Distributed Gradient-Regularized Newton Method: Scheduled Consensus and $O(\varepsilon^{-1})$ Global Iteration Complexity

MOA 2026 International Workshop on Modern Optimization and Applications

Wei Hu, Pengcheng Xie, Ya-Xiang Yuan, Li Zhang
Academy of Mathematics and Systems Science, Chinese Academy of Sciences

Problem setup

Introduction of distributed optimization.

We want to solve

$$\min_{x\in\mathbb{R}^d} f(x) := \frac{1}{N}\sum_{i=1}^{N} f_i(x), \qquad \text{node } i \text{ only knows } f_i.$$

Equivalently, over a connected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$,

$$\min_{x_1,\dots,x_N} \frac{1}{N}\sum_{i=1}^{N} f_i(x_i), \qquad \text{s.t. } x_i = x_j, \quad (i,j)\in\mathcal{E}.$$

Why distributed?

- Split large problems into local subproblems.
- Use distributed compute and storage.
- Match distributed data structures.

Distributed algorithms.

- Local computation uses partial objectives.
- Neighbor communication controls consensus.

Both computation and communication costs matter.

Motivation: why practical second-order distributed algorithms?

First-order drawbacks.

- Slow convergence on ill-conditioned problems.
- High accuracy needs many communication rounds.

Second-order challenges.

1. **Averaging–inversion gap:**

$$\frac{1}{N}\sum_{i=1}^{N} A_i^{-1} g_i \neq \left(\frac{1}{N}\sum_{i=1}^{N} A_i\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} g_i\right).$$

   Local Newton averages may miss the centralized step.
2. **Initialization and step-size sensitivity:** decentralized line search is hard.
3. **Indefinite Hessian surrogates:** tracking errors can destroy PSD structure.

Our approach

1. Two-stage scheduled mixing
Pre-mixing moves local inputs toward averages; post-mixing controls disagreement.
Depths scale as

$$O\Big((1-\rho)^{-1}\log(k+2)\Big).$$

2. Gradient-regularized solve
The tracker approximates the global gradient, $\tilde{g}_{i,k} \approx \nabla f(\bar{x}_k)$. Node i uses

$$\lambda_{i,k} = \sqrt{M_{i,k}\,\|\tilde{g}_{i,k}\|}.$$

No global norm or line search is required.

3. Eigenvalue-shift stabilizer
Set

$$\delta_{i,k} = \max\{0, -\lambda_{\min}(\tilde{H}_{i,k})\},$$

so that $\tilde{H}_{i,k} + \delta_{i,k} I \succeq 0$.
With $\lambda_{i,k} > 0$, the local system is positive definite.

step-size-free + robust + efficient

Algorithm and practical variants

Notation. Let $\mathrm{Mix}_t(Z) := W^t Z$, $\rho = \|W - J\|_2 < 1$, and $J = \frac{1}{N}11^\top$.
Rows of $X_k, G_k, \mathcal{H}_k$ stack local variables; $\nabla F(X_k)$ and $\nabla^2 F(X_k)$ stack derivatives rowwise.

Pre-mix → Newton step → Post-mix → Trackers

One outer iteration. The four stages are

$$\begin{aligned}
\textbf{Pre-mix:}\quad & (\tilde{X}_k, \tilde{G}_k, \tilde{\mathcal{H}}_k) = \mathrm{Mix}_{\tau_k}(X_k, G_k, \mathcal{H}_k),\\
\textbf{Regularize:}\quad & \lambda_{i,k} = \sqrt{M_{i,k}\,\|\tilde{g}_{i,k}\|}, \quad \delta_{i,k} = \max\{0, -\lambda_{\min}(\tilde{H}_{i,k})\},\\
\textbf{Newton step:}\quad & s_{i,k} = -\big[\tilde{H}_{i,k} + (\lambda_{i,k}+\delta_{i,k})I\big]^{-1}\tilde{g}_{i,k},\\
& S_k = (s_{1,k},\dots,s_{N,k})^\top,\\
\textbf{Post-mix:}\quad & X_{k+1} = \mathrm{Mix}_{t_k}(\tilde{X}_k + S_k),\\
\textbf{Trackers:}\quad & D^g_{k+1} = \nabla F(X_{k+1}) - \nabla F(X_k), \quad D^H_{k+1} = \nabla^2 F(X_{k+1}) - \nabla^2 F(X_k),\\
& G_{k+1} = \mathrm{Mix}_{t_k}(\tilde{G}_k + D^g_{k+1}), \quad \mathcal{H}_{k+1} = \mathrm{Mix}_{t_k}(\tilde{\mathcal{H}}_k + D^H_{k+1}).
\end{aligned}$$

If $\tilde{g}_{i,k} = 0$, use the convention $s_{i,k} = 0$.

Variants. Adaptive–M variants choose $M_{i,k}$ from local Hessian secants:

$$\widehat{L}_{i,k} = \frac{\|\nabla^2 f_i(x_{i,k}) - \nabla^2 f_i(x_{i,k-1})\|}{\|x_{i,k} - x_{i,k-1}\|}, \qquad \lambda_{i,k} = \sqrt{\widehat{M}_{i,k}\,\|\tilde{g}_{i,k}\|}.$$

Communication-efficient variants send compressed Hessian increments $C(\Delta H_{i,k})$ by Top-k or low-rank symmetric compression. Lazy updates are also considered.

References.
[1] W. Hu, P. Xie, Y.-X. Yuan, and L. Zhang, *Distributed Gradient-Regularized Newton Method*, arXiv:2605.19396, 2026.
[2] K. Mishchenko, *Regularized Newton Method with Global $O(1/k^2)$ Convergence*, SIAM J. Optim., 33(3):1440–1462, 2023.

Code available at: `https://github.com/huwei0121/DisGRem`

Theoretical results

Assumption set.

Each f_i is convex and C^2, with L_1-Lipschitz gradient and L_2-Lipschitz Hessian; $\mathcal{G}$ is connected; W is symmetric doubly stochastic with $\rho < 1$; the generated trajectory is bounded; trackers start from exact local gradients/Hessians; and $M \geq L_2$.

Theorem 1: global complexity.

Run DisGrem with $p \geq 3$ and logarithmic two-stage mixing. Let

$$K(\varepsilon) = \min\{k : \|\nabla f(\bar{x}_k)\| \leq \varepsilon\}, \qquad 0 < \varepsilon \leq 1.$$

There exist $K_0(\varepsilon) < \infty$ and C_K independent of ε such that

$$K(\varepsilon) \leq K_0(\varepsilon) + C_K \varepsilon^{-1}, \qquad K_0(\varepsilon) = O(\varepsilon^{-2/(p-1)}).$$

$$K(\varepsilon) = O(\varepsilon^{-1}) \quad \textbf{for scheduled mixing with } p \geq 3$$

The neighbor-round complexity is

$$\sum_{k\leq K(\varepsilon)} (\tau_k + 2t_k) = O\Big((1-\rho)^{-1}\varepsilon^{-1}\log(1/\varepsilon)\Big).$$

Theorem 2: local superlinear tail.

Assume additionally that f is locally strongly convex near $x_\star$. If, along the local tail, consensus and tracking errors are relatively accurate with respect to the current gradient, then for all sufficiently large k,

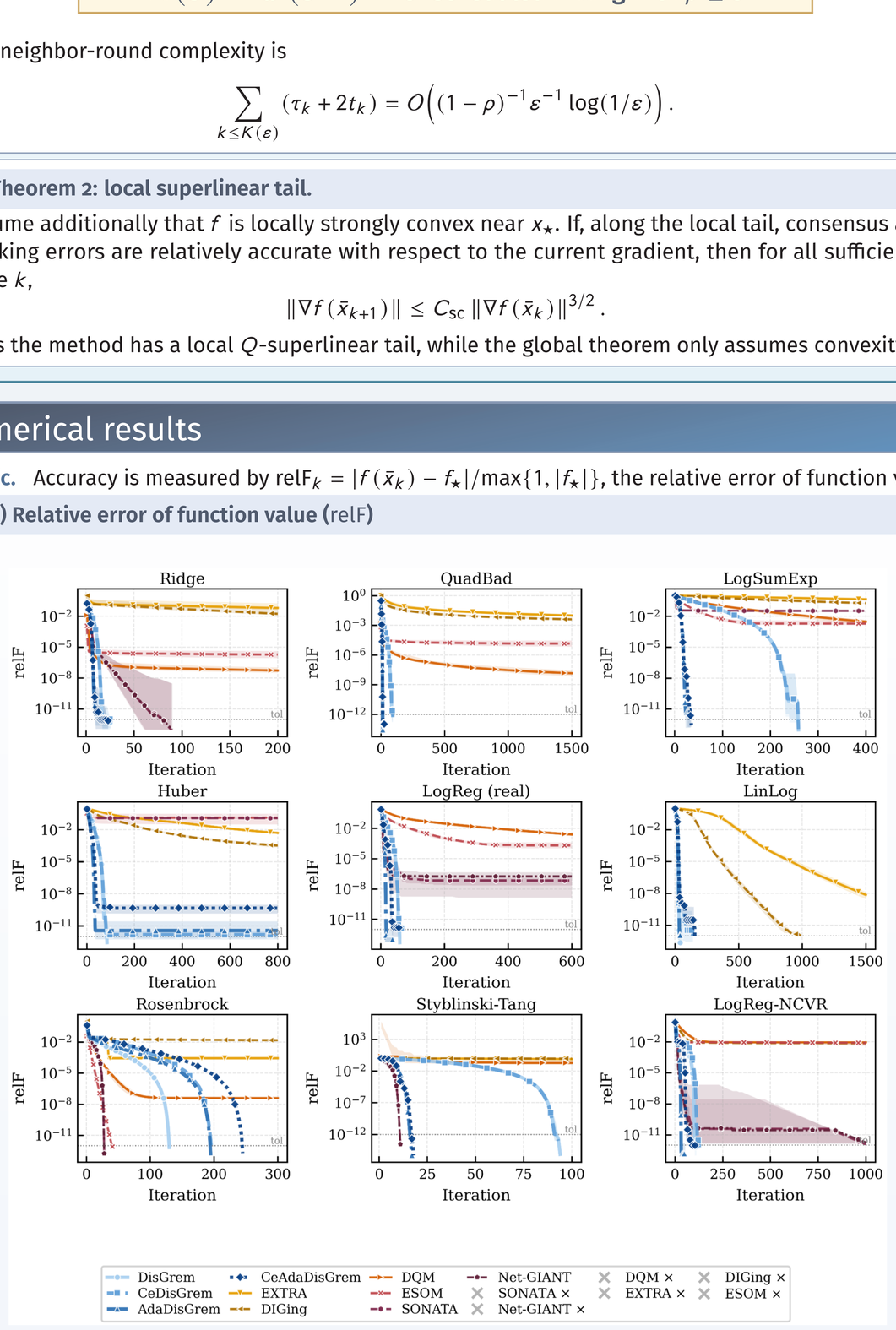
$$\|\nabla f(\bar{x}_{k+1})\| \leq C_{\mathrm{sc}}\,\|\nabla f(\bar{x}_k)\|^{3/2}.$$

Thus the method has a local Q-superlinear tail, while the global theorem only assumes convexity.

Numerical results

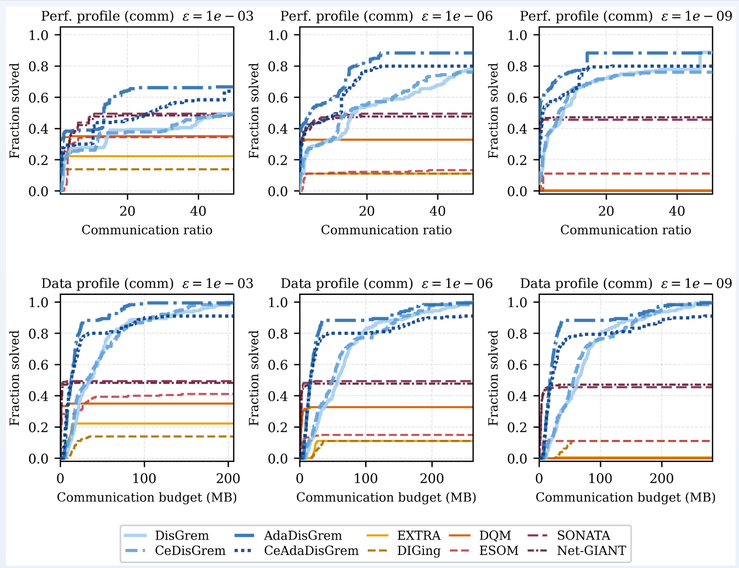
Metric. Accuracy is measured by $\mathrm{relF}_k = |f(\bar{x}_k) - f_\star|/\max\{1, |f_\star|\}$, the relative error of function value.

(a) Relative error of function value (relF)



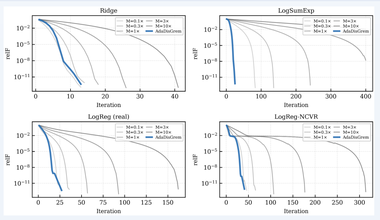
Representative convergence histories on nine benchmark problems; the vertical axis is relF.

(b) Communication profiles



Profiles measured by communication ratio and cumulative MB.

(c) Adaptive mechanism



Fixed-M choices versus AdaDisGrem.