On the Convergence to a Global Solution of Shuffling-Type Gradient Algorithms

Lam M. Nguyen

IBM Research, Thomas J. Watson Research Center
Yorktown Heights, New York

joint work with
Trang H. Tran (Cornell)

37th Conference on Neural Information Processing Systems (NeurIPS 2023)

IBM Research

Empirical Risk Minimization

We consider the following finite-sum minimization:

$$\min_{w \in \mathbb{R}^d} \left\{ F(w) := \frac{1}{n} \sum_{i=1}^{n} f(w; i) \right\}, \tag{1}$$

where $f(\cdot; i) : \mathbb{R}^d \to \mathbb{R}$ is Lipschitz smooth and possibly non-convex for $i \in [n] := \{1, \ldots, n\}$.

Basic Assumptions

Assumption 1

Suppose that $f_i^* := \min_{w \in \mathbb{R}^d} f(w;i) > -\infty$, $i \in \{1, \dots, n\}$.

Assumption 2

Suppose that $f(\cdot; i)$ *is* L*-smooth for all* $i \in \{1, \dots, n\}$, *i.e. there exists a constant* $L \in (0, +\infty)$ *such that:*

$$\|\nabla f(w;i) - \nabla f(w';i)\| \leq L\|w - w'\|, \quad \forall w, w' \in \mathbb{R}^d. \tag{2}$$

Some Other Assumptions

Assumption 3

Suppose that $f(\cdot; i)$ satisfies average PL inequality for some constant $\mu > 0$ such that

$$\frac{1}{n}\sum_{i=1}^{n}\|\nabla f(w; i)\|^2 \geq 2\mu \frac{1}{n}\sum_{i=1}^{n}[f(w; i) - f_i^*], \quad \forall w \in \mathbb{R}^d. \tag{3}$$

where $f_i^* := \min_{w \in \mathbb{R}^d} f(w; i)$.

Assumption 4

Suppose that the best variance at w_* is small, that is, for $\varepsilon > 0$ and for some $P > 0$

$$\inf_{w_* \in \mathcal{W}_*} \left(\frac{1}{n}\sum_{i=1}^{n}\|\nabla f(w_*; i)\|^2 \right) \leq P\varepsilon, \tag{4}$$

Assumption 5

Using Algorithm 1, let us assume that there exist some constants $M > 0$ and $N > 0$ such that at each epoch $t = 1, \dots, T$, we have for $i = 1, \dots, n$:

$$\begin{aligned} &\|\nabla f(w_{i-1}^{(t)}; \pi^{(t)}(i)) - \nabla f(w_*; \pi^{(t)}(i))\|^2 \\ &\leq M\langle \nabla f(w_{i-1}^{(t)}; \pi^{(t)}(i)) - \nabla f(w_*; \pi^{(t)}(i)), w_{i-1}^{(t)} - w_* \rangle + N\frac{1}{n}\sum_{i=1}^{n}\|w_i^{(t)} - w_0^{(t)}\|^2, \end{aligned} \tag{5}$$

Shuffling-Type Gradient Algorithm

Algorithm (Shuffling-Type Gradient Algorithm for Solving (1))

Initialization: Choose an initial point $\tilde{w}_0 \in \mathrm{dom}\,(F)$.
for $t = 1, 2, \dots, T$ **do**
    Set $w_0^{(t)} := \tilde{w}_{t-1}$;
    Generate any permutation $\pi^{(t)}$ of $[n]$ (either deterministic or random);
    for $i = 1, \dots, n$ **do**
        Update $w_i^{(t)} := w_{i-1}^{(t)} - \eta_i^{(t)} \nabla f(w_{i-1}^{(t)}; \pi^{(t)}(i))$;
    end for
    Set $\tilde{w}_t := w_n^{(t)}$;
end for

New Framework for Convergence to a Global Solution

Theorem 1

Assume that Assumptions 1, 2, 3, and 5 hold. Let $\{\tilde{w}_t\}_{t=1}^{T}$ be the sequence generated by Algorithm 1 with the learning rate $\eta_i^{(t)} = \frac{\eta_t}{n}$ where $0 < \eta_t \leq \min\left\{\frac{n}{2M}, \frac{1}{2L}\right\}$. Let the number of iterations $T = \frac{\lambda}{\varepsilon^{3/2}}$ for some $\lambda > 0$ and $\varepsilon > 0$. Constants C_1, C_2, and C_3 are defined in (7) for any $\gamma > 0$. We further define $K = 1 + C_1 D^3 \varepsilon^{3/2}$ and specify the learning rate $\eta_t = K\eta_{t-1} = K^t \eta_0$ and $\eta_0 = \frac{D\sqrt{\varepsilon}}{K\exp(\lambda C_1 D^3)}$ such that $\frac{D\sqrt{\varepsilon}}{K} \leq \min\left\{\frac{n}{2M}, \frac{1}{2L}\right\}$ for some constant $D > 0$. Then we have

$$\frac{1}{T}\sum_{t=1}^{T}[F(\tilde{w}_{t-1}) - F_*] \leq \frac{K\exp(\lambda C_1 D^3)}{C_3 D \lambda}\|\tilde{w}_0 - w_*\|^2 \cdot \varepsilon + \frac{C_2}{C_3}\sigma_*^2, \tag{6}$$

where $F_* = \min_{w\in\mathbb{R}^d} F(w)$, σ_*^2 is the variance at w_*, w_* is a global solution of F, and

$$\begin{cases} C_1 = \frac{8L^2}{3} + \frac{14NL^2}{M} + \frac{4\gamma L^4}{6M}, \\ C_2 = \frac{2}{M} + 1 + \frac{5}{6L^2} + \frac{8N}{3ML^2} + \frac{5\gamma}{12M}, \\ C_3 = \frac{\gamma}{\gamma+1}\frac{\mu}{M}. \end{cases} \tag{7}$$

New Framework for Convergence to a Global Solution

Corollary 1

Suppose that the conditions in Theorem 1 and Assumption 4 hold. Choose $C_1 D\lambda = 1$ *and* $\varepsilon = \hat{\varepsilon}/G$ *such that* $0 < \hat{\varepsilon} \leq G$ *with the constants*

$$G = \frac{2C_1 D^2 e}{C_3}\|\tilde{w}_0 - w_*\|^2 + \frac{C_2 P}{C_3}, \ \textit{where}$$

$$\begin{cases} C_1 = \frac{8L^2}{3} + \frac{14NL^2}{M} + \frac{4L^2}{3M}, \\ C_2 = \frac{2}{M} + 1 + \frac{5}{6L^2} + \frac{8N}{3ML^2} + \frac{5}{12ML}, \\ C_3 = \frac{1}{L^2+1}\frac{\mu}{M}. \end{cases}$$

Then, the we need $T = \frac{\lambda G^{3/2}}{\hat{\varepsilon}^{3/2}}$ *epochs to guarantee*

$$\min_{1\leq t\leq T}[F(\tilde{w}_{t-1}) - F_*] \leq \frac{1}{T}\sum_{t=1}^{T}[F(\tilde{w}_{t-1}) - F_*] \leq \hat{\varepsilon}.$$

Computational Complexity

Table: Comparisons of computational complexity (the number of individual gradient evaluations) needed by SGD algorithm to reach an $\hat{\varepsilon}$-accurate solution w that satisfies $F(w) - F(w_*) \leq \hat{\varepsilon}$ (or $\|\nabla F(w)\|^2 \leq \hat{\varepsilon}$ in the non-convex case). **SS: Shuffling Schemes; GS: Global Solution**.

Settings	References	Complexity	SS	GS
Convex	[Nemirovski et al., 2009, Shamir and Zhang, 2013] (1)	$\mathcal{O}\left(\frac{\Delta_0^2+G^2}{\hat{\varepsilon}^2}\right)$	✗	✓
	[Mishchenko et al., 2020, Nguyen et al., 2021] (2)	$\mathcal{O}\left(\frac{n}{\hat{\varepsilon}^{3/2}}\right)$	✓	✓
PL condition	[Nguyen et al., 2021]	$\tilde{\mathcal{O}}\left(\frac{n\sigma^2}{\hat{\varepsilon}^{1/2}}\right)$	✓	✓
Star-convex related	[Gower et al., 2021] (3)	$\mathcal{O}\left(\frac{1}{\hat{\varepsilon}^2}\right)$	✗	✓
Non-convex	[Ghadimi and Lan, 2013] (5)	$\mathcal{O}\left(\frac{\sigma^2}{\hat{\varepsilon}^2}\right)$	✗	✗
	[Nguyen et al., 2021, Mishchenko et al., 2020] (5)	$\mathcal{O}\left(\frac{n\sigma}{\hat{\varepsilon}^{3/2}}\right)$	✓	✗
Our setting (non-convex)	**This paper**(4)	$\mathcal{O}\left(\frac{n(N\vee 1)^{3/2}}{\hat{\varepsilon}^{3/2}}\right)$	✓	✓

Numerical Experiments

Table: Datasets used in our experiments

Data name	# Samples	# Features	Networks layers	Sources
Diabetes	442	10	300-100	[Efron et al., 2004]
Life Expectancy	1649	19	900-300-100	[Repository, 2016]
California Housing	16514	8	900-300-100	[Repository, 1997]

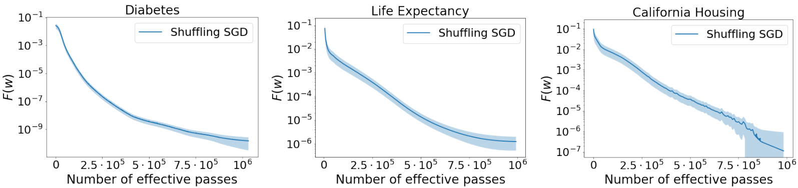

Figure: The train loss produced by Shuffling SGD algorithm for three datasets: Diabetes, Life Expectancy and California Housing.

Our Contributions

- We investigate a new framework for the convergence of a shuffling-type gradient algorithm to a global solution. We consider a relaxed set of assumptions and discuss their relations with previous settings. We show that our average-PL inequality holds for a wide range of neural networks equipped with squared loss function.
- Our analysis generalizes the class function called star-M-smooth-convex. This class contains non-convex functions and is more general than the class of star-convex smooth functions with respect to the minimizer (in the over-parameterized settings). In addition, our analysis does not use any bounded gradient or bounded weight assumptions.
- We show the total complexity of $\mathcal{O}(\frac{n}{\hat{\varepsilon}^{3/2}})$ for a class of non-convex functions to reach an $\hat{\varepsilon}$-accurate global solution. This result matches the same gradient complexity to a stationary point for unified shuffling methods in non-convex settings, however, we are able to show the convergence to a global minimizer.

Related Work

- **General.** [Brutzkus et al., 2018, Soudry et al., 2018, Arora et al., 2019, Du et al., 2019b, Du et al., 2019a, Allen-Zhu et al., 2019, Zou and Gu, 2019, Zou et al., 2018].
- **Polyak-Lojasiewicz (PL) condition and related assumptions.** [Polyak, 1964], [Karimi et al., 2016, Nesterov and Polyak, 2006, De et al., 2017, Gower et al., 2021, Haochen and Sra, 2019, Ahn et al., 2020, Nguyen et al., 2021, Schmidt and Roux, 2013, Vaswani et al., 2019, Sankararaman et al., 2020]
- **Over-paramaterized settings for neural networks.** [Schmidt and Roux, 2013], [Ma et al., 2018, Meng et al., 2020, Loizou et al., 2021, Zhou et al., 2019]
- **Star-convexity and related conditions.** [Nesterov and Polyak, 2006, Lee and Valiant, 2016, Bjorck et al., 2021, Zhou et al., 2019, Hinder et al., 2020, Hardt et al., 2018, Jin, 2020, Gower et al., 2021]

Our Poster:

Poster Session 1 - Great Hall & Hall B1+B2 #1101

References I

Ahn, K., Yun, C., and Sra, S. (2020).

Sgd with shuffling: optimal rates without component convexity and large epoch requirements.

In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17526–17535. Curran Associates, Inc.

Allen-Zhu, Z., Li, Y., and Song, Z. (2019).

A convergence theory for deep learning via over-parameterization.

In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 242–252. PMLR.

Arora, S., Du, S., Hu, W., Li, Z., and Wang, R. (2019).

Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks.

In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 322–332. PMLR.

Bjorck, J., Kabra, A., Weinberger, K. Q., and Gomes, C. (2021).

Characterizing the loss landscape in non-negative matrix factorization.

Proceedings of the AAAI Conference on Artificial Intelligence, 35(8):6768–6776.

Brutzkus, A., Globerson, A., Malach, E., and Shalev-Shwartz, S. (2018).

SGD learns over-parameterized networks that provably generalize on linearly separable data.

In *International Conference on Learning Representations*.

References II

De, S., Yadav, A., Jacobs, D., and Goldstein, T. (2017).

Automated Inference with Adaptive Batches.

In Singh, A. and Zhu, J., editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1504–1513. PMLR.

Du, S., Lee, J., Li, H., Wang, L., and Zhai, X. (2019a).

Gradient descent finds global minima of deep neural networks.

In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1675–1685. PMLR.

Du, S. S., Zhai, X., Poczos, B., and Singh, A. (2019b).

Gradient descent provably optimizes over-parameterized neural networks.

In *International Conference on Learning Representations*.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004).

Diabetes dataset.

Ghadimi, S. and Lan, G. (2013).

Stochastic first-and zeroth-order methods for nonconvex stochastic programming.

SIAM J. Optim., 23(4):2341–2368.

Gower, R., Sebbouh, O., and Loizou, N. (2021).

Sgd for structured nonconvex functions: Learning rates, minibatching and interpolation.

In *International Conference on Artificial Intelligence and Statistics*, pages 1315–1323. PMLR.

References III

Haochen, J. and Sra, S. (2019).
Random shuffling beats sgd after finite epochs.
In *International Conference on Machine Learning*, pages 2624–2633. PMLR.

Hardt, M., Ma, T., and Recht, B. (2018).
Gradient descent learns linear dynamical systems.
Journal of Machine Learning Research, 19(29):1–44.

Hinder, O., Sidford, A., and Sohoni, N. (2020).
Near-optimal methods for minimizing star-convex functions and beyond.
In *Conference on Learning Theory*, pages 1894–1938. PMLR.

Jin, J. (2020).
On the convergence of first order methods for quasar-convex optimization.
arXiv preprint arXiv:2010.04937.

Karimi, H., Nutini, J., and Schmidt, M. (2016).
Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition.
In Frasconi, P., Landwehr, N., Manco, G., and Vreeken, J., editors, *Machine Learning and Knowledge Discovery in Databases*, pages 795–811, Cham. Springer International Publishing.

Lee, J. C. and Valiant, P. (2016).
Optimizing star-convex functions.
In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 603–614.

References IV

Loizou, N., Vaswani, S., Hadj Laradji, I., and Lacoste-Julien, S. (2021).

Stochastic polyak step-size for sgd: An adaptive learning rate for fast convergence.

In Banerjee, A. and Fukumizu, K., editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1306–1314. PMLR.

Ma, S., Bassily, R., and Belkin, M. (2018).

The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning.

In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3325–3334. PMLR.

Meng, S. Y., Vaswani, S., Laradji, I. H., Schmidt, M., and Lacoste-Julien, S. (2020).

Fast and furious convergence: Stochastic second order methods under interpolation.

In Chiappa, S. and Calandra, R., editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1375–1386. PMLR.

Mishchenko, K., Khaled Ragab Bayoumi, A., and Richtárik, P. (2020).

Random reshuffling: Simple analysis with vast improvements.

Advances in Neural Information Processing Systems, 33.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. (2009).

Robust stochastic approximation approach to stochastic programming.

SIAM J. on Optimization, 19(4):1574–1609.

Nesterov, Y. and Polyak, B. T. (2006).

Cubic regularization of Newton method and its global performance.

Mathematical Programming, 108(1):177–205.

References V

Nguyen, L. M., Tran-Dinh, Q., Phan, D. T., Nguyen, P. H., and van Dijk, M. (2021).
A unified convergence analysis for shuffling-type gradient methods.
Journal of Machine Learning Research, 22(207):1–44.

Polyak, B. T. (1964).
Some methods of speeding up the convergence of iteration methods.
USSR Computational Mathematics and Mathematical Physics, 4(5):1–17.

Repository, G. H. O. D. (2016).
Life expectancy and healthy life expectancy.

Repository, S. (1997).
California housing.

Sankararaman, K. A., De, S., Xu, Z., Huang, W. R., and Goldstein, T. (2020).
The impact of neural network overparameterization on gradient confusion and stochastic gradient descent.
In *International conference on machine learning*, pages 8469–8479. PMLR.

Schmidt, M. and Roux, N. L. (2013).
Fast convergence of stochastic gradient descent under a strong growth condition.

References VI

Shamir, O. and Zhang, T. (2013).

Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes.

In Dasgupta, S. and McAllester, D., editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 71–79, Atlanta, Georgia, USA. PMLR.

Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., and Srebro, N. (2018).

The implicit bias of gradient descent on separable data.

J. Mach. Learn. Res., 19(1):2822–2878.

Vaswani, S., Bach, F., and Schmidt, M. (2019).

Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron.

In Chaudhuri, K. and Sugiyama, M., editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1195–1204. PMLR.

Zhou, Y., Yang, J., Zhang, H., Liang, Y., and Tarokh, V. (2019).

Sgd converges to global minimum in deep learning via star-convex path.

arXiv preprint arXiv:1901.00451.

Zou, D., Cao, Y., Zhou, D., and Gu, Q. (2018).

Stochastic gradient descent optimizes over-parameterized deep relu networks.

References VII

Zou, D. and Gu, Q. (2019).

An improved analysis of training over-parameterized deep neural networks.

In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 2053–2062.

THANK YOU!

Lam M. Nguyen
LamNguyen.MLTD@ibm.com
www.lamnguyen.org