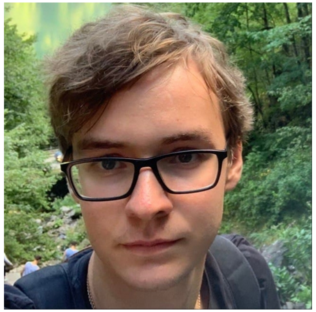
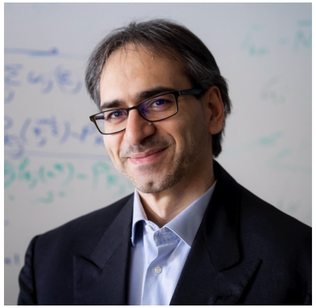
Optimal Gradient Sliding and its Application to Distributed Optimization Under Similarity

Dmitry Kovalev
KAUST



Aleksandr Beznosikov
MIPT, HSE and Yandex

Ekaterina Borodich
MIPT and HSE



Gesualdo Scutari
Purdue University

Alexander Gasnikov
MIPT, HSE and IITP

NeurIPS 2022

Distributed Optimization Problem

$$\min_{x \in \mathbb{R}^d} f(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x)$$

Distributed Optimization Problem

$$\min_{x \in \mathbb{R}^d} f(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x)$$

μ-strongly convex

L-smooth and convex

Distributed Optimization Problem

$$\min_{x \in \mathbb{R}^d} f(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x)$$

μ-strongly-convex

L-smooth and convex

f_i on local devices

Distributed Optimization Problem

$$\min_{x \in \mathbb{R}^d} f(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x)$$

μ-strongly-convex

L-smooth and convex

f_i on local devices → **Communication bottleneck!**

Distributed Optimization Problem

$$\min_{x \in \mathbb{R}^d} f(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x)$$

μ-strongly-convex

L-smooth and convex

f_i on local devices $\longrightarrow$ **Communication bottleneck!**

Use similarity of local functions!

Similarity

$$\|\nabla^2 f_i(x) - \nabla^2 f(x)\| \leq \delta$$

local

global

Similarity

$$\|\nabla^2 f_i(x) - \nabla^2 f(x)\| \leq \delta$$

local

global

For uniform data similarity parameter is **small**

$$\delta = \tilde{O}(1/\sqrt{n})$$

n – number of local samples

Similarity

$$\|\nabla^2 f_i(x) - \nabla^2 f(x)\| \leq \delta$$

local global

For uniform data similarity parameter is **small**

$$\delta = \tilde{O}(1/\sqrt{n})$$

n – number of local samples

Long similarity story

		Reference	Communication complexity	Local gradient complexity	Order	Limitations
Minimization	Upper	DANE [42]	$\mathcal{O}\left(\frac{\delta^2}{\mu^2}\log\frac{1}{\varepsilon}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\sqrt{\frac{\delta^3}{\mu^3}}\log^2\frac{1}{\varepsilon}\right)$ (2)	1st	quadratic
		DiSCO [51]	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}(\log\frac{1}{\varepsilon} + C^2\Delta F_0)\log\frac{L}{\mu}\right)$	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}(\log\frac{1}{\varepsilon} + C^2\Delta F_0)\log\frac{L}{\mu}\right)$	2nd	C - self-concordant (3)
		AIDE [40]	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\log\frac{1}{\varepsilon}\log\frac{L}{\delta}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\sqrt{\frac{\delta}{\mu}}\log\frac{1}{\varepsilon}\log\frac{L}{\delta}\right)$ (4)	1st	quadratic
		DANE-LS [50]	$\mathcal{O}\left(\frac{\delta}{\mu}\log\frac{1}{\varepsilon}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\frac{\delta^{3/2}}{\mu^{3/2}}\log\frac{1}{\varepsilon}\right)$ (5)	1st/2nd	quadratic (6)
		DANE-HB [50]	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\log\frac{1}{\varepsilon}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\frac{\delta}{\mu}\log\frac{1}{\varepsilon}\right)$ (5)	1st/2nd	quadratic (6)
		SONATA [45]	$\mathcal{O}\left(\frac{\delta}{\mu}\log\frac{1}{\varepsilon}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\sqrt{\frac{\delta}{\mu}}\log^2\frac{1}{\varepsilon}\right)$ (2)	1st	decentralized
		SPAG [21]	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}\right)$ (1)	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\sqrt{\frac{L}{\delta}}\log^2\frac{1}{\varepsilon}\right)$ (1,2)	1st	M - Lipshitz hessian
		DiRegINA [12]	$\mathcal{O}\left(\frac{\delta}{\mu}\log\frac{1}{\varepsilon} + \sqrt{\frac{M\delta R_0}{\mu}}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\sqrt{\frac{\delta}{\mu}}\log^2\frac{1}{\varepsilon} + \sqrt{\frac{MLR_0}{\mu}}\log\frac{1}{\varepsilon}\right)$ (2)	2nd	M -Lipshitz hessian
		ACN [1]	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\log\frac{1}{\varepsilon} + \sqrt[3]{\frac{M\delta R_0}{\mu}}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log^2\frac{1}{\varepsilon} + \sqrt[3]{\frac{M\delta R_0}{\mu}}\sqrt{\frac{L}{\delta}}\log\frac{1}{\varepsilon}\right)$ (2)	2nd	M -Lipshitz hessian
		AccSONATA [46]	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\log\frac{1}{\varepsilon}\log\frac{L}{\mu}\right)$	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\log^2\frac{1}{\varepsilon}\log\frac{\delta}{\mu}\right)$ (2)	1st	decentralized
		This paper	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\log\frac{1}{\varepsilon}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}\right)$	1st	
	Lower	[4]	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\log\frac{1}{\varepsilon}\right)$	—		
		[37]	—	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}\right)$		non-distributed

Similarity

$$\|\nabla^2 f_i(x) - \nabla^2 f(x)\| \leq \delta$$

local global

For uniform data similarity parameter is **small**

$$\delta = \tilde{O}(1/\sqrt{n})$$

n – number of local samples

Long similarity story

		Reference	Communication complexity	Local gradient complexity	Order	Limitations
Minimization	Upper	DANE [42]	$\mathcal{O}\left(\frac{\delta^2}{\mu^2}\log\frac{1}{\varepsilon}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\sqrt{\frac{\delta^3}{\mu^3}}\log^2\frac{1}{\varepsilon}\right)$ (2)	1st	quadratic
		DiSCO [51]	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}(\log\frac{1}{\varepsilon} + C^2\Delta F_0)\log\frac{L}{\mu}\right)$	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}(\log\frac{1}{\varepsilon} + C^2\Delta F_0)\log\frac{L}{\mu}\right)$	2nd	C - self-concordant (3)
		AIDE [40]	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\log\frac{1}{\varepsilon}\log\frac{L}{\delta}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\sqrt{\frac{\delta}{\mu}}\log\frac{1}{\varepsilon}\log\frac{L}{\delta}\right)$ (4)	1st	quadratic
		DANE-LS [50]	$\mathcal{O}\left(\frac{\delta}{\mu}\log\frac{1}{\varepsilon}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\frac{\delta^{3/2}}{\mu^{3/2}}\log\frac{1}{\varepsilon}\right)$ (5)	1st/2nd	quadratic (6)
		DANE-HB [50]	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\log\frac{1}{\varepsilon}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\frac{\delta}{\mu}\log\frac{1}{\varepsilon}\right)$ (5)	1st/2nd	quadratic (6)
		SONATA [45]	$\mathcal{O}\left(\frac{\delta}{\mu}\log\frac{1}{\varepsilon}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\sqrt{\frac{\delta}{\mu}}\log^2\frac{1}{\varepsilon}\right)$ (2)	1st	decentralized
		SPAG [21]	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}\right)$ (1)	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\sqrt{\frac{L}{\delta}}\log^2\frac{1}{\varepsilon}\right)$ (1,2)	1st	M - Lipshitz hessian
		DiRegINA [12]	$\mathcal{O}\left(\frac{\delta}{\mu}\log\frac{1}{\varepsilon} + \sqrt{\frac{M\delta R_0}{\mu}}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\sqrt{\frac{\delta}{\mu}}\log^2\frac{1}{\varepsilon} + \sqrt{\frac{MLR_0}{\mu}}\log\frac{1}{\varepsilon}\right)$ (2)	2nd	M -Lipshitz hessian
		ACN [1]	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\log\frac{1}{\varepsilon} + \sqrt[3]{\frac{M\delta R_0}{\mu}}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log^2\frac{1}{\varepsilon} + \sqrt[3]{\frac{M\delta R_0}{\mu}}\sqrt{\frac{L}{\delta}}\log\frac{1}{\varepsilon}\right)$ (2)	2nd	M -Lipshitz hessian
		AccSONATA [46]	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\log\frac{1}{\varepsilon}\log\frac{L}{\mu}\right)$	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\log^2\frac{1}{\varepsilon}\log\frac{\delta}{\mu}\right)$ (2)	1st	decentralized
		This paper	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\log\frac{1}{\varepsilon}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}\right)$	1st	
	Lower	[4]	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\log\frac{1}{\varepsilon}\right)$	—		
		[37]	—	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}\right)$		non-distributed

2014 — 1st method

Similarity

$$\|\nabla^2 f_i(x) - \nabla^2 f(x)\| \leq \delta$$

local global

For uniform data similarity parameter is **small**

$$\delta = \tilde{O}(1/\sqrt{n})$$

n – number of local samples

Long similarity story

		Reference	Communication complexity	Local gradient complexity	Order	Limitations
Minimization	Upper	DANE [42]	$\mathcal{O}\left(\frac{\delta^2}{\mu^2}\log\frac{1}{\varepsilon}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\sqrt{\frac{\delta^3}{\mu^3}}\log^2\frac{1}{\varepsilon}\right)$ (2)	1st	quadratic
		DiSCO [51]	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}(\log\frac{1}{\varepsilon} + C^2\Delta F_0)\log\frac{L}{\mu}\right)$	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}(\log\frac{1}{\varepsilon} + C^2\Delta F_0)\log\frac{L}{\mu}\right)$	2nd	C - self-concordant (3)
		AIDE [40]	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\log\frac{1}{\varepsilon}\log\frac{L}{\delta}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\sqrt{\frac{\delta}{\mu}}\log\frac{1}{\varepsilon}\log\frac{L}{\delta}\right)$ (4)	1st	quadratic
		DANE-LS [50]	$\mathcal{O}\left(\frac{\delta}{\mu}\log\frac{1}{\varepsilon}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\frac{\delta^{3/2}}{\mu^{3/2}}\log\frac{1}{\varepsilon}\right)$ (5)	1st/2nd	quadratic (6)
		DANE-HB [50]	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\log\frac{1}{\varepsilon}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\frac{\delta}{\mu}\log\frac{1}{\varepsilon}\right)$ (5)	1st/2nd	quadratic (6)
		SONATA [45]	$\mathcal{O}\left(\frac{\delta}{\mu}\log\frac{1}{\varepsilon}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\sqrt{\frac{\delta}{\mu}}\log^2\frac{1}{\varepsilon}\right)$ (2)	1st	decentralized
		SPAG [21]	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}\right)$ (1)	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\sqrt{\frac{L}{\delta}}\log^2\frac{1}{\varepsilon}\right)$ (1,2)	1st	M - Lipshitz hessian
		DiRegINA [12]	$\mathcal{O}\left(\frac{\delta}{\mu}\log\frac{1}{\varepsilon} + \sqrt{\frac{M\delta R_0}{\mu}}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\sqrt{\frac{\delta}{\mu}}\log^2\frac{1}{\varepsilon} + \sqrt{\frac{MLR_0}{\mu}}\log\frac{1}{\varepsilon}\right)$ (2)	2nd	M -Lipshitz hessian
		ACN [1]	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\log\frac{1}{\varepsilon} + \sqrt[3]{\frac{M\delta R_0}{\mu}}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log^2\frac{1}{\varepsilon} + \sqrt[3]{\frac{M\delta R_0}{\mu}}\sqrt{\frac{L}{\delta}}\log\frac{1}{\varepsilon}\right)$ (2)	2nd	M -Lipshitz hessian
		AccSONATA [46]	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\log\frac{1}{\varepsilon}\log\frac{L}{\mu}\right)$	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\log^2\frac{1}{\varepsilon}\log\frac{\delta}{\mu}\right)$ (2)	1st	decentralized
		This paper	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\log\frac{1}{\varepsilon}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}\right)$	1st	
	Lower	[4]	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\log\frac{1}{\varepsilon}\right)$	—		
		[37]	—	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}\right)$		non-distributed

2014 — 1st method

2015 — lower bounds

Similarity

$$\|\nabla^2 f_i(x) - \nabla^2 f(x)\| \leq \delta$$

local   global

For uniform data similarity parameter is **small**

$$\delta = \tilde{O}(1/\sqrt{n})$$

n – number of local samples

Long similarity story

		Reference	Communication complexity	Local gradient complexity	Order	Limitations
Minimization	Upper	DANE [42]	$\mathcal{O}\left(\frac{\delta^2}{\mu^2}\log\frac{1}{\varepsilon}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\sqrt{\frac{\delta^3}{\mu^3}}\log^2\frac{1}{\varepsilon}\right)$ (2)	1st	quadratic
		DiSCO [51]	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}(\log\frac{1}{\varepsilon} + C^2\Delta F_0)\log\frac{L}{\mu}\right)$	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}(\log\frac{1}{\varepsilon} + C^2\Delta F_0)\log\frac{L}{\mu}\right)$	2nd	C - self-concordant (3)
		AIDE [40]	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\log\frac{1}{\varepsilon}\log\frac{L}{\delta}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\sqrt{\frac{\delta}{\mu}}\log\frac{1}{\varepsilon}\log\frac{L}{\delta}\right)$ (4)	1st	quadratic
		DANE-LS [50]	$\mathcal{O}\left(\frac{\delta}{\mu}\log\frac{1}{\varepsilon}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\frac{\delta^{3/2}}{\mu^{3/2}}\log\frac{1}{\varepsilon}\right)$ (5)	1st/2nd	quadratic (6)
		DANE-HB [50]	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\log\frac{1}{\varepsilon}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\frac{\delta}{\mu}\log\frac{1}{\varepsilon}\right)$ (5)	1st/2nd	quadratic (6)
		SONATA [45]	$\mathcal{O}\left(\frac{\delta}{\mu}\log\frac{1}{\varepsilon}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\sqrt{\frac{\delta}{\mu}}\log^2\frac{1}{\varepsilon}\right)$ (2)	1st	decentralized
		SPAG [21]	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}\right)$ (1)	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\sqrt{\frac{L}{\delta}}\log^2\frac{1}{\varepsilon}\right)$ (1,2)	1st	M - Lipshitz hessian
		DiRegINA [12]	$\mathcal{O}\left(\frac{\delta}{\mu}\log\frac{1}{\varepsilon} + \sqrt{\frac{M\delta R_0}{\mu}}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\sqrt{\frac{\delta}{\mu}}\log^2\frac{1}{\varepsilon} + \sqrt{\frac{MLR_0}{\mu}}\log\frac{1}{\varepsilon}\right)$ (2)	2nd	M -Lipshitz hessian
		ACN [1]	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\log\frac{1}{\varepsilon} + \sqrt[3]{\frac{M\delta R_0}{\mu}}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log^2\frac{1}{\varepsilon} + \sqrt[3]{\frac{M\delta R_0}{\mu}}\sqrt{\frac{L}{\delta}}\log\frac{1}{\varepsilon}\right)$ (2)	2nd	M -Lipshitz hessian
		AccSONATA [46]	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\log\frac{1}{\varepsilon}\log\frac{L}{\mu}\right)$	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\log^2\frac{1}{\varepsilon}\log\frac{\delta}{\mu}\right)$ (2)	1st	decentralized
		This paper	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\log\frac{1}{\varepsilon}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}\right)$	1st	
	Lower	[4]	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\log\frac{1}{\varepsilon}\right)$	—		
		[37]	—	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}\right)$		non-distributed

2014 — 1st method

2015 - 2022 — no optimal methods in general

2015 — lower bounds

Similarity

$$\|\nabla^2 f_i(x) - \nabla^2 f(x)\| \leq \delta$$

local   global

For uniform data similarity parameter is **small**

$$\delta = \tilde{O}(1/\sqrt{n})$$

n – number of local samples

Long similarity story

		Reference	Communication complexity	Local gradient complexity	Order	Limitations
Minimization	Upper	DANE [42]	$\mathcal{O}\left(\frac{\delta^2}{\mu^2}\log\frac{1}{\varepsilon}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\sqrt{\frac{\delta^3}{\mu^3}}\log^2\frac{1}{\varepsilon}\right)$ (2)	1st	quadratic
		DiSCO [51]	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}(\log\frac{1}{\varepsilon} + C^2\Delta F_0)\log\frac{L}{\mu}\right)$	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}(\log\frac{1}{\varepsilon} + C^2\Delta F_0)\log\frac{L}{\mu}\right)$	2nd	C - self-concordant (3)
		AIDE [40]	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\log\frac{1}{\varepsilon}\log\frac{L}{\delta}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\sqrt{\frac{\delta}{\mu}}\log\frac{1}{\varepsilon}\log\frac{L}{\delta}\right)$ (4)	1st	quadratic
		DANE-LS [50]	$\mathcal{O}\left(\frac{\delta}{\mu}\log\frac{1}{\varepsilon}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\frac{\delta^{3/2}}{\mu^{3/2}}\log\frac{1}{\varepsilon}\right)$ (5)	1st/2nd	quadratic (6)
		DANE-HB [50]	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\log\frac{1}{\varepsilon}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\frac{\delta}{\mu}\log\frac{1}{\varepsilon}\right)$ (5)	1st/2nd	quadratic (6)
		SONATA [45]	$\mathcal{O}\left(\frac{\delta}{\mu}\log\frac{1}{\varepsilon}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\sqrt{\frac{\delta}{\mu}}\log^2\frac{1}{\varepsilon}\right)$ (2)	1st	decentralized
		SPAG [21]	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}\right)$ (1)	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\sqrt{\frac{L}{\delta}}\log^2\frac{1}{\varepsilon}\right)$ (1,2)	1st	M - Lipshitz hessian
		DiRegINA [12]	$\mathcal{O}\left(\frac{\delta}{\mu}\log\frac{1}{\varepsilon} + \sqrt{\frac{M\delta R_0}{\mu}}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\sqrt{\frac{\delta}{\mu}}\log^2\frac{1}{\varepsilon} + \sqrt{\frac{MLR_0}{\mu}}\log\frac{1}{\varepsilon}\right)$ (2)	2nd	M -Lipshitz hessian
		ACN [1]	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\log\frac{1}{\varepsilon} + \sqrt[3]{\frac{M\delta R_0}{\mu}}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log^2\frac{1}{\varepsilon} + \sqrt[3]{\frac{M\delta R_0}{\mu}}\sqrt{\frac{L}{\delta}}\log\frac{1}{\varepsilon}\right)$ (2)	2nd	M -Lipshitz hessian
		AccSONATA [46]	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\log\frac{1}{\varepsilon}\log\frac{L}{\mu}\right)$	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\log^2\frac{1}{\varepsilon}\log\frac{\delta}{\mu}\right)$ (2)	1st	decentralized
		This paper	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\log\frac{1}{\varepsilon}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}\right)$	1st	
	Lower	[4]	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\log\frac{1}{\varepsilon}\right)$	—		
		[37]	—	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}\right)$		non-distributed

2014 — 1st method

We present optimal method in communications and local computations!

2015 — lower bounds

Idea

$$\min_{x\in\mathbb{R}^d} r(x) = \underbrace{f_1(x)}_{:=q(x)} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}[f_i(x) - f_1(x)]}_{:=p(x)}$$

Idea

$$\min_{x \in \mathbb{R}^d} r(x) = \underbrace{f_1(x)}_{:=q(x)} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}[f_i(x) - f_1(x)]}_{:=p(x)}$$

μ -strongly-convex

L -smooth and convex

δ -smooth and non-convex

Idea

only local computations

communications

$$\min_{x \in \mathbb{R}^d} r(x) = \underbrace{f_1(x)}_{:=q(x)} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}[f_i(x) - f_1(x)]}_{:=p(x)}$$

μ -strongly-convex

L -smooth and convex

δ -smooth and non-convex

Algorithm

Algorithm 1 Accelerated Extragradient

Input: $x^0 = x_f^0 \in \mathbb{R}^d$
Parameters: $\tau \in (0, 1], \eta, \theta, \alpha > 0, K \in \{1, 2, \ldots\}$
for $k = 0, 1, 2, \ldots, K-1$ **do**
  $x_g^k = \tau x^k + (1-\tau) x_f^k$
  $x_f^{k+1} \approx \arg\min_{x \in \mathbb{R}^d} \left[A_\theta^k(x) := p(x_g^k) + \langle \nabla p(x_g^k), x - x_g^k \rangle + \frac{1}{2\theta} \|x - x_g^k\|^2 + q(x) \right]$
  $x^{k+1} = x^k + \eta\alpha(x_f^{k+1} - x^k) - \eta \nabla r(x_f^{k+1})$
end for
Output: x^K

Algorithm

Algorithm 1 Accelerated Extragradient

Input: $x^0 = x_f^0 \in \mathbb{R}^d$
Parameters: $\tau \in (0,1], \eta, \theta, \alpha > 0, K \in \{1, 2, \ldots\}$
for $k = 0, 1, 2, \ldots, K-1$ **do**
  $x_g^k = \tau x^k + (1-\tau) x_f^k$
  $x_f^{k+1} \approx \arg\min_{x \in \mathbb{R}^d} \left[A_\theta^k(x) := p(x_g^k) + \langle \nabla p(x_g^k), x - x_g^k \rangle + \frac{1}{2\theta}\|x - x_g^k\|^2 + q(x) \right]$
  $x^{k+1} = x^k + \eta\alpha(x_f^{k+1} - x^k) - \eta \nabla r(x_f^{k+1})$
end for
Output: x^K

3 ideas:

- Extragradient: 2 steps per iteration

Algorithm

Algorithm 1 Accelerated Extragradient

Input: $x^0 = x_f^0 \in \mathbb{R}^d$
Parameters: $\tau \in (0, 1], \eta, \theta, \alpha > 0, K \in \{1, 2, \dots\}$
for $k = 0, 1, 2, \dots, K-1$ **do**
$\quad x_g^k = \tau x^k + (1-\tau) x_f^k$
$\quad x_f^{k+1} \boxed{\approx \arg\min_{x\in\mathbb{R}^d}} \left[A_\theta^k(x) := p(x_g^k) + \boxed{\langle \nabla p(x_g^k),} x - x_g^k\rangle + \frac{1}{2\theta}\|x - x_g^k\|^2 + \boxed{q(x)}\right]$
$\quad x^{k+1} = x^k + \eta\alpha(x_f^{k+1} - x^k) - \eta\nabla r(x_f^{k+1})$
end for
Output: x^K

3 ideas:

- Extragradient: 2 steps per iteration
- Sliding (inexact prox)

Algorithm

Algorithm 1 Accelerated Extragradient

1: **Input:** $x^0 = x_f^0 \in \mathbb{R}^d$
2: **Parameters:** $\tau \in (0, 1], \eta, \theta, \alpha > 0, K \in \{1, 2, \ldots\}$
3: **for** $k = 0, 1, 2, \ldots, K-1$ **do**
4: $\quad \boxed{x_g^k = \tau x^k + (1-\tau) x_f^k}$
5: $\quad x_f^{k+1} \approx \arg\min_{x \in \mathbb{R}^d} \left[A_\theta^k(x) := p(x_g^k) + \langle \nabla p(x_g^k), x - x_g^k \rangle + \frac{1}{2\theta}\|x - x_g^k\|^2 + q(x) \right]$
6: $\quad x^{k+1} = x^k + \boxed{\eta\alpha(x_f^{k+1} - x^k)} - \eta \nabla r(x_f^{k+1})$
7: **end for**
8: **Output:** x^K

3 ideas:

- Extragradient: 2 steps per iteration
- Sliding (inexact prox)
- Acceleration

Convergence

Theorem. Let assumptions from previous slides be satisfied with $\mu \leq \delta \leq L$. Then, to find ε-solution of the distributed optimization problem Algorithm 1 requires

$$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}}\log\frac{1}{\varepsilon}\right) \text{ communication rounds and } \mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}\right) \text{ local gradient computations.}$$

Convergence

Theorem. Let assumptions from previous slides be satisfied with $\mu \leq \delta \leq L$. Then, to find ε-solution of the distributed optimization problem Algorithm 1 requires

$$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}} \log \frac{1}{\varepsilon}\right) \text{ communication rounds and } \mathcal{O}\left(\sqrt{\frac{L}{\mu}} \log \frac{1}{\varepsilon}\right) \text{ local gradient computations.}$$

Optimal estimates in terms of communications and local computations

Thank you!