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

Lam M. Nguyen*[1] · Trang H. Tran*[2]

[1] IBM Research, Thomas J. Watson Research Center [2] Cornell University, School of Operations Research and Information Engineering

* Equally contributed. Correspondence to Lam M. Nguyen, LamNguyen.MLTD@ibm.com.

NEURAL INFORMATION PROCESSING SYSTEMS

Problem Statement

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 smooth and possibly non-convex for $i\in[n]:=\{1,\cdots,n\}$.

Shuffling-Type Gradient Algorithm

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

```
Initialization: Choose an initial point w̃_0 ∈ dom(F).
for t = 1, 2, ..., T do
  Set w_0^(t) := w̃_{t-1};
  Generate any permutation π^(t) of [n] (either deterministic or random);
  for i = 1, ..., n do
    Update w_i^(t) := w_{i-1}^(t) − η_i^(t) ∇f(w_{i-1}^(t); π^(t)(i));
  end for
  Set w̃_t := w_n^(t);
end for
```

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:*

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

Average PL condition

Definition 1 (Polyak-Lojasiewicz condition). *We say that* f *satisfies Polyak-Lojasiewicz (PL) inequality for some constant* $\mu>0$ *if*

$$\|\nabla f(w)\|^2\ge 2\mu[f(w)-f^*],\quad \forall w\in\mathbb{R}^d, \tag{3}$$

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

Assumption 3. *Suppose that* $f(\cdot;i)$ *satisfies average PL inequality for some constant* $\mu>0$ *such that*

$$\frac{1}{n}\sum_{i=1}^{n}\|\nabla f(w;i)\|^2\ge 2\mu\frac{1}{n}\sum_{i=1}^{n}[f(w;i)-f_i^*],\quad \forall w\in\mathbb{R}^d. \tag{4}$$

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

Theorem 1. *Let* $\{(x^{(i)},y^{(i)})\}_{i=1}^n$ *is a training data set where* $x^{(i)}\in\mathbb{R}^m$ *is the input data and* $y^{(i)}\in\mathbb{R}^c$ *is the output data for* $i=1,\dots,n$. *We consider an architecture* $h(w;i)$ *with* w *be the vectorized weight and*

$$h(w;i)=W^Tz(\theta;i)+b,$$

where $w=\mathbf{vec}(\{\theta,W,b\})$ *and* $z(\theta;i)$ *are some inner architectures, which can be chosen arbitrarily. Next, we consider the squared loss* $f(w;i)=\frac{1}{2}\|h(w;i)-y^{(i)}\|^2$. *Then*

$$\|\nabla f(w;i)\|^2\ge 2[f(w;i)-f_i^*],\quad \forall w\in\mathbb{R}^d, \tag{5}$$

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

Generalized Star-Smooth-Convex Condition

For any global solution w_* of F, let us define

$$\sigma_*^2:=\inf_{w_*\in\mathcal{W}_*}\left(\frac{1}{n}\sum_{i=1}^{n}\|\nabla f(w_*;i)\|^2\right). \tag{6}$$

Assumption 4. *Suppose that the best variance at* w_* *is small, that is, for* $\varepsilon>0$

$$\sigma_*^2\le P\varepsilon, \tag{7}$$

for some $P>0$.

Definition 2. *The function* g *is star-M-smooth-convex with respect to a reference point* $\hat{w}\in\mathbb{R}^d$ *if*

$$\|\nabla g(w)-\nabla g(\hat{w})\|^2\le M\langle\nabla g(w)-\nabla g(\hat{w}),w-\hat{w}\rangle,\quad \forall w\in\mathbb{R}^d. \tag{8}$$

For the analysis of shuffling type algorithm in this paper, we consider the general assumption called the *generalized star-smooth-convex condition for shuffling algorithms*:

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* $i=1,\dots,n$:

$$\begin{aligned}&\|\nabla f(w_{i-1}^{(t)};\pi^{(t)}(i))-\nabla f(w_*;\pi^{(t)}(i))\|^2\\ &\le M\langle\nabla f(w_{i-1}^{(t)};\pi^{(t)}(i))-\nabla f(w_*;\pi^{(t)}(i)),w_{i-1}^{(t)}-w_*\rangle+N\frac{1}{n}\sum_{i=1}^{n}\|w_i^{(t)}-w_0^{(t)}\|^2,\end{aligned} \tag{9}$$

where w_* *is a global solution of* F.

Main results

Theorem 2. *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* $0<\eta_t\le\min\left\{\frac{n}{2M},\frac{1}{2L}\right\}$. *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 as below for any* $\gamma>0$. *We further define* $K=1+C_1D^3\varepsilon^{3/2}$ *and specify the learning rate* $\eta_t=K\eta_{t-1}=K^t\eta_0$ *and* $\eta_0=\frac{D\sqrt{\varepsilon}}{K\exp(\lambda C_1D^3)}$ *such that* $\frac{D\sqrt{\varepsilon}}{K}\le\min\left\{\frac{n}{2M},\frac{1}{2L}\right\}$ *for some constant* $D>0$. *Then we have*

$$\frac{1}{T}\sum_{t=1}^{T}[F(\tilde{w}_{t-1})-F_*]\le\frac{K\exp(\lambda C_1D^3)}{C_3D\lambda}\|\tilde{w}_0-w_*\|^2\cdot\varepsilon+\frac{C_2}{C_3}\sigma_*^2, \tag{10}$$

where $F_*=\min_{w\in\mathbb{R}^d}F(w)$, σ_*^2 *is defined in (6),* w_* *is a global solution of* F, *and*

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

Corollary 1. *Suppose that the conditions in Theorem 2 and Assumption 4 hold. Choose* $C_1D\lambda=1$ *and* $\varepsilon=\hat{\varepsilon}/G$ *such that* $0<\hat{\varepsilon}\le G$ *with the constant* $G=\frac{2C_1D^2e}{C_3}\|\tilde{w}_0-w_*\|^2+\frac{C_2P}{C_3}$ *where*

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

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

$$\min_{1\le t\le T}[F(\tilde{w}_{t-1})-F_*]\le\frac{1}{T}\sum_{t=1}^{T}[F(\tilde{w}_{t-1})-F_*]\le\hat{\varepsilon}. \tag{11}$$

Computational Complexity

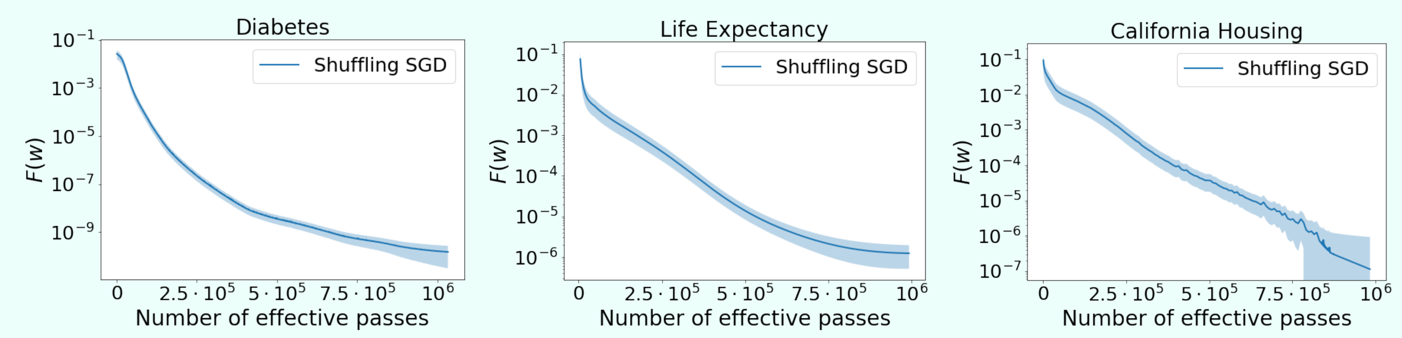
Comparisons of computational complexity (the number of individual gradient evaluations) needed by SGD algorithm to reach an $\hat{\varepsilon}$-accurate solution w that satisfies $F(w)-F(w_*)\le\hat{\varepsilon}$ (or $\|\nabla F(w)\|^2\le\hat{\varepsilon}$ in the non-convex case)

Settings	**Complexity**	**Shuffling Schemes**	**Global Solution**
Convex	$\mathcal{O}\left(\frac{\Delta_0^2+G^2}{\hat{\varepsilon}^2}\right)$	✗	✓
	$\mathcal{O}\left(\frac{n}{\hat{\varepsilon}^{3/2}}\right)$	✓	✓
PL condition	$\tilde{\mathcal{O}}\left(\frac{n\sigma^2}{\hat{\varepsilon}^{1/2}}\right)$	✓	✓
Star-convex related	$\mathcal{O}\left(\frac{1}{\hat{\varepsilon}^2}\right)$	✗	✓
Non-convex	$\mathcal{O}\left(\frac{\sigma^2}{\hat{\varepsilon}^2}\right)$	✗	✗
	$\mathcal{O}\left(\frac{n\sigma}{\hat{\varepsilon}^{3/2}}\right)$	✓	✗
Our setting (non-convex)	$\mathcal{O}\left(\frac{n(N\vee 1)^{3/2}}{\hat{\varepsilon}^{3/2}}\right)$	✓	✓

Experiments

We modify the initial data lightly to guarantee the over-parameterized setting in our experiment. We first use GD algorithm to find a weight w that yields a sufficiently small function value, then change the label data to $\hat{y}_i$ such that the weight w yields ϵ loss function

Below is the train loss produced by Shuffling SGD algorithm for three datasets: Diabetes, Life Expectancy and California Housing.



Contributions

- We investigate a new framework for the convergence of a shuffling-type gradient algorithm to a global solution.
- Our analysis generalizes the class function called star-M-smooth-convex.
- 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.

Key References

[1] Nesterov, Y., and Polyak, B. Cubic regularization of Newton method and its global performance. *Mathematical Programming, 108(1):177–205, 2006*

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

[3] Zhou, Y., Yang, J., Zhang, H., Liang, Y., and Tarokh, V. Sgd converges to global minimum in deep learning via star-convex path *The 7th International Conference on Learning Representations, 2019*