# Centralized convex optimization with similarity

**Alexander Gasnikov**

**gasnikov.av@mipt.ru**

*Based on the joint works with A. Beznosikov, A. Agafonov, A. Scutari, D. Kamzolov. P. Dvurechensky et al.*

**arXiv:2103.14392 arXiv:2102.08246 arXiv:2102.06780**

Sirius; July 14, 2021

# Structure of the Lecture

**1. Preliminaries from Machine Learning and Convex optimization**

**2. Monte Carlo approach and sum-type optimization problems**

**3. Variance reduction (VR) vs Statistical Similarity**

**4. Gradient descent with relative smoothness and strong convexity and relation with Similarity setup**

**5. Lower bound for Similarity and the problem of acceleration**

**6. Partial acceleration is possible via second-order methods**

**7. Sum-type Saddle-point problems (SPP) - no need of acceleration. Optimal method**

**8. Open problem: Distributed VR for SaSPP**

# Sum type target convex function

$$f(x) = \mathbb{E}[f(x,\xi)] \to \min_{x \in Q \subseteq \mathbb{R}^n}.$$

How to choose $m$ in:

$$\min_{x \in Q \subseteq \mathbb{R}^n} \frac{1}{m}\sum_{k=1}^{m} f(x,\xi^k) + \frac{\varepsilon}{2R^2}\|x - x^0\|_2^2$$

$\|x^0 - x_*\|_2$

Answer (up to a log-factor):

$$m = \min\left\{ O\left(\frac{M^2R^2}{\varepsilon^2}\right), O\left(\frac{M^2}{\mu\varepsilon}\right) \right\}$$

Where:

Strong convexity constant of $f$

$$\mathbb{E}[\|\nabla f(x,\xi)\|_2^2] \le M^2$$

S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Stochastic convex optimization. In *COLT*, 2009.

# Too many terms!

$$\min_{x\in Q\subseteq\mathbb{R}^n} \frac{1}{m}\sum_{k=1}^{m} f(x,\xi^k) + \frac{\varepsilon}{2R^2}\|x-x^0\|_2^2$$

We skip this term for simplicity

Problem: How to store data in memory?

Answer: To use distributed approach

Problem: Too many communications and oracle calls required

Answer: To store at each node a lot of data. And rewrite optimization problem

$$\min_{x\in Q} F(x) := \frac{1}{M}\sum_{k=1}^{M} \boxed{\frac{1}{r}\sum_{i=1}^{r} f\left(x,\xi^{k,i}\right)}$$

$$F_k(x) = \frac{1}{r}\sum_{i=1}^{r} f\left(x,\xi^{k,i}\right)$$

This problem has specific structure

1) $F_k(x)$ has sum-type structure and variance reduction is possible

2) Statistical Similarity $\|\nabla^2 F_k(x) - \nabla^2 F(x)\| = O\left(\sqrt{\frac{L_2^2 n}{r}}\right)$

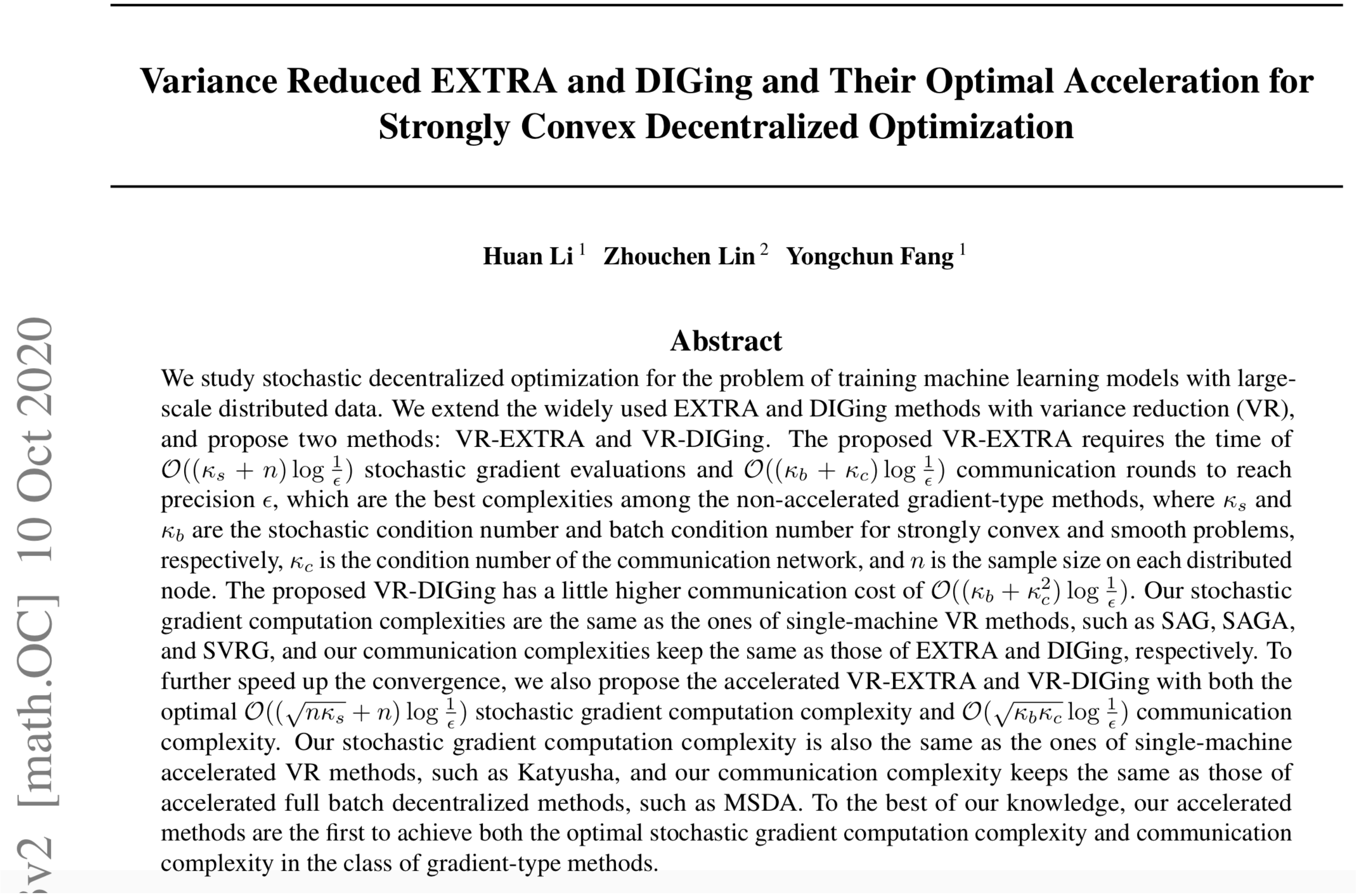
# Variance reduction

$$\min_{x\in Q} F(x) := \frac{1}{M}\sum_{k=1}^{M}\boxed{\frac{1}{r}\sum_{i=1}^{r} f\left(x,\xi^{k,i}\right)}$$

This problem has specific structure

$$F_k(x) = \frac{1}{r}\sum_{i=1}^{r} f\left(x,\xi^{k,i}\right)$$

1) $F_k(x)$ has sum-type structure and variance reduction is possible



**Variance Reduced EXTRA and DIGing and Their Optimal Acceleration for Strongly Convex Decentralized Optimization**

Huan Li[1] Zhouchen Lin[2] Yongchun Fang[1]

### Abstract

We study stochastic decentralized optimization for the problem of training machine learning models with large-scale distributed data. We extend the widely used EXTRA and DIGing methods with variance reduction (VR), and propose two methods: VR-EXTRA and VR-DIGing. The proposed VR-EXTRA requires the time of $\mathcal{O}((\kappa_s + n)\log\frac{1}{\epsilon})$ stochastic gradient evaluations and $\mathcal{O}((\kappa_b + \kappa_c)\log\frac{1}{\epsilon})$ communication rounds to reach precision $\epsilon$, which are the best complexities among the non-accelerated gradient-type methods, where $\kappa_s$ and $\kappa_b$ are the stochastic condition number and batch condition number for strongly convex and smooth problems, respectively, $\kappa_c$ is the condition number of the communication network, and $n$ is the sample size on each distributed node. The proposed VR-DIGing has a little higher communication cost of $\mathcal{O}((\kappa_b + \kappa_c^2)\log\frac{1}{\epsilon})$. Our stochastic gradient computation complexities are the same as the ones of single-machine VR methods, such as SAG, SAGA, and SVRG, and our communication complexities keep the same as those of EXTRA and DIGing, respectively. To further speed up the convergence, we also propose the accelerated VR-EXTRA and VR-DIGing with both the optimal $\mathcal{O}((\sqrt{n\kappa_s} + n)\log\frac{1}{\epsilon})$ stochastic gradient computation complexity and $\mathcal{O}(\sqrt{\kappa_b\kappa_c}\log\frac{1}{\epsilon})$ communication complexity. Our stochastic gradient computation complexity is also the same as the ones of single-machine accelerated VR methods, such as Katyusha, and our communication complexity keeps the same as those of accelerated full batch decentralized methods, such as MSDA. To the best of our knowledge, our accelerated methods are the first to achieve both the optimal stochastic gradient computation complexity and communication complexity in the class of gradient-type methods.

# Variance reduction

$$\min_{x \in Q} F(x) := \frac{1}{M}\sum_{k=1}^{M} \boxed{\frac{1}{r}\sum_{i=1}^{r} f\left(x, \xi^{k,i}\right)}$$

This problem has specific structure

1) $F_k(x)$ has sum-type structure and variance reduction is possible

Assumptions:

$f(x, \xi)$ is $L$-smooth in $x$ for all $\xi$

$f(x, \xi)$ is $\lambda$-strongly convex in $x$ for all $\xi$

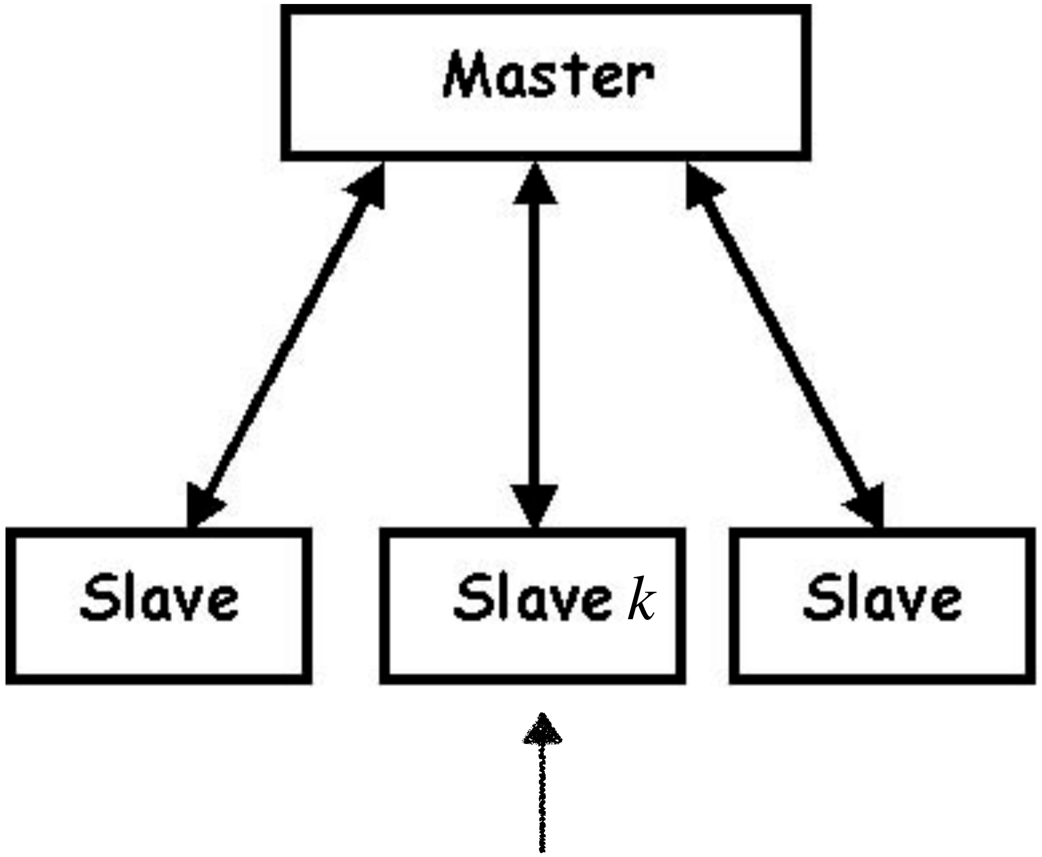
Optimal bounds for general $f_{k,i}(x) \neq f(x, \xi^{k,i})$:

$$O\left(\sqrt{\frac{L}{\lambda}} \log\left(\frac{LR^2}{\varepsilon}\right)\right)$$

Communication rounds

$$O\left(\left(r + \sqrt{r\frac{L}{\lambda}}\right) \log\left(\frac{LR^2}{\varepsilon}\right)\right)$$

Oracle calls per node



$$F_k(x) = \frac{1}{r}\sum_{i=1}^{r} f\left(x, \xi^{k,i}\right)$$

# Statistical Similarity

$$\min_{x\in Q} F(x) := \frac{1}{M}\sum_{k=1}^{M}\frac{1}{r}\sum_{i=1}^{r} f\left(x, \xi^{k,i}\right)$$

This problem has specific structure

1) $F_k(x)$ has sum-type structure and variance reduction is possible

Assumptions:

$f(x,\xi)$ is $L$-smooth in $x$ for all $\xi$

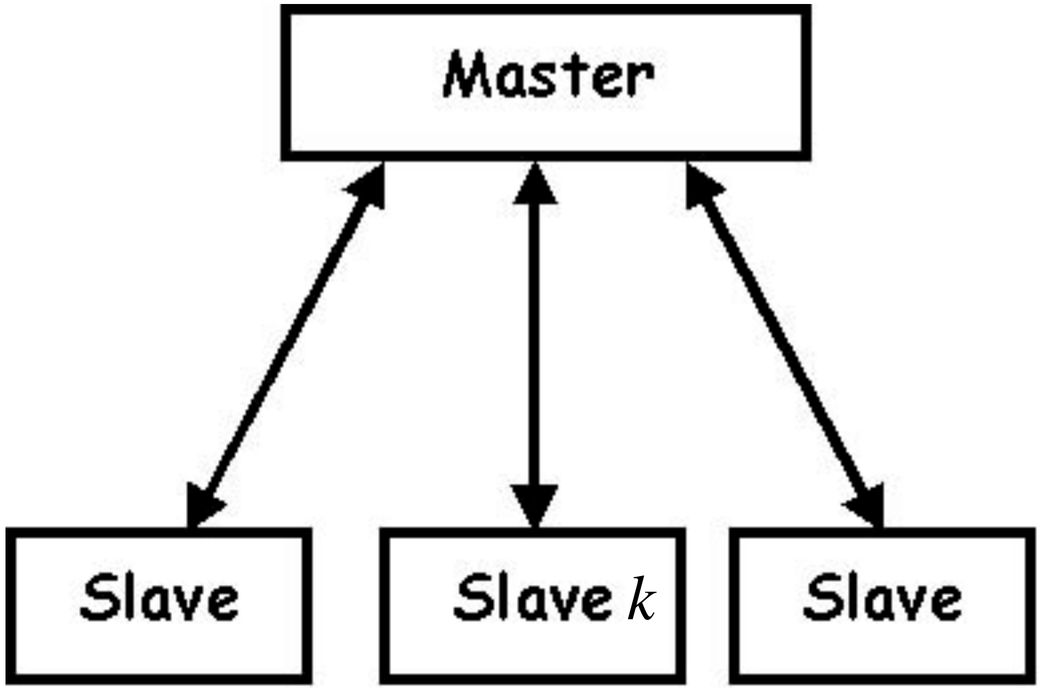
$f(x,\xi)$ is $\lambda$-strongly convex in $x$ for all $\xi$

Optimal bound ?:

$$O\left(\sqrt{\frac{L}{\lambda}}\log\left(\frac{LR^2}{\varepsilon}\right)\right)$$

Communication rounds

Is this bound optimal?



$$F_k(x) = \frac{1}{r}\sum_{i=1}^{r} f\left(x, \xi^{k,i}\right)$$

Answer: No, if we use similarity: $f_{k,i}(x) = f(x,\xi^{k,i})$, $\xi^{k,i}$ i.i.d.

# Statistical Similarity

$$\min_{x \in Q} F(x) := \frac{1}{M}\sum_{k=1}^{M} \frac{1}{r}\sum_{i=1}^{r} f\left(x, \xi^{k,i}\right)$$

$$F_k(x) = \frac{1}{r}\sum_{i=1}^{r} f\left(x, \xi^{k,i}\right)$$

This problem has specific structure

1) $F_k(x)$ has sum-type structure and variance reduction is possible

Assumptions:

$f(x,\xi)$ is $L$-smooth in $x$ for all $\xi \Rightarrow \|\nabla^2 F_k(x)\| \le L$

$f(x,\xi)$ is $\lambda$-strongly convex in $x$ for all $\xi$

Communication rounds

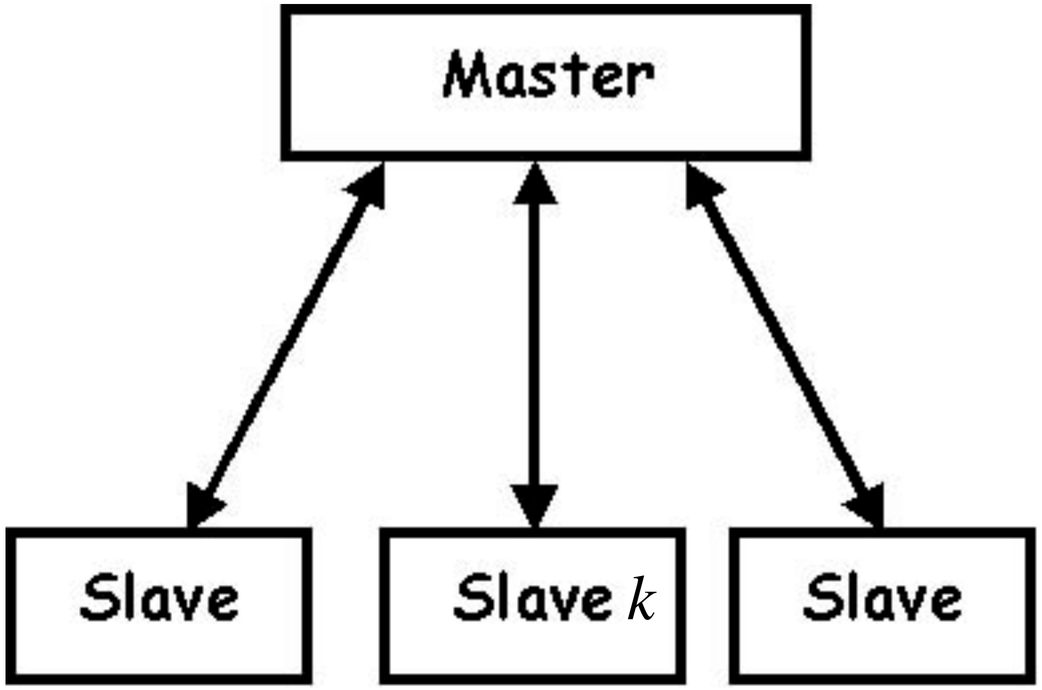
$$O\left(\sqrt{\frac{L}{\lambda}}\log\left(\frac{LR^2}{\varepsilon}\right)\right)$$

Variance reduction

$$O\left(\sqrt{\frac{\delta}{\lambda}}\log\left(\frac{LR^2}{\varepsilon}\right)\right)$$

$\|\nabla^2 F_k(x) - \nabla^2 F(x)\| \le \delta$

$\delta$-Similarity



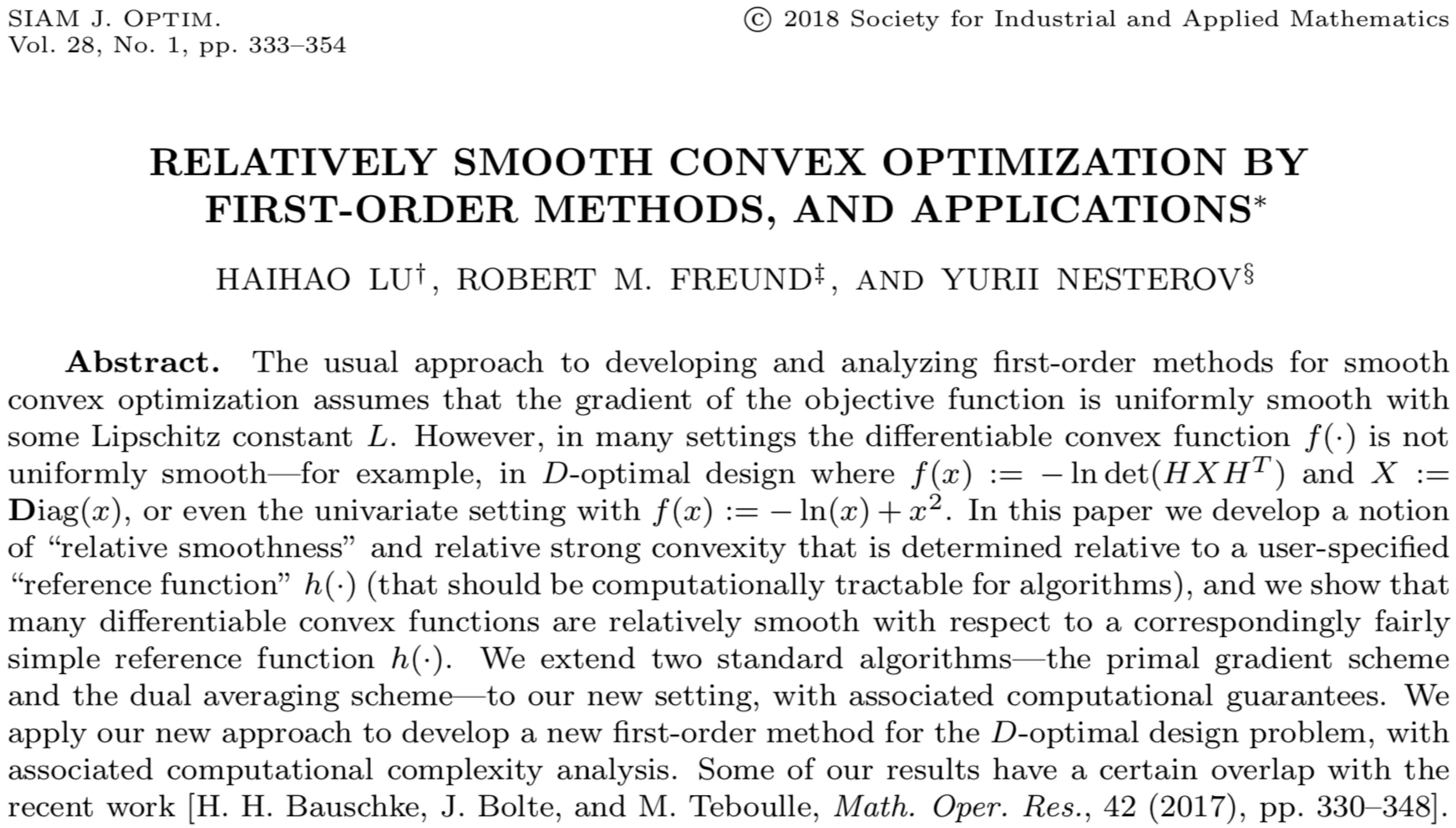
$$F_k(x) = \frac{1}{r}\sum_{i=1}^{r} f\left(x, \xi^{k,i}\right)$$

# Gradient method with relative smoothness and strong convexity

$$\min_{x \in Q} F(x) := \frac{1}{M} \sum_{k=1}^{M} F_k(x)$$

$d(x)$ - Smooth convex function

$V(y, x) = d(y) - d(x) - \langle \nabla d(x), y - x \rangle$ - Bregman divergence



SIAM J. Optim.
Vol. 28, No. 1, pp. 333–354

## RELATIVELY SMOOTH CONVEX OPTIMIZATION BY FIRST-ORDER METHODS, AND APPLICATIONS*

HAIHAO LU†, ROBERT M. FREUND‡, AND YURII NESTEROV§

**Abstract.** The usual approach to developing and analyzing first-order methods for smooth convex optimization assumes that the gradient of the objective function is uniformly smooth with some Lipschitz constant $L$. However, in many settings the differentiable convex function $f(\cdot)$ is not uniformly smooth—for example, in $D$-optimal design where $f(x) := -\ln\det(HXH^T)$ and $X := \mathbf{D}\mathrm{iag}(x)$, or even the univariate setting with $f(x) := -\ln(x) + x^2$. In this paper we develop a notion of "relative smoothness" and relative strong convexity that is determined relative to a user-specified "reference function" $h(\cdot)$ (that should be computationally tractable for algorithms), and we show that many differentiable convex functions are relatively smooth with respect to a correspondingly fairly simple reference function $h(\cdot)$. We extend two standard algorithms—the primal gradient scheme and the dual averaging scheme—to our new setting, with associated computational guarantees. We apply our new approach to develop a new first-order method for the $D$-optimal design problem, with associated computational complexity analysis. Some of our results have a certain overlap with the recent work [H. H. Bauschke, J. Bolte, and M. Teboulle, *Math. Oper. Res.*, 42 (2017), pp. 330–348].

# Gradient method with relative smoothness and strong convexity

$$\min_{x \in Q} F(x) := \frac{1}{M} \sum_{k=1}^{M} F_k(x)$$

$$\lambda \nabla^2 d(x) \prec \nabla^2 F(x) \prec L \nabla^2 d(x)$$

$$V(y, x) = d(y) - d(x) - \langle \nabla d(x), y - x \rangle$$

$$x^{k+1} = \arg\min_{x \in Q} \left\{ F(x^k) + \langle \nabla F(x^k), x - x^k \rangle + LV(x, x^k) \right\}$$

$$F(x^N) - \min_{x \in Q} F(x) \leq \varepsilon$$

$$N = O\left( \frac{L}{\lambda} \log \left( \frac{\Delta F}{\varepsilon} \right) \right)$$

# Gradient method with relative smoothness and strong convexity and Similarity

$$\min_{x\in Q} F(x) := \frac{1}{M}\sum_{k=1}^{M} F_k(x)$$

$$d(x) = F_1(x) + \frac{\delta}{2}\|x\|^2$$

Available at master node (1)

$$\frac{\lambda}{\lambda+2\delta}\nabla^2 d(x) \prec \nabla^2 F(x) \prec \nabla^2 d(x)$$

$$x^{k+1} = \arg\min_{x\in Q}\left\{F(x^k) + \langle \nabla F(x^k), x - x^k\rangle + V(x, x^k)\right\}$$

Available at master node (1) via communications

$$F(x^N) - \min_{x\in Q} F(x) \le \varepsilon$$

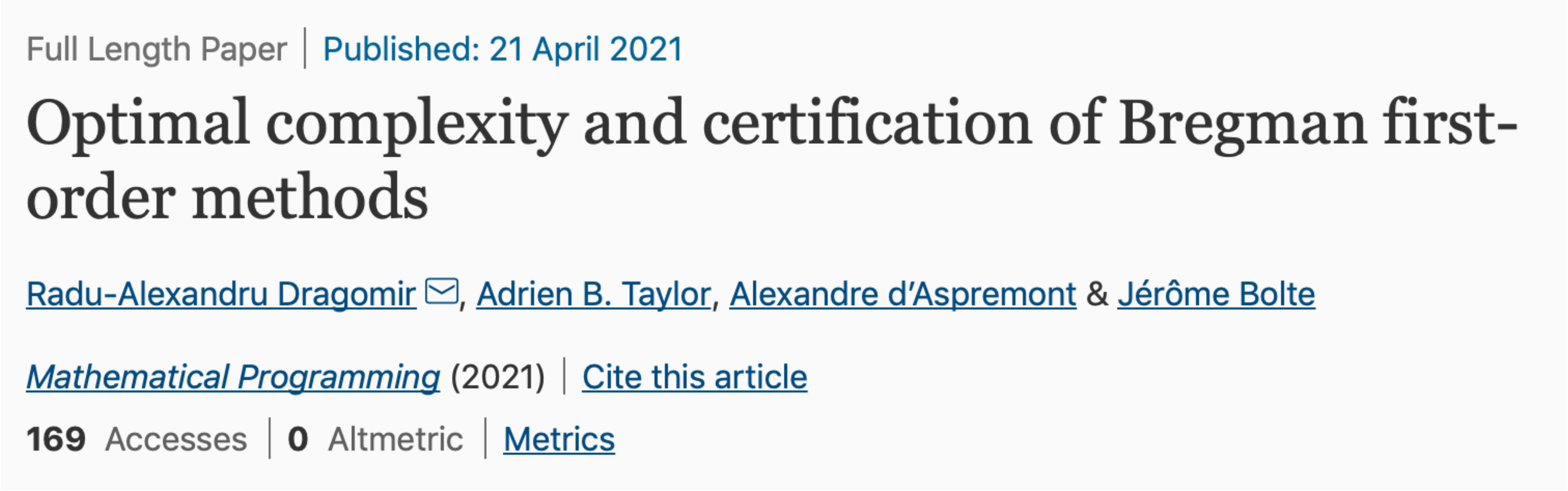
$$N = O\left(\frac{\max\{\delta,\lambda\}}{\lambda}\log\left(\frac{\Delta F}{\varepsilon}\right)\right)$$

# Gradient method with relative smoothness and strong convexity and Similarity

$$N = O\left(\frac{\max\{\delta, \lambda\}}{\lambda}\log\left(\frac{\Delta F}{\varepsilon}\right)\right) = O\left(\frac{\delta}{\lambda}\log\left(\frac{\Delta F}{\varepsilon}\right)\right)$$

Warning: Unfortunately, this rate is not optimal (accelerated)!

But! Accelerated method with relative smoothness and strong convexity is in principle impossible in general set up.



Full Length Paper | Published: 21 April 2021

Optimal complexity and certification of Bregman first-order methods

Radu-Alexandru Dragomir, Adrien B. Taylor, Alexandre d'Aspremont & Jérôme Bolte

*Mathematical Programming* (2021) | Cite this article

169 Accesses | 0 Altmetric | Metrics

# Statistical Similarity

$$\min_{x \in Q} F(x) := \frac{1}{M}\sum_{k=1}^{M} \frac{1}{r}\sum_{i=1}^{r} f\left(x, \xi^{k,i}\right)$$

This problem has specific structure

$$F_k(x) = \frac{1}{r}\sum_{i=1}^{r} f\left(x, \xi^{k,i}\right)$$

1) $F_k(x)$ has sum-type structure and variance reduction is possible

Assumptions:

$f(x, \xi)$ is $L$-smooth in $x$ for all $\xi \Rightarrow \|\nabla^2 F_k(x)\| \leq L$

$f(x, \xi)$ is $\lambda$-strongly convex

Communication rounds

$$O\left(\sqrt{\frac{L}{\lambda}} \log\left(\frac{LR^2}{\varepsilon}\right)\right)$$

Variance reduction

$$\boxed{O\left(\sqrt{\frac{\delta}{\lambda}} \log\left(\frac{LR^2}{\varepsilon}\right)\right)}$$

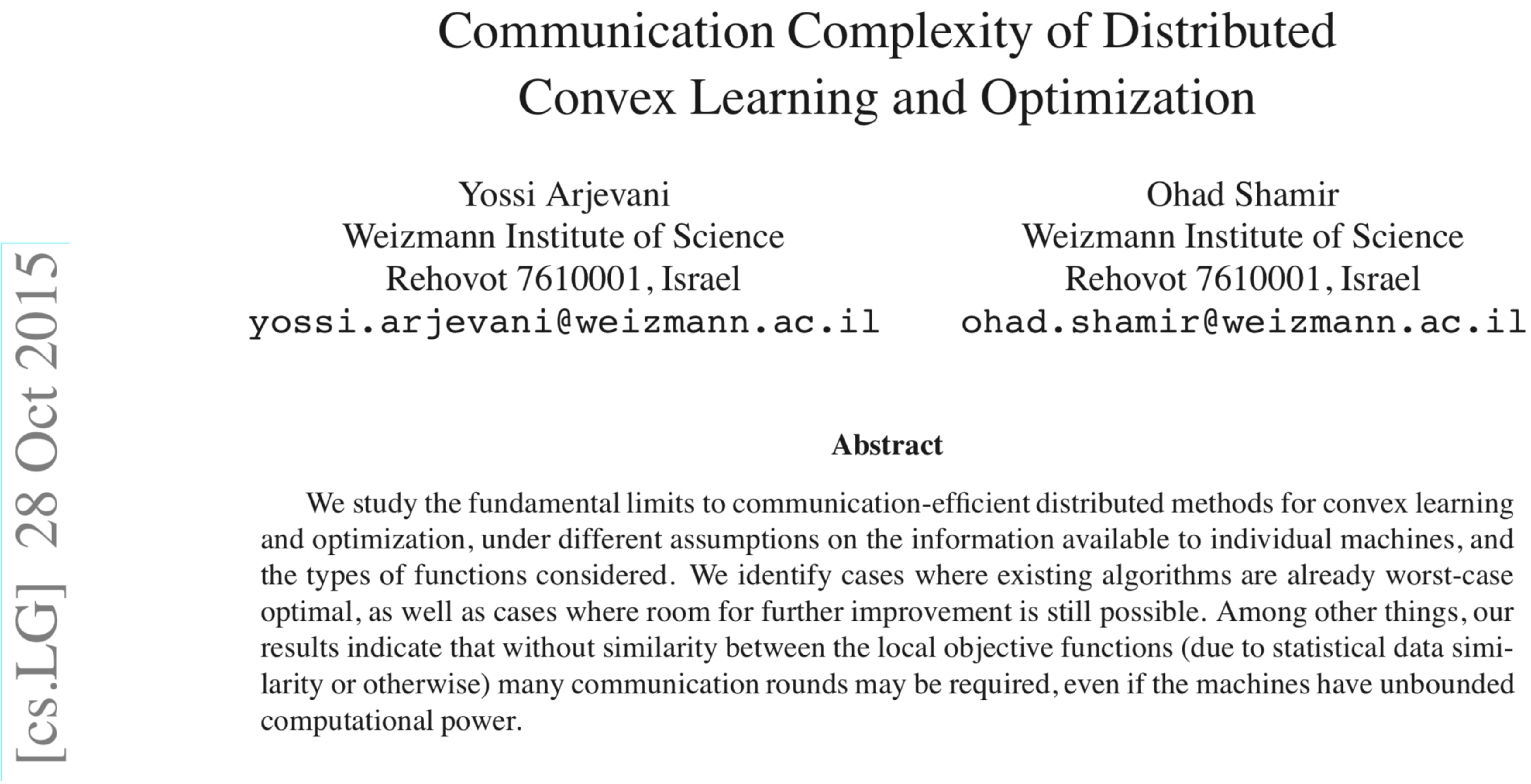
$$\|\nabla^2 F_k(x) - \nabla^2 F(x)\| \leq \delta$$

$\delta$-Similarity

Is this bound tight? That is, can we propose such algorithm that works according to this bound? What is a lower bound?

The answer is - this is the lower bound (up to a log-factors), but is this bound tight or not? - is still an open question in general.

# Lower bound for distributed convex optimization under similarity



## Communication Complexity of Distributed Convex Learning and Optimization

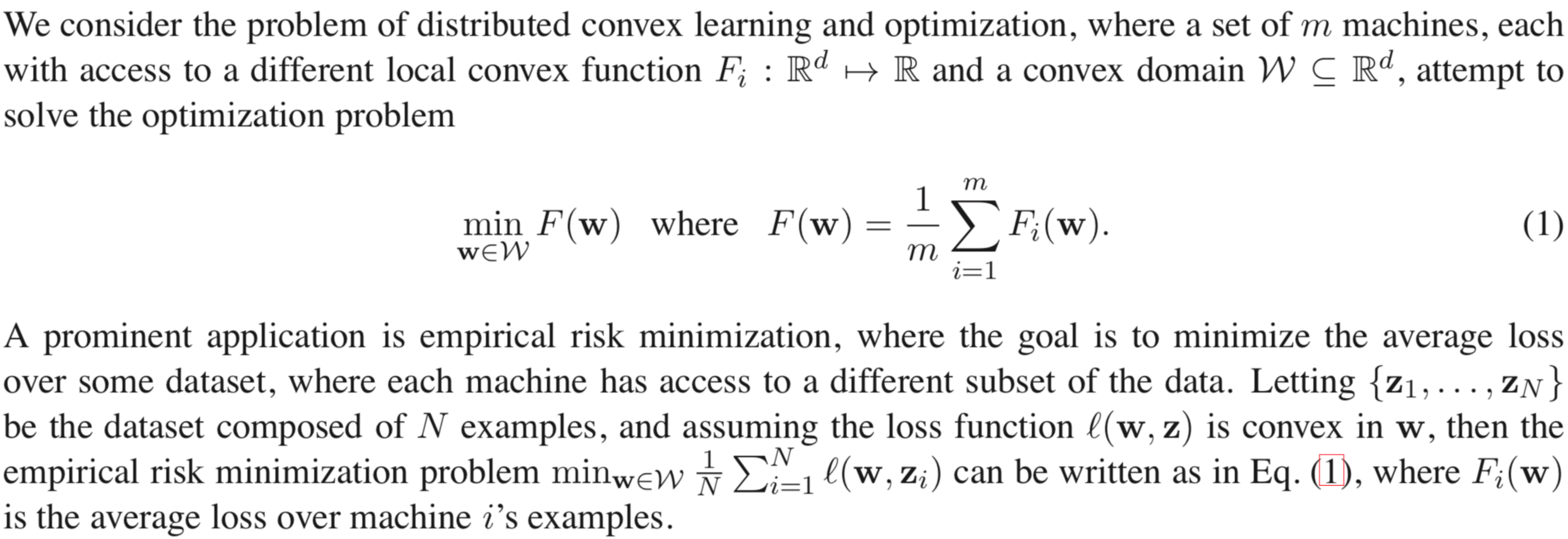
Yossi Arjevani
Weizmann Institute of Science
Rehovot 7610001, Israel
yossi.arjevani@weizmann.ac.il

Ohad Shamir
Weizmann Institute of Science
Rehovot 7610001, Israel
ohad.shamir@weizmann.ac.il

**Abstract**

We study the fundamental limits to communication-efficient distributed methods for convex learning and optimization, under different assumptions on the information available to individual machines, and the types of functions considered. We identify cases where existing algorithms are already worst-case optimal, as well as cases where room for further improvement is still possible. Among other things, our results indicate that without similarity between the local objective functions (due to statistical data similarity or otherwise) many communication rounds may be required, even if the machines have unbounded computational power.

# Lower communication rounds bound for distributed convex optimization under similarity

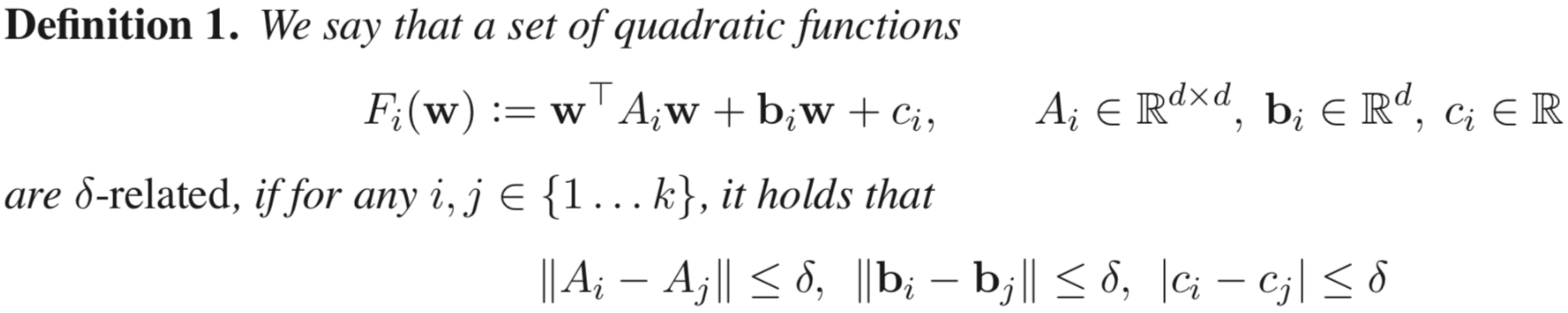


We consider the problem of distributed convex learning and optimization, where a set of $m$ machines, each with access to a different local convex function $F_i : \mathbb{R}^d \mapsto \mathbb{R}$ and a convex domain $\mathcal{W} \subseteq \mathbb{R}^d$, attempt to solve the optimization problem

$$\min_{\mathbf{w} \in \mathcal{W}} F(\mathbf{w}) \quad \text{where} \quad F(\mathbf{w}) = \frac{1}{m} \sum_{i=1}^{m} F_i(\mathbf{w}). \tag{1}$$

A prominent application is empirical risk minimization, where the goal is to minimize the average loss over some dataset, where each machine has access to a different subset of the data. Letting $\{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$ be the dataset composed of $N$ examples, and assuming the loss function $\ell(\mathbf{w}, \mathbf{z})$ is convex in $\mathbf{w}$, then the empirical risk minimization problem $\min_{\mathbf{w} \in \mathcal{W}} \frac{1}{N} \sum_{i=1}^{N} \ell(\mathbf{w}, \mathbf{z}_i)$ can be written as in Eq. (1), where $F_i(\mathbf{w})$ is the average loss over machine $i$'s examples.

# Lower communication rounds bound for distributed convex optimization under similarity

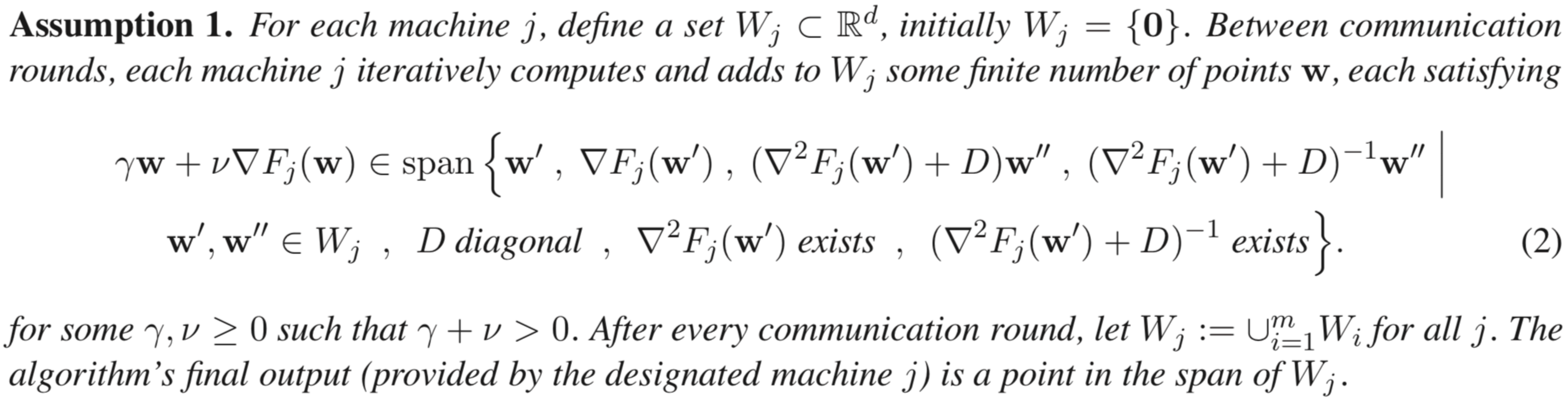


**Definition 1.** *We say that a set of quadratic functions*

$$F_i(\mathbf{w}) := \mathbf{w}^\top A_i \mathbf{w} + \mathbf{b}_i \mathbf{w} + c_i, \qquad A_i \in \mathbb{R}^{d\times d},\ \mathbf{b}_i \in \mathbb{R}^d,\ c_i \in \mathbb{R}$$

*are* $\delta$-related, *if for any* $i, j \in \{1 \ldots k\}$, *it holds that*

$$\|A_i - A_j\| \le \delta,\ \ \|\mathbf{b}_i - \mathbf{b}_j\| \le \delta,\ \ |c_i - c_j| \le \delta$$

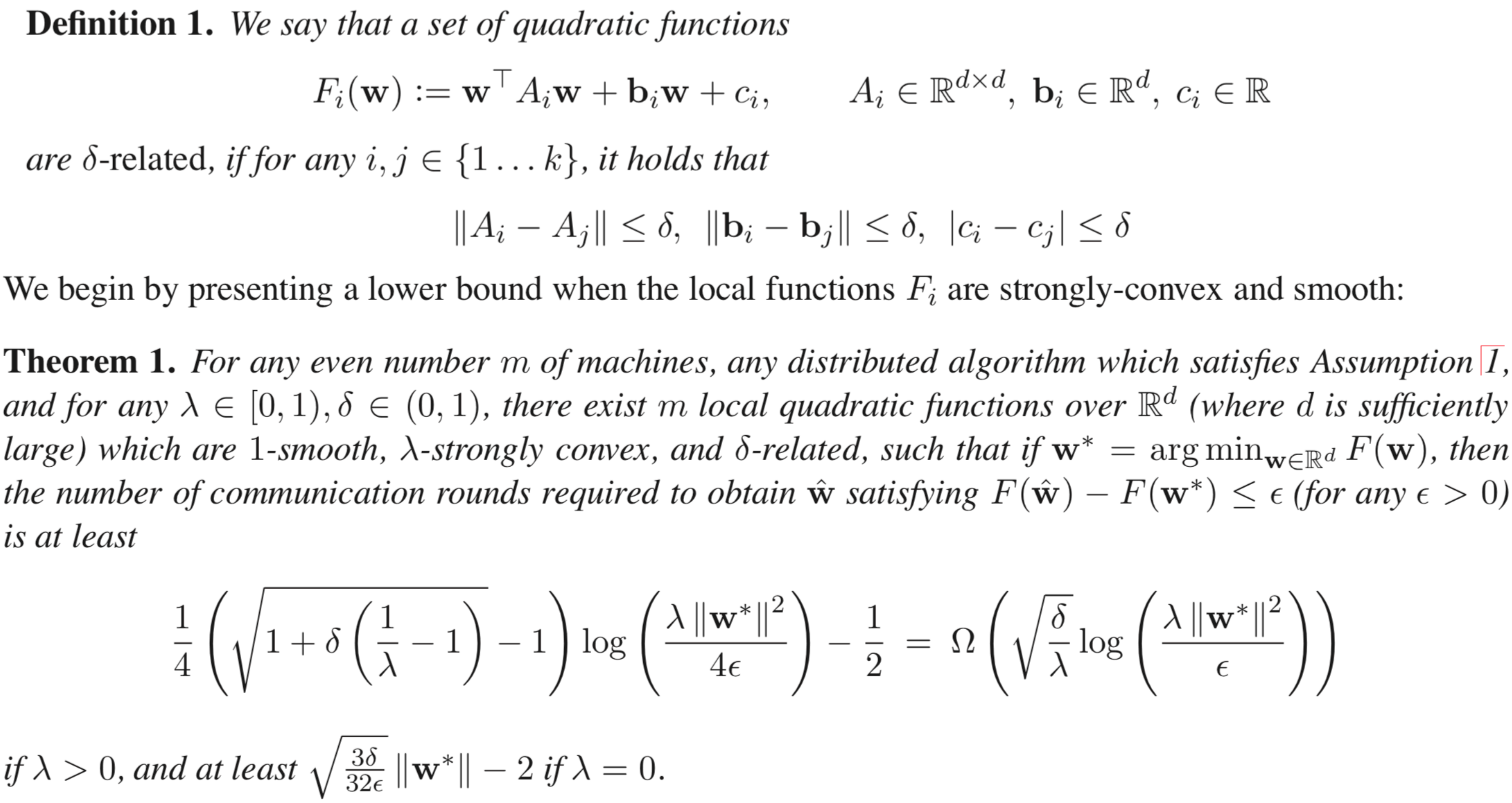


**Assumption 1.** *For each machine* $j$, *define a set* $W_j \subset \mathbb{R}^d$, *initially* $W_j = \{\mathbf{0}\}$. *Between communication rounds, each machine* $j$ *iteratively computes and adds to* $W_j$ *some finite number of points* $\mathbf{w}$, *each satisfying*

$$\gamma\mathbf{w} + \nu\nabla F_j(\mathbf{w}) \in \operatorname{span}\Big\{\mathbf{w}'\ ,\ \nabla F_j(\mathbf{w}')\ ,\ (\nabla^2 F_j(\mathbf{w}') + D)\mathbf{w}''\ ,\ (\nabla^2 F_j(\mathbf{w}') + D)^{-1}\mathbf{w}''\ \Big|\ \mathbf{w}', \mathbf{w}'' \in W_j\ ,\ D \textit{ diagonal}\ ,\ \nabla^2 F_j(\mathbf{w}') \textit{ exists}\ ,\ (\nabla^2 F_j(\mathbf{w}') + D)^{-1} \textit{ exists}\Big\}. \tag{2}$$

*for some* $\gamma, \nu \ge 0$ *such that* $\gamma + \nu > 0$. *After every communication round, let* $W_j := \cup_{i=1}^m W_i$ *for all* $j$. *The algorithm's final output (provided by the designated machine* $j$*) is a point in the span of* $W_j$.

# Lower communication rounds bound for distributed convex optimization under similarity



**Definition 1.** *We say that a set of quadratic functions*

$$F_i(\mathbf{w}) := \mathbf{w}^\top A_i \mathbf{w} + \mathbf{b}_i \mathbf{w} + c_i, \qquad A_i \in \mathbb{R}^{d \times d},\ \mathbf{b}_i \in \mathbb{R}^d,\ c_i \in \mathbb{R}$$

*are* $\delta$-related, *if for any* $i, j \in \{1 \dots k\}$, *it holds that*

$$\|A_i - A_j\| \le \delta,\ \ \|\mathbf{b}_i - \mathbf{b}_j\| \le \delta,\ \ |c_i - c_j| \le \delta$$

We begin by presenting a lower bound when the local functions $F_i$ are strongly-convex and smooth:

**Theorem 1.** *For any even number* $m$ *of machines, any distributed algorithm which satisfies Assumption 1, and for any* $\lambda \in [0, 1), \delta \in (0, 1)$, *there exist* $m$ *local quadratic functions over* $\mathbb{R}^d$ *(where* $d$ *is sufficiently large) which are* 1*-smooth,* $\lambda$*-strongly convex, and* $\delta$*-related, such that if* $\mathbf{w}^* = \arg\min_{\mathbf{w} \in \mathbb{R}^d} F(\mathbf{w})$, *then the number of communication rounds required to obtain* $\hat{\mathbf{w}}$ *satisfying* $F(\hat{\mathbf{w}}) - F(\mathbf{w}^*) \le \epsilon$ *(for any* $\epsilon > 0$*) is at least*

$$\frac{1}{4}\left(\sqrt{1 + \delta\left(\frac{1}{\lambda} - 1\right)} - 1\right) \log\left(\frac{\lambda \|\mathbf{w}^*\|^2}{4\epsilon}\right) - \frac{1}{2} = \Omega\left(\sqrt{\frac{\delta}{\lambda}} \log\left(\frac{\lambda \|\mathbf{w}^*\|^2}{\epsilon}\right)\right)$$

*if* $\lambda > 0$, *and at least* $\sqrt{\frac{3\delta}{32\epsilon}}\, \|\mathbf{w}^*\| - 2$ *if* $\lambda = 0$.

# Lower communication rounds bound for distributed convex optimization under similarity (two machines)

$$F_1(\mathbf{w}) = \frac{\delta(1-\lambda)}{4}\mathbf{w}^\top A_1 \mathbf{w} - \frac{\delta(1-\lambda)}{2}\mathbf{e}_1^\top \mathbf{w} + \frac{\lambda}{2}\left\|\mathbf{w}\right\|^2$$

$$F_2(\mathbf{w}) = \frac{\delta(1-\lambda)}{4}\mathbf{w}^\top A_2 \mathbf{w} + \frac{\lambda}{2}\left\|\mathbf{w}\right\|^2, \quad \text{where}$$

$$A_1 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 1 & -1 & 0 & 0 & 0 & \dots \\ 0 & -1 & 1 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 1 & -1 & 0 & \dots \\ 0 & 0 & 0 & -1 & 1 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}, \quad A_2 = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & \dots \\ -1 & 1 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 1 & -1 & 0 & 0 & \dots \\ 0 & 0 & -1 & 1 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 1 & -1 & \dots \\ 0 & 0 & 0 & 0 & -1 & 1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

# Upper communication rounds bound for distributed convex optimization under similarity

Lower bound (Arjevani-Shamir, 2015):

$$\Omega\left(\sqrt{\frac{\delta}{\lambda}}\log\left(\frac{\lambda R^2}{\varepsilon}\right)\right)$$

Upper bound (for DISCO, 2019):

http://proceedings.mlr.press/v37/zhangb15.pdf

$$\tilde{O}\left(\sqrt{\frac{\delta}{\lambda}}\left(\log\left(\frac{\lambda R^2}{\varepsilon}\right)+\frac{L_2^2\Delta F}{\lambda^3}\right)\right)$$

Upper bound (for SONATA, 2019):

arXiv:1905.02637v2

$$\tilde{O}\left(\frac{\delta}{\lambda}\log\left(\frac{\lambda R^2}{\varepsilon}\right)\right)$$

Upper bound (for SPAG, 2020):

http://proceedings.mlr.press/v119/hendrikx20a/hendrikx20a.pdf

$$\tilde{O}\left(\frac{\delta}{\lambda}\log\left(\frac{\lambda R^2}{\varepsilon}\right)\right)$$

Upper bound (Agafonov et al., 2021):

arXiv:2103.14392

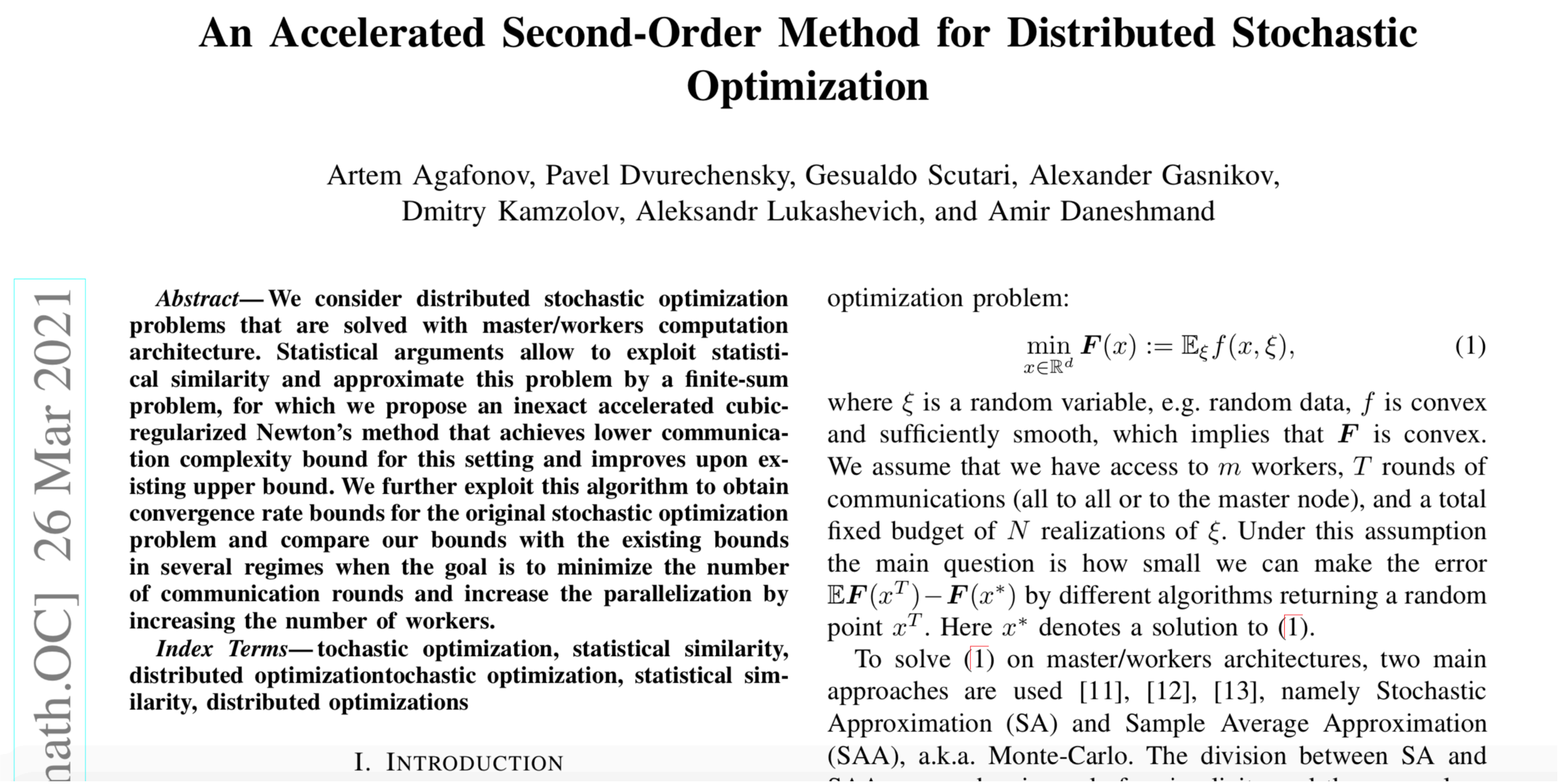
$$\tilde{O}\left(\sqrt{\frac{\delta}{\lambda}}\log\left(\frac{\lambda R^2}{\varepsilon}\right)+\left(\frac{L_2^2\Delta F}{\lambda^3}\right)^{1/6}\right)$$

# Upper communication rounds bound for distributed convex optimization under similarity

Upper bound (Agafonov et al., 2021):

arXiv:2103.14392

$$\tilde{O}\left(\sqrt{\frac{\delta}{\lambda}}\log\left(\frac{\lambda R^2}{\varepsilon}\right)+\left(\frac{L_2^2\Delta F}{\lambda^3}\right)^{1/6}\right)$$



## An Accelerated Second-Order Method for Distributed Stochastic Optimization

Artem Agafonov, Pavel Dvurechensky, Gesualdo Scutari, Alexander Gasnikov,
Dmitry Kamzolov, Aleksandr Lukashevich, and Amir Daneshmand

***Abstract*— We consider distributed stochastic optimization problems that are solved with master/workers computation architecture. Statistical arguments allow to exploit statistical similarity and approximate this problem by a finite-sum problem, for which we propose an inexact accelerated cubic-regularized Newton's method that achieves lower communication complexity bound for this setting and improves upon existing upper bound. We further exploit this algorithm to obtain convergence rate bounds for the original stochastic optimization problem and compare our bounds with the existing bounds in several regimes when the goal is to minimize the number of communication rounds and increase the parallelization by increasing the number of workers.**

***Index Terms*— tochastic optimization, statistical similarity, distributed optimizationtochastic optimization, statistical similarity, distributed optimizations**

### I. Introduction

optimization problem:

$$\min_{x\in\mathbb{R}^d} \boldsymbol{F}(x) := \mathbb{E}_\xi f(x,\xi), \qquad (1)$$

where $\xi$ is a random variable, e.g. random data, $f$ is convex and sufficiently smooth, which implies that $\boldsymbol{F}$ is convex. We assume that we have access to $m$ workers, $T$ rounds of communications (all to all or to the master node), and a total fixed budget of $N$ realizations of $\xi$. Under this assumption the main question is how small we can make the error $\mathbb{E}\boldsymbol{F}(x^T)-\boldsymbol{F}(x^*)$ by different algorithms returning a random point $x^T$. Here $x^*$ denotes a solution to (1).

To solve (1) on master/workers architectures, two main approaches are used [11], [12], [13], namely Stochastic Approximation (SA) and Sample Average Approximation (SAA), a.k.a. Monte-Carlo. The division between SA and

# Upper communication rounds bound for distributed convex optimization under similarity

Upper bound (A. Agafonov et al., 2021): $\tilde{O}\left(\sqrt{\frac{\delta}{\lambda}}\log\left(\frac{\lambda R^2}{\varepsilon}\right)+\left(\frac{L_2^2\Delta F}{\lambda^3}\right)^{1/6}\right)$

arXiv:2103.14392

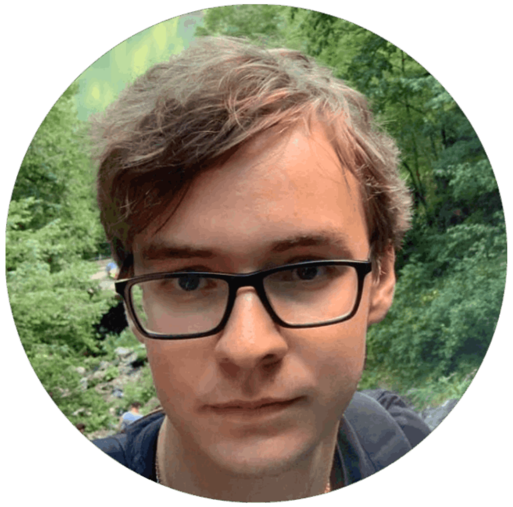
Main idea: To use (Accelerated) Cubic Regularized Newton method (Nesterov-Polyak, 2006; Nesterov 2008)

$$\min_x F(x) := \frac{1}{m}\sum_{i=1}^{m} F_i(x), \qquad \|\nabla^2 F_i(x) - \nabla^2 F(x)\| \le \delta$$
$$\|\nabla^2 F_i(y) - \nabla^2 F_i(x)\| \le L_2\|y - x\|.$$

$$x^{k+1} = \arg\min_x\{F(x^k) + \langle \nabla F(x^k), x - x^k\rangle +$$
$$+\frac{1}{2}\langle x - x^k, (\nabla^2 F_1(x^k) + 2\delta I)(x - x^k)\rangle + \frac{L_2}{6}\|x - x^k\|^3\}$$

Available at master node (1) via communications

In Cubic Regularized Newton method we have $\nabla^2 F(x^k)$ instead of $\nabla^2 F_1(x^k) + 2\delta I$



A. Beznosikov

# Saddle-point problems

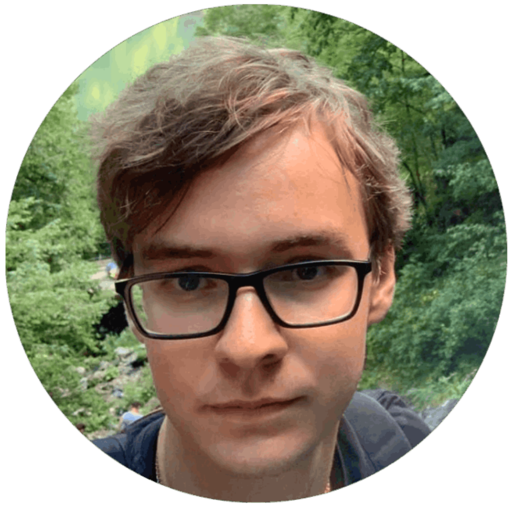
We study smooth (strongly-)convex-(strongly-)concave SPPs over a network of $M$ agents:

$$\min_{x\in X}\max_{y\in Y} f(x,y) := \frac{1}{M}\sum_{m=1}^{M} f_m(x,y), \tag{P}$$

Let us stack the $x$- and $y$-variables in the tuple $z=(x,y)$; accordingly, define $\mathcal{Z}=\mathcal{X}\times\mathcal{Y}$ and the operators $F_m, F:\mathcal{Z}\to\mathcal{R}^{2d}$:

$$F_m(z) := \begin{pmatrix} \nabla_x f_m(x,y) \\ -\nabla_y f_m(x,y) \end{pmatrix}, \quad \text{and} \quad F(z) := \frac{1}{M}\sum_{m=1}^{M} F_m(z). \tag{1}$$

The following conditions are standard for strongly convex-strongly concave SPPs.



# Saddle-point problems

**Assumption 1** *Given* (P)*, the following hold:*

*(i)* $\emptyset \neq \mathcal{Z}$ *is a convex and compact set;*

*(ii)* *Each* $f_m : \mathcal{R}^{2d} \to \mathcal{R}$ *is twice differentiable on (an open set containing)* $\mathcal{Z}$*, with* $L$*-smooth gradient:* $\|F_m(z_1) - F_m(z_2)\| \leq L\|z_1 - z_2\|$*, for all* $z_1, z_2 \in \mathcal{Z}$*;*

*(iii)* $f(z)$ *is* $\mu$*-strongly convex-strongly concave on* $\mathcal{Z}$*, i.e.,* $\langle F(z_1) - F(z_2), z_1 - z_2\rangle \geq \mu\|z_1 - z_2\|^2$*, for all* $z_1, z_2 \in \mathcal{Z}$*;*

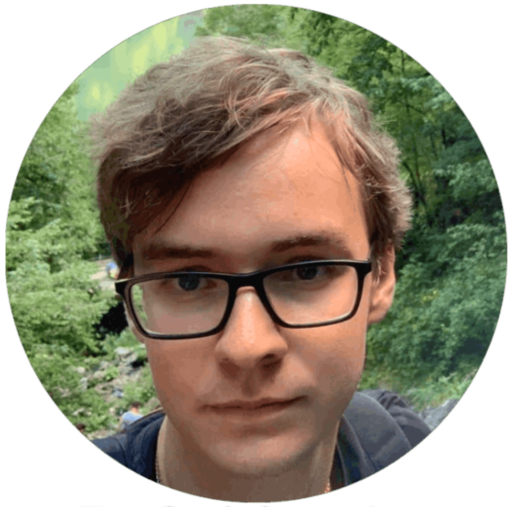
*(iv)* *Each* $f_m(z)$ *is convex-concave on* $\mathcal{Z}$*, i.e.* $0$*-strongly convex-strongly concave.*

We are interested in studying Problem (P) under function similarity.

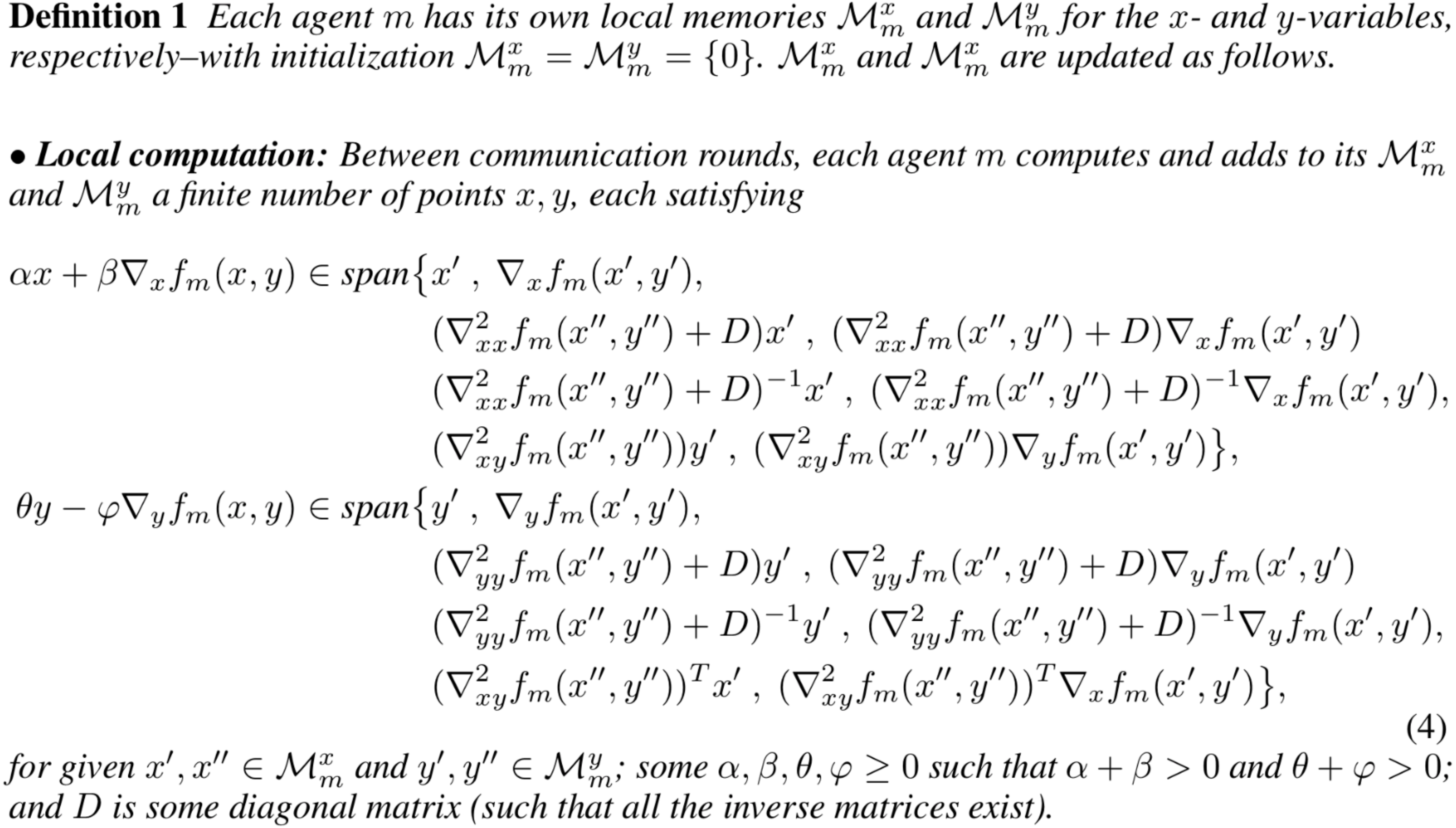
**Assumption 2 ($\delta$-related $f_m$'s)** *The local functions are* $\delta$*-related: for all* $(x, y) \in \mathcal{Z}$*,*

$$\|\nabla^2_{xx} f_m(x, y) - \nabla^2_{xx} f(x, y)\| \leq \delta,$$
$$\|\nabla^2_{xy} f_m(x, y) - \nabla^2_{xy} f(x, y)\| \leq \delta,$$
$$\|\nabla^2_{yy} f_m(x, y) - \nabla^2_{yy} f(x, y)\| \leq \delta.$$

The interesting case is when $1 + \delta/\mu \ll L/\mu$.



# Saddle-point problems. Lower complexity bounds

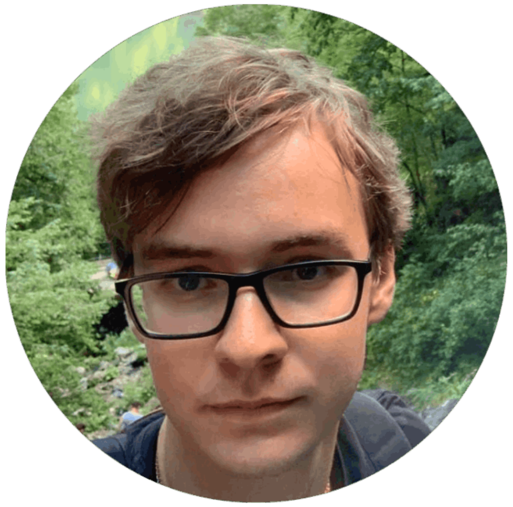


**Definition 1** *Each agent $m$ has its own local memories $\mathcal{M}_m^x$ and $\mathcal{M}_m^y$ for the $x$- and $y$-variables, respectively–with initialization $\mathcal{M}_m^x = \mathcal{M}_m^y = \{0\}$. $\mathcal{M}_m^x$ and $\mathcal{M}_m^x$ are updated as follows.*

- ***Local computation:*** *Between communication rounds, each agent $m$ computes and adds to its $\mathcal{M}_m^x$ and $\mathcal{M}_m^y$ a finite number of points $x, y$, each satisfying*

$$\begin{aligned}
\alpha x + \beta \nabla_x f_m(x,y) \in \textit{span}\{&x' \ , \ \nabla_x f_m(x',y'), \\
&(\nabla^2_{xx} f_m(x'',y'') + D)x' \ , \ (\nabla^2_{xx} f_m(x'',y'') + D)\nabla_x f_m(x',y') \\
&(\nabla^2_{xx} f_m(x'',y'') + D)^{-1}x' \ , \ (\nabla^2_{xx} f_m(x'',y'') + D)^{-1}\nabla_x f_m(x',y'), \\
&(\nabla^2_{xy} f_m(x'',y''))y' \ , \ (\nabla^2_{xy} f_m(x'',y''))\nabla_y f_m(x',y')\}, \\
\theta y - \varphi \nabla_y f_m(x,y) \in \textit{span}\{&y' \ , \ \nabla_y f_m(x',y'), \\
&(\nabla^2_{yy} f_m(x'',y'') + D)y' \ , \ (\nabla^2_{yy} f_m(x'',y'') + D)\nabla_y f_m(x',y') \\
&(\nabla^2_{yy} f_m(x'',y'') + D)^{-1}y' \ , \ (\nabla^2_{yy} f_m(x'',y'') + D)^{-1}\nabla_y f_m(x',y'), \\
&(\nabla^2_{xy} f_m(x'',y''))^T x' \ , \ (\nabla^2_{xy} f_m(x'',y''))^T \nabla_x f_m(x',y')\},
\end{aligned} \tag{4}$$

*for given $x', x'' \in \mathcal{M}_m^x$ and $y', y'' \in \mathcal{M}_m^y$; some $\alpha, \beta, \theta, \varphi \geq 0$ such that $\alpha + \beta > 0$ and $\theta + \varphi > 0$; and $D$ is some diagonal matrix (such that all the inverse matrices exist).*



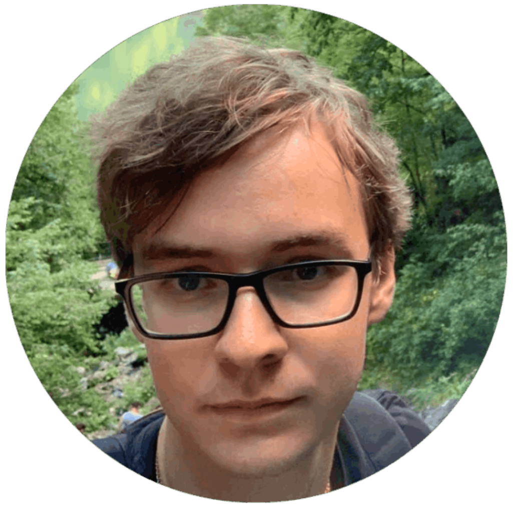
# Saddle-point problems.
# Lower complexity bounds

- ***Communication:*** *Based upon communication rounds among neighbouring nodes, $\mathcal{M}_m^x$ and $\mathcal{M}_m^y$ are updated according to*

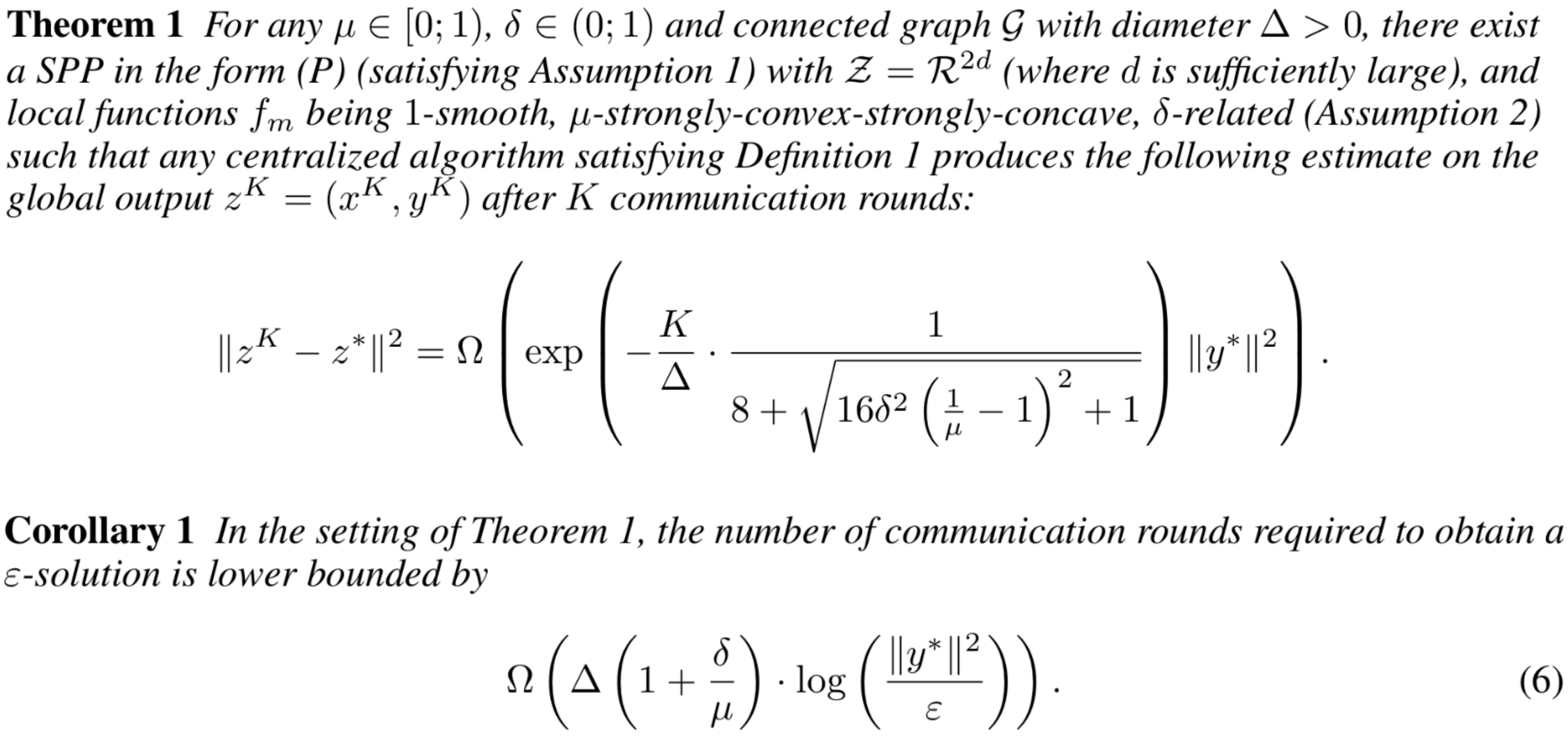
$$\mathcal{M}_m^x := span\left\{ \bigcup_{(i,m)\in\mathcal{E}} \mathcal{M}_i^x \right\}, \quad \mathcal{M}_m^y := span\left\{ \bigcup_{(i,m)\in\mathcal{E}} \mathcal{M}_i^y \right\}. \tag{5}$$

- ***Output:*** *The final global output is calculated as:*

$$x^K \in span\left\{ \bigcup_{m=1}^{M} \mathcal{M}_m^x \right\}, \; y^K \in span\left\{ \bigcup_{m=1}^{M} \mathcal{M}_m^y \right\}.$$



# Saddle-point problems. Lower complexity bounds

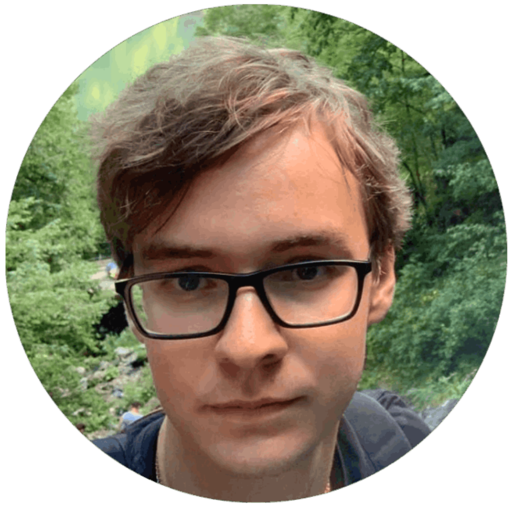


**Theorem 1** *For any $\mu \in [0; 1)$, $\delta \in (0; 1)$ and connected graph $\mathcal{G}$ with diameter $\Delta > 0$, there exist a SPP in the form (P) (satisfying Assumption 1) with $\mathcal{Z} = \mathcal{R}^{2d}$ (where $\mathrm{d}$ is sufficiently large), and local functions $f_m$ being $1$-smooth, $\mu$-strongly-convex-strongly-concave, $\delta$-related (Assumption 2) such that any centralized algorithm satisfying Definition 1 produces the following estimate on the global output $z^K = (x^K, y^K)$ after $K$ communication rounds:*

$$\|z^K - z^*\|^2 = \Omega\left(\exp\left(-\frac{K}{\Delta}\cdot\frac{1}{8+\sqrt{16\delta^2\left(\frac{1}{\mu}-1\right)^2+1}}\right)\|y^*\|^2\right).$$

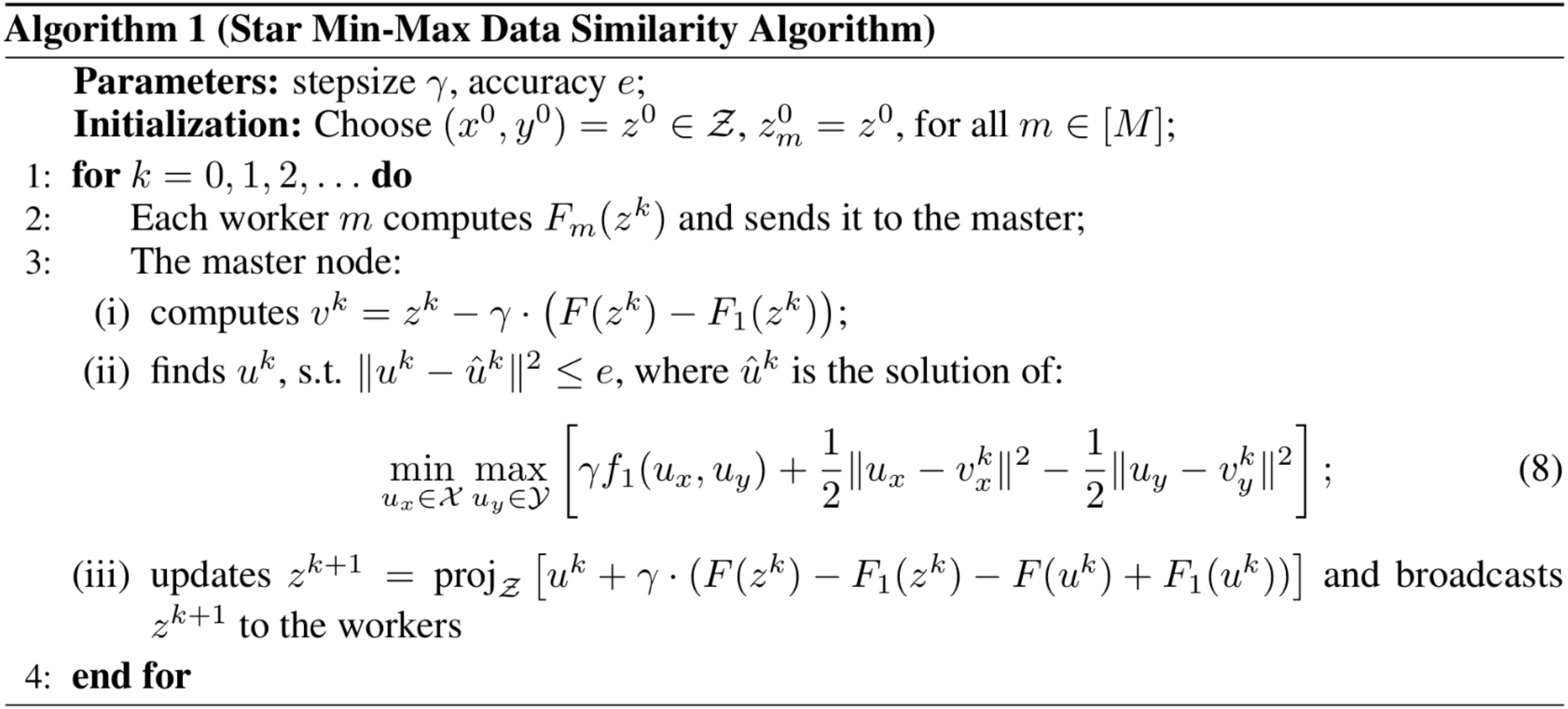
**Corollary 1** *In the setting of Theorem 1, the number of communication rounds required to obtain a $\varepsilon$-solution is lower bounded by*

$$\Omega\left(\Delta\left(1+\frac{\delta}{\mu}\right)\cdot\log\left(\frac{\|y^*\|^2}{\varepsilon}\right)\right). \tag{6}$$



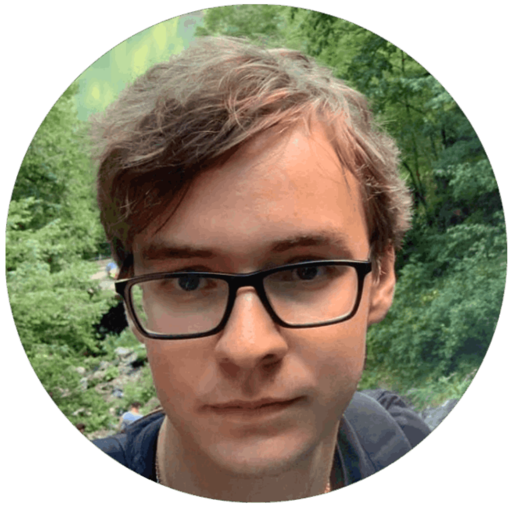
# Saddle-point problems. Optimal algorithm

$$\min_{x\in X}\max_{y\in Y} f(x,y) := \frac{1}{M}\sum_{m=1}^{M} f_m(x,y), \quad F_m(z) := \begin{pmatrix} \nabla_x f_m(x,y) \\ -\nabla_y f_m(x,y) \end{pmatrix}, \quad \text{and} \quad F(z) := \frac{1}{M}\sum_{m=1}^{M} F_m(z).$$

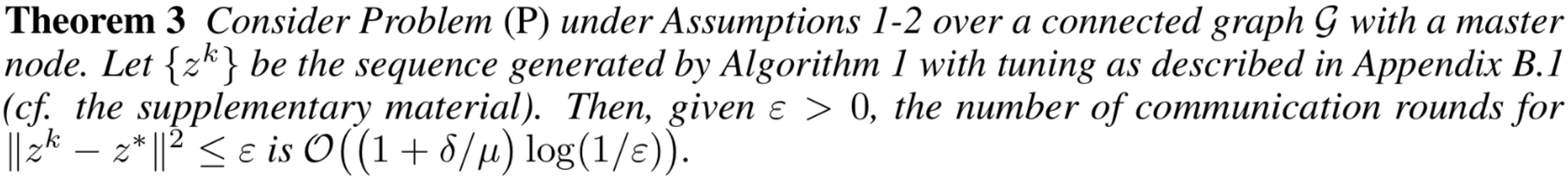


**Algorithm 1 (Star Min-Max Data Similarity Algorithm)**

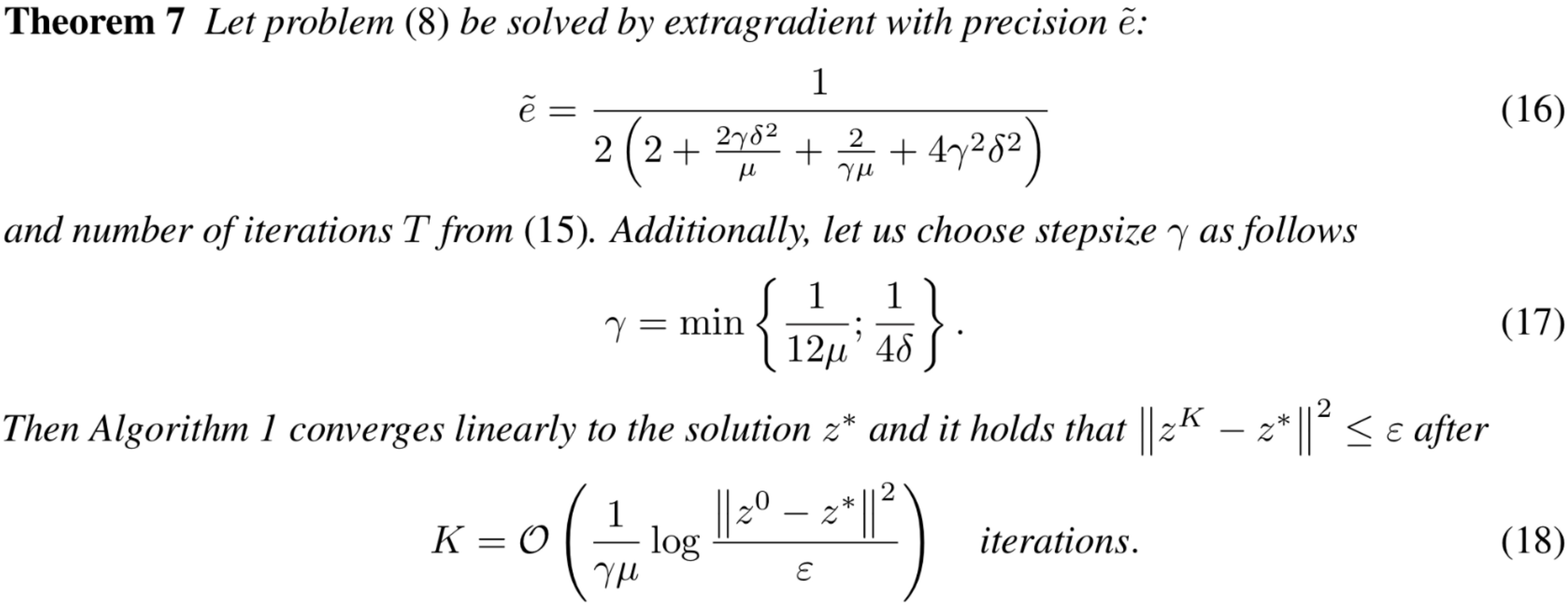
```
Parameters: stepsize $\gamma$, accuracy $e$;
Initialization: Choose $(x^0, y^0) = z^0 \in \mathcal{Z}$, $z_m^0 = z^0$, for all $m \in [M]$;
for $k = 0, 1, 2, \ldots$ do
    Each worker $m$ computes $F_m(z^k)$ and sends it to the master;
    The master node:
  (i) computes $v^k = z^k - \gamma \cdot \big(F(z^k) - F_1(z^k)\big)$;
  (ii) finds $u^k$, s.t. $\|u^k - \hat{u}^k\|^2 \le e$, where $\hat{u}^k$ is the solution of:
          $$\min_{u_x\in\mathcal{X}}\max_{u_y\in\mathcal{Y}} \left[\gamma f_1(u_x,u_y) + \frac{1}{2}\|u_x - v_x^k\|^2 - \frac{1}{2}\|u_y - v_y^k\|^2\right];  \qquad (8)$$
  (iii) updates $z^{k+1} = \mathrm{proj}_{\mathcal{Z}}\left[u^k + \gamma \cdot (F(z^k) - F_1(z^k) - F(u^k) + F_1(u^k))\right]$ and broadcasts
        $z^{k+1}$ to the workers
end for
```



# Saddle-point problems. Optimal algorithm



**Theorem 3** *Consider Problem* (P) *under Assumptions 1-2 over a connected graph $\mathcal{G}$ with a master node. Let $\{z^k\}$ be the sequence generated by Algorithm 1 with tuning as described in Appendix B.1 (cf. the supplementary material). Then, given $\varepsilon > 0$, the number of communication rounds for $\|z^k - z^*\|^2 \leq \varepsilon$ is $\mathcal{O}\big((1+\delta/\mu)\log(1/\varepsilon)\big)$.*



**Theorem 7** *Let problem* (8) *be solved by extragradient with precision $\tilde{e}$:*

$$\tilde{e} = \frac{1}{2\left(2 + \frac{2\gamma\delta^2}{\mu} + \frac{2}{\gamma\mu} + 4\gamma^2\delta^2\right)} \tag{16}$$

*and number of iterations $T$ from* (15). *Additionally, let us choose stepsize $\gamma$ as follows*

$$\gamma = \min\left\{\frac{1}{12\mu}; \frac{1}{4\delta}\right\}. \tag{17}$$

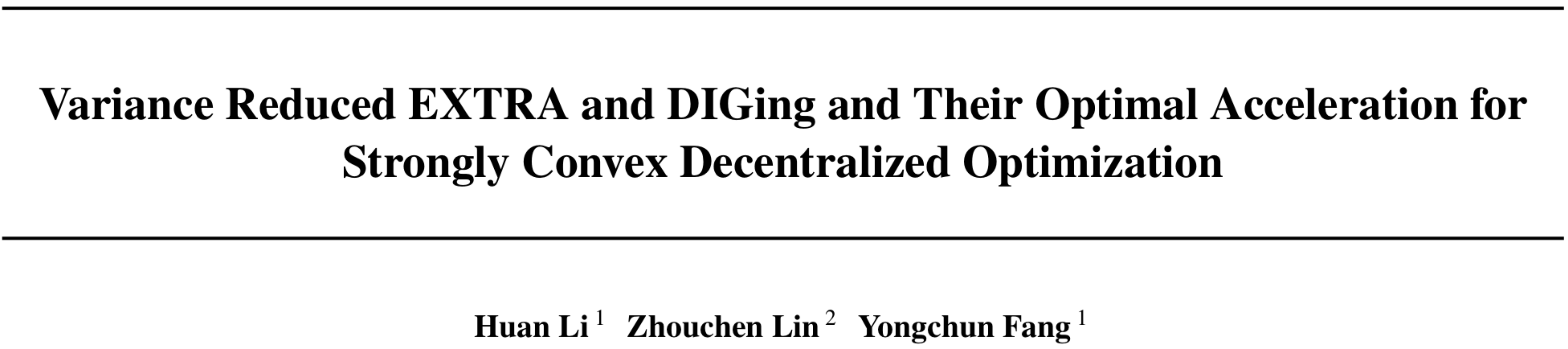
*Then Algorithm 1 converges linearly to the solution $z^*$ and it holds that $\left\|z^K - z^*\right\|^2 \leq \varepsilon$ after*

$$K = \mathcal{O}\left(\frac{1}{\gamma\mu}\log\frac{\left\|z^0 - z^*\right\|^2}{\varepsilon}\right) \quad \textit{iterations.} \tag{18}$$

# Saddle-point problems.
# Variance reduction

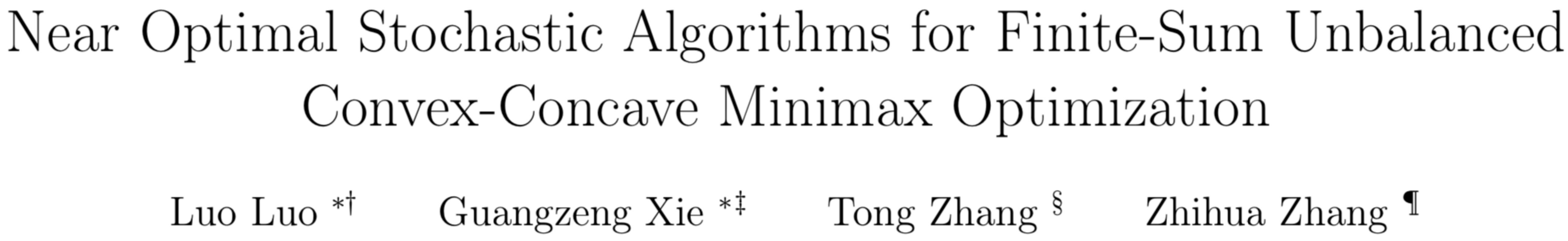
How to obtain optimal bound for distributed saddle-point problems with variance reduction?

Hypothesis: to combine ideas of arXiv:2009.04373, http://proceedings.mlr.press/v130/gorbunov21a/gorbunov21a.pdf and arXiv:2102.08352, arXiv:2106.01761



**Variance Reduced EXTRA and DIGing and Their Optimal Acceleration for Strongly Convex Decentralized Optimization**

**Huan Li** [1] **Zhouchen Lin** [2] **Yongchun Fang** [1]



Near Optimal Stochastic Algorithms for Finite-Sum Unbalanced Convex-Concave Minimax Optimization

Luo Luo [*†] Guangzeng Xie [*‡] Tong Zhang [§] Zhihua Zhang [¶]