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

The Talk is Based on Three Papers

- Gorbunov, E., Danilova, M., & Gasnikov, A. (2020). *Stochastic optimization with heavy-tailed noise via accelerated gradient clipping*. NeurIPS 2020
- Gorbunov, E., Danilova, M., Shibaev, I., Dvurechensky, P., & Gasnikov, A. (2021). *Near-optimal high probability complexity bounds for non-smooth stochastic optimization with heavy-tailed noise*. arXiv:2106.05958
- Gorbunov, E., Danilova, M., Dobre, D., Dvurechensky, P., Gasnikov, A., & Gidel, G. (2022). *Clipped Stochastic Methods for Variational Inequalities with Heavy-Tailed Noise*. arXiv:2206.01095

Stochastic Gradient Descent (SGD)

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

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

Clipped Stochastic Gradient Descent (clipped-SGD)

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

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

Origin of Clipping

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

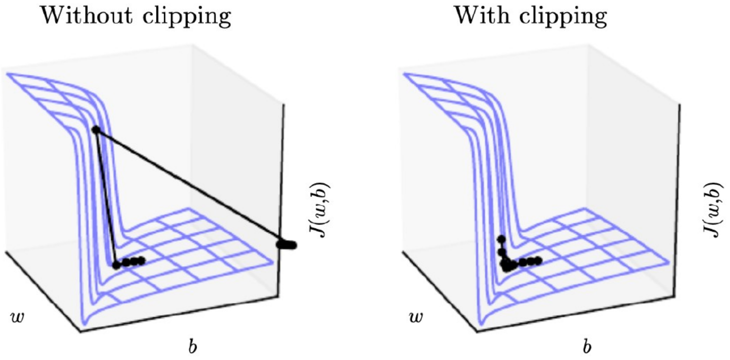


Figure: from [Goodfellow et al., 2016]

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?

Heavy-Tailed Noise in Stochastic Gradients

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

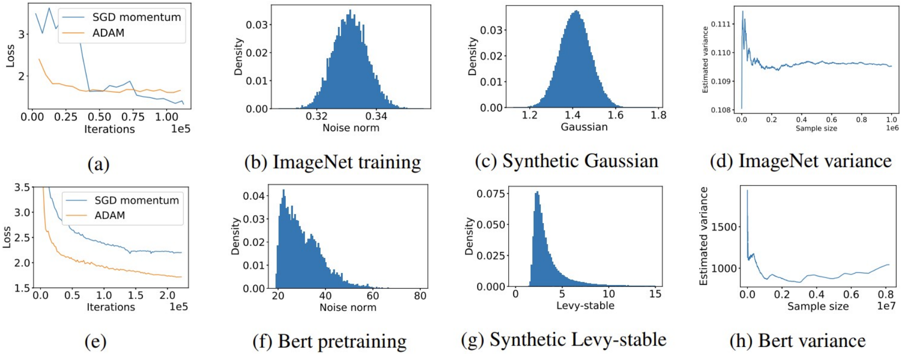


Figure: from [Zhang et al., 2020]

Definition of Heavy-Tailed Noise in Stochastic Gradients

- Random vector X has light tails if

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

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

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

Definition of Heavy-Tailed Noise in Stochastic Gradients

- Random vector X has light tails if

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

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

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

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

$$\mathbb{E}\left[\|X - \mathbb{E}[X]\|^2\right] \leq \sigma^2 \tag{5}$$

Problem and Assumptions

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

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

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

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

Problem and Assumptions

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

- $f : \mathbb{R}^n \to \mathbb{R}^n$ 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., $\forall x \in \mathbb{R}^n$

$$\mathbb{E}_\xi \left[\nabla f(x, \xi) \right] = \nabla f(x), \quad \mathbb{E}_\xi \left[\|\nabla f(x, \xi) - \nabla f(x)\|^2 \right] \leq \sigma^2. \tag{9}$$

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:

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

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

In-Expectation Guarantees 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:

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

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

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

In-Expectation Guarantees are Less Sensitive to Distribution

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

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

In-Expectation Guarantees are Less Sensitive to Distribution

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

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

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

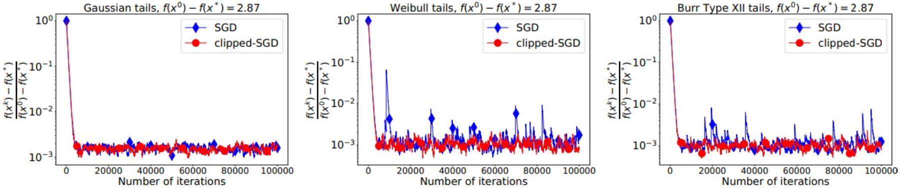


Figure: from [Gorbunov et al., 2020]

High-Probability Results under Light-Tails Assumption

Light-tails assumption (classical one):

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

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 SGD finds $\hat{x}$ such that $f(\hat{x}) - f(x^*) \leq \varepsilon$ with probability at least $1 - \beta$ using

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

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

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

- 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

$$\mathcal{O}\left(\max\left\{\sqrt{\frac{LR_0^2}{\varepsilon}}, \frac{\sigma^2 R_0^2}{\varepsilon^2}\ln^2\left(\frac{1}{\beta}\right)\right\}\right) \quad \text{oracle calls}$$

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

High-Probability Results under Heavy-Tails Assumption

- 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`

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

- $\nabla f(x^k, \boldsymbol{\xi}^k) = \frac{1}{m_k}\sum\limits_{i=1}^{m_k} \nabla f(x^k, \xi_i^k)$, where $\xi_1^k, \ldots, \xi_{m_k}^k$ are i.i.d. samples

Key Challenge in the Analysis of clipped-SGD

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

- $\nabla f(x^k, \boldsymbol{\xi}^k) = \frac{1}{m_k} \sum\limits_{i=1}^{m_k} \nabla f(x^k, \xi_i^k)$, where $\xi_1^k, \ldots, \xi_{m_k}^k$ are i.i.d. samples
- Key challenge: $\mathbb{E}\left[\widetilde{\nabla} f(x^k, \boldsymbol{\xi}^k) \mid x^k\right] \neq \nabla f(x^k)$

Analysis of `clipped-SGD`: Key Idea

- We start the proof classically:

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

Analysis of `clipped-SGD`: Key Idea

- We start the proof classically:

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

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

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

How to upper bound the sums in red?

Bernstein Inequality for Martingale Differences

Lemma 1 [Bennett, 1962, Dzhaparidze and Van Zanten, 2001, Freedman et al., 1975]

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

Bernstein Inequality for Martingale Differences

Lemma 1 [Bennett, 1962, Dzhaparidze and Van Zanten, 2001, Freedman et al., 1975]

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

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

Bernstein Inequality for Martingale Differences

Lemma 1 [Bennett, 1962, Dzhaparidze and Van Zanten, 2001, Freedman et al., 1975]

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

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

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

- upper bound bias, variance, and distortion of θ_k
- have upper bounds for $\|x^k - x^*\|$ and $\|\theta_k\|$ that hold with large probability

Magnitude, Bias, Variance, Distortion

Lemma 2

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

$$\left\|\widetilde{X} - \mathbb{E}[\widetilde{X}]\right\| \leq 2\lambda. \tag{12}$$

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

$$\left\|\mathbb{E}[\widetilde{X}] - x\right\| \leq \frac{4\sigma^2}{\lambda}, \tag{13}$$

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

$$\mathbb{E}\left[\left\|\widetilde{X} - \mathbb{E}[\widetilde{X}]\right\|^2\right] \leq 18\sigma^2. \tag{15}$$

Bound on the Distance to the Solution

Inequality

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

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 `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 $\ln(LR_0^2/\varepsilon\beta) \geq 2$ there exists a choice of γ such that `clipped-SGD` with clipping level $\lambda \sim 1/\gamma$ and batchsize $m_k = 1$ finds $\bar{x}^N$ satisfying $f(\bar{x}^N) - f(x^*) \leq \varepsilon$ with probability at least $1 - \beta$ using

High-Probability Convergence of `clipped-SGD`

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 γ such that `clipped-SGD` with clipping level $\lambda \sim {1}/{\gamma}$ and batchsize $m_k = 1$ finds $\bar{x}^N$ satisfying $f(\bar{x}^N) - f(x^*) \leq \varepsilon$ with probability at least $1 - \beta$ using

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

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 γ such that `clipped-SGD` with clipping level $\lambda \sim 1/\gamma$ and batchsize $m_k = 1$ finds $\bar{x}^N$ satisfying $f(\bar{x}^N) - f(x^*) \leq \varepsilon$ with probability at least $1 - \beta$ using

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

- Same result (up to the difference in logarithmic factors) as for `SGD` in the light-tailed case

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 γ such that `clipped-SGD` with clipping level $\lambda \sim 1/\gamma$ and batchsize $m_k = 1$ finds $\bar{x}^N$ satisfying $f(\bar{x}^N) - f(x^*) \leq \varepsilon$ with probability at least $1 - \beta$ using

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

- Same result (up to the difference in logarithmic factors) as for `SGD` in the light-tailed case
- Same result (up to the difference in logarithmic factors) as for `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:

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

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

$$z^{k+1} = z^k - \alpha_{k+1} \underbrace{\widetilde{\nabla} f(x^{k+1}, \xi^k)}_{\texttt{clip}(\nabla f(x^{k+1}, \xi^k), \lambda_{k+1})}$$

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

Accelerated `clipped-SGD`: `clipped-SSTM`

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

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

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

$$z^{k+1} = z^k - \alpha_{k+1} \underbrace{\widetilde{\nabla} f(x^{k+1}, \xi^k)}_{\texttt{clip}(\nabla f(x^{k+1}, \xi^k), \lambda_{k+1})}$$

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

- Why factor a is needed?
- Why λ_{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 $\widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k)$

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 $\widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k)$
  - For deterministic SSTM (i.e., STM) one can prove $\|\nabla f(x^{k+1})\| = \mathcal{O}(1/\alpha_{k+1})$
  - This hints to choose $\lambda_{k+1} \sim 1/\alpha_{k+1}$ (in the hope that $\|\nabla f(x^{k+1})\| = \mathcal{O}(1/\alpha_{k+1})$ 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 $\widetilde{\nabla} f(x^{k+1}, \xi^k)$
  - For deterministic SSTM (i.e., STM) one can prove $\|\nabla f(x^{k+1})\| = \mathcal{O}(1/\alpha_{k+1})$
  - This hints to choose $\lambda_{k+1} \sim 1/\alpha_{k+1}$ (in the hope that $\|\nabla f(x^{k+1})\| = \mathcal{O}(1/\alpha_{k+1})$ 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 $\ln(\sqrt{LR_0}/\sqrt{\varepsilon}\beta) \geq 2$ 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 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

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

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 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

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

- Same result (up to the difference in logarithmic factors) as for `AC-SA` in the light-tailed case

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 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

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

- 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[1]:

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

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

Numerical Experiments: Noise Distribution

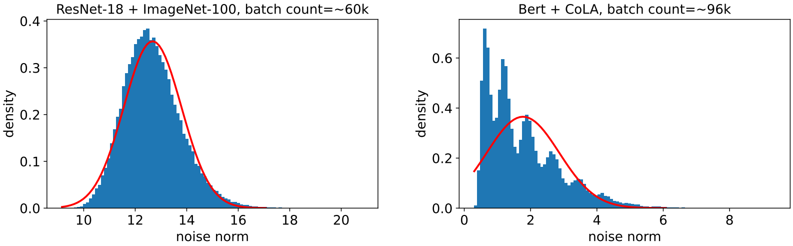


Figure: Noise distribution of the stochastic gradients for `ResNet-18` on `ImageNet-100` and `BERT` fine-tuning on the `CoLA` dataset before the training. Red lines: probability density functions of normal distributions with means and variances empirically estimated by the samples. Batch count is the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

Evolution of the Noise Distribution, Image Classification

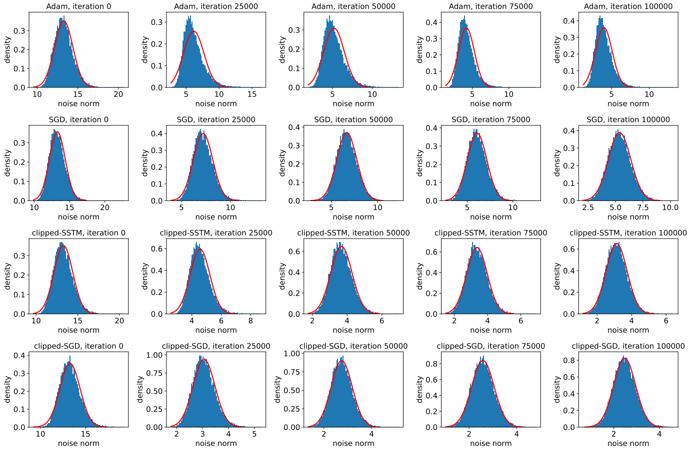


Figure: Evolution of the noise distribution for `ResNet-18` + `ImageNet-100` task.

Evolution of the Noise Distribution, Text Classification

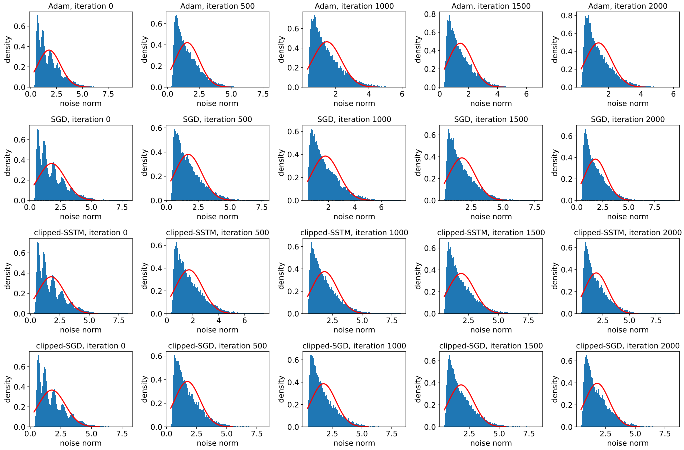


Figure: Evolution of the noise distribution for BERT + CoLA task.

Evolution of the Noise Distribution, Text Classification

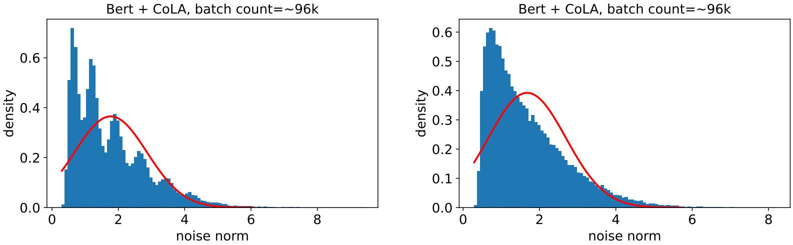


Figure: Evolution of the noise distribution for BERT + CoLA task, from iteration 0 (before the training) to iteration 500.

Numerical Results, Image Classification

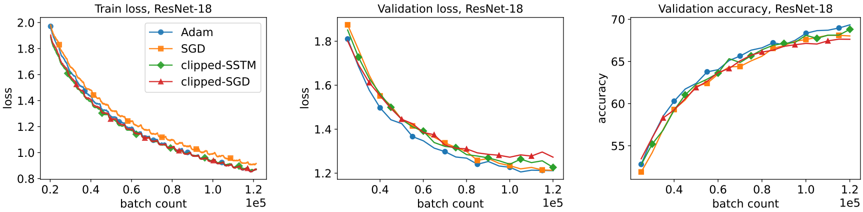


Figure: Train and validation loss + accuracy for different optimizers on `ResNet-18` + `ImageNet-100` problem. Here, "batch count" denotes the total number of used stochastic gradients. The noise distribution is almost Gaussian even vanilla `SGD` performs well, i.e., gradient clipping is not required.

Numerical Results, Text Classification

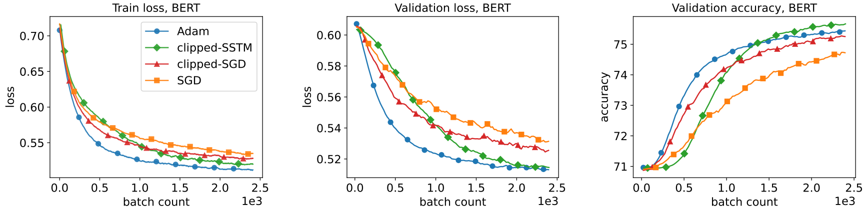


Figure: Train and validation loss + accuracy for different optimizers on BERT + CoLA problem. The noise distribution is heavy-tailed, the methods with clipping outperform SGD by a large margin.

Variational Inequality Problem

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

Variational Inequality Problem

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

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

$$\|F(x) - F(y)\| \leq L\|x - y\| \tag{16}$$

Variational Inequality Problem

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

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

$$\|F(x) - F(y)\| \leq L\|x - y\| \tag{16}$$

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

$$\langle F(x) - F(y), x - y \rangle \geq 0 \tag{17}$$

Variational Inequality Problem: Examples

- Min-max problems:

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

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

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

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 $Q = U \times V$, $x = (u^\top, v^\top)^\top$, and

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

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 $Q = U \times V$, $x = (u^\top, v^\top)^\top$, 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

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

In this talk, we focus on (40) rather than (VIP-C)

Gradient Descent-Ascent (GDA) and Extragradient (EG)

- GDA [Krasnosel'skiı, 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'skiı, 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)

Stochastic VIP

We consider with

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

- We have access to F_ξ such that for all $x \in \mathbb{R}^n$

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

Stochastic VIP

We consider with

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

- We have access to F_ξ such that for all $x \in \mathbb{R}^n$

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

- For `GDA`-based methods we assume ℓ-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_ξ such that for all $x \in \mathbb{R}^n$

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

- For `GDA`-based methods we assume ℓ-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:

$$x^{k+1} = x^k - \gamma_2 F_{\xi_2^k}\left(x^k - \gamma_1 F_{\xi_1^k}(x^k)\right)$$

Stochastic GDA (SGDA) and Stochastic EG (SEG)

- SGDA:

$$x^{k+1} = x^k - \gamma F_{\xi^k}(x^k)$$

- SEG:

$$x^{k+1} = x^k - \gamma_2 F_{\xi_2^k}\left(x^k - \gamma_1 F_{\xi_1^k}(x^k)\right)$$

  - ξ_1^k, ξ_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

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

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

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

[2] $\texttt{Gap}_D(y) = \max_{x:\|x-x^*\|\leq D}\langle F(x), y - x\rangle$

clipped-SGDA and clipped-SEG

- SGDA:

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

- SEG:

$$x^{k+1} = x^k - \gamma_2 \cdot \texttt{clip}\left(F_{\xi_2^k}(\widetilde{x}^k), \lambda_{2,k}\right), \quad \widetilde{x}^k = x^k - \gamma_1 \cdot \texttt{clip}\left(F_{\xi_1^k}(x^k), \lambda_{1,k}\right)$$

  - ξ_1^k, ξ_2^k are i.i.d. samples
  - $\gamma_2 \leq \gamma_1$

clipped-SGDA and clipped-SEG

- SGDA:

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

- SEG:

$$x^{k+1} = x^k - \gamma_2 \cdot \texttt{clip}\left(F_{\xi_2^k}(\widetilde{x}^k), \lambda_{2,k}\right), \quad \widetilde{x}^k = x^k - \gamma_1 \cdot \texttt{clip}\left(F_{\xi_1^k}(x^k), \lambda_{1,k}\right)$$

  - ξ_1^k, ξ_2^k 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 $B_{4R}(x^*)$, $R \geq R_0$.

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 $\ln(6LR_0^2/\varepsilon\beta) \geq 1$ there exists a choice of $\gamma_1 = \gamma_2 = \gamma$ such that `clipped-SEG` with clipping level $\lambda \sim 1/\gamma$ finds $\hat{x}$ satisfying $\mathtt{Gap}_R(\hat{x}) \leq \varepsilon$ 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 $\ln(6LR_0^2/\varepsilon\beta) \geq 1$ there exists a choice of $\gamma_1 = \gamma_2 = \gamma$ such that `clipped-SEG` with clipping level $\lambda \sim 1/\gamma$ finds $\hat{x}$ satisfying $\texttt{Gap}_R(\hat{x}) \leq \varepsilon$ with probability at least $1 - \beta$ using

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

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 $\beta \in (0,1)$, $\varepsilon \geq 0$ such that $\ln({}^{6LR_0^2}/_{\varepsilon\beta}) \geq 1$ there exists a choice of $\gamma_1 = \gamma_2 = \gamma$ such that clipped-SEG with clipping level $\lambda \sim {}^1/_\gamma$ finds $\hat{x}$ satisfying $\mathtt{Gap}_R(\hat{x}) \leq \varepsilon$ with probability at least $1 - \beta$ using

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

- 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 ℓ-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 ℓ-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 γ such that `clipped-SGDA` with clipping level $\lambda \sim 1/\gamma$ finds $\hat{x}$ satisfying

$\frac{1}{K+1}\sum\limits_{k=0}^{K}\|F(x^k)\|^2 \leq \varepsilon$ 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 ℓ-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 γ such that `clipped-SGDA` with clipping level $\lambda \sim {}^1/_\gamma$ finds $\hat{x}$ satisfying $\frac{1}{K+1}\sum\limits_{k=0}^{K}\|F(x^k)\|^2 \leq \varepsilon$ with probability at least $1-\beta$ using

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

High-Probability Convergence of `clipped-SGDA`

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

Theorem 4

Let F be ℓ-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 γ such that `clipped-SGDA` with clipping level $\lambda \sim 1/\gamma$ finds $\hat{x}$ satisfying $\frac{1}{K+1}\sum\limits_{k=0}^{K}\|F(x^k)\|^2 \leq \varepsilon$ with probability at least $1-\beta$ using

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

- 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`

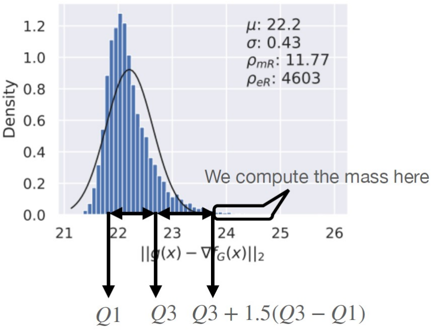
Numerical Experiments

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

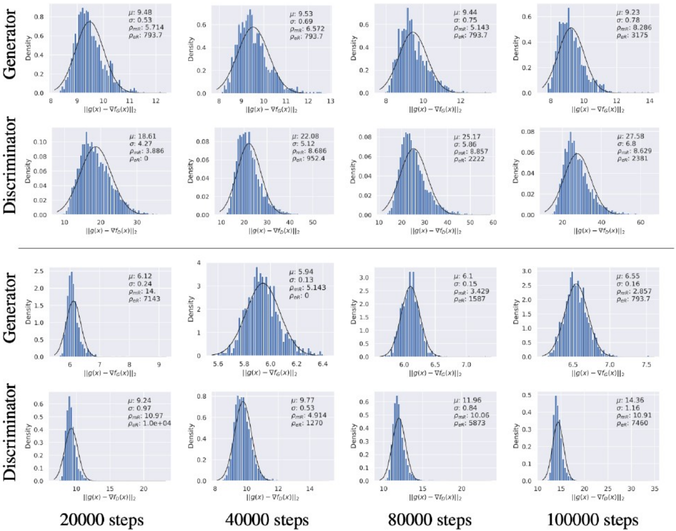
- `clipped-SGDA` with alternating updates
- `Coord-clipped-SGDA` – `clipped-SGDA` with coordinate-wise clipping and alternating updates
- `clipped-SEG`
- `Coord-clipped-SEG`

WGAN-GP on CIFAR10 Has Heavy-Tailed Gradients

- ρ_{mR}: relative fraction of mass after $Q_3 + 1.5 \cdot (Q_3 - Q_1)$
  - For normal distribution there is $\approx .35\%$ of the mass
  - In this plot: ≈ 12 times more
- ρ_{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: ≈ 4603 times more

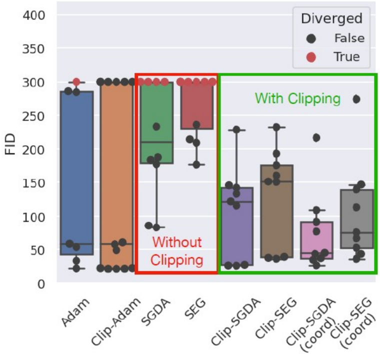


WGAN-GP on CIFAR10 Has Heavy-Tailed Gradients

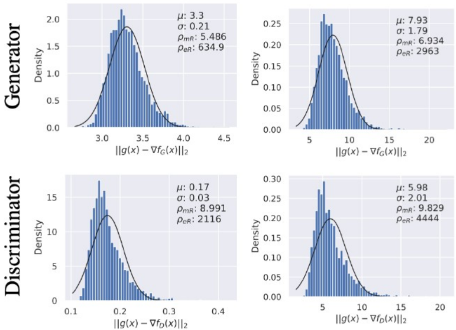


Clipping Helps for WGAN-GP on CIFAR10

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

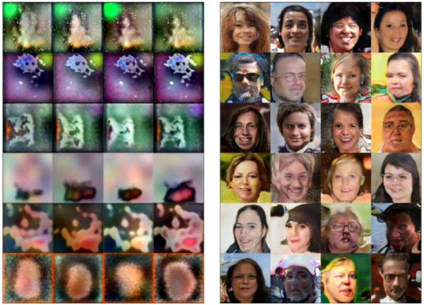


StyleGAN2 on FFHQ Has Heavy-Tailed Gradients



(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 I

A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust optimization*. Princeton university press, 2009.

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

D. Davis, D. Drusvyatskiy, L. Xiao, and J. Zhang. From low probability to high confidence in stochastic convex optimization. *Journal of machine learning research*, 22(49), 2021.

O. Devolder et al. Stochastic first order methods in smooth convex optimization. Technical report, CORE, 2011.

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

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

A. Gasnikov and Y. Nesterov. Universal fast gradient method for stochastic composit optimization problems. *arXiv preprint arXiv:1604.05275*, 2016.

References II

S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.

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

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

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

I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *ICLR 2015*, 2015.

References III

E. Gorbunov, M. Danilova, and A. Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. *Advances in Neural Information Processing Systems*, 33:15042–15053, 2020. URL https://proceedings.neurips.cc/paper/2020/file/abd1c782880cc59759f4112fda0b8f98-Supplemental.pdf.

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

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

References IV

M. Hast, K. J. Åström, B. Bernhardsson, and S. Boyd. Pid design by convex-concave optimization. In *2013 European Control Conference (ECC)*, pages 4460–4465. IEEE, 2013.

A. Juditsky, A. Nemirovski, and C. Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

M. Krasnosel'skiı. Two remarks on the method of successive approximations, uspehi mat. *Nauk*, 10:123–127, 1955.

A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *ICLR 2018*, 2018.

W. R. Mann. Mean value methods in iteration. *Proceedings of the American Mathematical Society*, 4(3):506–510, 1953.

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

References V

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

A. V. Nazin, A. S. Nemirovsky, A. B. Tsybakov, and A. B. Juditsky. Algorithms of robust stochastic optimization based on mirror descent method. *Automation and Remote Control*, 80(9):1607–1627, 2019.

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

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

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