Decentralized Riemannian gradient descent on Stiefel manifold

Shixiang Chen[1]
(with Alfredo Garcia[1], Mingyi Hong[2], Shahin Shahrampour[1])

[1]Wm Michael Barnes '64 Department of Industrial and Systems Engineering, Texas A&M University
[2] The Department of Electrical and Computer Engineering, University of Minnesota

TEXAS A&M UNIVERSITY

UNIVERSITY OF MINNESOTA

ICML | 2021
Thirty-eighth International Conference on Machine Learning

Decentralized Optimization Problem

$$\min \frac{1}{n}\sum_{i=1}^{n} f_i(x_i)$$

consensus constraint

$$s.t. \quad x_1 = x_2 = \ldots = x_n$$

$$x_i \in \mathcal{M} \quad \forall i = 1, \ldots, n$$

$$\Longleftrightarrow$$

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

$$s.t. \quad x \in \mathcal{M}$$

(centralized problem)

$$\boldsymbol{x}^\top \boldsymbol{x} = \boldsymbol{I_r}$$

Here $\mathcal{M} = \{x \in \mathbf{R}^{d\times r} : x^\top x = I_r\}$ is the Stiefel manifold.

1. We assume that $f_i(x_i)$ is Lipschitz smooth.

2. The network is associated with a doubly stochastic matrix W.

Motivations

1. Privacy

The datasets are collected, stored in distributed manner. To protect users' privacy, the central server is not allowed.

The deterministic decentralized method is treated as a compromise. Sparser/larger network, lower convergence rate.

2. Acceleration in stochastic algorithms

The decentralized setting is more communication-efficient

- For decentralized stochastic gradient descent(SGD), each node takes the same computation complexity as that of the centralized SGD (Lian et al., 2017)

Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. In Advances in Neural Information Processing Systems, pages 5330–5340, 2017.

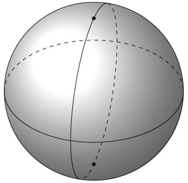
Challenges

1. The Stiefel manifold is a nonconvex set in Euclidean space. Previous results do not apply…

   Nedic et al., 2010; Shi et al., 2015; Di Lorenzo & Scutari, 2016; Qu & Li, 2017; Nedic et al., 2017; Lian et al., 2017;...

   To achieve local linear consensus on Stiefel manifold, variables should stay in a local region, denoted by $\mathcal{N}$. (Chen, et al. 2021)

   For example, on the sphere, $\mathcal{N}$ is the hemisphere.



Sphere $S^2 \cong St(3,1) \subset R^3$

2. The manifold is nonlinear space…

   We need to use Riemannian optimization tools

   Euclidean consensus $\sum_{i=1}^{n} W_{i,j} x_i$ is not feasible.

Shixiang Chen, Alfredo Garcia, Mingyi Hong, and Shahin Shahrampour. On the local linear rate of consensus on the stiefel manifold. arXiv preprint arXiv:2101.09346, 2021.

Contributions

1. We propose a Decentralized Riemannian stochastic gradient descent algorithm(DRSGD). We show

   (i) DRSGD can achieve linear speedup w.r.t the nodes number n. The convergence rate to stationary point is $\mathcal{O}(1/\sqrt{nk})$ for sufficiently large k.

   (ii) DRSGD is faster than the corresponding centralized Riemannian SGD.

2. We propose the first Decentralized Riemannian gradient tracking algorithm (DRGTA).

   (i) DRGTA can use constant stepsize

   (ii) The convergence rate is $\mathcal{O}(1/k)$

Algorithm 1: Decentralized Riemannian Stochastic gradient descent(DRSGD)

DRSGD: stepsize $\alpha > 0$, $\beta > 0$, an integer $t \geq \log_{\sigma_2}(\frac{1}{2n})$

At each node i :

1. Choose $v_{i,k}$ s.t. $\mathrm{E}v_{i,k} = grad f_i(x_{i,k})$

2. $$x_{i,k+1} = \mathcal{R}_{x_{i,k}}(\underbrace{\alpha P_{T_{x_{i,k}}\mathcal{M}}(\textstyle\sum_{j=1}^{n} W_{ij}^t x_{j,k})}_{\text{Multi-step Consensus; also preserve } x_{i,k+1} \in \mathcal{N}} - \beta v_{i,k}).$$

(Chen, et al. 2021)

minimize the function

$\mathcal{R}_{x_{i,k}}$: retraction mapping helps preserve feasibility

$P_{T_{x_{i,k}}\mathcal{M}}$: orthogonal projection onto the tangent space $T_{x_{i,k}}\mathcal{M}$

Assumption:

1. $v_{i,k}$ and $v_{j,k}$ are independent for any i, j
2. unbiased and bounded variance:

$$\mathrm{E} v_{i,k} = grad f_i(x_{i,k})$$

$$\mathrm{E}\|v_{i,k} - grad f_i(x_{i,k})\|^2 \leq \Xi$$

3. uniform bound: $\|v_{i,k}\| \leq D$ for all i, k.

Convergence:

If K is sufficiently large, $\beta_k \equiv \beta = \frac{1}{2L_G + \Xi\sqrt{(K+1)/n}}$, $t \geq \lceil \log_{\sigma_2}(\frac{1}{2\sqrt{n}}) \rceil$, $\mathbf{x}_0 \in \mathcal{N}$ one has $\mathbf{x}_k \in \mathcal{N}$

$$\min_{k=0,...,K} \mathrm{E}\|grad f(\bar{x}_k)\|_F^2 = O(\frac{1}{\sqrt{nK}})$$

the linear speedup w.r.t n

Algorithm 2: Decentralized Riemannian Gradient Tracking (DRGTA)

Key idea of DRGTA: auxiliary sequence $\{y_{1,k}, \ldots, y_{n,k}\}$ estimates the Riemannian gradient

At each node i :

projection onto tangent space, better estimation

1. $v_{i,k} = P_{T_{x_{i,k}}\mathcal{M}} y_{i,k}$

2. $x_{i,k+1} = \mathcal{R}_{x_{i,k}}(\alpha P_{T_{x_{i,k}}\mathcal{M}}\overbrace{(\sum_{j=1}^{n} W_{ij}^{t} x_{j,k})}^{\text{Multi-step Consensus}} - \beta v_{i,k}).$

3. $y_{i,k+1} = \sum_{j=1}^{n} W_{ij}^{t} y_{j,k} + \mathrm{grad} f_i(x_{i,k+1}) - \mathrm{grad} f_i(x_{i,k}).$

Riemannian gradient tracking step

Extension of the DIGing algorithm (Qu & Li, 2017; Nedic et al., 2017)

Convergence of DRGTA

If $\beta = O(\frac{(1-\rho_t)^2}{L_G})$, $\mathbf{x}_0 \in \mathcal{N}$, $t \geq \lceil \log_{\sigma_2}(\frac{1}{2\sqrt{n}}) \rceil$ then $\mathbf{x}_k \in \mathcal{N}$. And the following holds

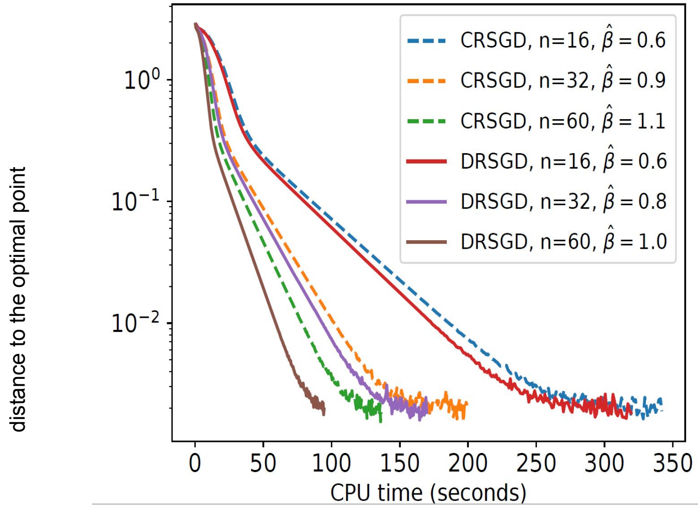
$$\min_{k \leq K} \frac{1}{n} \|\mathbf{x}_k - \bar{\mathbf{x}}_k\|_F^2 = O(\frac{1}{K})$$

consensus error

$$\min_{k=0,\ldots,K} \mathrm{E}\|grad f(\bar{x}_k)\|_F^2 = O(\frac{1}{K})$$

stationarity

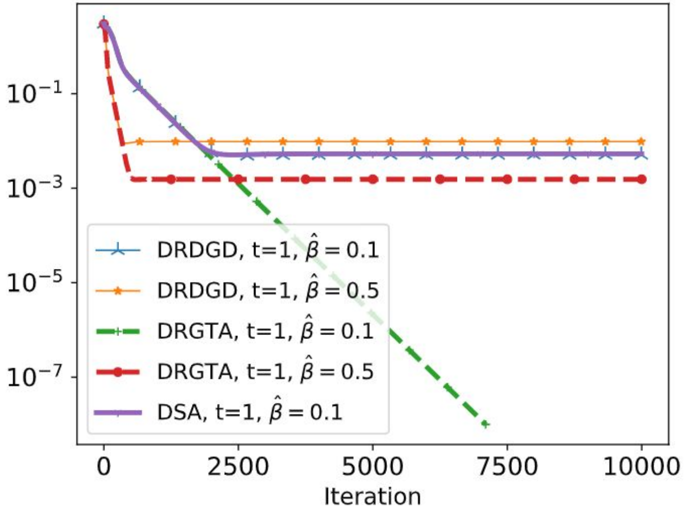
Numerical results: PCA on MNIST dataset



$$\beta_k = \frac{\sqrt{n}}{10000\sqrt{300}}\hat{\beta}.$$

DRSGD v.s Centralized Riemannain stochastic gradient descent (CRSGD)

Convergence of DRGTA



(a) MNIST, $n = 20$, ring graph

$$\beta_k = \frac{\hat{\beta}}{60000}.$$

DRGTA

References

David Kempe and Frank McSherry. A decentralized algorithm for spectral analysis. Journal of Computer and System Sciences, 74(1):70–83, 2008.

Haroon Raja and Waheed U Bajwa. Cloud k-svd: A collaborative dictionary learning algorithm for big, distributed data. IEEE Transactions on Signal Processing, 64(1):173–188, 2015.

Martin Arjovsky, Amar Shah, and Yoshua Bengio. Unitary evolution recurrent neural networks. In International Conference on Machine Learning, pages 1120–1128. PMLR, 2016.

Eugene Vorontsov, Chiheb Trabelsi, Samuel Kadoury, and Chris Pal. On orthogonality and learning recurrent networks with long term dependencies. In International Conference on Machine Learning, pages 3570–3578. PMLR, 2017.

Lei Huang, Xianglong Liu, Bo Lang, Adams Yu, Yongliang Wang, and Bo Li. Orthogonal weight normalization: Solution to optimization over multiple dependent stiefel manifolds in deep neural networks. In Proceedings of the AAAI Conference on Artificial Intelligence, 2018.

Angelia Nedic, Asuman Ozdaglar, and Pablo A Parrilo. Constrained consensus and optimization in multi-agent networks. IEEE Transactions on Automatic Control, 55(4):922–938, 2010.

Wei Shi, Qing Ling, Gang Wu, and Wotao Yin. Extra: An exact first-order algorithm for decentralized consensus optimization. SIAM Journal on Optimization, 25(2):944–966, 2015.

Guannan Qu and Na Li. Harnessing smoothness to accelerate distributed optimization. IEEE Transactions on Control of Network Systems, 5(3):1245–1260, 2017.

Paolo Di Lorenzo and Gesualdo Scutari. Next: In-network nonconvex optimization. IEEE Transactions on Signal and Information Processing over Networks, 2(2):120–136, 2016.

Angelia Nedic, Alex Olshevsky, and Wei Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. SIAM Journal on Optimization, 27(4):2597–2633, 2017.

Thank you!