When Does Stochastic Gradient Algorithm Work Well?

Katya Scheinberg

Joint work with L. M. Nguyen, N. H. Nguyen, D. T. Phan and J. R. Kalagnanam

Industrial and Systems Engineering Department

LEHIGH UNIVERSITY

US & Mexico Workshop on
Optimization and its Applications

January 12, 2018

Outline

- Basic setting
- New look at convergence rates
- Numerical evidence
- Conclusions

The usual SGD algorithm

Expected risk minimization

$$\min_{\mathbf{w}\in\mathbb{R}^d} \{F(\mathbf{w}) = \mathbb{E}[f(\mathbf{w};\xi)]\},$$

where ξ is a random variable obeying some distribution, or empirical risk minimization (ERM):

$$\min_{\mathbf{w}\in\mathbb{R}^d} \left\{F(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{w})\right\}.$$

Basic SG algorthms

Algorithm 1 Stochastic Gradient Method with Fixed Stepsize

Initialize $\mathbf{w}_0$, choose stepsize $\eta > 0$, and batch size b.
for $i = 1, 2, \cdots$ **do**
  Generate random variables $\{\xi_{t,i}\}_{i=1}^{b}$ i.i.d.
  Compute a stochastic gradient

$$\mathbf{g}_t = \frac{1}{b} \sum_{i=1}^{b} \nabla f(\mathbf{w}_t; \xi_{t,i}).$$

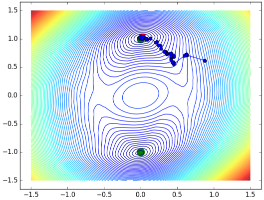
  Update the new iterate $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \mathbf{g}_t$.

Two examples of behavior

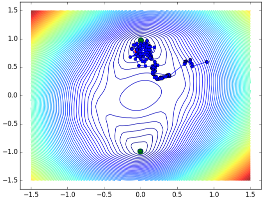
$$\min_{\mathbf{w}} \left\{ F(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - (\mathbf{a}_i^\top \mathbf{w})^2)^2 \right\}.$$

(i) *All* components $f_i(\mathbf{w}) = (y_i - (\mathbf{a}_i^\top \mathbf{w})^2)^2$ are small at $\mathbf{w}_*$

(ii) *Many* components $f_i(\mathbf{w}) = (y_i - (\mathbf{a}_i^\top \mathbf{w})^2)^2$ are large at $\mathbf{w}_*$



(i)



(ii)

Define quantities

Definition 1

Let $\mathbf{w}_*$ *be a stationary point of the objective function* $F(\mathbf{w})$*. For any given threshold* $\epsilon > 0$*, define*

$$p_\epsilon := \mathbb{P}\left\{\|\mathbf{g}_*\|^2 \leq \epsilon\right\},$$

where $\mathbf{g}_* = \frac{1}{b}\sum_{i=1}^{b} \nabla f(\mathbf{w}_*;\xi_i)$*.*
We also define

$$M_\epsilon := \mathbb{E}\left[\|\mathbf{g}_*\|^2 \mid \|\mathbf{g}_*\|^2 > \epsilon\right].$$

Remark 1

p_ϵ *decreases as* ϵ *increases. There exists an* ϵ *such that* $p_\epsilon \approx 1-\epsilon$*.*

Outline

Theorem 1

Suppose that $F(\mathbf{w})$ is μ-strongly convex and $f(\mathbf{w};\xi)$ is L-smooth and convex for every realization of ξ. Consider the fixed step SGD algorithm with $\eta \le \frac{1}{L}$. Then, for any $\epsilon > 0$

$$\begin{aligned}\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_*\|^2] &\le (1-\mu\eta(1-\eta L))^t \|\mathbf{w}_0 - \mathbf{w}_*\|^2 \\ &+ \frac{2\eta}{\mu(1-\eta L)} p_\epsilon \epsilon + \frac{2\eta}{\mu(1-\eta L)}(1-p_\epsilon) M_\epsilon,\end{aligned}$$

where $\mathbf{w}_* = \arg\min_{\mathbf{w}} F(\mathbf{w})$.

Corollary 1

For any ϵ such that $1 - p_\epsilon \leq \epsilon$, and for $\eta \leq \frac{1}{2L}$, we have

$$\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_*\|^2] \leq (1 - \mu\eta)^t \|\mathbf{w}_0 - \mathbf{w}_*\|^2 + \frac{2\eta}{\mu}(1 + M_\epsilon)\,\epsilon.$$

If $t \geq T$ for $T = \frac{1}{\mu\eta} \log\left(\frac{\mu\|\mathbf{w}_0 - \mathbf{w}_*\|^2}{2\eta(1+M_\epsilon)\epsilon}\right)$, then

$$\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_*\|^2] \leq \frac{4\eta}{\mu}(1 + M_\epsilon)\,\epsilon.$$

Theorem 2

Suppose that $f(\mathbf{w};\xi)$ is L-smooth and convex for every realization of ξ. Let $\eta < \frac{1}{L}$. Then for any $\epsilon > 0$, we have

$$\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}_*)] \leq \frac{\|\mathbf{w}_0 - \mathbf{w}_*\|^2}{2\eta(1-\eta L)t} + \frac{\eta}{(1-\eta L)} p_\epsilon \epsilon + \frac{\eta M_\epsilon}{(1-\eta L)}(1-p_\epsilon),$$

where $\mathbf{w}_*$ is any optimal solution of $F(\mathbf{w})$.

Corollary 2

For any ϵ *such that* $1 - p_\epsilon \leq \epsilon$, *and* $\eta \leq \frac{1}{2L}$, *it holds that*

$$\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}_*)] \leq \frac{\|\mathbf{w}_0 - \mathbf{w}_*\|^2}{\eta t} + 2\eta\,(1 + M_\epsilon)\,\epsilon.$$

Hence, if $t \geq T$ *for* $T = \frac{\|\mathbf{w}_0 - \mathbf{w}_*\|^2}{(2\eta^2)(1+M_\epsilon)\epsilon}$, *we have*

$$\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}_*)] \leq 4\eta\,(1 + M_\epsilon)\,\epsilon.$$

Assumption 1

$\exists\ N > 0$, *such that for any sequence of iterates* $\mathbf{w}_0$, $\mathbf{w}_1$, $\ldots$, $\mathbf{w}_t$ *of any realization of SDG, there exists a stationary point* $\mathbf{w}_*$ *of* $F(\mathbf{w})$ *(possibly dependent on that sequence) such that*

$$\frac{1}{t+1}\sum_{k=0}^{t}\left(\mathbb{E}\left[\left\|\frac{1}{b}\sum_{i=1}^{b}\nabla f(\mathbf{w}_k;\xi_{k,i}) - \frac{1}{b}\sum_{i=1}^{b}\nabla f(\mathbf{w}_*;\xi_{k,i})\right\|^2 \middle| \mathcal{F}_k\right]\right) \le N\frac{1}{t+1}\sum_{k=0}^{t}\|\nabla F(\mathbf{w}_k)\|^2,$$

where the expectation is taken over random variables $\xi_{k,i}$. *Let* $\mathcal{W}_*$ *denote the set of all such stationary points* $\mathbf{w}_*$, *determined by the constant* N *and by realizations* $\mathbf{w}_0$, $\mathbf{w}_1$, $\ldots$, $\mathbf{w}_t$.

Definition 2

For any given threshold $\epsilon > 0$, define

$$p_\epsilon := \inf_{\mathbf{w}_*} \in \mathcal{W}_* \mathbb{P}\left\{\|\mathbf{g}_*\|^2 \leq \epsilon\right\},$$

where $\mathbf{g}_* = \frac{1}{b}\sum_{i=1}^{b} \nabla f(\mathbf{w}_*; \xi_i)$.
Similarly,

$$M_\epsilon := \sup_{\mathbf{w}_* \in \mathcal{W}_*} \mathbb{E}\left[\|\mathbf{g}_*\|^2 \mid \|\mathbf{g}_*\|^2 > \epsilon\right].$$

Theorem 3

Let Assumption 1 hold for some $N > 0$. Suppose that F is L-smooth and let $\eta < \frac{1}{LN}$. Then, for any $\epsilon > 0$, we have

$$\begin{aligned} \frac{1}{t+1}\sum_{k=0}^{t}\mathbb{E}[\|\nabla F(\mathbf{w}_k)\|^2] &\leq \frac{[F(\mathbf{w}_0) - F^*]}{\eta\,(1 - L\eta N)\,(t+1)} \\ &+ \frac{L\eta}{(1 - L\eta N)}\epsilon + \frac{L\eta M_\epsilon}{(1 - L\eta N)}(1 - p_\epsilon), \end{aligned}$$

where F^* is any lower bound of F; and p_ϵ and M_ϵ are as defined.

Corollary 3

For any ϵ *such that* $1 - p_\epsilon \leq \epsilon$, *and for* $\eta \leq \frac{1}{2LN}$, *we have*

$$\frac{1}{t+1}\sum_{k=0}^{t}\mathbb{E}[\|\nabla F(\mathbf{w}_k)\|^2] \leq \frac{2[F(\mathbf{w}_0) - F^*]}{\eta(t+1)} + 2L\eta(1 + M_\epsilon)\epsilon.$$

Hence, if $t \geq T$ *for* $T = \frac{[F(\mathbf{w}_0) - F^*]}{(L\eta^2)(1+M_\epsilon)\epsilon}$, *we have*

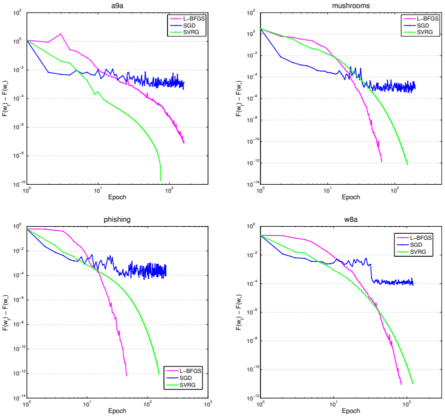
$$\frac{1}{t+1}\sum_{k=0}^{t}\mathbb{E}[\|\nabla F(\mathbf{w}_k)\|^2] \leq 4L\eta(1 + M_\epsilon)\epsilon.$$

Outline

Percentage of f_i with small gradient value for different threshold ϵ

Datasets	$F(\mathbf{w}_{SGD}) - F(\mathbf{w}_*)$	$\epsilon = 10^{-2}$	$\epsilon = 10^{-3}$	$\epsilon = 10^{-4}$	$\epsilon = 10^{-5}$	$\epsilon = 10^{-6}$
covtype	$5 \cdot 10^{-4}$	100%	100%	100%	99.9995%	54.9340%
ijcnn1 (91701)	$1 \cdot 10^{-4}$	100%	100%	100%	96.8201%	89.0197%
ijcnn2	$1 \cdot 10^{-4}$	100%	100%	100%	99.2874%	90.4565%
w8a	$1 \cdot 10^{-4}$	100%	99.9899%	99.4231%	98.3557%	92.7818%
a9a	$1 \cdot 10^{-3}$	100%	100%	84.0945%	58.5824%	40.0909%
mushrooms	$6 \cdot 10^{-5}$	100%	100%	99.9261%	98.7568%	94.4239%
phishing	$2 \cdot 10^{-4}$	100%	100%	100%	89.9231%	73.8128%
skin_nonskin	$6 \cdot 10^{-5}$	100%	100%	100%	99.6331%	91.3730%

SGD behavior for logistic regression



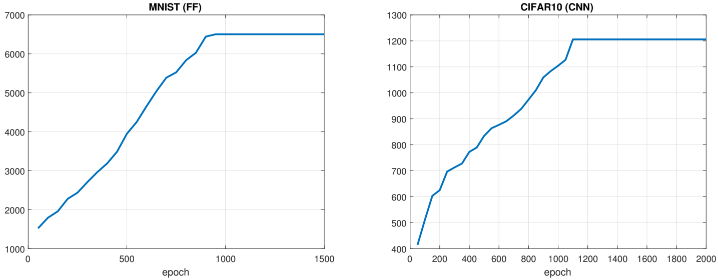
The convergence comparisons of SGD, SVRG, and L-BFGS

Percentage of f_i with small gradient value for different threshold ϵ (Neural Networks)

Datasets	Architecture	$\|\nabla F(\mathbf{w}_*)\|^2$	$\epsilon = 10^{-3}$	$\epsilon = 10^{-5}$	$\epsilon = 10^{-7}$	N	M
MNIST	**FF**	$1.3 \cdot 10^{-15}$	100%	100%	99.99%	6500	10^{-8}
SVHN	**FF**	$3.5 \cdot 10^{-3}$	99.94%	99.92%	99.91%	12000	500
MNIST	**CNN**	$1.6 \cdot 10^{-17}$	100%	100%	100%	6083	10^{-8}
SVHN	**CNN**	$8.1 \cdot 10^{-7}$	99.99%	99.98%	99.96%	8068	0.18
CIFAR10	**CNN**	$5.1 \cdot 10^{-20}$	100%	100%	100%	1205	10^{-14}
CIFAR100	**CNN**	$5.5 \cdot 10^{-2}$	99.50%	99.45%	99.42%	984	3000

Existence of constant N

$$r_t = \frac{\frac{1}{t+1}\sum_{k=0}^{t}\left(\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(\mathbf{w}_k) - \nabla f_i(\mathbf{w}_*)\|^2\right)}{\frac{1}{t+1}\sum_{k=0}^{t}\|F(\mathbf{w}_k)\|^2}$$



The behaviors of r_t

Outline

Conclusion

New view of complexity of SGD

Methods	Strongly convex	General convex	Nonconvex
GD	$\mathcal{O}\left(n\frac{L}{\mu}\log\left(\frac{1}{\epsilon}\right)\right)$	$\mathcal{O}\left(\frac{n}{\epsilon}\right)$	$\mathcal{O}\left(\frac{n}{\epsilon}\right)$
SVRG	$\mathcal{O}\left((n+\frac{L}{\mu})\log\left(\frac{1}{\epsilon}\right)\right)$	$\mathcal{O}\left(n+\frac{\sqrt{n}}{\epsilon}\right)$	$\mathcal{O}\left(n+\frac{n^{2/3}}{\epsilon}\right)$
SARAH	$\mathcal{O}\left((n+\frac{L}{\mu})\log\left(\frac{1}{\epsilon}\right)\right)$	$\mathcal{O}\left((n+\frac{1}{\epsilon})\log\left(\frac{1}{\epsilon}\right)\right)$	$\mathcal{O}\left(n+\frac{1}{\epsilon^2}\right)$
SGD	$\mathcal{O}\left(\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$	$\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$
SGD $1-p_\epsilon \leq \epsilon$	$\mathcal{O}\left(\frac{L}{\mu}\log\left(\frac{1}{\epsilon}\right)\right)$	$\mathcal{O}\left(\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\frac{1}{\epsilon}\right)$

Thank you, Don!

On the way back from Huatulco, 2007.