Dual Averaging for Distributed Optimization

Presentation by: Mojtaba Kadkhodaie

University of Minnesota

March 10, 2014

Reference

- J. Duchi, A. Agarwal, and M. Wainwright “Dual Averaging for Distributed Optimization: Convergence Analysis and Network Scaling,” *IEEE Transactions on Automatic control, 57:3, 592-606, 2012.*

Agenda

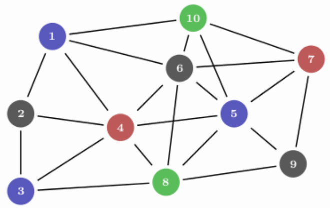
- Introduction
- Standard (Centralized) Dual Averaging Algorithm
- Distributed Dual Averaging
- Convergence Analysis
- Simulation Results

Motivation

Network-structured optimization problems arise in various areas.

- Machine Learning:
  - Large training dataset
  - Distribute the data between processors
  - Minimize empirical loss over the i-th dataset

- Multi-agent coordination
- Sensor network estimation



Problem Setup

- Undirected graph: $G = (V, E)$
  - $G = \{1, 2, \cdots, n\}$: Vertex set
  - $E \subset V \times V$: Edge set

$$\boxed{\min_x \frac{1}{n} \sum_{i=1}^{n} f_i(x) \ \text{ subject to } x \in \mathcal{X}} \tag{1}$$

  - $f_i : \mathbb{R}^d \to \mathbb{R}$: convex objective associated with agent $i \in V$
  - $\mathcal{X}$: closed and convex set

- Agent i
  - maintains its own parameter vector x_i.
  - has local access to f_i.
  - directly communicates with its neighbors

$$j \in N(i) = \{j \in V | (i, j) \in E\}.$$

Basic Tools and Assumptions

- $\phi : \mathcal{X} \to \mathbb{R}$: proximal function
  - 1-strongly convex w.r.t. the norm $\|\cdot\|$

$$\phi(y) \geq \phi(x) + \langle \nabla\phi(x), y - x\rangle + \frac{1}{2}\|y - x\|^2 \;\; \forall x, y \in \mathcal{X}$$

  - $\phi(x) = \frac{1}{2}\|x\|_2^2$ and ℓ_2-norm
  - $\phi(x) = \sum_{i=1}^{d}(x_i \log x_i - x_i)$ and ℓ_1-norm
- Proximity operator

$$\Pi_{\mathcal{X}}^{\phi}(z, \alpha) = \arg\min_{x \in \mathcal{X}} \left\{ \langle z, x\rangle + \frac{1}{\alpha}\phi(x) \right\}$$

- f_i : $L-$Lipschitz continuous w.r.t. $\|\cdot\|$

$$|f_i(x) - f_i(y)| \leq L\|x - y\|, \;\; \forall x, y, \in \mathcal{X}$$

Standard Dual Averaging

- Generates a primal-dual sequence $\{x(t), z(t)\}_{t=0}^{\infty}$ as

$$\text{Dual Update: } z(t+1) = z(t) + g(t)$$
$$\text{Primal Update: } x(t+1) = \Pi_{\mathcal{X}}^{\phi}(z(t+1), \alpha(t))$$

- $g(t) \in \partial f(x(t))$
- $z(t+1)$: accumulated gradient at $x(t)$
- $\{\alpha(t)\}_{t=0}^{\infty}$: non-increasing step-size sequence

Distributed Dual Averaging (DDA)

At iteration t, each node $i \in V$

- Computes a sub-gradient $g_i(t) \in \partial f_i(x_i(t))$
- Receives dual variables $\{z_j(t), j \in N(i)\}$ from its neighbors
- Performs the updates

Dual Update: $z_i(t+1) = \sum_{j \in N(i)} P_{ji} z_i(t) + g_i(t)$

Primal Update: $x_i(t+1) = \Pi_{\mathcal{X}}^{\phi}(z_i(t+1), \alpha(t))$

- Estimates the optimum via the running local average

$$\hat{x}_i(T) = \frac{1}{T} \sum_{t=1}^{T} x_i(t).$$

Weighting Matrix

- $P \in \mathbb{R}_{+}^{n\times n}$ respects the graph structure, i.e. when $i \neq j$

$$P_{i\,j} > 0 \text{ only if } (i,j) \in E.$$

- P is doubly stochastic,

$$P\,\mathbf{1}_n = \mathbf{1}_n \text{ and } \mathbf{1}_n^T P = \mathbf{1}_n^T.$$

Laplacian Matrix

Let

- $A \in \mathbb{R}^{n \times n}$ be the graph adjacency matrix

$$A_{ij} = \begin{cases} 1 & \text{if } (i,j) \in E \\ 0 & \text{otherwise} \end{cases}$$

- $D = \text{diag}\{\delta_1, \cdots, \delta_n\}$, where $\delta_i = |N(i)|$.
- $\mathcal{L}(G)$ be the Normalized graph Laplacian

$$\boxed{\mathcal{L}(G) = I - D^{-1/2} A D^{-1/2}}$$

Then, a particular choice for P is

$$\boxed{P_n(G) = I - \frac{1}{\delta_{\max}}(D - A) = I - \frac{1}{\delta_{\max} + 1} D^{1/2} \mathcal{L} D^{1/2}.}$$

P is doubly stochastic since $\mathcal{L} D^{1/2} \mathbf{1}_n = 0$.

Theorem 1

For any $x^* \in \mathcal{X}$ and for each $i \in V$, we have

$$f(\hat{x}_i(T)) - f(x^*) \leq \text{OPT} + \text{NET} \tag{2}$$

where

$$\text{OPT} = \frac{1}{T\alpha(T)}\phi(x^*) + \frac{L^2}{2T}\sum_{t=1}^{T}\alpha(t-1) \tag{3}$$

and

$$\text{NET} = \frac{L}{T}\sum_{t=1}^{T}\alpha(t)\left[\frac{2}{n}\sum_{j=1}^{n}\|\bar{z}(t) - z_j(t)\|_* + \|\bar{z}(t) - z_i(t)\|_*\right] \tag{4}$$

with $\bar{z}(t)$ denoting the averaged dual variable

$$\bar{z}(t) = (1/n)\sum_{i=1}^{n} z_i(t).$$

Sketch of the Proof

- Step 1: $\bar{z}(t)$ evolves in a very simple way:

$$\bar{z}(t+1) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(P_{ji}(z_j(t) - \bar{z}(t))\right) + \bar{z}(t) + \frac{1}{n}\sum_{j=1}^{n} g_j(t)$$
$$= \bar{z}(t) + \frac{1}{n}\sum_{j=1}^{n} g_j(t)$$

  Similar update as in the centralized case

- Step2: Define $y(t) = \Pi_{\mathcal{X}}^{\phi}(\bar{z}(t), \alpha(t-1))$. Then

$$\sum_{t=1}^{T} f(x_i(t)) - f(x^*) \leq \sum_{t=1}^{T} f(y(t)) - f(x^*) + L\sum_{t=1}^{T}\alpha(t)\|\bar{z}(t) - z_i(t)\|_*$$

  which is due to the Lipschitz continuity of the proximity operator.

Sketch of the Proof (Cont.)

- Step 3: $L-$ Lipschitz continuity of f_i implies

$$n\sum_{t=1}^{T} f(y(t)) - f(x^*) \leq \sum_{t=1}^{T}\sum_{i=1}^{n}[f_i(x_i(t)) - f_i(x^*) + L\|y(t) - x_i(t)\|]$$

- Step 4:

$$\begin{aligned}\sum_{i=1}^{n} f_i(x_i(t)) - f_i(x^*) &\leq \sum_{i=1}^{n}\langle g_i(t), x_i(t) - x^*\rangle \\ &\leq \frac{1}{2}\sum_{t=1}^{T}\alpha(t-1)\|g(t)\|_*^2 + \frac{1}{\alpha(T)}\phi(x^*)\end{aligned}$$

- Lipschitz continuity of the proximity operator can be used to bound

$$\|y(t) - x_i(t)\|$$

Theorem 2

- Effects of network topology on convergence rates.

Let

- $\gamma(P) = 1 - \sigma_2(P)$ be the spectral gap of P,
- $\phi(x^*) \leq R^2$
- $\alpha(t) = R\sqrt{\gamma(P)}/(4L\sqrt{t})$.

Then

$$f(\hat{x}_i(T)) - f(x^*) \leq \frac{R\,L}{\sqrt{T}} \cdot \frac{\log(T\sqrt{n})}{\sqrt{\gamma(P)}} \tag{5}$$

for all $i \in V$.

- Information propagation through the network depends on the spectral gap.

Interesting Network Topologies

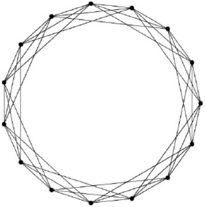


Figure: 3-Connected cycle

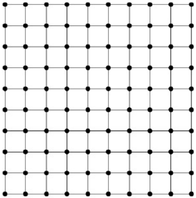


Figure: 1-Connected Grid Graph

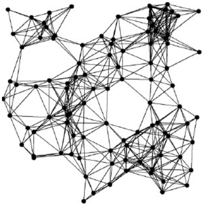


Figure: Random Geometric Graph

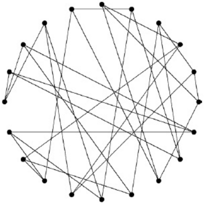


Figure: 3-regular Expande

Convergence Rate

Network Topology	$f(\hat{x}_i(T)) - f(x^*)$	Comments
$k-$connected cycles and pathes	$\mathcal{O}\left(\frac{R\,L}{\sqrt{T}} \cdot \frac{n\,\log(T\,n)}{k}\right)$	Poorly Connected for small k
$k-$connected $\sqrt{n} \times \sqrt{n}$ grids	$\mathcal{O}\left(\frac{R\,L}{\sqrt{T}} \cdot \frac{\sqrt{n}\,\log(T\,n)}{k}\right)$	
Random Geometric Graph with connectivity radius $r = \Omega\left(\sqrt{\log^{1+\epsilon} n/n}\right)$	$\mathcal{O}\left(\frac{R\,L}{\sqrt{T}} \cdot \sqrt{\frac{n}{\log n}}\,\log(T\,n)\right)$	Bound holds with high probability
Expanders with Bounded $\frac{\delta_{\max}}{\delta_{\min}}$	$\mathcal{O}\left(\frac{R\,L}{\sqrt{T}} \cdot \log(T\,n)\right)$	Highly Connected

Iteration Complexity Analysis

- $T_G(\epsilon; n)$: number of iterations to achieve error ϵ for G
- Theorem 2 implies that

$$T_G(\epsilon;) = \mathcal{O}\left(\frac{1}{\epsilon^2} \cdot \frac{1}{1 - \sigma_2(P_n(G))}\right). \quad (6)$$

Single Cycle Graph	Two-Dimensional Grid	Bounded Degree Expander
$\mathcal{O}(n^2/\epsilon^2)$	$\mathcal{O}(n/\epsilon^2)$	$\mathcal{O}(1/\epsilon^2)$

The bound (6) is sharp:

- Sub-gradient methods achieve $\epsilon-$ accuracy in $\Omega(1/\epsilon^2)$ iterations.

- Let $\phi(x) = \frac{1}{2}\|x\|_2^2$.
- For any graph G with n nodes, DDA achieves $\epsilon-$accuracy if

$$T_G(c; n) = \Omega\left(\frac{1}{1 - \sigma_2(P_n(G))}\right).$$

Stochastic Communication Links

- Time-varying communication matrix $P(t)$
- Example 1: Random edge selection in dense networks
  - Reduces network congestion
- Example 2: Link failures in real networks

Theorem 3: Let

- $\{P(t)\}_{t=0}^{\infty}$ be an i.i.d. sequence of doubly stochastic matrices.
- $\lambda_2(G) = \lambda_2(\mathbb{E}[P(t)^T P(t)])$.
- $\alpha(t) \propto R\sqrt{1-\lambda_2}/(L\sqrt{t})$.

Then with probability at least $1-(1/T)$

$$f(\hat{x}_i(T)) - f(x^*) \le c\frac{R\,L}{\sqrt{T}} \cdot \frac{\log(Tn)}{\sqrt{1-\lambda_2(G)}}. \tag{7}$$

Stochastic Gradient Algorithm

- Gradients corrupted with zero-mean and bounded-variance noise
- Let $\mathcal{F}_{t-1}$ be the $\sigma-$field containing all the information up to time $t-1$, i.e.

$$g_i(1), \cdots, g_i(t-1) \in \mathcal{F}_{t-1}$$
$$x_i(1), \cdots, x_i(t) \in \mathcal{F}_{t-1}$$

  for all $i \in V$.
- A stochastic oracle provides gradients estimates satisfying

$$\boxed{\mathbb{E}[\hat{g}_i(t)|\mathcal{F}_{t-1}] \in \partial f_i(x_i(t)) \text{ and } \mathbb{E}\left[\|\hat{g}_i(t)\|_*^2|\mathcal{F}_{t-1}\right] \leq L^2} \quad (8)$$

- The model includes the additive noise oracle.

Theorem 4

Assume

- $\hat{g}_i(t)$ is provided by the stochastic oracle ($\|\hat{g}_i(t)\|_* \leq L$),
- $\mathcal{X}$ has finite radius $R = \sup_{x \in \mathcal{X}} \|x - x^*\|$.

Then with probability $1 - \delta$ we have

$$f(\hat{x}_i(T)) - f(x^*) \leq \mathsf{OPT} + \mathsf{NET} + 8LR\sqrt{\frac{\log \frac{1}{\delta}}{T}}$$

where

$$\mathsf{OPT} = \frac{1}{T\alpha(T)}\phi(x^*) + \frac{8L^2}{T}\sum_{t=1}^{T}\alpha(t-1)$$

and

$$\mathsf{NET} = \frac{3L^2}{T} \cdot \frac{\log(T\sqrt{n})}{1 - \sigma_2(P)}\sum_{t=1}^{T}\alpha(t).$$

Simulation Setup

- Sum of ℓ_1-regression loss functions:

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} |y_i - \langle b_i, x\rangle| = \frac{1}{n}\|y - Bx\|_1$$

  where $(b_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ is a training data point.

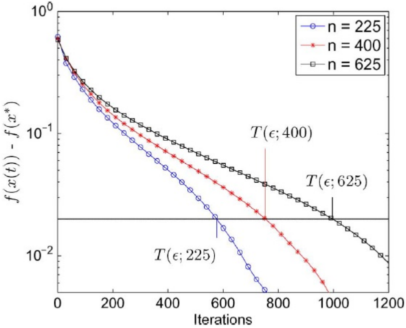
- f is $L-$Lipschitz with $L = \max_i \|b_i\|_2$.
- $\mathcal{X} = \{x \in \mathbb{R}^d | \|x\|_2 \leq 5\}$
- Graph size $= n =$ size of dataset
- Three different graph structures:
  - Single cycle
  - Two dimensional Grid
  - $5-$regular expanders

Simulation Results 1

- Grid graph: $n = 225, 400, 625$
- Error function:

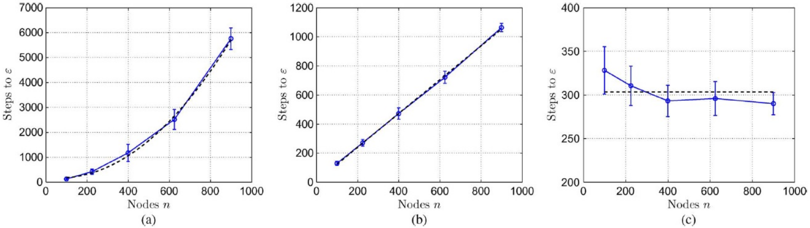
$$\max_i [f(\hat{x}_i(t)) - f(x^*)]$$

- Convergence time $T_G(\epsilon; n)$ scales with n.



Simulation Results 2

- $T_G(\epsilon; n)$ with $\epsilon = 0.1$ versus the graph size n



- Three graph structures:
  - Panel (a): Single cycle: $T_G(\epsilon; n) = \mathcal{O}(n^2)$
  - Panel (b): Grid Graph: $T_G(\epsilon; n) = \mathcal{O}(n)$
  - Panel (c): 5-regular Expander: $T_G(\epsilon; n) = \mathcal{O}(1)$
- Blue curves: Average of 20 trials
- Dashed curves: Theoretical predictions