

Optimal Gradient Sliding and its Application to Distributed Optimization Under Similarity



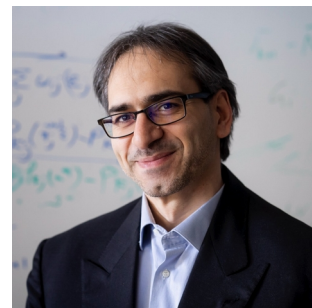
Dmitry Kovalev
KAUST



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

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

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

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

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

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

The diagram illustrates the decomposition of a minimization problem. The equation $\min_{x \in \mathbb{R}^d} r(x) = f_1(x) + \frac{1}{n} \sum_{i=1}^n [f_i(x) - f_1(x)]$ is shown. An orange arrow points from the text 'only local computations' to the term $f_1(x)$, which is underlined and labeled $:=q(x)$. Another orange arrow points from the text 'communications' to the summation term $\frac{1}{n} \sum_{i=1}^n [f_i(x) - f_1(x)]$, which is underlined and labeled $:=p(x)$. A third orange arrow points from the text ' μ -strongly-convex' to the domain $x \in \mathbb{R}^d$. A fourth orange arrow points from the text ' L -smooth and convex' to the underlined $f_1(x)$. A fifth orange arrow points from the text ' δ -smooth and non-convex' to the underlined summation term.

Algorithm

Algorithm 1 Accelerated Extragradient

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

Algorithm

Algorithm 1 Accelerated Extragradient

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

3 ideas:

- Extragradient: 2 steps per iteration

Algorithm

Algorithm 1 Accelerated Extragradient

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

3 ideas:

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

Algorithm

Algorithm 1 Accelerated Extragradient

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

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

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

**Optimal estimates in terms of
communications and local computations**

Thank you!