Communication-Efficient Distributed Second-Order Optimization Methods for Generalized Convex Problems

Rixon Crane (PhD Student, School of Mathematics and Physics, The University of Queensland)
Fred Roosta (Lecturer, School of Mathematics and Physics, The University of Queensland)

Every Minute There Are:

~500,000 Tweets Sent

~50,000 Photos Posted

~100,000 Hours Streamed

~4,000,000 Searches

~750,000 Songs Streamed

~2,000,000 Snaps

The Problem

Centralized Computing Environment

Worker node i with $f_i : \mathbb{R}^d \to \mathbb{R}$

Driver node

Goal

Over m workers, solve:

$$\min_{w \in \mathbb{R}^d} \left\{ f(w) \stackrel{\text{def}}{=} \frac{1}{m} \sum_{i=1}^{m} f_i(w) \right\}.$$

Use Case: Big Data Regimes

Distributively Working With a Very Large Dataset $\{x_i\}_{i=1}^n$

Worker i stores data indexed by set $S_i \subseteq \{1, \ldots, n\}$, i.e., stores $\{x_j\}_{j \in S_i}$

Why use Second-Order Methods?

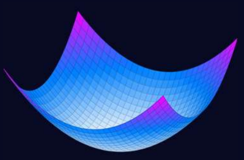
Second-order methods employ curvature (Hessian matrix) information to transform the gradient so that it is a more suitable direction to follow.

Benefits

- Perform more computations per iteration
- May take full advantage of available distributed computational resources
- May require significantly less communication costs
- Often require far fewer iterations to achieve similar results

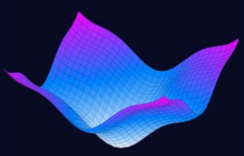
Related Work

Method	Applicable to Non-Convex Functions	Arbitrary Data Distribution	Arbitrary Form of f_i	Simple Sub-Problems	Not Sensitive to Hyper-Parameters
GIANT	✗	✗	✗	✓	✓
DiSCO	✗	✓	✓	✓	✓
DANE	✓	✓	✓	✗	✗
InexactDANE	✓	✓	✓	✗	✗
AIDE	✓	✓	✓	✗	✗
DINGO	✓	✓	✓	✓	✓



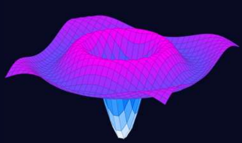
Convex

- Hessian is positive semidefinite.
- Local minima are global minima.



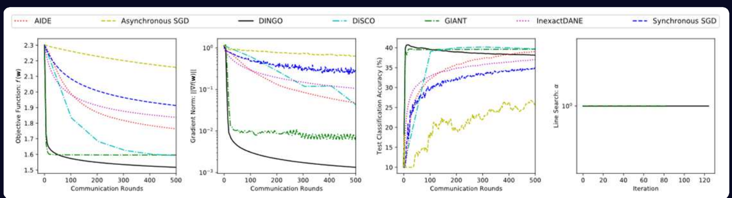
Invex

- Hessian can be indefinite and singular.
- Local minima are global minima.



Non-Convex

- Hessian can be indefinite and singular.
- Not all local minima are global minima.



Softmax regression, with regularization, problem on the CIFAR10 dataset.

Our Method: DINGO

Derived by optimization of the gradient's norm as a surrogate function, i.e.,

$$\min_{w \in \mathbb{R}^d} \left\{ \frac{1}{2} \|\nabla f(w)\|^2 = \frac{1}{2m^2} \left\| \sum_{i=1}^{m} \nabla f_i(w) \right\|^2 \right\}.$$

DINGO is for "**DI**stributed **N**ewton-type method for **G**radient-norm **O**ptimization". DINGO is particularly suitable for invex objectives. A strict linear-rate reduction in the gradient norm is always guaranteed.

Each Iteration of DINGO

w_t

$g_{t,i}$

$g_t = \frac{1}{m}\sum_{i=1}^m g_{t,i}$

$H_{t,i} g_t, H_{t,i}^\dagger g_t, \tilde{H}_{t,i}^\dagger \tilde{g}_t$

$\frac{1}{m}\sum_{i=1}^m H_{t,i}^\dagger g_t, \quad \frac{1}{m}\sum_{i=1}^m \tilde{H}_{t,i}^\dagger \tilde{g}_t, \quad H_t g_t = \frac{1}{m}\sum_{i=1}^m H_{t,i} g_t$

Update Direction: p_t

Case 1

If: $\langle \frac{1}{m}\sum_{i=1}^m H_{t,i}^\dagger g_t, H_t g_t \rangle \geq \theta \|g_t\|^2$ then $p_t = \frac{1}{m}\sum_{i=1}^m p_{t,i}$ with $p_{t,i} = -H_{t,i}^\dagger g_t$

Case 2

Else If: $\langle \frac{1}{m}\sum_{i=1}^m \tilde{H}_{t,i}^\dagger \tilde{g}_t, H_t g_t \rangle \geq \theta \|g_t\|^2$ then $p_t = \frac{1}{m}\sum_{i=1}^m p_{t,i}$ with $p_{t,i} = -\tilde{H}_{t,i}^\dagger \tilde{g}_t$

Case 3

Else: $p_{t,i} = -\tilde{H}_{t,i}^\dagger \tilde{g}_t$ for all $i \notin I_t \stackrel{\text{def}}{=} \{ i = 1, \ldots, m \mid \langle \tilde{H}_{t,i}^\dagger \tilde{g}_t, H_t g_t \rangle < \theta \|g_t\|^2 \}$

Send $H_t g_t$ to all workers $i \in I_t$

$p_{t,i} = -\tilde{H}_{t,i}^\dagger \tilde{g}_t - \lambda_{t,i} \left(\tilde{H}_{t,i}^T \tilde{H}_{t,i}\right)^{-1} H_t g_t, \qquad \lambda_{t,i} = \frac{-g_t^T H_t \tilde{H}_{t,i}^\dagger \tilde{g}_t + \theta \|g_t\|^2}{g_t^T H_t \left(\tilde{H}_{t,i}^T \tilde{H}_{t,i}\right)^{-1} H_t g_t}$

$p_t = \frac{1}{m}\sum_{i=1}^m p_{t,i}$

Line Search: α_t

Distributively choose largest $\alpha_t \in (0,1]$ such that $\|\nabla f(w_t + \alpha_t p_t)\|^2 \leq \|g_t\|^2 + 2\alpha_t \rho \langle p_t, H_t g_t \rangle$.

Example: Backtracking Line Search

p_t

Compute $\nabla f_i(w_t + \alpha p_t)$ $\forall \alpha \in \{1, \omega, \ldots, \omega^k\}$, where $\omega \in (0,1), k \in \mathbb{N}$

Let $\alpha_t = \max\{1, \ldots, \omega^k\}$ such that $\|\nabla f(w_t + \alpha_t p_t)\|^2 \leq \|g_t\|^2 + 2\alpha_t \rho \langle p_t, H_t g_t \rangle$

Update

$w_{t+1} = w_t + \alpha_t p_t$

The constants $\theta, \phi > 0$ and $\rho \in (0,1)$ are hyper-parameters. The vector $w_t \in \mathbb{R}^d$ denotes the point at iteration t. For notational convenience, we denote $g_{t,i} \stackrel{\text{def}}{=} \nabla f_i(w_t)$, $H_{t,i} \stackrel{\text{def}}{=} \nabla^2 f_i(w_t)$, $g_t \stackrel{\text{def}}{=} \nabla f(w_t)$, $H_t \stackrel{\text{def}}{=} \nabla^2 f(w_t)$. We also let

$$\tilde{H}_{t,i} \stackrel{\text{def}}{=} \begin{bmatrix} H_{t,i} \\ \phi I \end{bmatrix} \in \mathbb{R}^{2d \times d}, \qquad \tilde{g}_t \stackrel{\text{def}}{=} \begin{bmatrix} g_t \\ 0 \end{bmatrix} \in \mathbb{R}^{2d},$$

where I is the identity matrix and 0 is the zero vector. Green and purple rectangles represent the driver node and worker nodes, respectively.

References

1. Crane, R., & Roosta, F. (2019). DINGO: Distributed Newton-Type Method for Gradient-Norm Optimization. *arXiv preprint arXiv:1901.05134*.
2. https://www.domo.com/blog/data-never-sleeps-6