Distributed Optimization in Undirected Graphs
Gradient and EXTRA Algorithms

Wotao Yin (UCLA Math)

Joint work with Qing Ling, Wei Shi, Kun Yuan (USTC Automation)

March 15, 2014 – SIAM CSE, Salt Lake City

Consensus optimization

- A connected network of n agents

- Each agent i has a private Lipschitz-differentiable function f_i
- Model: find a *consensus solution* $x^* \in \mathbb{R}^p$ to

$$\underset{x \in \mathbb{R}^p}{\text{minimize}}\, f(x) := \sum_{i=1}^{n} f_i(x)$$

- x^* can mean average reading, common knowledge, joint decision ......
- Assume *bi-directional links* and *synchronized* iterations in this talk

This talk

- Analyze DGD[1] (gradient descent with neighborhood averaging)
  - Lyapunov function, inexact gradient
  - Relation step size and speed/accuracy
- Introduce EXTRA (DGD along with correction steps, much faster)
  - Motivation
  - Convergence properties
- Extension: PG-EXTRA for smooth+nonsmooth objectives

[1] Nedic-Ozdaglar'09

Challenges

Conventional algorithms don't apply

- No long-distance communication
- No center or lead agent
- Must optimize and exchange information at the same time

Applications:

- sensor networks: wireless and under-water sensors, security cameras
- groups robots and UAVs
- collaborative machine learning
- understanding social networks and predator-prey group behaviors

Compact notation

- Each agent i: local variable $x_{(i)} \in \mathbb{R}^p$, placed on the i**th row** of $\mathbf{x}$.

$$\mathbf{x} \triangleq \begin{pmatrix} — & x_{(1)}^{\mathrm{T}} & — \\ — & x_{(2)}^{\mathrm{T}} & — \\ & \vdots & \\ — & x_{(n)}^{\mathrm{T}} & — \end{pmatrix} \in \mathbb{R}^{n \times p}$$

- $\mathbf{x}$ is consensual if all its rows are equal: $x_{(i)}^T = x_{(j)}^T$, $\forall i \neq j$.

$$\mathbf{f}(\mathbf{x}) \triangleq \begin{pmatrix} f(x_{(1)}) \\ f(x_{(2)}) \\ \vdots \\ f(x_{(n)}) \end{pmatrix} \in \mathbb{R}^n, \quad \nabla \mathbf{f}(\mathbf{x}) \triangleq \begin{pmatrix} — & \nabla f_1(x_{(1)})^{\mathrm{T}} & — \\ — & \nabla f_2(x_{(2)})^{\mathrm{T}} & — \\ & \vdots & \\ — & \nabla f_n(x_{(n)})^{\mathrm{T}} & — \end{pmatrix} \in \mathbb{R}^{n \times p}.$$

- original problem $\Longleftrightarrow$

$$\text{minimize } \mathbf{1}^T \mathbf{f}(\mathbf{x}), \text{ subject to } x_{(i)} = x_{(j)}, \ \forall i \neq j.$$

Decentralized gradient descent (DGD)

Introduce the communication matrix $W = [w_{ij}]$:

- $w_{ij} = 0$, $i \neq j$, if nodes i and j are not neighbors
- **this talk:** *symmetric, doubly stochastic* $W = W^T$, $W\mathbf{1} = \mathbf{1}$.

Nedic-Ozdaglar'09: neighborhood averaging + local gradient descent

Original form: $x_{(i)}^{k+1} = \sum_j w_{ij} x_{(j)}^k - \alpha \nabla f_i(x_{(i)}^k)$, by agents $i = 1, 2, \ldots, n$.

Compact form: $\boxed{\mathbf{x}^{k+1} = W\mathbf{x}^k - \alpha \nabla \mathbf{f}(\mathbf{x}^k)}$

Convergence requires diminishing step sizes. Some recent works:

- $\alpha_k = \alpha / k^{1/3}$: Jakovetic-Xavier-Moura'14
- $\alpha_k = \alpha / k^{1/2}$: I-An Chen'12

Original problem is equivalent to:

$$\underset{\mathbf{x}}{\text{minimize}}\ \mathbf{1}^T\mathbf{f}(\mathbf{x}) \quad \text{subject to } \mathbf{x} = W\mathbf{x}.$$

DGD interpretation 1: the unit-step gradient descent applied to

$$\xi_\alpha(\mathbf{x}) := \underbrace{\frac{1}{2}\mathrm{tr}(\mathbf{x}^T(I-W)\mathbf{x})}_{\text{quad penalty of } \mathbf{x}=W\mathbf{x}} + \alpha\mathbf{1}^T\mathbf{f}(\mathbf{x}).$$

Interpretation 2: multiply $\frac{1}{n}\mathbf{1}^T \times$ (DGD updateformula):

$$\bar{x}^{k+1} = \bar{x}^k - \alpha\left[\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x_{(i)}^k)\right].$$

Replacing $x_{(i)}^k$ by $\bar{x}^k$ gives the gradient descent applied to

$$\underset{\bar{x}}{\text{minimize}}\ \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\bar{x}).$$

Fixed-step size results

Theorem (Yuan-Ling-Y'14)

Assume ∇f_i is L_i-Lipschitz and fix $\alpha \leq (1+\lambda_n(W))/\max_i L_i$. Then

- *the DGD iterates remain bounded*
- *any local solution is bounded from the mean up to $O(\frac{\alpha}{1-\beta})$, where β is the 2nd largest absolute eigenvalue of W*
- *progress stagnates at $O(\frac{\alpha}{1-\beta})$*
- *objective error reduces at $O(\frac{1}{\alpha k})$ until stagnation*
- *if all f_i strongly convex, objective error and solution converge linearly until stagnation*

Proposed algorithm: EXTRA

Introduce

$$\overline{W} := (W + I)/2.$$

Taking the difference between two DGD iterations

$$\mathbf{x}^{k+1} = \overline{W}\mathbf{x}^{k} \quad - \alpha\nabla\mathbf{f}(\mathbf{x}^{k}), \tag{1}$$

$$\mathbf{x}^{k+2} = W\mathbf{x}^{k+1} - \alpha\nabla\mathbf{f}(\mathbf{x}^{k+1}), \tag{2}$$

gives a 3-point iteration: "EXTRA"

$$\boxed{\mathbf{x}^{k+2} - \mathbf{x}^{k+1} = W\mathbf{x}^{k+1} - \overline{W}\mathbf{x}^{k} - \alpha\nabla\mathbf{f}(\mathbf{x}^{k+1}) + \alpha\nabla\mathbf{f}(\mathbf{x}^{k}).} \tag{3}$$

If $\mathbf{x}^k \to \bar{\mathbf{x}}$, then easy to show

- $\bar{\mathbf{x}} = W\bar{\mathbf{x}}$,
- $\mathbf{1}^T\nabla\mathbf{f}(\bar{\mathbf{x}}) = 0$,

so $\bar{\mathbf{x}}$ is an optimal consensual solution.

Interpretation

Equivalent iteration:

$$\mathbf{x}^{k+1} = W\mathbf{x}^k - \alpha\nabla\mathbf{f}(\mathbf{x}^k) + \underbrace{\sum_{i=0}^{k-1}(W - \overline{W})\mathbf{x}^i}_{\text{correction}}.$$

- suppose that $\mathbf{x}^{k+1} = W\mathbf{x}^k$ asymptotically
- we need $\mathbf{1}^T\nabla\mathbf{f}(\mathbf{x}^k) \to 0$ (optimality)
- $\sum_{i=0}^{k-1}(W - \overline{W})\mathbf{x}^i$ is the simplest term we found that cancels $\nabla\mathbf{f}(\mathbf{x}^k)$ over $\mathrm{span}\{\mathbf{1}\}^\perp$
- is a nonstandard primal-dual algorithm for

$$\underset{\mathbf{x}}{\text{minimize}}\ \mathbf{1}^T\mathbf{f}(\mathbf{x}) \quad \text{subject to } \mathbf{x} = W\mathbf{x}.$$

Convergence

Theorem (sublinear convergence)

Assume (i) convex Lipschitz differentiable objectives, (ii) consensus solution exists, (iii) symmetric doubly stochastic W and $\overline{W}$ obeying

$$\overline{W} \succ 0 \quad \text{and} \quad \frac{I+W}{2} \succeq \overline{W} \succeq W.$$

If step size $\alpha < 2\lambda_{\min}(\overline{W})/\max L_i$, EXTRA has $O(1/k)$ ergodic convergence.

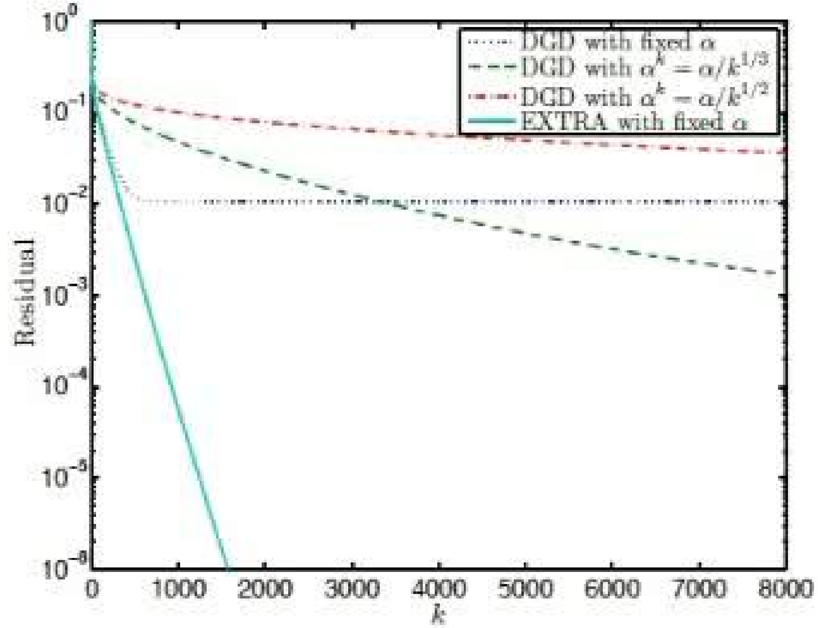
Theorem (linear convergence)

In addition, if

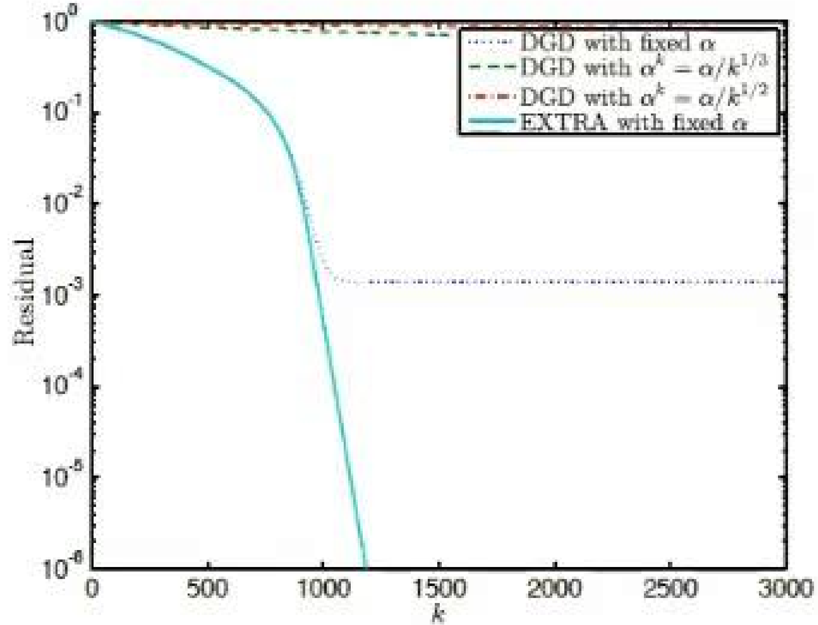
$$\sum_{i=1}^{n} f_i(x)$$

is (restrict) strongly convex, then $\|\mathbf{x}^k - \mathbf{x}^*\|_{W}$ linearly converges to 0.

Example: decentralized least squares



Example: decentralized sum of Huber functions



PG-EXTRA for composite objectives

Consider

$$\underset{\mathbf{x}}{\text{minimize}}\, \mathbf{1}^T(\mathbf{f}(\mathbf{x})+\mathbf{r}(\mathbf{x})) \quad \text{subject to } \mathbf{x}=W\mathbf{x},$$

where r_i are **possibly non-differentiable** functions with simple proximal maps.

Applications: geometric mean, compressed sensing, machine learning

Proposed:

$$\mathbf{x}^{k+1+\frac{1}{2}} = W\mathbf{x}^{k+1} + \mathbf{x}^{k+\frac{1}{2}} - \widetilde{W}\mathbf{x}^{k} - \alpha[\nabla\mathbf{f}(\mathbf{x}^{k+1}) - \nabla\mathbf{f}(\mathbf{x}^{k})],$$

$$\mathbf{x}^{k+2} = \arg\min_{\mathbf{x}}\ \mathbf{r}(\mathbf{x}) + \frac{1}{2\alpha}\|\mathbf{x}-\mathbf{x}^{k+1+\frac{1}{2}}\|_{\mathrm{F}}^2.$$

A nontrivial extension to EXTRA since $\{\mathbf{x}^k\}$ and $\{\mathbf{x}^{k+1/2}\}$ are interlaced.

Convergence and performance: similar to EXTRA's

Summary and future work

This talk

- analyzed decentralized gradient descent (DGD) method
- introduced methods that perform nearly as well as centralized gradient methods

Future work

- obtain optimal $1/k^2$ algorithms for Lipschitz differentiable problems
- extension to directed / asymmetric networks
- extension to asynchronous communication