Linear Convergence Rate in Convex Setup is Possible! First- and Zero-Order Algorithms under Generalized Smoothness

Aleksandr Lobanov

Moscow Institute of Physics and Technology, Dolgoprudny, Russia
Skolkovo Institute of Science and Technology, Moscow, Russia
ISP RAS Research Center for Trusted Artificial Intelligence, Moscow, Russia

lobbsasha@mail.ru

Sirius Mathematics Center

Today's plan

Problem Statement and Background

This work focuses on a stochastic optimization problem:

$$\min_{x \in \mathbb{R}^d} \left\{ f(x) := \mathbb{E}_{\xi \sim \mathcal{D}} \left[f(x, \xi) \right] \right\}, \tag{1}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is a smooth convex, possibly stochastic function.

Gradient Descent (Cauchy, 1847) [1]

$$\boxed{x^{k+1} = x^k - \eta_k \nabla f(x^k)} \qquad \rightarrow \qquad f(x^N) - f^* \lesssim \mathcal{O}\left(\frac{LR^2}{N}\right)$$

Accelerated Gradient Descent (Nesterov, 1983) [2]

$$\boxed{x^{k+1} = x^k - \eta_k \nabla f\left(x^k + \beta_k(x^k - x^{k-1})\right) + \beta_k(x^k - x^{k-1})} \quad \rightarrow \quad f(x^N) - f^* \lesssim \mathcal{O}\left(\frac{LR^2}{N^2}\right)$$

Problem Statement and Background

This work focuses on a stochastic optimization problem:

$$\min_{x \in \mathbb{R}^d} \left\{ f(x) := \mathbb{E}_{\xi \sim \mathcal{D}} \left[f(x, \xi) \right] \right\}, \tag{1}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is a smooth convex, possibly stochastic function.

Gradient Descent (Cauchy, 1847) [1]

$$\boxed{x^{k+1} = x^k - \eta_k \nabla f(x^k)} \quad \rightarrow \quad f(x^N) - f^* \lesssim \mathcal{O}\left(\frac{LR^2}{N}\right)$$

Accelerated Gradient Descent (Nesterov, 1983) [2]

$$\boxed{x^{k+1} = x^k - \eta_k \nabla f\left(x^k + \beta_k (x^k - x^{k-1})\right) + \beta_k (x^k - x^{k-1})} \quad \rightarrow \quad f(x^N) - f^* \lesssim \mathcal{O}\left(\frac{LR^2}{N^2}\right)$$

Problem Statement and Background

This work focuses on a stochastic optimization problem:

$$\min_{x \in \mathbb{R}^d} \left\{ f(x) := \mathbb{E}_{\xi \sim \mathcal{D}} \left[f(x, \xi) \right] \right\}, \tag{1}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is a smooth convex, possibly stochastic function.

Function f is L-smooth if the following inequality is satisfied for any $x, y \in \mathbb{R}^d$:

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

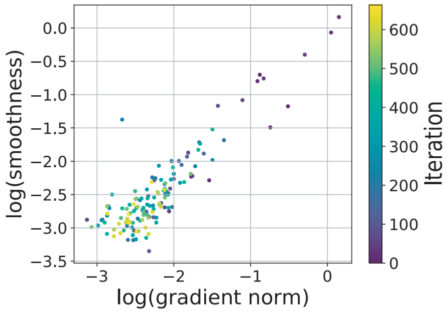
Gradient Descent (Cauchy, 1847) [1]

$$\boxed{x^{k+1} = x^k - \eta_k \nabla f(x^k)} \quad \to \quad f(x^N) - f^* \lesssim \mathcal{O}\left(\frac{LR^2}{N}\right)$$

Accelerated Gradient Descent (Nesterov, 1983) [2]

$$\boxed{x^{k+1} = x^k - \eta_k \nabla f\left(x^k + \beta_k (x^k - x^{k-1})\right) + \beta_k (x^k - x^{k-1})} \quad \to \quad f(x^N) - f^* \lesssim \mathcal{O}\left(\frac{LR^2}{N^2}\right)$$

Motivation of Generalized Smoothness



Relaxed Smoothness Condition [3]

A second order differentiable function f is (L_0, L_1)-smooth $\forall x \in \mathbb{R}^d$ if

$$\left\|\nabla^2 f(x)\right\| \leq L_0 + L_1 \left\|\nabla f(x)\right\|$$

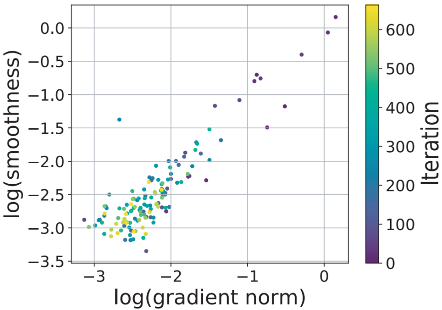
More Relaxed Smoothness Condition [4]

A function f is (L_0, L_1)-smooth $\forall x, y \in \mathbb{R}^d$ with $\|y - x\| \leq \frac{1}{L_1}$ if:

$$\left\|\nabla f(y) - \nabla f(x)\right\| \leq \left(L_0 + L_1 \left\|\nabla f(x)\right\|\right) \left\|y - x\right\|$$

Function examples [5]

Motivation of Generalized Smoothness



Relaxed Smoothness Condition [3]

A second order differentiable function f is (L_0, L_1)-smooth $\forall x \in \mathbb{R}^d$ if

$$\left\|\nabla^2 f(x)\right\| \leq L_0 + L_1 \left\|\nabla f(x)\right\|$$

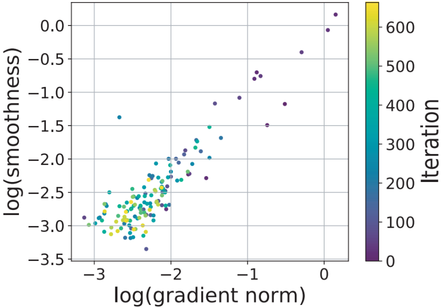
More Relaxed Smoothness Condition [4]

A function f is (L_0, L_1)-smooth $\forall x, y \in \mathbb{R}^d$ with $\|y - x\| \leq \frac{1}{L_1}$ if:

$$\left\|\nabla f(y) - \nabla f(x)\right\| \leq \left(L_0 + L_1 \left\|\nabla f(x)\right\|\right) \|y - x\|$$

Function examples [5]

Motivation of Generalized Smoothness



Relaxed Smoothness Condition [3]

A second order differentiable function f is (L_0, L_1)-smooth $\forall x \in \mathbb{R}^d$ if

$$\left\|\nabla^2 f(x)\right\| \leq L_0 + L_1 \left\|\nabla f(x)\right\|$$

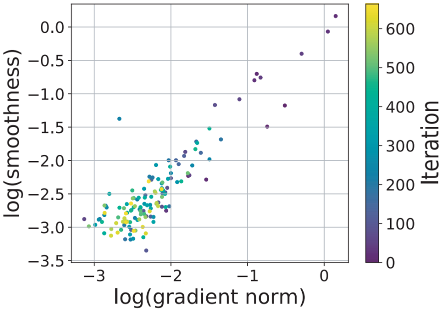
More Relaxed Smoothness Condition [4]

A function f is (L_0, L_1)-smooth $\forall x, y \in \mathbb{R}^d$ with $\|y - x\| \leq \frac{1}{L_1}$ if:

$$\|\nabla f(y) - \nabla f(x)\| \leq (L_0 + L_1 \|\nabla f(x)\|) \|y - x\|$$

Function examples [5]

Motivation of Generalized Smoothness



Relaxed Smoothness Condition [3]

A second order differentiable function f is (L_0, L_1)-smooth $\forall x \in \mathbb{R}^d$ if

$$\left\|\nabla^2 f(x)\right\| \leq L_0 + L_1 \left\|\nabla f(x)\right\|$$

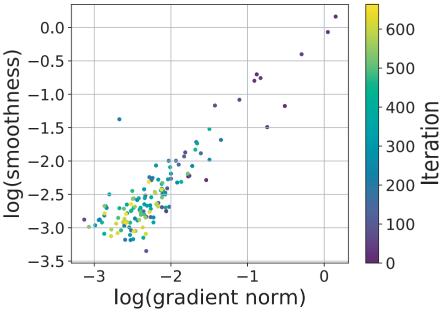
More Relaxed Smoothness Condition [4]

A function f is (L_0, L_1)-smooth $\forall x, y \in \mathbb{R}^d$ with $\|y - x\| \leq \frac{1}{L_1}$ if:

$$\left\|\nabla f(y) - \nabla f(x)\right\| \leq \left(L_0 + L_1 \left\|\nabla f(x)\right\|\right)\|y - x\|$$

Function examples [5]

① : $f(x) = \|x\|^{2n}$, where $n \in \mathbb{N}$. $f(x)$ is $(2n, 2n-1)$-smooth, but is not L-smooth for $n \geq 2$.



Relaxed Smoothness Condition [3]

A second order differentiable function f is (L_0, L_1)-smooth $\forall x \in \mathbb{R}^d$ if

$$\left\|\nabla^2 f(x)\right\| \leq L_0 + L_1 \left\|\nabla f(x)\right\|$$

More Relaxed Smoothness Condition [4]

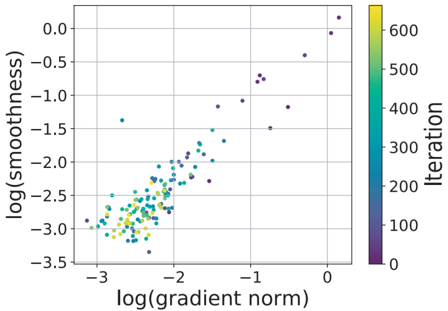
A function f is (L_0, L_1)-smooth $\forall x, y \in \mathbb{R}^d$ with $\|y - x\| \leq \frac{1}{L_1}$ if:

$$\|\nabla f(y) - \nabla f(x)\| \leq (L_0 + L_1 \|\nabla f(x)\|) \|y - x\|$$

Function examples [5]

① : $f(x) = \|x\|^{2n}$, where $n \in \mathbb{N}$. $f(x)$ is $(2n, 2n-1)$-smooth, but is not L-smooth for $n \geq 2$.
② : $f(x) = \exp(a^\top x)$, where $a \in \mathbb{R}^d$. $f(x)$ is $(0, \|a\|)$-smooth, but is not L-smooth for $a \neq 0$.

Motivation of Generalized Smoothness



Relaxed Smoothness Condition [3]

A second order differentiable function f is (L_0, L_1)-smooth $\forall x \in \mathbb{R}^d$ if

$$\left\|\nabla^2 f(x)\right\| \leq L_0 + L_1 \left\|\nabla f(x)\right\|$$

More Relaxed Smoothness Condition [4]

A function f is (L_0, L_1)-smooth $\forall x, y \in \mathbb{R}^d$ with $\|y - x\| \leq \frac{1}{L_1}$ if:

$$\left\|\nabla f(y) - \nabla f(x)\right\| \leq \left(L_0 + L_1 \left\|\nabla f(x)\right\|\right) \|y - x\|$$

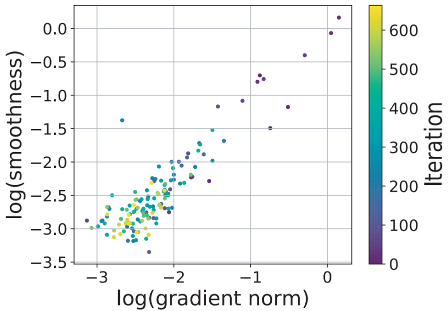
Function examples [5]

① : $f(x) = \|x\|^{2n}$, where $n \in \mathbb{N}$. $f(x)$ is $(2n, 2n-1)$-smooth, but is not L-smooth for $n \geq 2$.

② : $f(x) = \exp(a^\top x)$, where $a \in \mathbb{R}^d$. $f(x)$ is $(0, \|a\|)$-smooth, but is not L-smooth for $a \neq 0$.

③ : $f(x) = \log(1 + \exp(-a^\top x))$, where $a \in \mathbb{R}^d$. $L = \|a\|^2$. However, $L_0 = 0$ and $L_1 = \|a\|$.

Motivation of Generalized Smoothness



Relaxed Smoothness Condition [3]

A second order differentiable function f is (L_0, L_1)-smooth $\forall x \in \mathbb{R}^d$ if

$$\left\|\nabla^2 f(x)\right\| \leq L_0 + L_1 \left\|\nabla f(x)\right\|$$

More Relaxed Smoothness Condition [4]

A function f is (L_0, L_1)-smooth $\forall x, y \in \mathbb{R}^d$ with $\|y - x\| \leq \frac{1}{L_1}$ if:

$$\|\nabla f(y) - \nabla f(x)\| \leq \left(L_0 + L_1 \|\nabla f(x)\|\right) \|y - x\|$$

Function examples [5]

① : $f(x) = \|x\|^{2n}$, where $n \in \mathbb{N}$. $f(x)$ is $(2n, 2n-1)$-smooth, but is not L-smooth for $n \geq 2$.
② : $f(x) = \exp(a^\top x)$, where $a \in \mathbb{R}^d$. $f(x)$ is $(0, \|a\|)$-smooth, but is not L-smooth for $a \neq 0$.
③ : $f(x) = \log(1 + \exp(-a^\top x))$, where $a \in \mathbb{R}^d$. $L = \|a\|^2$. However, $L_0 = 0$ and $L_1 = \|a\|$.

Algorithm 1 Stochastic Gradient Descent Method (SGD)

Input: initial point $x_0 \in \mathbb{R}^d$, iterations number N, batch size B, step size $\eta_k > 0$

for $k = 0$ **to** $N-1$ **do**

1. Draw fresh i.i.d. samples $\xi_1^k, ..., \xi_B^k$
2. $\nabla f(x^k, \boldsymbol{\xi}^k) = \frac{1}{B}\sum_{i=1}^B \nabla f(x^k, \xi_i^k)$
3. $x^{k+1} \leftarrow x^k - \eta_k \cdot \nabla f(x^k, \boldsymbol{\xi}^k)$

end for

Return: x^N

Ass: $f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y-x\|^2$

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

① Step size $\boxed{\eta_k = \eta \leq (L_0 + L_1 M)^{-1}}$:

$$\mathbb{E}\left[f(x^N)\right] - f^* \leq \frac{R^2}{2\eta N} + \frac{\sigma^2 \eta}{B}$$

② Step size $\boxed{\eta_k = \eta \leq (L_0 + L_1 M)^{-1}}$:

$$\mathbb{E}\left[f(x^N)\right] - f^* \leq (1-\eta\mu)^N F_0 + \frac{\sigma^2}{2\mu B}$$

(Stochastic) Gradient Descent Method

Algorithm 1 Stochastic Gradient Descent Method (SGD)

Input: initial point $x_0 \in \mathbb{R}^d$, iterations number N, batch size B, step size $\eta_k > 0$

for $k = 0$ **to** $N-1$ **do**

1. Draw fresh i.i.d. samples $\xi_1^k, ..., \xi_B^k$
2. $\nabla f(x^k, \boldsymbol{\xi}^k) = \frac{1}{B}\sum_{i=1}^{B} \nabla f(x^k, \xi_i^k)$
3. $x^{k+1} \leftarrow x^k - \eta_k \cdot \nabla f(x^k, \boldsymbol{\xi}^k)$

end for

Return: x^N

Ass: $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2$

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

① Step size $\boxed{\eta_k = \eta \leq (L_0 + L_1 M)^{-1}}$:

$$\mathbb{E}\left[f(x^N)\right] - f^* \leq \frac{R^2}{2\eta N} + \frac{\sigma^2 \eta}{B}$$

② Step size $\boxed{\eta_k = \eta \leq (L_0 + L_1 M)^{-1}}$:

$$\mathbb{E}\left[f(x^N)\right] - f^* \leq (1 - \eta\mu)^N F_0 + \frac{\sigma^2}{2\mu B}$$

(Stochastic) Gradient Descent Method

Algorithm 1 Stochastic Gradient Descent Method (SGD)

Input: initial point $x_0 \in \mathbb{R}^d$, iterations number N, batch size B, step size $\eta_k > 0$

for $k = 0$ **to** $N - 1$ **do**

1. Draw fresh i.i.d. samples $\xi_1^k, ..., \xi_B^k$
2. $\nabla f(x^k, \boldsymbol{\xi}^k) = \frac{1}{B}\sum_{i=1}^{B} \nabla f(x^k, \xi_i^k)$
3. $x^{k+1} \leftarrow x^k - \eta_k \cdot \nabla f(x^k, \boldsymbol{\xi}^k)$

end for

Return: x^N

③ Step size $\boxed{\eta_k = \min\left\{\left(L_0 + L_1 \left\|\nabla f(x^k)\right\|\right)^{-1}\right\}}$:

$$\mathbb{E}\left[f(x^N)\right] - f^* \leq \left(1 - \frac{1}{4L_1R}\right)^K F_0 + \frac{L_0R^2}{N-K}$$

Ass: $f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y - x\|^2$

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

① Step size $\boxed{\eta_k = \eta \leq (L_0 + L_1 M)^{-1}}$:

$$\mathbb{E}\left[f(x^N)\right] - f^* \leq \frac{R^2}{2\eta N} + \frac{\sigma^2\eta}{B}$$

② Step size $\boxed{\eta_k = \eta \leq (L_0 + L_1 M)^{-1}}$:

$$\mathbb{E}\left[f(x^N)\right] - f^* \leq (1 - \eta\mu)^N F_0 + \frac{\sigma^2}{2\mu B}$$

Normalized Stochastic Gradient Descent

Algorithm 2 Normalized Stochastic Gradient Descent Method (NSGD)

Input: initial point $x_0 \in \mathbb{R}^d$, iterations number N, batch size B, step size $\eta_k > 0$ and hyperparameter $\lambda > 0$
for $k = 0$ **to** $N-1$ **do**
1. Draw fresh i.i.d. samples $\xi_1^k, \dots, \xi_B^k$
2. $\nabla f(x^k, \boldsymbol{\xi}^k) = \frac{1}{B}\sum_{i=1}^{B} \nabla f(x^k, \xi_i^k)$
3. $x^{k+1} \leftarrow x^k - \eta_k \cdot \frac{\nabla f(x^k, \boldsymbol{\xi}^k)}{\|\nabla f(x^k, \boldsymbol{\xi}^k)\|}$

end for
Return: x^N

Step size $\boxed{\eta_k = \eta \leq \lambda / \left[2(L_0 + L_1\lambda)\right]}$; $F_k = \mathbb{E}\left[f(x^k)\right] - f^*$

$$\mathbb{E}\left[f(x^N)\right] - f^* \lesssim \left(1 - \frac{\eta}{R}\right)^N F_0 + \frac{\sigma^2 M R}{B\lambda^2} + \lambda R$$

Clipped Stochastic Gradient Descent

Algorithm 3 Clipped Stochastic Gradient Descent Method (ClipSGD)

Input: initial point $x_0 \in \mathbb{R}^d$, iterations number N, batch size B, step size $\eta_k > 0$ and clipping radius $c > 0$

for $k = 0$ **to** $N-1$ **do**

1. Draw fresh i.i.d. samples $\xi_1^k, \ldots, \xi_B^k$
2. $\nabla f(x^k, \boldsymbol{\xi}^k) = \frac{1}{B}\sum_{i=1}^{B} \nabla f(x^k, \xi_i^k)$
3. $\text{clip}_c(\nabla f(x^k, \boldsymbol{\xi}^k)) = \min\{1, \frac{c}{\|\nabla f(x^k, \boldsymbol{\xi}^k)\|}\} \nabla f(x^k, \boldsymbol{\xi}^k)$
4. $x^{k+1} \leftarrow x^k - \eta_k \cdot \text{clip}_c(\nabla f(x^k, \boldsymbol{\xi}^k))$

end for

Return: x^N

Step size $\boxed{\eta_k = \eta \leq [4(L_0 + L_1 c)]^{-1}}$; $F_k = \mathbb{E}\left[f(x^k)\right] - f^*$; $\mathcal{R} = \left(\eta + \frac{MR}{c^2} + \frac{R}{c}\right)$

$$F_N \lesssim \left(1 - \frac{\eta c}{R}\right)^K F_0 + \frac{R^2}{\eta(N-K)} + \frac{\sigma^2 \mathcal{R}}{B}$$

Summary of results

Table 1: Comparison of iteration complexity of SGD (Algorithm 1), NSGD (Algorithm 2) and ClipSGD (Algorithm 3) under strong growth condition for smoothness ((L_0, L_1)-smoothness with $L_0 = 0$). Notation: $\eta_k > 0$ – step size; $c > 0$ – clipping radius; $M = \max_k\{\|\nabla f(x^k)\|\}$; $R = \|x^0 - x^*\|$; ε = desired accuracy; LCR = linear convergence rate; CSS = constant step size.

Reference	Algorithm	Iteration Complexity #N	Step Size	Convex? ($\mu = 0$)	LCR?	CSS?
Theorem 3.1	SGD	$\mathcal{O}\left(\frac{L_1 M R^2}{\varepsilon}\right)$	$\eta_k = \eta \leq (L_1 M)^{-1}$	✓	✗	✓
Theorem 3.3	SGD	$\mathcal{O}\left(L_1 R \log\frac{1}{\varepsilon}\right)$	$\eta_k = (L_1 \|\nabla f(x^k, \xi^k)\|)^{-1}$	✓	✓	✗
Theorem 3.4	SGD	$\mathcal{O}\left(\frac{L_1 M}{\mu} \log\frac{1}{\varepsilon}\right)$	$\eta_k = \eta \leq (L_1 M)^{-1}$	✗	✓	✓
Theorem 3.5	NSGD	$\mathcal{O}\left(L_1 R \log\frac{1}{\varepsilon}\right)$	$\eta_k = \eta \leq (2L_1)^{-1}$	✓	✓	✓
Theorem 4.1	ClipSGD	$\mathcal{O}\left(L_1 R \log\frac{1}{\varepsilon} + \frac{L_1 c R^2}{\varepsilon}\right)$	$\eta_k = \eta \leq (4L_1 c)^{-1}$	✓	✓	✓

Numerical experiments

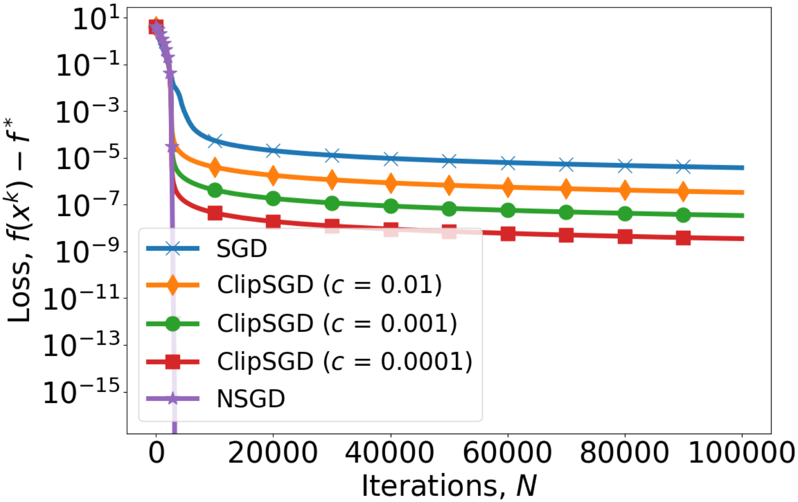


Figure: Comparison of convergence of SGD, NSGD and ClipSGD on w1a dataset $(B = 1000)$

Useful links

Where were the materials sourced from?

- *Linear Convergence Rate in Convex Setup is Possible! Gradient Descent Method Variants under (L_0, L_1)-Smoothness*
- *Power of (L_0, L_1)-Smoothness in Stochastic Convex Optimization: First- and Zero-Order Algorithms*

Thank you for your attention!

Figure: Contact me

Reference I

[1] Augustin Cauchy et al. “Méthode générale pour la résolution des systemes d’équations simultanées”. In: *Comp. Rend. Sci. Paris* 25.1847 (1847), pp. 536–538.

[2] Yurii Nesterov. “A method for unconstrained convex minimization problem with the rate of convergence O (1/k2)”. In: *Dokl. Akad. Nauk. SSSR*. Vol. 269. 3. 1983, p. 543.

[3] Jingzhao Zhang et al. “Why Gradient Clipping Accelerates Training: A Theoretical Justification for Adaptivity”. In: *International Conference on Learning Representations*.

[4] Bohang Zhang et al. “Improved analysis of clipping algorithms for non-convex optimization”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 15511–15521.

[5] Eduard Gorbunov et al. “Methods for convex (l_0, l_1)-smooth optimization: Clipping, acceleration, and adaptivity”. In: *arXiv preprint arXiv:2409.14989* (2024).