Tackling Data Heterogeneity: A New Unified Framework for Decentralized SGD with Sample-induced Topology

Yan Huang[1], Ying Sun[2], Zehan Zhu[1], Changzhi Yan[1], Jinming Xu[1,*]

[1]Zhejiang University, Hangzhou 310027, China
[2]The Pennsylvania State University, PA 16802, USA

*jimmyxu@zju.edu.cn

The 39th International Conference on Machine Learning (ICML 2022)

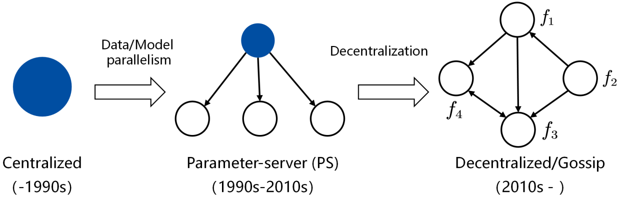
Motivation & Background

- Collaborative Empirical Risk Minimization (ERM)

$$\min_{x \in \mathbb{R}^d} f(x) = \frac{1}{n} \sum_{i=1}^{n} \left(f_i(x) := \frac{1}{m} \sum_{j=1}^{m} \underbrace{f(x, \xi_{i,j})}_{f_{ij}(x)} \right)$$

  - $\xi_{i,j} \sim \mathcal{D}_i$ is the j-th samples of the local dataset $\mathcal{D}_i$ of device i

- Typical topology structures



Motivation & Background

- Challenges:
  - Sampling variance within devices: $f_{ij} \neq f_i$, for all device i.
  - Data heterogeneity across devices: $f_i \neq f$, for any device i.
- Existing first-order optimization methods
  - **SGD-based methods**: DSGD (Ram et al., 2009), Local-SGD (Konevcnỳ et al., 2016), Gossip-PGA (Chen et al., 2021); $\rightarrow$ efficiency
  - **Variance-Reduction (VR)**: SAGA (Defazio et al., 2014), (L-)SVRG (Qian et al., 2021), SARAH (Nguyen et al., 2017); $\rightarrow$ ~~inner variance~~
  - **Gradient-Tracking (GT)**: DSGT (Pu and Nedić, 2020), DSA (Mokhtari and Ribeiro, 2016), GT-VR (Xin et al., 2020); $\rightarrow$ ~~external variance~~
- **Question**: Can we unify these above methods and beyond?

Related Work

- The existing state-of-the-art framework

	Schemes				Structures	
Framework	VR	GT	Local	PGA	PS	Gossip
Hu et al. (2017)	✓	✗	✗	✗	✗	✗
Cooperative SGD (Wang and Joshi, 2021)	✗	✗	✓	✗	✓	✓
Decentralized (Gossip) SGD (Koloskova et al., 2020)	✗	✗	✓	✗	✓	✓
GT-VR (Xin et al., 2020)	✓	✓	✗	✗	✗	✓
Gorbunov et al. (2021)	✓	✓	✓	✗	✓	✗
SPP **(Ours)**	✓	✓	✓	✓	✓	✓

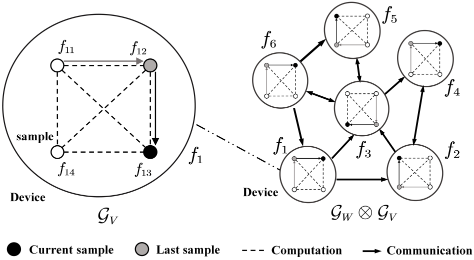
Contribution: Unify all these schemes both in PS and Gossip structures with rate guarantee showing clear dependency on these schemes.

The Proposed Framework

Sample-wise Push-Pull (SPP) Framework $\mathcal{A}(\Gamma_k, R_k, C_k)$

$$X_{k+1} = R_k X_k - \alpha \Gamma_k Y_k,$$
$$Y_{k+1} = C_k Y_k + \nabla F(X_{k+1}) - \nabla F(X_k),$$

– $\Gamma_k, R_k, C_k \in \mathbb{R}^{M \times M}$, $M = nm$ are matrices to be properly designed



An illustration of a two-level augmented graph with $n = 6, m = 4$

The Proposed Framework

Sample-wise Push-Pull (SPP) Framework $\mathcal{A}(\Gamma_k, R_k, C_k)$

$$X_{k+1} = R_k X_k - \alpha \Gamma_k Y_k,$$
$$Y_{k+1} = C_k Y_k + \nabla F(X_{k+1}) - \nabla F(X_k),$$

– $\Gamma_k, R_k, C_k \in \mathbb{R}^{M \times M}$, $M = nm$ are matrices to be properly designed

- Sampling on augmented graph:

$$\Gamma_k := \Lambda_{k+1} \left(W_k \otimes \mathbf{1}\mathbf{1}^T \right) \frac{\Lambda_k}{b_k},$$

– $\Lambda_k = \mathbf{diag}(\boldsymbol{e}_k)$; $\boldsymbol{e}_k = [\boldsymbol{e}_{1,k}^T, \cdots, \boldsymbol{e}_{n,k}^T]^T$, $\boldsymbol{e}_{i,k} \in \{0,1\}^{m \times 1}$; $b_k := \mathbf{1}^T \boldsymbol{e}_{i,k}$

- Intra and inter consensus guarantee

$$R_k := \underbrace{\mathbf{I}_M - \Lambda_{k+1}}_{\text{unelected part}} + \Gamma_k,$$

- Accurate gradient estimation tackling data heterogeneity

$$C_k = G_k \otimes V_k.$$

The Proposed Framework

Sample-wise Push-Pull (SPP) Framework $\mathcal{A}(\Gamma_k, R_k, C_k)$

$$X_{k+1} = R_k X_k - \alpha \Gamma_k Y_k,$$

$$Y_{k+1} = C_k Y_k + \nabla F(X_{k+1}) - \nabla F(X_k),$$

– $\Gamma_k, R_k, C_k \in \mathbb{R}^{M \times M}$, $M = nm$ are matrices to be properly designed

- Sampling on augmented graph:

$$\Gamma_k := \Lambda_{k+1} \left(W_k \otimes \mathbf{1}\mathbf{1}^T \right) \frac{\Lambda_k}{b_k},$$

– $\Lambda_k = \mathbf{diag}(\boldsymbol{e}_k)$; $\boldsymbol{e}_k = [\boldsymbol{e}_{1,k}^T, \cdots, \boldsymbol{e}_{n,k}^T]^T$, $\boldsymbol{e}_{i,k} \in \{0,1\}^{m \times 1}$; $b_k := \mathbf{1}^T \boldsymbol{e}_{i,k}$

- Intra and inter consensus guarantee

$$R_k := \underbrace{\mathbf{I}_M - \Lambda_{k+1}}_{\text{unelected part}} + \Gamma_k,$$

- Accurate gradient estimation tackling data heterogeneity

$$C_k = G_k \otimes V_k.$$

The Proposed Framework

Sample-wise Push-Pull (SPP) Framework $\mathcal{A}(\Gamma_k, R_k, C_k)$

$$X_{k+1} = R_k X_k - \alpha \Gamma_k Y_k,$$
$$Y_{k+1} = C_k Y_k + \nabla F(X_{k+1}) - \nabla F(X_k),$$

- $\Gamma_k, R_k, C_k \in \mathbb{R}^{M \times M}$, $M = nm$ are matrices to be properly designed

- Sampling on augmented graph:

$$\Gamma_k := \Lambda_{k+1} \left(W_k \otimes \mathbf{1}\mathbf{1}^T \right) \frac{\Lambda_k}{b_k},$$

  - $\Lambda_k = \mathbf{diag}(\mathbf{e}_k)$; $\mathbf{e}_k = [\mathbf{e}_{1,k}^T, \cdots, \mathbf{e}_{n,k}^T]^T$, $\mathbf{e}_{i,k} \in \{0,1\}^{m \times 1}$; $b_k := \mathbf{1}^T \mathbf{e}_{i,k}$

- Intra and inter consensus guarantee

$$R_k := \underbrace{\mathbf{I}_M - \Lambda_{k+1}}_{\text{unelected part}} + \Gamma_k,$$

- Accurate gradient estimation tackling data heterogeneity

$$C_k = G_k \otimes V_k.$$

Recovering Existing Algorithms and Beyond

- Connections to well-known VR methods

Algorithms	SAGA	L-SVRG	SARAH
b_k	b	$\{b,\ m\}$	$\{b,\ m\}$
V_k	$\mathbf{J}_m$	$\{\mathbf{I}_m, \mathbf{J}_m\}$	$\mathbf{J}_m$

- Recovery of other existing schemes and new algorithms

Algorithms	W_k	V_k	G_k
SAGA / L-SVRG	1	$\{\mathbf{I}_m, \mathbf{J}_m\}$	1
DSGD/Gossip-SGD	$\{W,\ \mathbf{J}_n\}$	$\mathbf{I}_m$	$\mathbf{I}_n$
Local SAGA†/Local-SVRG	$\{\mathbf{I}_n,\ \mathbf{J}_n\}$	$\{\mathbf{I}_m, \mathbf{J}_m\}$	$\mathbf{I}_n$
GT-SAGA/PGA-GT-SAGA†	$\{W,\ \mathbf{J}_n\}$	$\mathbf{J}_m$	$\{W,\ \mathbf{J}_n\}$

"$\dagger$": New algorithms obtained from the framework

New Insights: An interesting connection among VR methods; A unifying perspective for GT- and VR-based methods.

Convergence Results

- Assumptions
  - Each f_i is μ-strongly convex[1] and f_{ij} is expected L-smooth;
  - Bounded stochastic gradient variance σ^* (inner variance), and data heterogeneity ζ^* (external variance) at optimum x^*;
  - The expected spectral norm of the doubly stochastic matrix W_k satisfies $\rho_{r,W} := \mathbb{E}\left[\|W_k - \mathbf{J}_n\|_2^2\right] < 1, \forall k \geqslant 0$.

Linear convergence without VR and GT

Consider algorithms $\mathcal{A}(\cdot, \cdot, C_k \equiv \mathbf{I}_M)$ with batch-size b. Suppose the above assumptions hold and $\mu > 0$. There exists a (constant) stepsize α such that

$$\mathbb{E}\left[\|\bar{x}_k - x^*\|^2\right] \leqslant \gamma^k T_0 + \frac{\alpha^3}{1-\gamma}\mathcal{O}\left(\frac{\rho_{r,W} L \zeta^*}{(1-\rho_{r,W})^2}\right) + \frac{\alpha^2}{1-\gamma}\mathcal{O}\left(\frac{\sigma^*}{nb}\right) + \frac{\alpha^3}{1-\gamma}\mathcal{O}\left(\frac{\rho_{r,W} L \sigma^*}{b(1-\rho_{r,W})}\right),$$

where $\gamma < 1$ is the linear rate; T_0 is the initialization error.

[1] Results for convex cases can be found in the paper.

Convergence Results

- Assumptions
  - Each f_i is μ-strongly convex[1] and f_{ij} is expected L-smooth;
  - Bounded stochastic gradient variance σ^* (inner variance), and data heterogeneity ζ^* (external variance) at optimum x^*;
  - The expected spectral norm of the doubly stochastic matrix W_k satisfies $\rho_{r,W} := \mathbb{E}\left[\|W_k - \mathbf{J}_n\|_2^2\right] < 1, \forall k \geqslant 0$.

Linear convergence with VR

Consider algorithms $\mathcal{A}(\cdot, \cdot, C_k \equiv \mathbf{I}_n \otimes V_k)$. Suppose the above assumptions hold and $\mu > 0$. There exists a (constant) stepsize α such that

$$\mathbb{E}\left[\|\bar{x}_k - x^*\|^2\right] \leqslant \gamma^k T_0 + \frac{\alpha^3}{1-\gamma}\mathcal{O}\left(\frac{\rho_{r,W} L \zeta^*}{(1-\rho_{r,W})^2}\right)$$
$$+ \frac{\alpha^2}{1-\gamma}\mathcal{O}\left(\frac{\sigma^*}{nb}\right) + \frac{\alpha^3}{1-\gamma}\mathcal{O}\left(\frac{\rho_{r,W} L \sigma^*}{b(1-\rho_{r,W})}\right),$$

where $\gamma < 1$ is the linear rate; T_0 is the initialization error.

[1] Results for convex cases can be found in the paper.

Convergence Results

- Assumptions
  - Each f_i is μ-strongly convex[1] and f_{ij} is expected L-smooth;
  - Bounded stochastic gradient variance σ^* (inner variance), and data heterogeneity ζ^* (external variance) at optimum x^*;
  - The expected spectral norm of the doubly stochastic matrix W_k satisfies $\rho_{r,W} := \mathbb{E}\left[\|W_k - \mathbf{J}_n\|_2^2\right] < 1, \forall k \geqslant 0$.

Linear convergence with VR and GT

Consider algorithms $\mathcal{A}(\cdot, \cdot, C_k = W_k \otimes \mathbf{J}_m)$. Suppose the above assumptions hold and $\mu > 0$. There exists a (constant) stepsize α such that

$$\mathbb{E}\left[\|\bar{x}_k - x^*\|^2\right] \leqslant \gamma^k T_0 + \frac{\alpha^3}{1-\gamma}\mathcal{O}\left(\frac{\rho_{r,W} L \zeta^*}{(1-\rho_{r,W})^2}\right) + \frac{\alpha^2}{1-\gamma}\mathcal{O}\left(\frac{\sigma^*}{nb}\right) + \frac{\alpha^3}{1-\gamma}\mathcal{O}\left(\frac{\rho_{r,W} L \sigma^*}{b(1-\rho_{r,W})}\right),$$

where $\gamma < 1$ is the linear rate; T_0 is the initialization error.

[1] Results for convex cases can be found in the paper.

Convergence Results

- Assumptions
  - Each f_i is μ-strongly convex and f_{ij} is expected L-smooth;
  - Bounded stochastic gradient variance σ^* (inner variance), and data heterogeneity ζ^* (external variance) at optimum x^*;
  - The expected spectral norm of the doubly stochastic matrix W_k satisfies $\rho_{r,W} := \mathbb{E}\left[\|W_k - \mathbf{J}_n\|_2^2\right] < 1, \forall k \geqslant 0$.

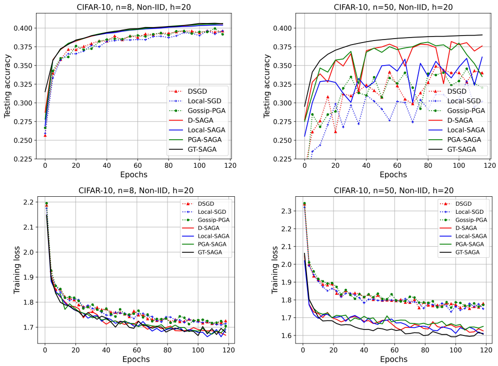
Algorithms	**Obtained Complexity**[2]
SAGA*/L-SVRG*	$\left(\frac{L}{\mu} + \frac{1}{pq}\right) \log \frac{1}{\varepsilon}$
DSGD*/Gossip-PGA	$\frac{L}{\mu\left(1-\rho_{r,W}\right)} + \frac{\sigma^*}{nb\mu^2\varepsilon} + \sqrt{\frac{L\left(b\zeta^*+\left(1-\rho_{r,W}\right)\sigma^*\right)}{\mu^3\left(1-\rho_{r,W}\right)^2 b\varepsilon}}$
Local SAGA†/Local-SVRG*	$\frac{L}{\mu\left(1-\rho_{r,W}\right)} + \frac{1}{pq} + \sqrt{\frac{L\zeta^*}{\mu^3\left(1-\rho_{r,W}\right)^2\varepsilon}}$
GT-SAGA†/PGA-GT-SAGA†	$\left(\frac{L}{\mu\left(1-\rho_{r,W}\right)^2} + \frac{m}{b}\right) \log \frac{1}{\varepsilon}$

"*" : obtain best-known rate; "†": obtain new algorithm or new rate

[2] $r := P\left(W_k = \mathbf{J}_n\right)$; $p := P\left(V_k = \mathbf{J}_m\right)$; $q := \mathbb{E}\left[b_k/m | V_k = \mathbf{J}_m\right]$

Numerical Experiments

- Regularized logistic regression for image classification



Directed ring graph with $n = 8$ Exponential graph with $n = 50$

Parameter Settings:

Dataset: CIFAR-10;

\# Training: 50000;

\# Testing: 10000;

\# Nodes (n): $\{8, 50\}$;

Batch-size (b): 200/n;

Step-size (α): 0.008;

Regularization (λ): 0.001.

Performance comparison of several SOTA algorithms on CIFAR-10 dataset with **unbalanced** label distribution

Conclusion and Future Work

- Conclusion
  - Propose a new framework that unifies GT, VR, Local and PGA schemes in both PS and Gossip structures;
  - Provide convergence results which show clear dependency of the convergence performance on these above schemes
  - Recover various existing algorithms with best-known/new rates and design new algorithms building on this framework;
- Future Work
  - Topology design with more efficient communication;
  - Improved convergence analysis taking into account the communication and computation trade-off.

Thank you !

References I

Chen, Y., Yuan, K., Zhang, Y., Pan, P., Xu, Y., and Yin, W. (2021). Accelerating gossip SGD with periodic global averaging. In *International Conference on Machine Learning*, pages 1791–1802. PMLR.

Defazio, A., Bach, F., and Lacoste-Julien, S. (2014). SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in neural information processing systems*, pages 1646–1654.

Gorbunov, E., Hanzely, F., and Richtárik, P. (2021). Local SGD: Unified theory and new efficient methods. In *International Conference on Artificial Intelligence and Statistics*, pages 3556–3564. PMLR.

Hu, B., Seiler, P., and Rantzer, A. (2017). A unified analysis of stochastic optimization methods using jump system theory and quadratic constraints. In *Conference on Learning Theory*, pages 1157–1189. PMLR.

Koloskova, A., Loizou, N., Boreiri, S., Jaggi, M., and Stich, S. (2020). A unified theory of decentralized SGD with changing topology and local updates. In *International Conference on Machine Learning*, pages 5381–5393. PMLR.

Konevcnỳ, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., and Bacon, D. (2016). Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*.

Mokhtari, A. and Ribeiro, A. (2016). DSA: Decentralized double stochastic averaging gradient algorithm. *The Journal of Machine Learning Research*, 17(1):2165–2199.

Nguyen, L. M., Liu, J., Scheinberg, K., and Takávc, M. (2017). SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *Proceedings of the 34th International Conference on Machine Learning*, pages 2613–2621.

Pu, S. and Nedić, A. (2020). Distributed stochastic gradient tracking methods. *Mathematical Programming*, pages 1–49.

Qian, X., Qu, Z., and Richtárik, P. (2021). L-SVRG and L-Katyusha with arbitrary sampling. *Journal of Machine Learning Research*, 22(112):1–47.

References II

Ram, S. S., Nedić, A., and Veeravalli, V. V. (2009). Asynchronous gossip algorithms for stochastic optimization. In *Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference*, pages 3581–3586. IEEE.

Wang, J. and Joshi, G. (2021). Cooperative SGD: A unified framework for the design and analysis of local-update SGD algorithms. *Journal of Machine Learning Research*, 22(213):1–50.

Xin, R., Khan, U. A., and Kar, S. (2020). Variance-reduced decentralized stochastic optimization with accelerated convergence. *IEEE Transactions on Signal Processing*, 68:6255–6271.