Stochastic Optimization with Heavy-Tailed Noise via Accelerated Gradient Clipping

Eduard Gorbunov

MIPT and HSE

Marina Danilova

MIPT and ICS RAS

Alexander Gasnilov

MIPT and HSE

1. The Problem

Smooth convex

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2$$
$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$$

Expectation

$$f(x) = \mathbf{E}_{\xi \sim \mathcal{D}}\left[f_\xi(x)\right]$$

$$\min_{x \in \mathbb{R}^n} f(x)$$

The most popular method? ⟶ SGD

2. Motivational Example

Stochastic gradient descent (SGD)

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

Stochastic gradient descent (SGD)

Iteration counter

Stepsize

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

Stochastic gradient

Stochastic gradient descent (SGD)

Iteration counter

Stepsize

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

Stochastic gradient

- $\mathbb{E}_\xi [\nabla f(x,\xi)] = \nabla f(x)$
- $\mathbb{E}_\xi \left[\|\nabla f(x,\xi) - \nabla f(x)\|_2^2 \right] \le \sigma^2$

Stochastic optimization problem

$$\min_{x \in \mathbb{R}^n} f(x)$$

Stochastic optimization problem

$$\min_{x \in \mathbb{R}^n} f(x)$$

Stochastic optimization problem

$$\min_{x \in \mathbb{R}^n} f(x)$$

$$f(x) = \frac{1}{2}\|x\|_2^2 = \mathbb{E}_\xi\left[f(x,\xi)\right]$$

$$f(x,\xi) = \frac{1}{2}\|x\|_2^2 + \langle \xi, x \rangle$$

Stochastic optimization problem

$$\min_{x \in \mathbb{R}^n} f(x)$$

$$f(x) = \frac{1}{2}\|x\|_2^2 = \mathbb{E}_\xi \left[f(x, \xi) \right]$$

$$f(x, \xi) = \frac{1}{2}\|x\|_2^2 + \langle \xi, x \rangle$$

- ξ — random vector with zero mean and bounded variance
- $f(x)$ — 1-strongly convex, L-smooth function
- $\nabla f(x, \xi) = x + \xi$ — stochastic gradient

Stochastic optimization problem

$$\min_{x \in \mathbb{R}^n} f(x)$$

$$f(x) = \frac{1}{2}\|x\|_2^2 = \mathbb{E}_\xi \left[f(x, \xi) \right]$$

$$f(x, \xi) = \frac{1}{2}\|x\|_2^2 + \langle \xi, x \rangle$$

- ξ — random vector with zero mean and bounded variance
- $f(x)$ — 1-strongly convex, L-smooth function
- $\nabla f(x, \xi) = x + \xi$ — stochastic gradient

Stochastic optimization problem

$$\min_{x \in \mathbb{R}^n} f(x)$$

$$f(x) = \frac{1}{2}\|x\|_2^2 = \mathbb{E}_\xi \left[f(x, \xi) \right]$$

$$f(x, \xi) = \frac{1}{2}\|x\|_2^2 + \langle \xi, x \rangle$$

Convergence in Expectation

$$\mathbb{E}\left[\|x^k - x^*\|_2^2 \right] \leq (1 - \gamma\mu)^k \|x^0 - x^*\|_2^2 + \frac{\gamma\sigma^2}{\mu}$$

Stochastic optimization problem

$$\min_{x \in \mathbb{R}^n} f(x)$$

$$f(x) = \frac{1}{2}\|x\|_2^2 = \mathbb{E}_\xi\left[f(x,\xi)\right]$$

$$f(x,\xi) = \frac{1}{2}\|x\|_2^2 + \langle \xi, x \rangle$$

Convergence in Expectation

$$\mathbb{E}\left[\|x^k - x^*\|_2^2\right] \leq (1-\gamma\mu)^k \|x^0 - x^*\|_2^2 + \frac{\gamma\sigma^2}{\mu}$$

After k iterations of SGD (State-of the art theory)

Stochastic optimization problem

$$\min_{x \in \mathbb{R}^n} f(x)$$

$$f(x) = \frac{1}{2}\|x\|_2^2 = \mathbb{E}_\xi \left[f(x,\xi) \right]$$

$$f(x,\xi) = \frac{1}{2}\|x\|_2^2 + \langle \xi, x \rangle$$

Convergence in Expectation

$$\mathbb{E}\left[\|x^k - x^*\|_2^2\right] \le (1-\gamma\mu)^k \|x^0 - x^*\|_2^2 + \frac{\gamma\sigma^2}{\mu}$$

$$f(x) = \frac{1}{2}\|x\|_2^2, \; f(x^*) = 0 \qquad \mu = 1 \qquad x^* = 0$$

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

Stochastic optimization problem

$$\min_{x \in \mathbb{R}^n} f(x)$$

$$f(x) = \frac{1}{2}\|x\|_2^2 = \mathbb{E}_\xi\left[f(x,\xi)\right]$$

$$f(x,\xi) = \frac{1}{2}\|x\|_2^2 + \langle \xi, x \rangle$$

Convergence in Expectation

$$\mathbb{E}\left[\|x^k - x^*\|_2^2\right] \le (1-\gamma\mu)^k \|x^0 - x^*\|_2^2 + \frac{\gamma\sigma^2}{\mu}$$

$$f(x) = \frac{1}{2}\|x\|_2^2,\ f(x^*) = 0 \qquad \mu = 1 \qquad x^* = 0$$

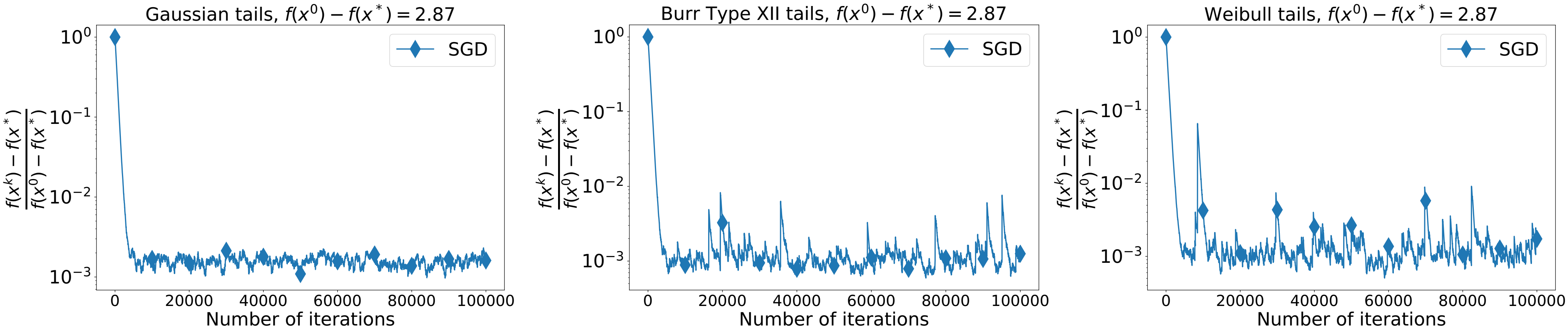
Our case

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

Convergence in Expectation

$$f(x,\xi) = \frac{1}{2}\|x\|_2^2 + \langle \xi, x \rangle$$

Gaussian | Burr Type XII | Weibull



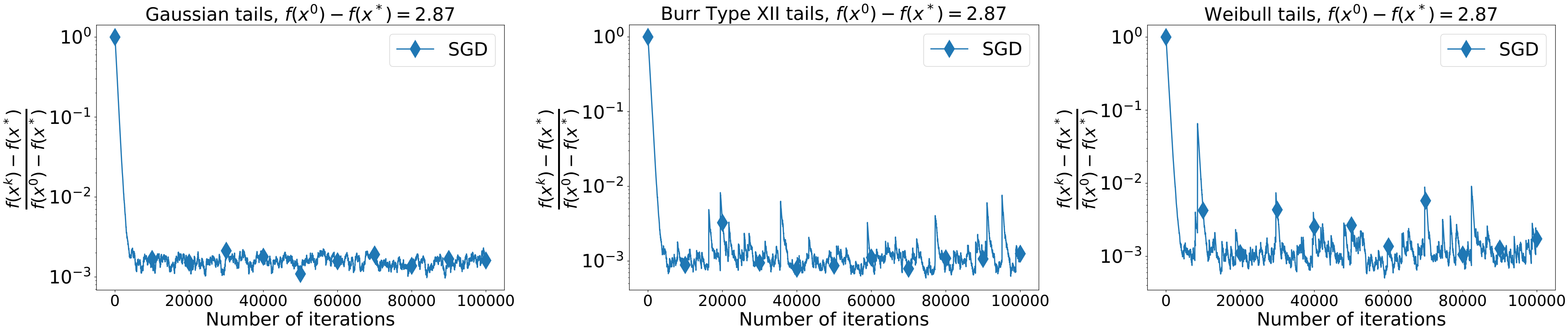
- 3 different distributions of ξ with the same σ

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

Problems

- **Heavy-Tailed Noise**

Gaussian | Burr Type XII | Weibull

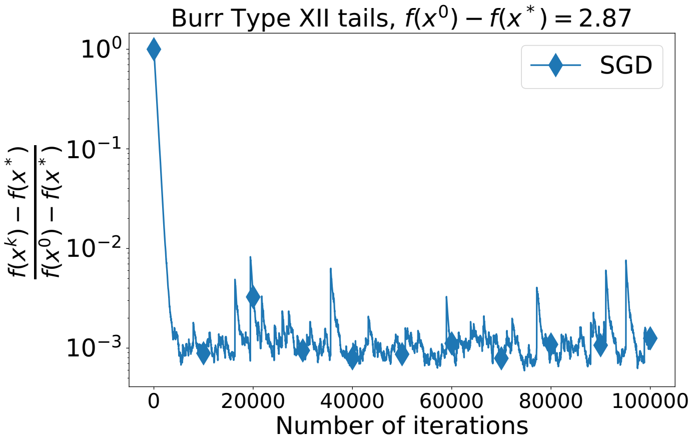


- How to obtain good accuracy of the solution with **high probability**?

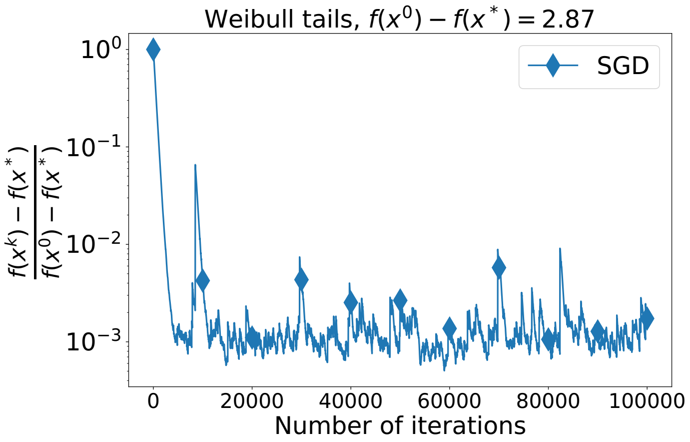
Problems

Heavy-Tailed Noise

Burr Type XII



Weibull

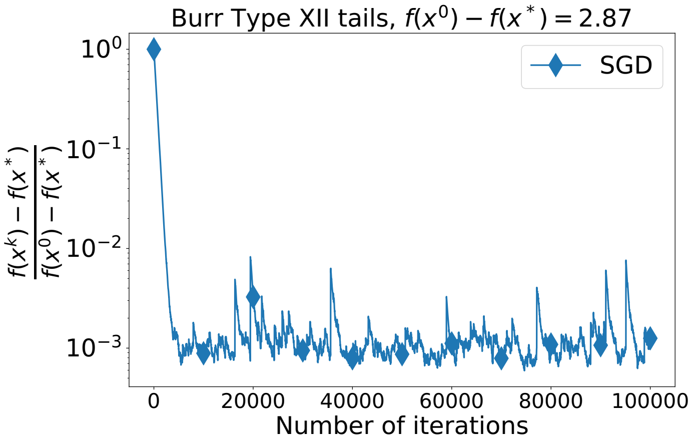


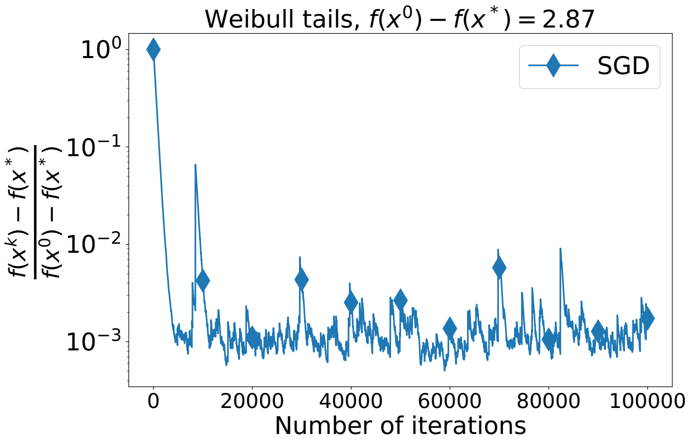
Problems

Heavy-Tailed Noise

Burr Type XII

Weibull





SGD

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

big even if we are close to the solution

Key idea

SGD

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

Key idea

SGD

Stochastic gradient

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

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

Clipped stochastic gradient $\widetilde{\nabla} f(x^k, \xi^k)$

Key idea

SGD

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

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

$$\text{clip}(\nabla f(x,\xi), \lambda) = \begin{cases} \nabla f(x,\xi), & \text{if } \|\nabla f(x,\xi)\|_2 \leq \lambda, \\ \frac{\lambda}{\|\nabla f(x,\xi)\|_2} \nabla f(x,\xi), & \text{otherwise} \end{cases}$$

Key idea

SGD

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

clipped-SGD

Clipping level

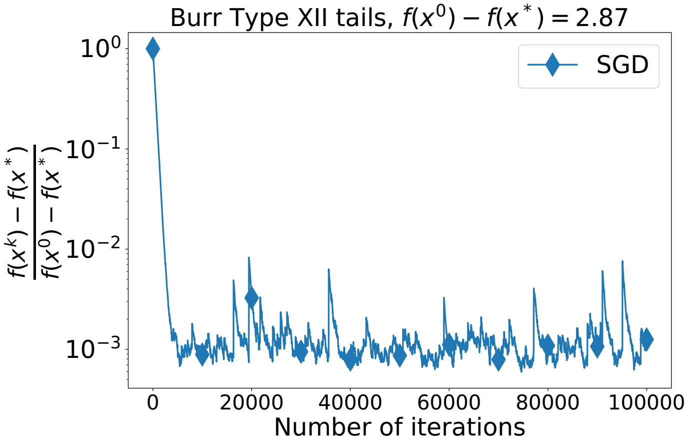
$$x^{k+1} = x^k - \gamma \text{clip}(\nabla f(x^k, \xi^k), \lambda)$$

$$\text{clip}(\nabla f(x,\xi), \lambda) = \begin{cases} \nabla f(x,\xi), & \text{if } \|\nabla f(x,\xi)\|_2 \le \lambda, \\ \frac{\lambda}{\|\nabla f(x,\xi)\|_2} \nabla f(x,\xi), & \text{otherwise} \end{cases}$$

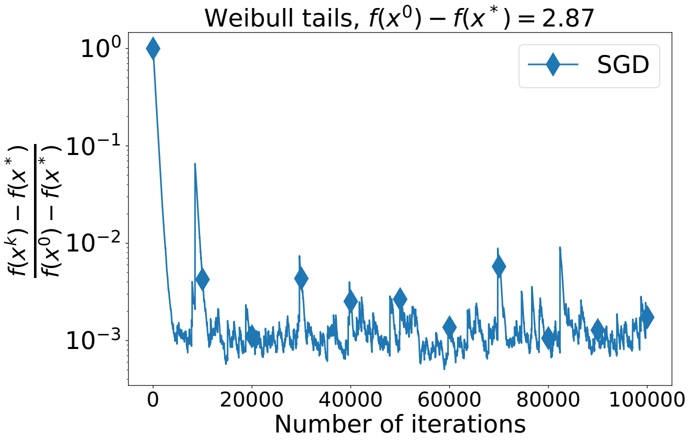
Key idea

SGD

Burr Type XII



Weibull



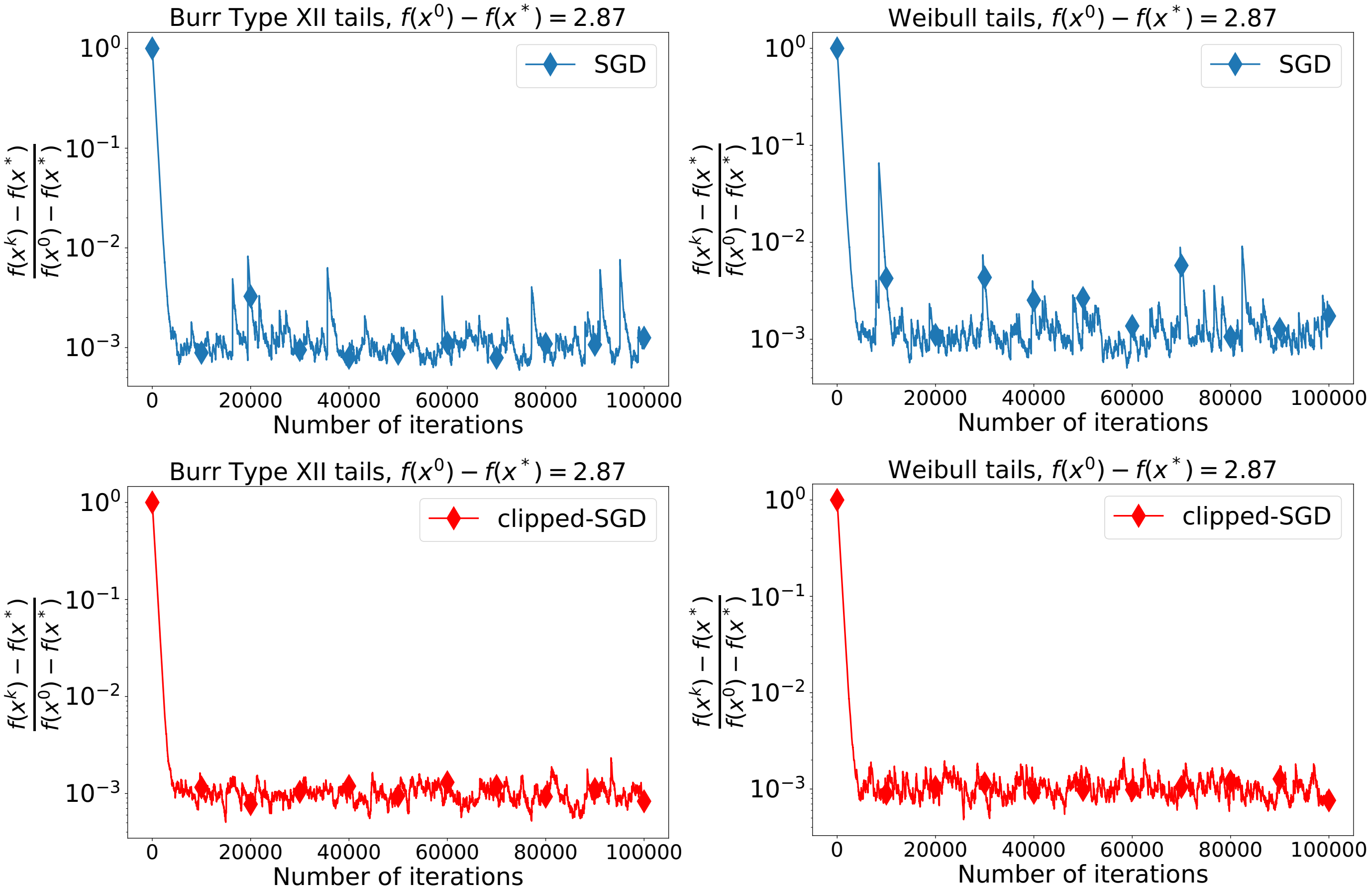
Key idea

SGD

clipped-SGD

Burr Type XII

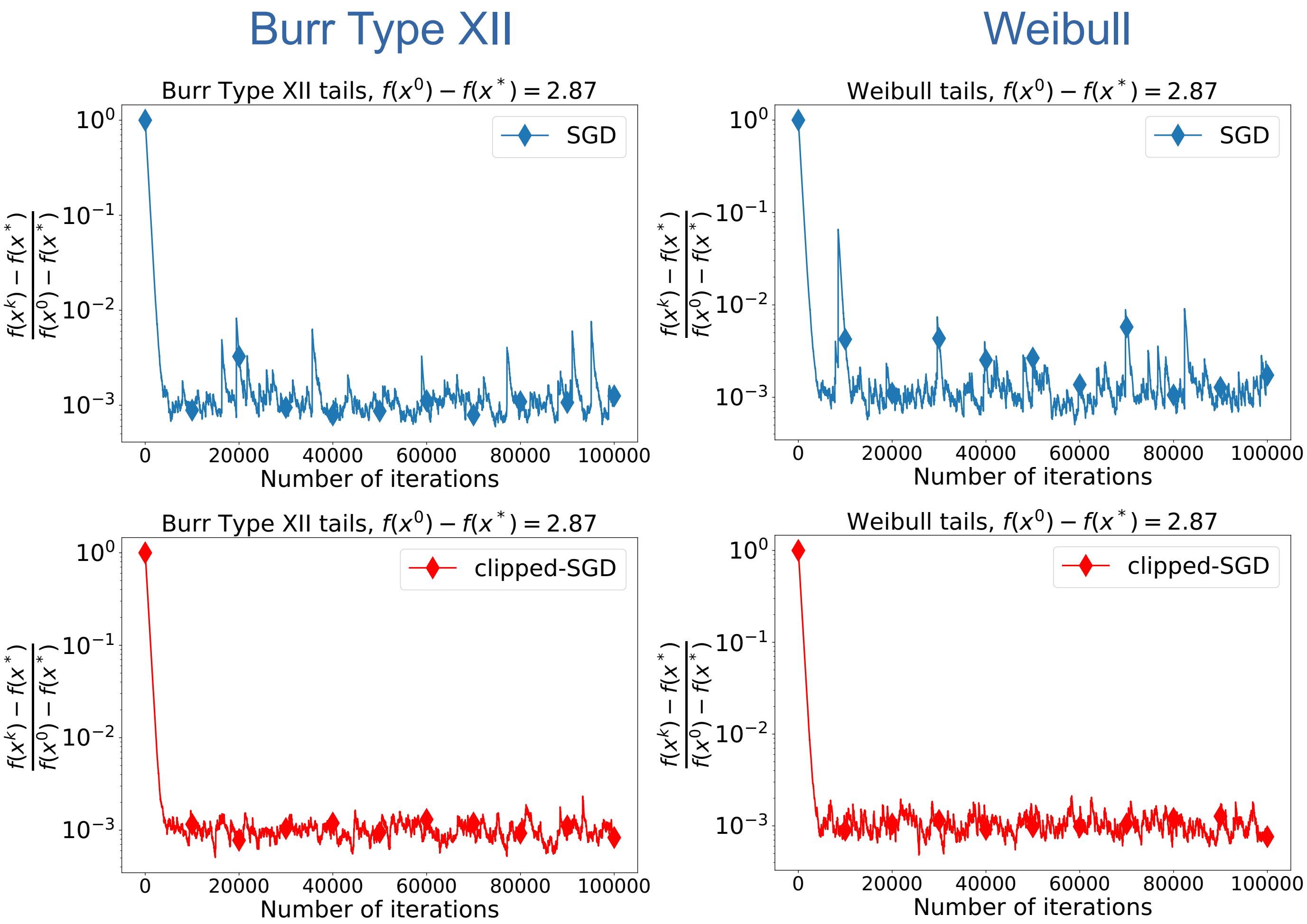
Weibull



Key idea

SGD

clipped-SGD

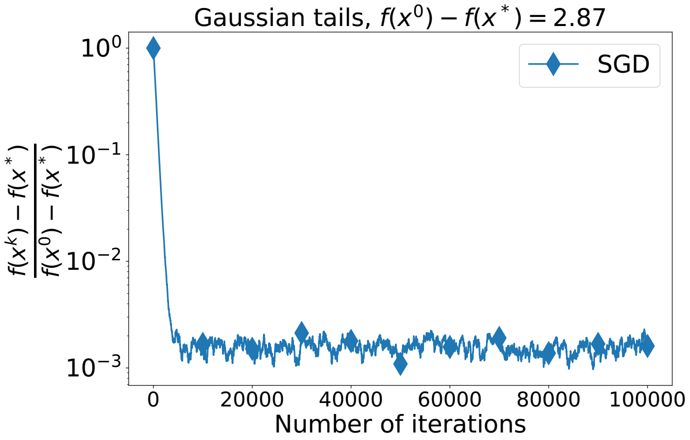


Small oscillations even for heavy-tailed distributions!

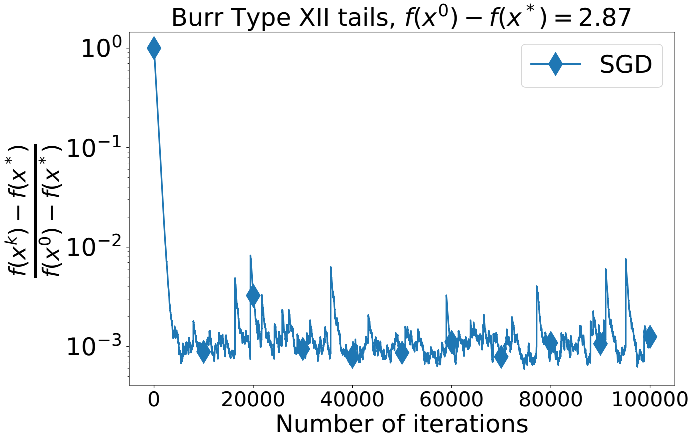
Problems

~~**Heavy-Tailed Noise**~~

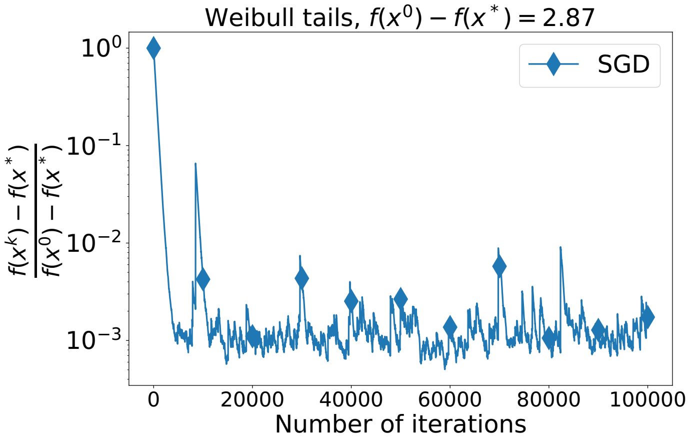
Gaussian



Burr Type XII



Weibull



- How to obtain good accuracy of the solution with **high probability**?

Problems

- How to obtain good accuracy of the solution with **high probability**?

Convergence in

Expectation

$$\mathbb{E}\left[f(x^N) - f(x^*)\right] \leq \varepsilon$$

How to obtain good accuracy of the solution with **high probability**?

Convergence in

Expectation

$$\mathbb{E}\left[f(x^N) - f(x^*)\right] \le \varepsilon$$

How to obtain good accuracy of the solution with **high probability**?

Convergence in

Expectation

$$\mathbb{E}\left[f(x^N) - f(x^*)\right] \leq \varepsilon$$

⇒

Probability

$$f(x^N) - f(x^*) \leq \varepsilon$$
$$\text{with probability} \geq 1 - \beta$$

- How to obtain good accuracy of the solution with **high probability**?

Convergence in

Expectation ⇒ **Probability**

$$\mathbb{E}\left[f(x^N) - f(x^*)\right] \leq \varepsilon$$

Desired accuracy

$$f(x^N) - f(x^*) \leq \varepsilon$$
$$\text{with probability} \geq 1 - \beta$$

Confidence level

- How to obtain good accuracy of the solution with **high probability**?

Convergence in

We focus on this situation!

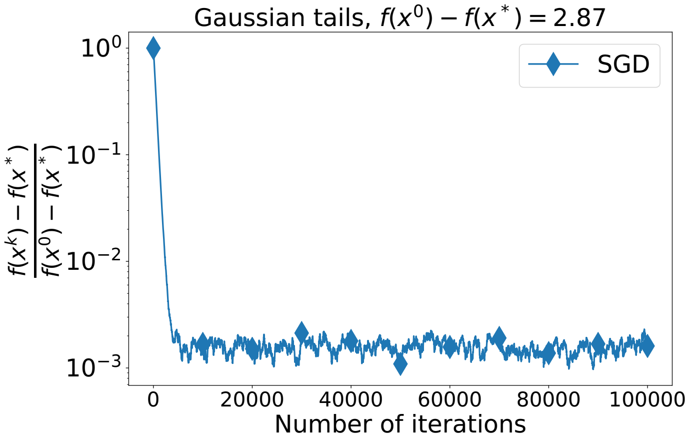
Expectation	Probability
$\mathbb{E}\left[f(x^N) - f(x^*)\right] \leq \varepsilon$	$f(x^N) - f(x^*) \leq \varepsilon$ with probability $\geq 1 - \beta$
Desired accuracy	Confidence level

- How to obtain good accuracy of the solution with **high probability**?

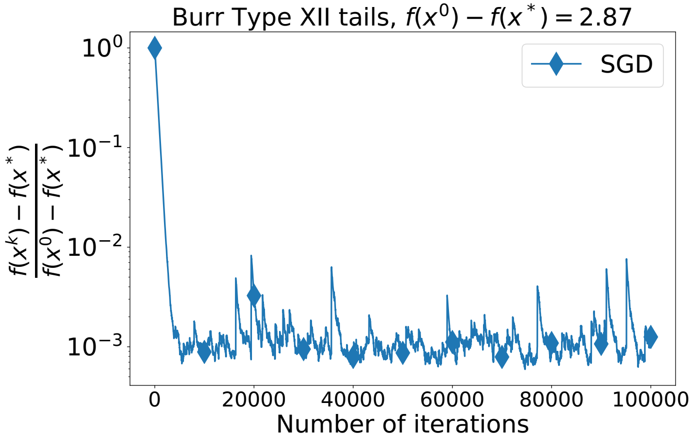
Problems

~~Heavy-Tailed Noise~~

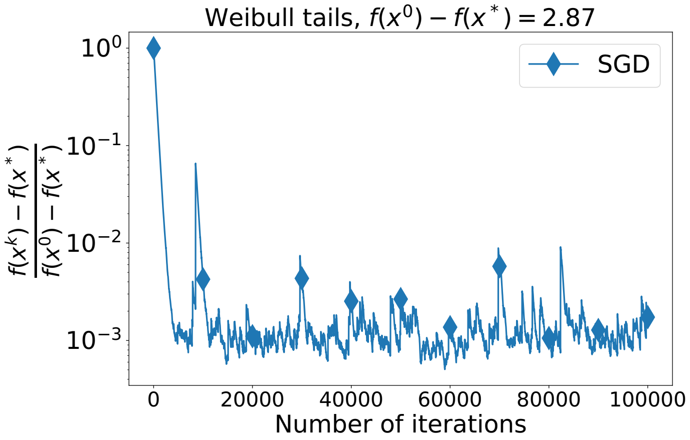
Gaussian



Burr Type XII



Weibull



~~How to obtain good accuracy of the solution with **high probability**?~~

3. Key Assumptions

Assumptions

$$\min_{x \in \mathbb{R}^n} f(x)$$

- $f(x) = \mathbf{E}_{\xi \sim \mathcal{D}} \left[f_\xi(x) \right]$ — expectation minimization
- $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$ — convexity
- $\|\nabla f(x) - \nabla f(y)\|_2 \leq L \|x - y\|_2$ — L-smoothness
- $\mathbb{E}_\xi [\nabla f(x, \xi)] = \nabla f(x)$ — unbiasedness
- $\mathbb{E}_\xi \left[\|\nabla f(x, \xi) - \nabla f(x)\|_2^2 \right] \leq \sigma^2$ — boundedness of the variance

Assumptions

$$\min_{x \in \mathbb{R}^n} f(x)$$

- $f(x) = \mathbf{E}_{\xi \sim \mathcal{D}}\left[f_\xi(x)\right]$ — expectation minimization
- $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$ — convexity
- $\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2$ — L-smoothness
- $\mathbb{E}_\xi[\nabla f(x, \xi)] = \nabla f(x)$ — unbiasedness
- $\mathbb{E}_\xi\left[\|\nabla f(x, \xi) - \nabla f(x)\|_2^2\right] \leq \sigma^2$ — boundedness of the variance

Assumptions

Light-tails assumption

$$\mathbb{E}\left[\exp\left(\frac{\|\nabla f(x,\xi)-\nabla f(x)\|_2^2}{\sigma^2}\right)\right]\leq \exp(1)$$

Heavy-tails assumption

$$\mathbb{E}_\xi\left[\|\nabla f(x,\xi)-\nabla f(x)\|_2^2\right]\leq \sigma^2$$

Assumptions

Light-tails assumption $\Longrightarrow$ **Heavy-tails assumption**

Light-tails assumption $\not\Longleftarrow$ **Heavy-tails assumption**

$$\mathbb{E}\left[\exp\left(\frac{\|\nabla f(x,\xi) - \nabla f(x)\|_2^2}{\sigma^2}\right)\right] \leq \exp(1)$$

$$\mathbb{E}_\xi\left[\|\nabla f(x,\xi) - \nabla f(x)\|_2^2\right] \leq \sigma^2$$

$\|\nabla f(x,\xi)$ → Sub-Gaussian distribution

Assumptions

Light-tails assumption

$$\mathbb{E}\left[\exp\left(\frac{\|\nabla f(x,\xi) - \nabla f(x)\|_2^2}{\sigma^2}\right)\right] \leq \exp(1)$$

Well understood

Heavy-tails assumption

$$\mathbb{E}_\xi\left[\|\nabla f(x,\xi) - \nabla f(x)\|_2^2\right] \leq \sigma^2$$

We focus on this situation!

3. Prior Works

High-probability convergence results

Method	Complexity	Tails	Domain	Batchsizes
SGD	$O\left(\max\left\{\frac{LR_0{}^2}{\varepsilon}, \frac{\sigma^2 R_0{}^2}{\varepsilon^2}\ln^2(\beta^{-1})\right\}\right)$	light	bounded	$O(1)$
SSTM	$O\left(\max\left\{\sqrt{\frac{LR_0^2}{\varepsilon}}, \frac{\sigma^2 R_0^2}{\varepsilon^2}\right\}\ln\frac{LR_0^2}{\varepsilon\beta}\right)$	light	$\mathbb{R}^n$	from $O(\varepsilon^{-1/2})$ to $O(\varepsilon^{-3/2})$
AC-SA	$O\left(\max\left\{\sqrt{\frac{LR_0^2}{\varepsilon}}, \frac{\sigma^2 R_0^2}{\varepsilon^2}\ln(\beta^{-1})\right\}\right)$	light	arbitrary	$O(1)$
RSMD	$O\left(\max\left\{\frac{L\Theta^2}{\varepsilon}, \frac{\sigma^2\Theta^2}{\varepsilon^2}\right\}\ln(\beta^{-1})\right)$	heavy	bounded	$O(1)$

High-probability convergence results

Heavy or light-tailed noise

How batchsizes grow during the optimization process

Method	Complexity	Tails	Domain	Batchsizes
SGD	$O\left(\max\left\{\frac{LR_0{}^2}{\varepsilon}, \frac{\sigma^2 R_0{}^2}{\varepsilon^2}\ln^2(\beta^{-1})\right\}\right)$	light	bounded	$O(1)$
SSTM	$O\left(\max\left\{\sqrt{\frac{LR_0^2}{\varepsilon}}, \frac{\sigma^2 R_0^2}{\varepsilon^2}\right\}\ln\frac{LR_0^2}{\varepsilon\beta}\right)$	light	$\mathbb{R}^n$	from $O(\varepsilon^{-1/2})$ to $O(\varepsilon^{-3/2})$
AC-SA	$O\left(\max\left\{\sqrt{\frac{LR_0^2}{\varepsilon}}, \frac{\sigma^2 R_0^2}{\varepsilon^2}\ln(\beta^{-1})\right\}\right)$	light	arbitrary	$O(1)$
RSMD	$O\left(\max\left\{\frac{L\Theta^2}{\varepsilon}, \frac{\sigma^2\Theta^2}{\varepsilon^2}\right\}\ln(\beta^{-1})\right)$	heavy	bounded	$O(1)$

stochastic first-order oracle calls

Set where optimization problem is defined

High-probability convergence results

Method	Complexity	Tails	Domain	Batchsizes
SGD	$O\left(\max\left\{\frac{L{R_0}^2}{\varepsilon}, \frac{\sigma^2 {R_0}^2}{\varepsilon^2}\ln^2(\beta^{-1})\right\}\right)$	light	bounded	$O(1)$
SSTM	$O\left(\max\left\{\sqrt{\frac{LR_0^2}{\varepsilon}}, \frac{\sigma^2 R_0^2}{\varepsilon^2}\right\}\ln\frac{LR_0^2}{\varepsilon\beta}\right)$	light	$\mathbb{R}^n$	from $O(\varepsilon^{-1/2})$ to $O(\varepsilon^{-3/2})$
AC-SA	$O\left(\max\left\{\sqrt{\frac{LR_0^2}{\varepsilon}}, \frac{\sigma^2 R_0^2}{\varepsilon^2}\ln(\beta^{-1})\right\}\right)$	light	arbitrary	$O(1)$
RSMD	$O\left(\max\left\{\frac{L\Theta^2}{\varepsilon}, \frac{\sigma^2\Theta^2}{\varepsilon^2}\right\}\ln(\beta^{-1})\right)$	heavy	bounded	$O(1)$

R_0 — initial distance to the optimum

High-probability convergence results

Method	Complexity	Tails	Domain	Batchsizes
SGD	$O\left(\max\left\{\frac{L{R_0}^2}{\varepsilon}, \frac{\sigma^2 {R_0}^2}{\varepsilon^2}\ln^2(\beta^{-1})\right\}\right)$	light	bounded	$O(1)$
SSTM	$O\left(\max\left\{\sqrt{\frac{LR_0^2}{\varepsilon}}, \frac{\sigma^2 R_0^2}{\varepsilon^2}\right\}\ln\frac{LR_0^2}{\varepsilon\beta}\right)$	light	$\mathbb{R}^n$	from $O(\varepsilon^{-1/2})$ to $O(\varepsilon^{-3/2})$
AC-SA	$O\left(\max\left\{\sqrt{\frac{LR_0^2}{\varepsilon}}, \frac{\sigma^2 R_0^2}{\varepsilon^2}\ln(\beta^{-1})\right\}\right)$	light	arbitrary	$O(1)$
RSMD	$O\left(\max\left\{\frac{L\Theta^2}{\varepsilon}, \frac{\sigma^2\Theta^2}{\varepsilon^2}\right\}\ln(\beta^{-1})\right)$	heavy	bounded	$O(1)$

Θ — a diameter of the set where the optimization problem is defined

High-probability convergence results

Method	Complexity	Tails	Domain	Batchsizes
SGD	$O\left(\max\left\{\frac{LR_0{}^2}{\varepsilon}, \frac{\sigma^2 R_0{}^2}{\varepsilon^2}\ln^2(\beta^{-1})\right\}\right)$	light	bounded	$O(1)$
SSTM	$O\left(\max\left\{\sqrt{\frac{LR_0^2}{\varepsilon}}, \frac{\sigma^2 R_0^2}{\varepsilon^2}\right\}\ln\frac{LR_0^2}{\varepsilon\beta}\right)$	light	$\mathbb{R}^n$	from $O(\varepsilon^{-1/2})$ to $O(\varepsilon^{-3/2})$
AC-SA	$O\left(\max\left\{\sqrt{\frac{LR_0^2}{\varepsilon}}, \frac{\sigma^2 R_0^2}{\varepsilon^2}\ln(\beta^{-1})\right\}\right)$	light	arbitrary	$O(1)$
RSMD	$O\left(\max\left\{\frac{L\Theta^2}{\varepsilon}, \frac{\sigma^2\Theta^2}{\varepsilon^2}\right\}\ln(\beta^{-1})\right)$	heavy	bounded	$O(1)$

$$\mathrm{Prob}\left\{f(x^N) - f(x^*) \leq \varepsilon\right\} \geq 1 - \beta$$

Desired accuracy   Confidence level

High-probability convergence results

Method	Complexity	Tails	Domain	Batchsizes
SGD	$O\left(\max\left\{\frac{L{R_0}^2}{\varepsilon}, \frac{\sigma^2{R_0}^2}{\varepsilon^2}\ln^2(\beta^{-1})\right\}\right)$	light	bounded	$O(1)$
SSTM	$O\left(\max\left\{\sqrt{\frac{LR_0^2}{\varepsilon}}, \frac{\sigma^2R_0^2}{\varepsilon^2}\right\}\ln\frac{LR_0^2}{\varepsilon\beta}\right)$	light	$\mathbb{R}^n$	from $O(\varepsilon^{-1/2})$ to $O(\varepsilon^{-3/2})$
AC-SA	$O\left(\max\left\{\sqrt{\frac{LR_0^2}{\varepsilon}}, \frac{\sigma^2R_0^2}{\varepsilon^2}\ln(\beta^{-1})\right\}\right)$	light	arbitrary	$O(1)$
RSMD	$O\left(\max\left\{\frac{L\Theta^2}{\varepsilon}, \frac{\sigma^2\Theta^2}{\varepsilon^2}\right\}\ln(\beta^{-1})\right)$	heavy	bounded	$O(1)$

Acceleration

4. Main Result

Stochastic Gradient Descent
SGD

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

Stochastic Gradient Descent
SGD

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

- $$\nabla f(x^k, \xi^k) = \frac{1}{m_k} \sum_{i=1}^{m_k} \nabla f(x^k, \xi_i^k)$$

Clipped Stochastic Gradient Descent
clipped-SGD

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

- $\nabla f(x^k, \boldsymbol{\xi}^k) = \frac{1}{m_k} \sum_{i=1}^{m_k} \nabla f(x^k, \xi_i^k)$

Clipped Stochastic Gradient Descent
clipped-SGD

- $\widetilde{\nabla} f(x^k, \xi^k) = \text{clip}(\nabla f(x^k, \xi^k), \lambda)$

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

- $\nabla f(x^k, \xi^k) = \frac{1}{m_k} \sum_{i=1}^{m_k} \nabla f(x^k, \xi_i^k)$

Stochastic Similar Triangles Method
SSTM

$$x^{k+1} = \left(A_k y^k + \alpha_{k+1} z^k\right) / A_{k+1}$$
$$z^{k+1} = z^k - \alpha_{k+1} \nabla f(x^{k+1}, \boldsymbol{\xi}^k)$$
$$y^{k+1} = \left(A_k y^k + \alpha_{k+1} z^{k+1}\right) / A_{k+1}$$

Stochastic Similar Triangles Method
SSTM

$$x^{k+1} = \left(A_k y^k + \alpha_{k+1} z^k\right) / A_{k+1}$$
$$z^{k+1} = z^k - \alpha_{k+1} \nabla f(x^{k+1}, \boldsymbol{\xi}^k)$$
$$y^{k+1} = \left(A_k y^k + \alpha_{k+1} z^{k+1}\right) / A_{k+1}$$

↓

Accelerated SGD

Stochastic Similar Triangles Method
SSTM

- Parameters

$$A_0 = \alpha_0 = 0$$
$$A_{k+1} = A_k + \alpha_{k+1} \qquad \alpha_{k+1} = \frac{k+2}{2L}$$

$$x^{k+1} = \left(A_k y^k + \alpha_{k+1} z^k\right) / A_{k+1}$$
$$z^{k+1} = z^k - \alpha_{k+1} \nabla f(x^{k+1}, \boldsymbol{\xi}^k)$$
$$y^{k+1} = \left(A_k y^k + \alpha_{k+1} z^{k+1}\right) / A_{k+1}$$

Clipped Stochastic Similar Triangles Method

clipped-SSTM

- Parameters
- Clipping

$$A_0 = \alpha_0 = 0$$
$$A_{k+1} = A_k + \alpha_{k+1}$$

$$\alpha_{k+1} = \frac{k+2}{2aL}$$

$$x^{k+1} = \left(A_k y^k + \alpha_{k+1} z^k\right) / A_{k+1}$$
$$z^{k+1} = z^k - \alpha_{k+1} \widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k)$$
$$y^{k+1} = \left(A_k y^k + \alpha_{k+1} z^{k+1}\right) / A_{k+1}$$

New method!

Clipped Stochastic Similar Triangles Method
clipped-SSTM

Parameters

Clipping

$$A_0 = \alpha_0 = 0$$
$$A_{k+1} = A_k + \alpha_{k+1}$$

$$\alpha_{k+1} = \frac{k+2}{2aL}$$

$$x^{k+1} = \left(A_k y^k + \alpha_{k+1} z^k\right) / A_{k+1}$$
$$z^{k+1} = z^k - \alpha_{k+1} \widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k)$$
$$y^{k+1} = \left(A_k y^k + \alpha_{k+1} z^{k+1}\right) / A_{k+1}$$

$$\widetilde{\nabla} f(x^{k+1}, \boldsymbol{\xi}^k) = \text{clip}\left(\nabla f(x^{k+1}, \boldsymbol{\xi}^k), \lambda_{k+1}\right)$$

$$\lambda_{k+1} = \frac{B}{\alpha_{k+1}}$$

High-probability convergence

Heavy-tailed noise

clipped-SGD [This work]	$O\left(\max\left\{\frac{L{R_0}^2}{\varepsilon}, \frac{\sigma^2{R_0}^2}{\varepsilon^2}\right\}\ln(\beta^{-1})\right)$	heavy	$\mathbb{R}^n$	$\widetilde{O}(\varepsilon^{-1})$
clipped-SSTM [This work]	$O\left(\frac{\sigma^2 R_0^2}{\varepsilon^2}\ln\frac{\sigma R_0}{\varepsilon\beta}\right)$, σ^2 is big	heavy	$\mathbb{R}^n$	$O(1)$
clipped-SSTM [This work]	$O\left(\max\left\{\frac{LR_0^2}{\varepsilon}, \frac{\sigma^2 R_0^2}{\varepsilon^2}\right\}\ln\frac{LR_0^2}{\varepsilon\beta}\right)$	heavy	$\mathbb{R}^n$	from $O(1)$ to $O(\varepsilon^{-1})$
clipped-SSTM [This work]	$O\left(\max\left\{\sqrt{\frac{LR_0^2}{\varepsilon}}, \frac{\sigma^2 R_0^2}{\varepsilon^2}\right\}\ln\frac{LR_0^2}{\varepsilon\beta}\right)$	heavy	$\mathbb{R}^n$	from $O(\varepsilon^{-1/2})$ to $O(\varepsilon^{-3/2})$

Nearly optimal

Unbounded

6. Experiments

Clipped-SSTM and Clipped-SGD
Logistic Regression

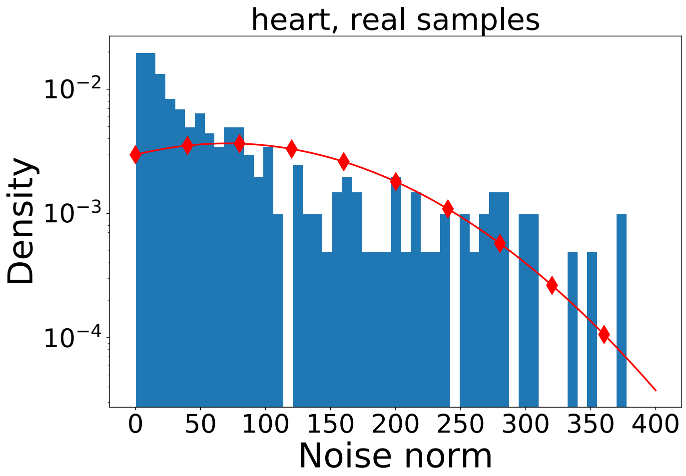
$$\min_{x\in\mathbb{R}^n} f(x) = \frac{1}{r}\sum_{i=1}^{r} \underbrace{\log\left(1+\exp\left(-y_i\cdot (Ax)_i\right)\right)}_{f_i(x)}$$

- $A \in \mathbb{R}^{r\times n}$ — matrix of instances
- $y \in \{0,1\}^m$ — vector of labels

Datasets from LIBSVM:

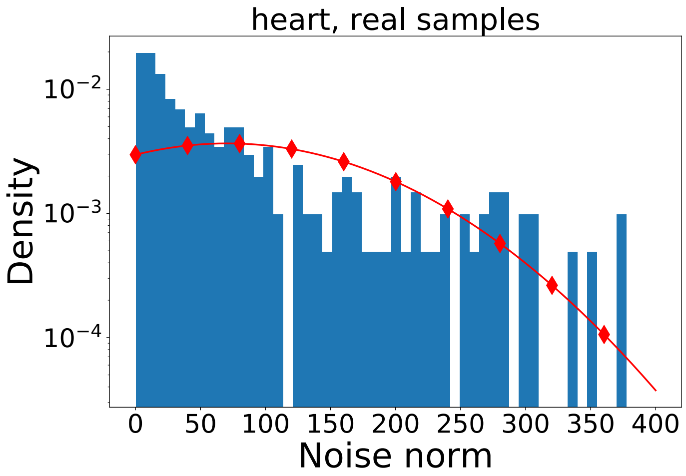
	heart	australian
Size	270	690
Dimension	13	13

Data analysis
Dataset - Heart



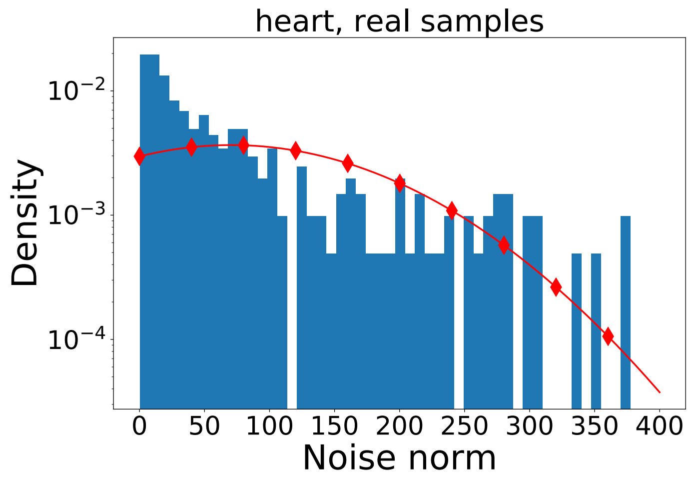
● Histogram of $\|\nabla f_i(x^*)\|_2$

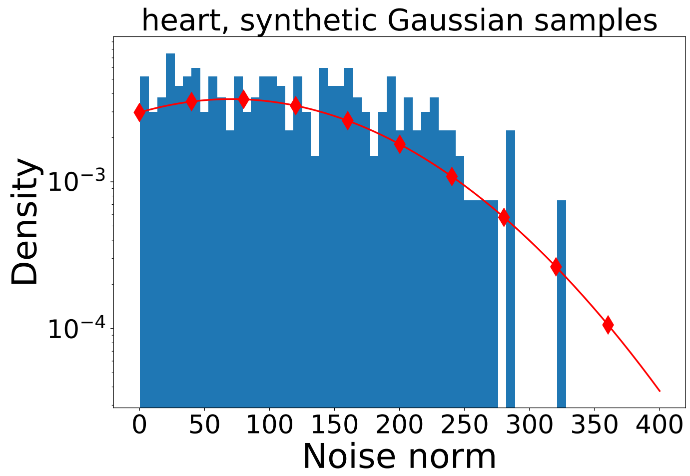
Data analysis
Dataset - Heart



- Histogram of $\|\nabla f_i(x^*)\|_2$
- Red lines correspond to probability density function of normal distribution with empirically estimated mean and variance

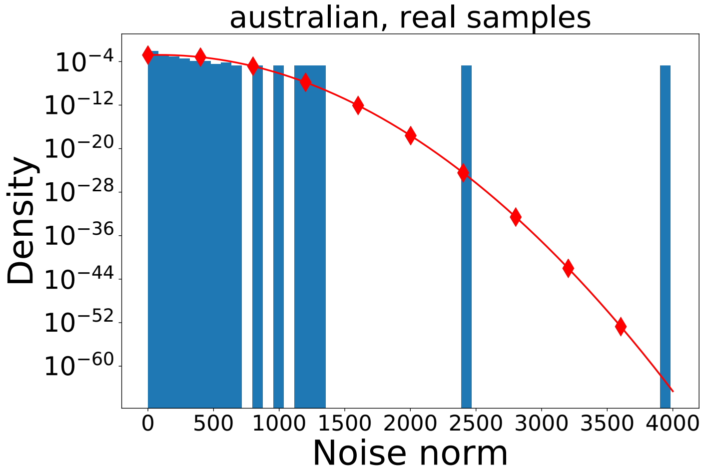
Data analysis
Dataset - Heart

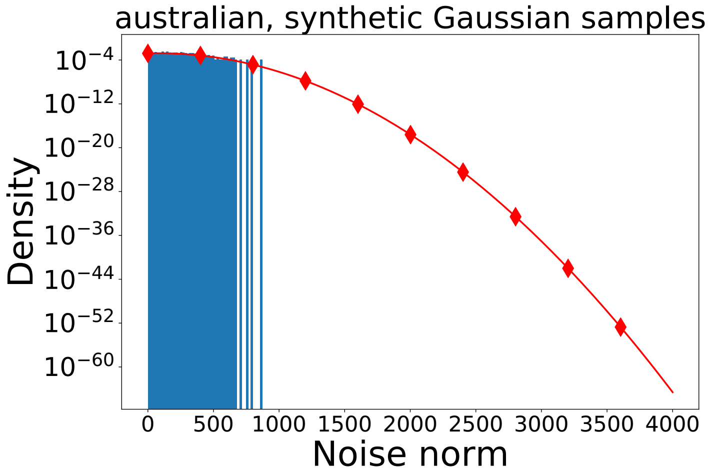




- Histogram of $\|\nabla f_i(x^*)\|_2$
- Red lines correspond to probability density function of normal distribution with empirically estimated mean and variance
- Histogram of synthetic Gaussian samples with mean and variance estimated via empirical mean and variance of real samples

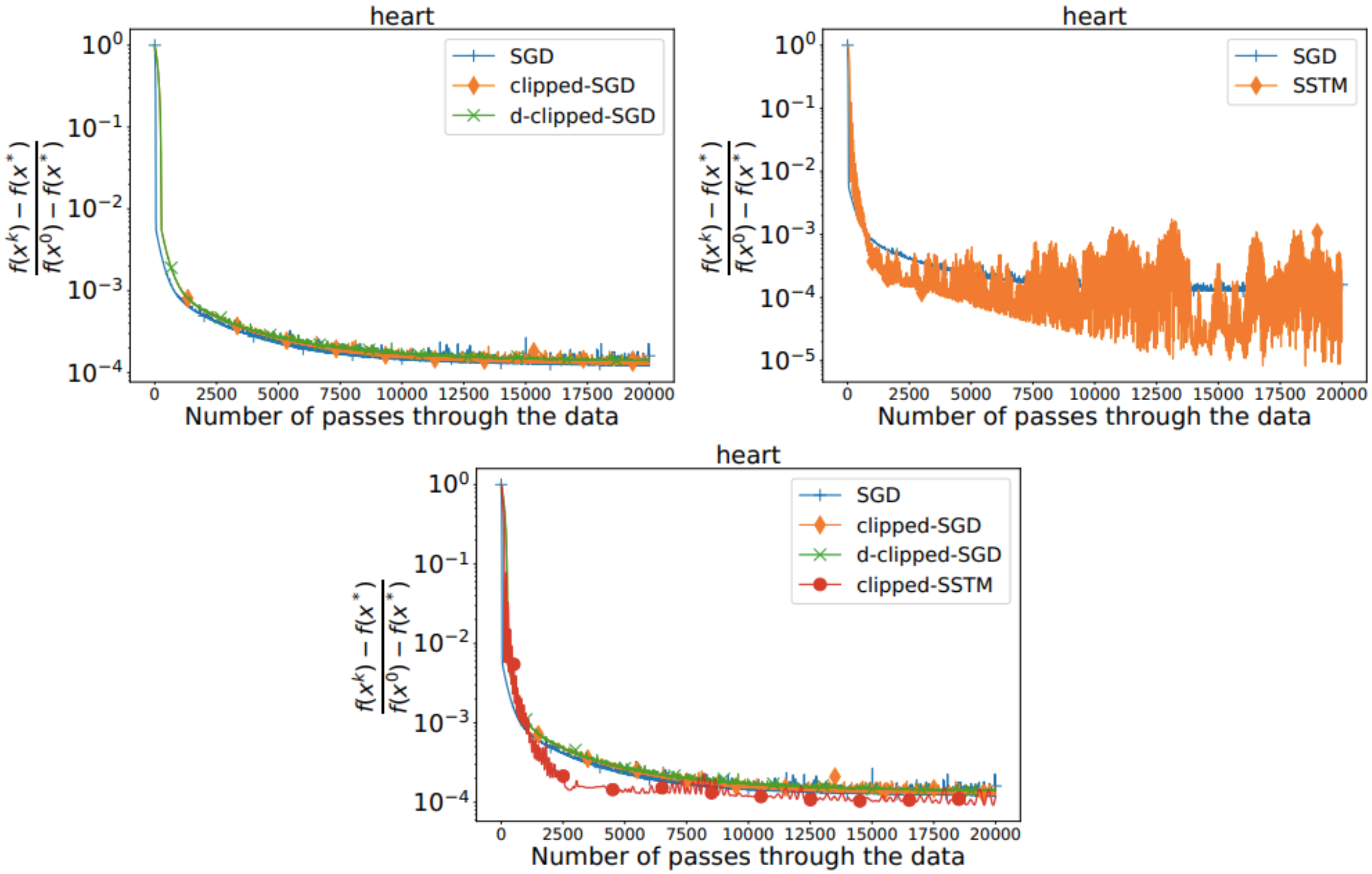
Data analysis
Dataset - Australian



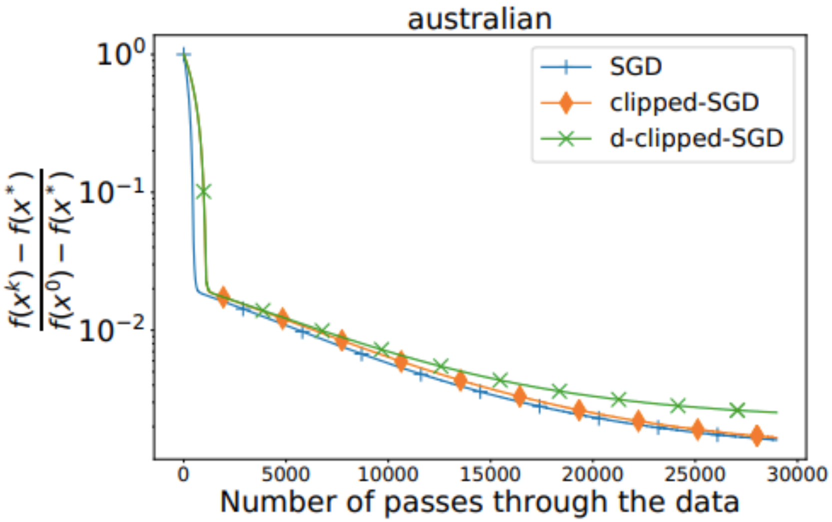


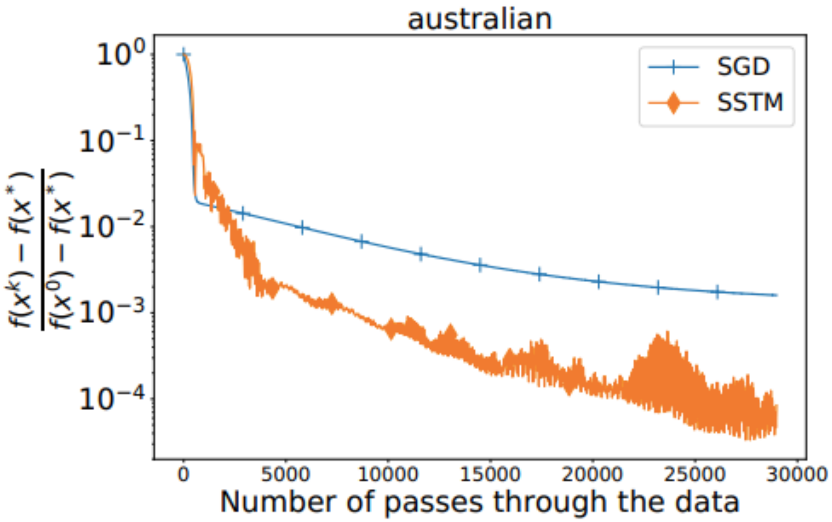
- Histogram of $\|\nabla f_i(x^*)\|_2$
- Red lines correspond to probability density function of normal distribution with empirically estimated mean and variance
- Histogram of synthetic Gaussian samples with mean and variance estimated via empirical mean and variance of real samples

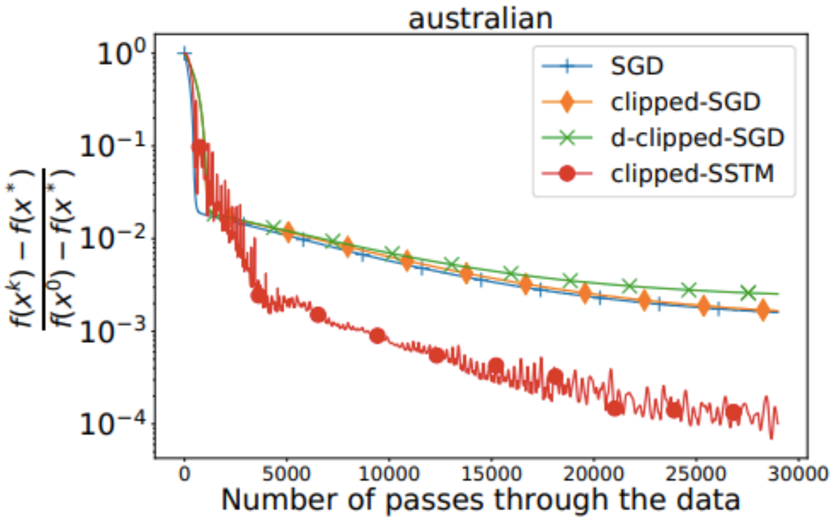
Trajectories - Heart



Trajectories - Australian







More details you could find in our work:

Gorbunov, Eduard, **Marina Danilova**, and Alexander Gasnikov. "**Stochastic Optimization with Heavy-Tailed Noise via Accelerated Gradient Clipping**." *arXiv preprint arXiv:2005.10785* (2020).

- strongly convex case
- more experiments

Thank you for your attention!

The End