Decentralized Optimization for Multi-Agent Networks

Qing Ling
Department of Automation, University of Science and Technology of China (USTC)

Joint work with Wotao Yin (UCLA), Wei Shi and Kun Yuan (USTC)

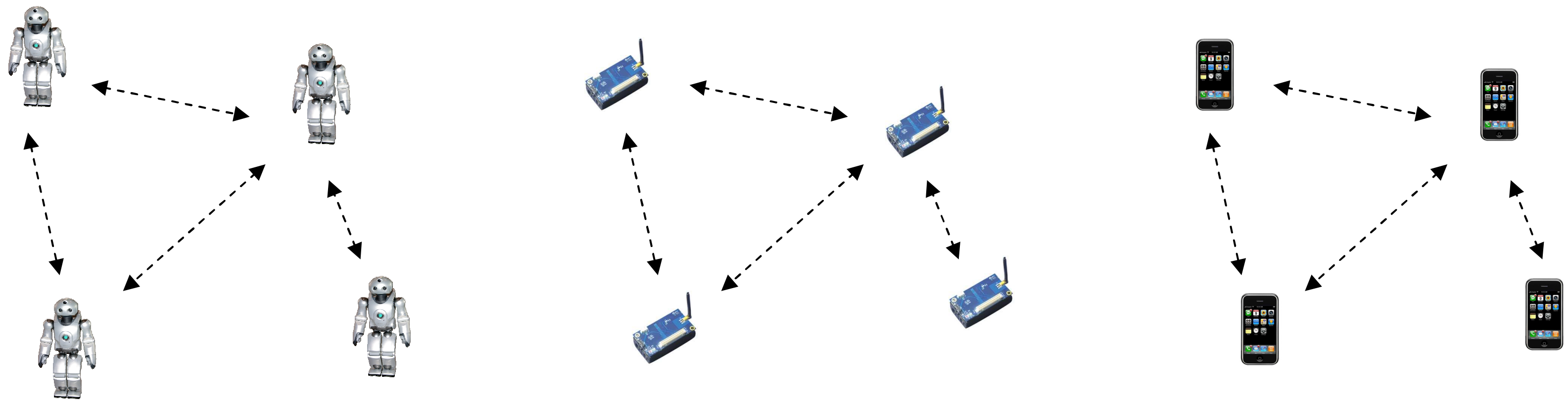
2014 Workshop on Optimization for Modern Computation
2014/09/02

Outline

- Background: multi-agent networks, decentralized optimization
- Decentralized gradient descent (DGD)
- Exact first-order algorithm (EXTRA)

Multi-agent networks

- A multi-agent network
  - A network of agents that are able to compute and communicate
  - Networks of computers, robots, wireless sensors, cognitive radios, etc



- In-network information processing, formulated as an optimization problem
  - Data transmission to fusion center is prohibitive (bandwidth, privacy)
  - Decentralized optimization through collaboration of neighboring agents

Decentralized consensus optimization

- A network of n agents solve

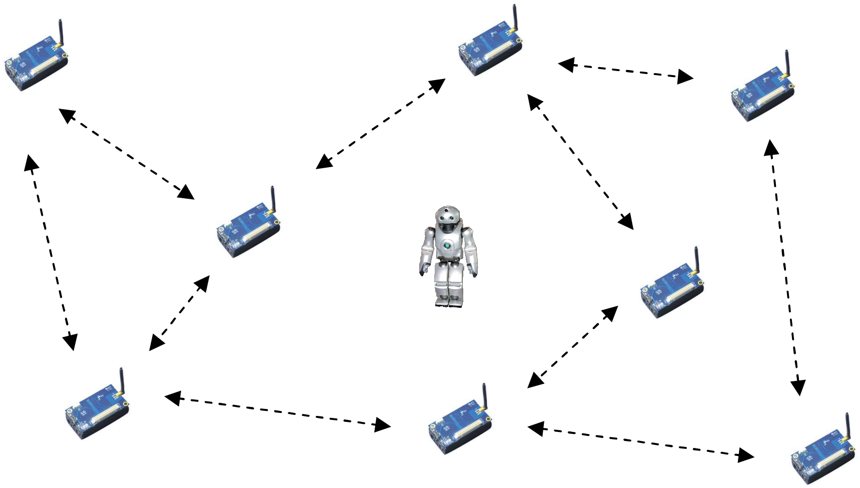
$$\min_{x} \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

  - $f_i : \mathbb{R}^p \to \mathbb{R}$ is local objective function at agent i
  - $x \in \mathbb{R}^p$ is common optimization variable to agents
  - $\mathcal{X}^*$ is optimal solution set

- In a decentralized optimization algorithm, each agent ...
  - Maintains a local iterate that can be shared with its neighbors
  - Is not allowed to exchange its local objective function
  - Is expected to eventually obtain a solution in $\mathcal{X}^*$ that is consensual

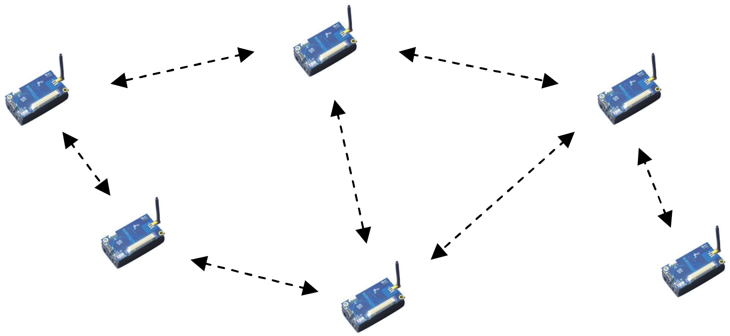
Example: target localization

- A network of n wireless sensors estimate position x of target
  - Position of sensor i is y_i
  - Distance measurement of sensor i is d_i
- Sensors collaboratively solves min $\frac{1}{n}\sum_{i=1}^{n}\left(d_i - \|y_i - x\|\right)^2$

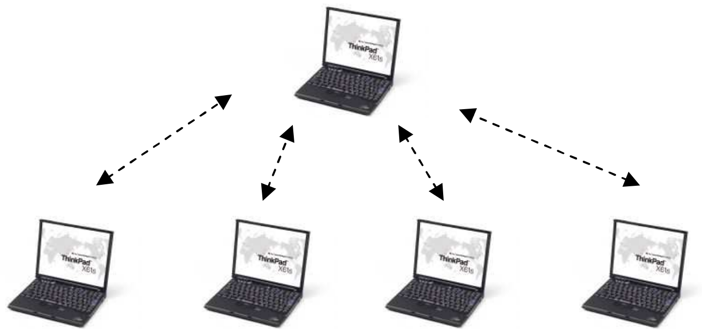


Decentralized versus distributed optimization

- Decentralized optimization



- Distributed optimization



- Designing decentralized and distributed optimization algorithms
  - Distributed is a special case of decentralized: a star topology
  - Utilize centralized controller for more efficient distributed algorithms

Related work

- Decentralized (sub)gradient descent [Nedic and Ozdaglar 2009]
  - Simple computation: mix neighboring solutions, descend locally
  - Slow or inaccurate convergence (as we will show)
- ADMM [Bertsekas and Tsitsiklis 1997, Schizas et al 2008]
  - Fast and accurate convergence in practice and theory [Shi et al 2014]
  - Complicated computation: solving an optimization problem
- Other algorithms: dual decomposition, dual averaging, etc
- This talk focuses on decentralized algorithms whose computations are simple

Assumptions

- Basic assumption on optimization problem

  f_i is differentiable and convex; optimal solution set $\mathcal{X}^*$ is nonempty

- Basic assumption on underlying network

  Network $(\mathcal{V}, \mathcal{A})$ is bidirectionally connected; communication is synchronized

- Assumption 1 (Lipschitz continuous gradient)

  ∇f_i is Lipschitz with constant L_{f_i}, $L_{max} = \max_i L_{f_i}$ and $L_{ave} = \frac{1}{n}\sum_{i=1}^{n} L_{f_i}$

- Assumption 2 (strong convexity)

  $\frac{1}{n}\sum_{i=1}^{n} f_i$ is strongly convex with constant μ_{ave}

Decentralized gradient descent (DGD)

- DGD: mix neighboring solutions, run local gradient descent

$$x_{(i)}^{k+1} = \sum_{j=1}^{n} w_{ij} x_{(j)}^{k} - \alpha^k \nabla f_i(x_{(i)}^k), \quad \forall i$$

  - Weight $w_{ij} = 0$ if $(i, j) \notin \mathcal{A}$ and $i \neq j \Rightarrow$ decentralized computation
  - Stepsize α^k: constant or diminishing

- Compare to centralized gradient descent

$$x^{k+1} = x^k - \frac{\alpha^k}{n} \sum_{i=1}^{n} \nabla f_i(x^k)$$

  - Maintain multiple local solutions, mix to keep closeness
  - Use local gradients to replace true average gradient

Mixing matrix

- Mixing matrix $\mathbf{W} = [w_{ij}] \in \mathbb{R}^{n \times n}$: belief on neighboring solutions
  - Nonnegative, symmetric, doubly stochastic ($\mathbf{W} = \mathbf{W}^T \geqslant 0, \mathbf{W}\mathbf{1} = \mathbf{1}$)
  - Eigenvalues of $\mathbf{W}$: $1 = \lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n \geqslant -1$
  - If connected, can design $\mathbf{W}$ such that second largest eigenvalue modulus

$$\rho = \max(|\lambda_2|, |\lambda_n|) < 1$$

- Metropolis-Hastings, maximum-degree, etc [Boyd et al 2004]

Existing convergence analysis

- $O(1/k)$ rate to neighborhood of $\mathcal{X}^*$ [Nedic & Ozdaglar 2009]
  - Bounded gradient/subgradient
  - Constant stepsize
- $O(1/k^{2/3})$ rate to $\mathcal{X}^*$ [Jakovetic et al 2014]
  - Bounded and Lipschitz continuous gradient
  - Diminishing stepsize $\sim O(1/k^{1/3})$
- We focus on DGD with constant stepsize
  - DGD is a centralized gradient descent to minimize a Lyapunov function
  - This equivalence enables deeper understanding and better results

Can we reach consensus?

- Suppose all local solutions eventually reach a consensual solution x^{con}

$$x^{con} = \sum_{j=1}^{n} w_{ij} x^{con} - \alpha \nabla f_i(x^{con}), \quad \forall i$$

  - $\mathbf{W}$ is doubly stochastic and $\alpha > 0 \Rightarrow \nabla f_i(x^{con}) = 0, \forall i$
  - $x^* \in \mathcal{X}^* \Rightarrow \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(x^*) = 0$
  - If such an x^{con} exists, then $x^{con} \in \mathcal{X}^*$; but it does not exist in general

- Dilemma of DGD
  - Constant stepsize $\longrightarrow$ inexact but fast (as we will show) convergence
  - Diminishing stepsize $\longrightarrow$ exact but slow convergence

Essence of DGD

- DGD with constant stepsize α

$$x_{(i)}^{k+1} = \sum_{j=1}^{n} w_{ij} x_{(j)}^{k} - \alpha \nabla f_i(x_{(i)}^{k}), \quad \forall i$$

is centralized gradient descent (stepsize 1) to minimize Lyapunov function

$$\min_{\{x_{(i)}\}} \quad \sum_{i=1}^{n} \left(\alpha f_i(x_{(i)}) + \frac{1}{2}\|x_{(i)}\|_2^2\right) - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} x_{(i)}^T x_{(j)}$$

- From the equivalence, we can show ...
  - When gradients are bounded, how fast convergence is
  - Where to converge

When gradients are bounded?

- Theorem: under Assumption 1 (Lipschitz continuous gradient), if

$$\alpha \leqslant \frac{1+\lambda_n}{L_{max}}$$

  then gradients are bounded

- Smaller L_{max} or larger λ_n (away from -1) $\Rightarrow$ larger critical stepsize
  - Critical stepsize is tight as we can show counterexamples
  - Same order as stepsize of centralized gradient descent $\frac{2}{L_{ave}}$
  - Have $L_{max} \in [L_{ave}, nL_{ave}]$; design $\mathbf{W}$ such that $\lambda_n > 0$

Where to converge and how fast?

- Theorem: under Assumption 1 (Lipschitz continuous gradient), if $\rho < 1$ and

$$\alpha \leqslant \min\{\frac{1+\lambda_n}{L_{max}}, \frac{1}{L_{ave}}\}$$

  then objective error decreases at a rate of $O(\frac{1}{\alpha k})$ until reaching $O(\frac{\alpha}{1-\rho})$

- Theorem: under Assumption 1 (Lipschitz continuous gradient) and Assumption 2 (strong convexity), if $\rho < 1$ and

$$\alpha \leqslant \min\{\frac{1+\lambda_n}{L_{max}}, \frac{1}{L_{ave}+\mu_{ave}}\}$$

  then point error decreases at a rate of $O(c^k)$ until reaching $O(\frac{\alpha}{1-\rho})$; here $c \in (0,1)$ is determined by α and ρ

- Large $\alpha \Rightarrow$ fast convergence and inaccurate solution

- Large ρ (achievable when network is dense) $\Rightarrow$ accurate solution

Concluding DGD

- Our contribution: establishing inexact convergence and rates of convergence
  - Lipschitz continuous gradient $\longrightarrow O(\frac{1}{k})$ rate
  - Lipschitz continuous gradient and strong convexity $\longrightarrow O(c^k)$ rate
  - Bounds of stepsizes are similar to those in centralized gradient descent
  - Tradeoff between speed and accuracy through tuning stepsize
- Can we improve DGD: exact convergence with large constant stepsize?

EXact firsT-ordeR Algorithm (EXTRA)

- EXTRA: mix neighboring solutions, run local gradient descent-ascent

$$x_{(i)}^{1} = \sum_{j=1}^{n} w_{ij} x_{(j)}^{0} - \alpha \nabla f_i(x_{(i)}^{0}), \quad \forall i$$

$$\begin{aligned} x_{(i)}^{k+2} = x_{(i)}^{k+1} &+ \sum_{j=1}^{n} w_{ij} x_{(j)}^{k+1} - \sum_{j=1}^{n} \tilde{w}_{ij} x_{(j)}^{k} \\ &- \alpha \left[\nabla f_i(x_{(i)}^{k+1}) - \nabla f_i(x_{(i)}^{k}) \right], \quad \forall i, \forall k \geqslant 0 \end{aligned}$$

  - Weights w_{ij} and $\tilde{w}_{ij} = 0$ if $(i,j) \notin \mathcal{A}$ and $i \neq j$
  - Stepsize α: constant

- Overheads comparing to DGD
  - Communication: same per iteration
  - Storage: storing previous neighboring solutions and local gradient

Mixing matrices

- Mixing matrices $\mathbf{W} = [w_{ij}]$ and $\tilde{\mathbf{W}} = [\tilde{w}_{ij}]$
  - (Symmetry) $\mathbf{W} = \mathbf{W}^T$ and $\tilde{\mathbf{W}} = \tilde{\mathbf{W}}^T$
  - (Null space) $\mathrm{null}\{\mathbf{W} - \tilde{\mathbf{W}}\} = \mathrm{span}\{\mathbf{1}\}$ and $\mathrm{null}\{\mathbf{I}_n - \tilde{\mathbf{W}}\} \subseteq \mathrm{span}\{\mathbf{1}\}$
  - (Spectral) $\tilde{\mathbf{W}} \succ 0$ and $\frac{\mathbf{I}_n + \mathbf{W}}{2} \succeq \tilde{\mathbf{W}} \succeq \mathbf{W}$
- Choose $\mathbf{W}$ as in DGD and set $\tilde{\mathbf{W}} = \frac{\mathbf{I}_n + \mathbf{W}}{2}$
  - Nonnegative, symmetric, doubly stochastic ($\mathbf{W} = \mathbf{W}^T \geqslant 0, \mathbf{W}\mathbf{1} = \mathbf{1}$)
  - Second largest eigenvalue modulus of $\mathbf{W}$: $\rho = \max(|\lambda_2|, |\lambda_n|) < 1$
  - Eigenvalues of $\mathbf{W}$: $1 = \lambda_1 > \lambda_2 \geqslant \cdots \geqslant \lambda_n > -1$
  - Eigenvalues of $\tilde{\mathbf{W}}$: $1 = \tilde{\lambda}_1 > \tilde{\lambda}_2 \geqslant \cdots \geqslant \tilde{\lambda}_n > 0$

Limit properties

- Suppose all local solutions eventually reach a consensual solution x^{con}

$$x^{con} = x^{con} + \sum_{j=1}^{n} w_{ij} x^{con} - \sum_{j=1}^{n} \tilde{w}_{ij} x^{con} - \alpha \left[\nabla f_i(x^{con}) - \nabla f_i(x^{con}) \right], \quad \forall i$$

  - $\text{null}\{\mathbf{W} - \tilde{\mathbf{W}}\} = \text{span}\{\mathbf{1}\} \Rightarrow \sum_{j=1}^{n} w_{ij} - \sum_{j=1}^{n} \tilde{w}_{ij} = 0, \forall i$
  - No contradiction, different to DGD that cannot stay at a consensual x^{con}

- If local solutions converge to $x_{(1)}^{\infty}, \cdots, x_{(n)}^{\infty}$, we have $x_{(1)}^{\infty} = \cdots = x_{(n)}^{\infty} \in \mathcal{X}^*$

Explanations of EXTRA

- EXTRA takes difference of two DGD updates and cancels steady-state error

$$x_{(i)}^{k+2} = \sum_{j=1}^{n} w_{ij} x_{(j)}^{k+1} - \alpha \nabla f_i(x_{(i)}^{k+1}) \quad \text{and} \quad x_{(i)}^{k+1} = \sum_{j=1}^{n} \tilde{w}_{ij} x_{(j)}^{k} - \alpha \nabla f_i(x_{(i)}^{k})$$

- Rewrite EXTRA as

$$x_{(i)}^{k+1} = \sum_{j=1}^{n} w_{ij} x_{(j)}^{k} - \alpha \nabla f_i(x_{(i)}^{k}) + \sum_{t=0}^{k-1} \sum_{j=1}^{n} (w_{ij} - \tilde{w}_{ij})\, x_{(j)}^{t}, \quad \forall i$$

  - EXTRA = DGD with constant stepsize + correction term
  - Corrected by weighted summation of all previous neighboring solutions

Sublinear convergence

- Theorem: under Assumption 1 (Lipschitz continuous gradient), if

$$\alpha < \frac{2\tilde{\lambda}_n}{L_{max}}$$

then $x_{(i)}^k$ converges to the same $x^* \in \mathcal{X}^*$ for all i and point progresses

$$\left\| x_{(i)}^{k+1} - x_{(i)}^k \right\|_2^2, \quad \forall i$$

decrease at a rate of $O(\frac{1}{k})$

- Remarks on the result
  - $O(\frac{1}{k})$ point progress convergence $\Rightarrow$ slower convergence of $x_{(i)}^k$ to x^*
  - $\tilde{\lambda}_n$ tunable in $(0,1)$ and $L_{max} \in [L_{ave}, nL_{ave}] \Rightarrow \frac{2\tilde{\lambda}_n}{L_{max}} \sim \frac{2}{L_{ave}}$

Linear convergence

- Theorem: under Assumption 1 (Lipschitz continuous gradient) and Assumption 2 (strong convexity), if

$$\alpha < \frac{2\mu_{ave}\tilde{\lambda}_n}{L_{max}^2}$$

then point errors

$$\left\| x_{(i)}^k - x^* \right\|_2^2, \quad \forall i$$

decrease at a rate of $O(c^k)$; here $c \in (0,1)$ and x^* is unique optimal solution

- Remarks on the result
  - $\frac{2\mu_{ave}\tilde{\lambda}_n}{L_{max}^2} \sim \frac{2}{L_{ave}+\mu_{ave}}$ when $L_{ave} \sim \mu_{ave}$
  - Allow larger stepsize in practice

Simulation settings

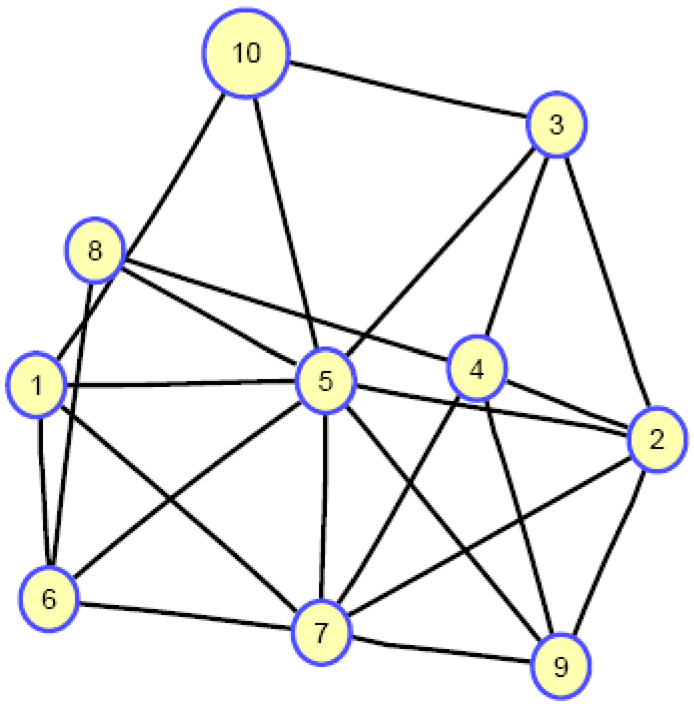
- Network of $n = 10$ agents, 21 random edges out of 45 are connected
- Decentralized consensus optimization problem

$$\min_x \frac{1}{n}\sum_{i=1}^{n} f_i(x) \quad \text{where} \quad f_i(x) = \frac{1}{2}\left\|\mathbf{A}_{(i)}x - y_{(i)}\right\|_2^2$$

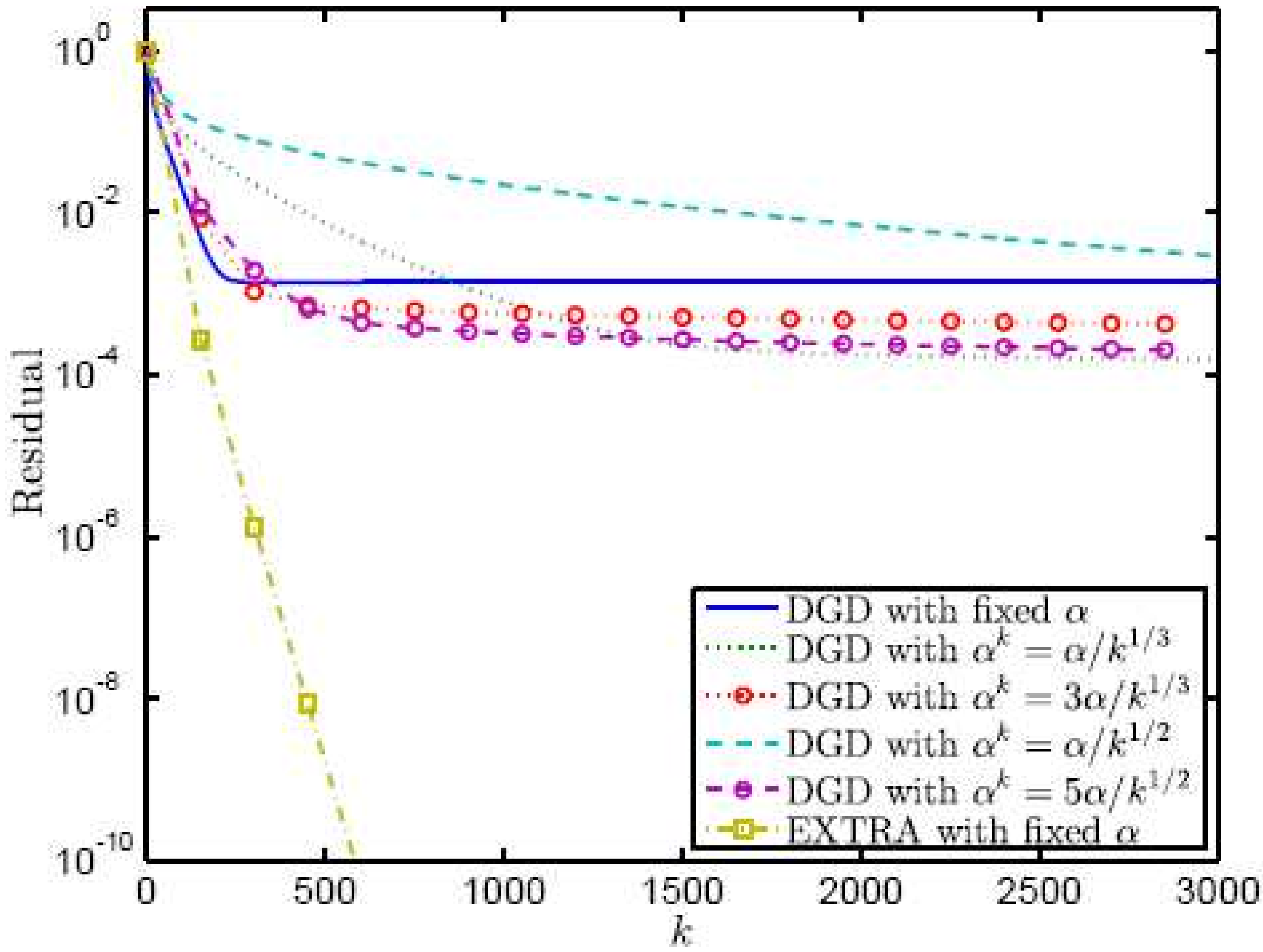
where $\mathbf{A}_{(i)} \in \mathbb{R}^{1\times 5}, y_{(i)} \in \mathbb{R}, x \in \mathbb{R}^5$

- Performance metric

$$\text{residual} \triangleq \frac{\sum_{i=1}^{n}\left\|x_{(i)}^k - x^*\right\|_2^2}{\sum_{i=1}^{n}\left\|x_{(i)}^0 - x^*\right\|_2^2}$$



Simulation of DGD and EXTRA



Concluding EXTRA

- EXTRA corrects steady-state error of DGD with one-step memory
- Communication cost remains the same as DGD
- Provable exact sublinear and linear rates of convergence
  - Lipschitz continuous gradient $\longrightarrow$ sublinear rate
  - Lipschitz continuous gradient and strong convexity $\longrightarrow$ $O(c^k)$ rate

Future research directions

- Differentiable local objectives $\longrightarrow$ differentiable plus nondifferentiable
- Synchronized network communication $\longrightarrow$ asynchronous
- Optimization with batch data $\longrightarrow$ streaming data

Thank you