 9.829$, $\rho_{eR} = 4444$

(a) Initialization  (b) clipped-SGD
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*!
• 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


