Enhancing the efficiency of the stochastic method by using non-smooth and non-convex optimization

Anjani Kumar Singha[1] and Swaleha Zubair [2]

[1] *Research scholar, Aligarh Muslim University*

[2] *Assistant Professor Aligarh 202002, Uttar Pradesh, India*

Abstract: *In this paper, an attempt has been made to analyses advanced stochastic methods for the optimization use of non-convex, non-smooth finite sum problems. Interestingly, despite the widespread use and importance of non-convex models, our standing of the non-smooth, non-convex counterpart is very limited. Knowingly, in this context, the non-convex part of smooth and the non-smooth part is convex. Surprisingly, it is not clear about the proximal stochastic gradient that it has probable convergence with constant mini-batches at a stationary point. Thus, this paper is instrumental in showing the development of fast stochastic algorithms that probably converge to a stationary point with constant mini-batches. Hence, it is helpful in minimizing a fundamental gap in our understanding of non-smooth, non-convex problems. For stochastic methods, the optimization techniques used in non-asymptotic convergence rates are applicable for non-convex, on-smooth problems with mini-batches. Perhaps, this is an extension of our analysis to mini-batch variants, showing (Theoretical) linear speed up due to mini batching in parallels settings. Comparatively, by using variants of these algorithms, the faster convergence rate has been induced than batch proximal gradient descent. Henceforth, this paper experimentally highlights an amazing subclass of non-smooth, non-convex functions for an extension of global linear convergence rates. Finally, the exposition of this advanced approach is concentrated around topics related to the experimental Ideas, although in certain aspects it is also pertinent to analogous issues in combinational optimization.*

Keywords**:** **Stochastic** gradient descent(SGD),Mini-Batch gradient descent(MBGD)

1. Introduction

We are studying non-convex, non-smooth finite-sum optimization problems of the form$y \in R^d F(y) := \frac{minimum}{f(y) + g(y)}$, where $f(\mathrm{y}) \coloneqq \frac{1}{n} \sum_{\mathrm{j=o}}^{\mathrm{n}-1} \mathrm{f_j}\ (\mathrm{y})$, (1)$y \in R^d$ is convex space, j is a lower limit of finite-sum of optimization problems and n is the optimal number of the stochastic process, wherever fj:R^d→ R is smooth for all j∈{0,1, . . . , n-1}≜ ⌊ŋ⌋while g: R^d→ R is non-smooth but relatively very simple and convex. The problems of such finite sum optimization are fundamental to deep learning. Thus, it arises generally inside the spectrum of regularization there is the minimization of empirical risk. There has been large-scale research for solving non-smooth convex finite sum problems,[4,16,32] (i.e. every $\mathrm{f_j}$ is convex for j∈ ⌊ŋ⌋)our knowledge of non-smooth, non-convex opponents are incidentally restricted —despite extensive use and signification of non-convex models. Thus, the authors the efficiency of enhancing the stochastic method for solving non-convex, non-smooth finite sum problems. An applicable approach to handle non-smooth through proximal operators(PO) [14,25]. For a function under closed convex g, the PO is defined as

$$\mathrm{prox}_{ŋg}{}^{(\mathrm{y})} = \frac{Argminimum}{y1 \epsilon R^d} (\mathrm{g(y1)} + \frac{1}{2ŋ} ||\mathrm{y1} - \mathrm{y}||^2), \text{ here,ŋ>0 is the step size} \qquad (2)$$

and g is convex where y1,is not linear in ŋ,so ŋ has the function as step size and is finite g(y1) .The power of proximal operators is instrumented in a generalization of projection. Still, if g is the function indicator $\mathrm{I_C(y)}$ at sets C of closed convex then $\mathrm{proxI_C(y)} \equiv \mathrm{proj_C(y)} \equiv \mathrm{Argminimum}_{\mathrm{y1 \in C}}||$ y1-y|.In this paper, we assess that computing proximal operator of g is a sample. This is evidently true for its applications in statistics and deep learning including regularization of sample constraints, box constraints amongst others [2,18]. Remarkably, we assess to a PO(Proximal oracle) that takes a point $\in R^d$ and returns the output of equation(2).More briefly, for explaining our complexity results, we experimentally use the IFFO(incremental first-order oracle).For function f = $\frac{1}{n} \sum_j \mathrm{f_j}$, $anIFOO$takes and index j ∈⌊ŋ⌋ at point $y \in R^d$ and returns the pairs ($\mathrm{f_j(y)}$,$\nabla \mathrm{f_j}$ (y)).The proximal gradient method ((ProxGD) is a standard (batch size) method for solving equation(1) [13] and it was first studied for non-convex problems in [5].The following iteration performs by this method.$\mathrm{y}^{\mathrm{t+1}} = \mathrm{prox}_{ŋg}\left(y^t - ŋ\nabla \mathrm{f}(y^t)\right)_{,}$if t=0,1,2,....., (3)

here,$ŋ > 0$,is a step size. In the recent past, this non-asymptotic rate of convergence results for the proximal gradient method has been proved.

2. Related Work

The exclusive research works have been done on finite sum problems. Therefore, we high light only a few closely related works. A vast study has been done on convex instance equation(1)[19,15,3] and these instances are fairly well known, the landmark newly progress for smooth convex instances of equation(1) in the creation of VR stochastic methods[26,28,4,8].In [32,4] a detail study on non-smooth proximal variance reduced stochastic method has been done, hence for strongly convex and non-strongly convex cases, the faster convergence rate has been proved. Lower bounds are studied in[1,10] whereas asynchronous variance reduced frameworks are developed in [12,21].Surprisingly, the non-convex instances of equation(1) are much lesser understood. The analysis of stochastic gradient for smooth non-convex problems have been computed in [6] and nearly in recent time, the results of convergence for variance reduced stochastic method have been obtained[22,23,33] for smooth non-convex problems. In the author's opinion, the variance reduced non-convex settings are different from ours, for instance, when at a point where hard thresholding is used the loss is convex. We created upon[20,22,23] and emphasis on handling the non-smooth convex regularization method($h\ 6\not\equiv 0$ in (1)). The incremental proximal gradient was also considered in [31] for this class, but asymptotic convergence was only highlighted. For our first analysis of version of a projection of non-convex proxSVRG is because of [24,27,29],which considers the peculiar problems PCA. The following works refer to [30].Probably, our work is closed to [7] in which convergence of mini-batch non-convex proxSGD method has been discussed. Hence, the convergence is slow for a stochastic gradient method. Furthermore, for constant mini-batch, no convergence is proved.

2.1 The main contributions

The relevant and important contributions are listed and therefore the outcomes are tabulated below.

(a)The non-convex proximal version of proposed stochastic algorithms SVRG and SAGA has been analyzed[4,8]. The confluence of the above algorithms is shown in mini-batches. In our understanding, this is the first work presenting non- asymptotic convergence rates for the SGD method applying non-convex problems in static mini-batches.

(b)By using the size of the mini-batch ($^1/_\epsilon$). Probably for faster convergence can be achieved in both proximal gradient and proximal stochastic gradient.

(c)In this context, we go through a non-convex subclass depending on Polyak-Łojasiewicz inequality[8,9]. The optimal result of this subclass is shown by PROXSVRG and ProxSAGA. This work is the first stochastic method for a subclass of problems with proven global linear convergence.

3. Preliminaries

Function g(x) is LSC and convex. Moreover domain(g) = $\{y \,\epsilon R^d \,|g(y) < +\infty\}$ is treated as closed. We say f is L-smooth if there is a constant L such that $\|\nabla f(y) - \nabla f(z)\| \leq K\|y - z\|$, $\forall$ y, z ϵR^d .In Table 1,the comparison of IFOO and PO complexity for different algorithms were analyzed in this study. The measurement of complexity can be achieved in terms of oracle calls and its count is required to get ϵ- an exact solution. The PO(PL) and IFOO(PL) complexity are shown by PL. This table shows the indication of a stochastic algorithm using a defined mini-batch size. According to our knowledge, a convergence of PROXSGD is not known on using mini-batch size for non-convex, non-smooth optimization. In the view of PL functions, there is not aware of specific convergence results for PROXSGD. In overall assumption, We find out particular functions f_j in (1) $\|\nabla f(y) - \nabla f(z)\| \leq L2\|y - z\|$ for all $j\epsilon\lfloor ŋ \rfloor$,where Lipschitz Continuous with Lipschitz factor. This kind of assumption' is typical the analysis of first-order approaches.

Algorithm	IFOO	PO	IFOO(PL)	PO(PL)	CM
ProxSGD	$\boldsymbol{\theta}(^{1}/_{\epsilon}\,2)$	$\theta(^{1}/_{\epsilon})$	$\boldsymbol{\theta}(^{1}/_{\epsilon}\,2)$	$\theta(^{1}/_{\epsilon})$	?
ProxGD	$\theta(^{n}/_{\epsilon})$	$\theta(^{1}/_{\epsilon})$	$\boldsymbol{\theta}$(nklog($^{1}/_{\epsilon}$)	$\boldsymbol{\theta}$(nklog($^{1}/_{\epsilon}$)	-
ProxSVRG	$\boldsymbol{\theta}$(kn+($kn^{2/3}/_{\epsilon}$))	$\theta(^{1}/_{\epsilon})$	$\boldsymbol{\theta}$(kn+$kn^{2/3}$)log($^{1}/_{\epsilon}$)	$\boldsymbol{\theta}$(nklog($^{1}/_{\epsilon}$)	✓
ProxSVGA	$\boldsymbol{\theta}$(kn+($kn^{2/3}/_{\epsilon}$))	$\theta(^{1}/_{\epsilon})$	$\boldsymbol{\theta}$(kn+$kn^{2/3}$)log($^{1}/_{\epsilon}$)	$\boldsymbol{\theta}$(nklog($^{1}/_{\epsilon}$)	✓

Table1:Comparision between IFOO and PO or PO(PL)

About convex problems, the particular optimality gap $F(Y)-F(Y^{*})$ is used as a criterion. It is logically not valid to apply such a criterion for general non-convex($g\equiv0$) problems because of their inapplicability. For instance, this is a suitable alternative to gradient mapping [17] but cannot be used for non-smooth problems.

$$G_{ŋ}:=\frac{1}{ŋ}[y-prox_{ŋg}\big(y-ŋ\nabla f(y)\big)]. \qquad (5)$$

When ($g\equiv0$) this mapping is certainly reduced $G_{ŋ}=\nabla f(y)_{=}\nabla F(y)$, the gradient of function F at y. The analysis of the algorithm is done using gradient descent of equation (5) as described In definition 1.

Definition 1. A point y output by SGD iteration algorithm intended for solving equation (1) is called a $\boldsymbol{\epsilon}$-actual solution, if $E[\|G_{ŋ}\|^{2}]\leq\boldsymbol{\epsilon}$ for there some $ŋ>0$.

4. Terminology

Table 2 Terminologies:

Notation	Description
GDS	Gradient Descent Stochastic
n	Optimal number of stochastic process
m	Mini-Batch size
MBGD	Mini-Batch Gradient Descent
IF	Indicators Function
IFOO	incremental first-order oracle
ProxGD	proximal-gradient descent
SC	Smooth-Convex
L1	L1 norm of a vector w(weight) respectively
y	Random variable
yϵR^d	A convex space
g(y)	Smooth function
f$_j$	Non-Smooth
SAGA	Stochastic Average Gradient Approach
SVRG	Stochastic Variance Reduce Gradient
L2	Lipschitz Continuous with Lipschitz factor
m1	epoch length
LSC	Lower Semi-Continuous(LSC)
CM	Constant Minibatch
PL	Polyak-Lojasiewicz (pl) inequality
I$_C$	Referred to as minibatch
g = 0	Proximal minimization algorithm.
h = I$_C$	Projected gradient descent
t,s	Convergent point
ŋ	Step size

5. Lemmatta

Few intermediate outcomes can be useful for each of our analyses. We proved throughout the change of perfect.

Lemma 1. y1= prox$_{ŋh}(y-ŋd)$. For some dϵR^d.Then y1, this inequality holds.

F1 (y1)]≤F1 (z)+$(y1-z,\nabla f(y)-d)+\left[\frac{L2}{2}-\frac{1}{2ŋ}\right]||y1-y||^2+[\frac{L2}{2}+\frac{1}{2ŋ}]||z-y||^2$-$\frac{1}{2\eta}$||y1-z||^2.

for all dϵR^d.

Lemma **2.**For iterates y_{t+1}^{s+1},$v1_t^{s+1}$,$\bar{y}^s$ where t∈{0,1, . . . , n-1} and sϵ{0,1,2,........,s1-1} in Algorithms 1,that inequality holds:

E[||∇f(y_t^{s+1})-$v1_t^{s+1}$||^2≤$\frac{L2^2}{m}$||y_t^{s+1}-$\bar{y}^s$||^2.

Lemma **3.**For iterates y^t,$v1$,and $\left\{\alpha 1_j^t\right\}_{j=1}^n$ where tϵ{0,1,2.........T1-1} in Algorithms 2, that inequality holds:

E[$||\nabla$f(y^t)-$v1^t||^2 \leq \frac{L2^2}{nm}\sum_{j=1}^{n} E||\, y^t$-α$1_j^t||^2$.

Lemma 4. Let function g:$R^d \rightarrow$ R is lower semi continuous and y1= $\text{prox}_{ŋg}^{(y)}$.Then following inequality holds:

g(y1)+$\frac{1}{2ŋ}||$y1-y$||^2 \leq$g(z)+$\frac{1}{2ŋ}||$z-y$||^2$-$\frac{1}{2ŋ}||$y1-z$||^2$, for all z$\boldsymbol{\epsilon} R^d$.

Lemma 5. Let function g:$R^d \rightarrow$ R is L2-smooth,then following inequality holds:
f(y1)+$\langle\nabla$f$(y1), y1 - y\rangle$-$\frac{L2}{2}||$y1-y$||^2 \leq$f(y) $\leq$f(y1)+$\langle\nabla$f$(y1), y1 - y\rangle$+$\frac{L2}{2}||$y-y1$||^2$, for all y,y1$\boldsymbol{\epsilon} R^d$.

Lemma 6. Any random variables z1,z2,..........,z_r are independent variable and mean is zero ,then we have the following :

E[||z1+z2+z3+.........+ $z_{r||}{}^2$] =E[$||$z1$||^2$+..........+$||$ $z_{r||}{}^2$] .

Lemma 7. Any random variables z1,z2,..........,zr, we have
E[||z1+z2+z3+.........+ $z_{r||}{}^2$] $\leq$r E[$||$z1$||^2$+..........+$||$ $z_{r||}{}^2$] .

6. Algorithms

there are two algorithms (a) ProxSVRG (b) ProxSVGA.

7.1. Non-convex Prox SVRG (Nonconvex Proximal SVRG)

At first, we consider a variety of ProxSVRG [32]; the pseudo-code of this variant is written in Algorithm 1.Whenever F is strongly convex,SVRG getsthe linear convergent rate as opposite to the sub linear convergent of SGD[11,8].Recalling this, when ProxSVRG is very difficult to start with m=1, all of us use its mini-batch alternative with batch size m. ProxSVRG is particular and most attractive because of its low memory requirement. For the requirement of proxSVRG low memory, it just needed **θ**(d) for additional memory in comparison to SGD for the conservation of mean gradient. and with the help of non-strongly convex composite Objectives(SAGA) by using fast incremental gradient method **θ**(nd) cost can be obtained. Furthermore, it is too strong assumptive results,SVRG is identified to better SGD experiment when a more powerful selection of phase size intended for problems of convex, ProxSVRG is recognized to inherently these upper arms of SVRG[32]. We introduce our analysis of non-convex proxSVRG, initials results with batch size m=1.

7.2 . Theorem

A theorem of Convergence Analysis:
We start the convergent rate associated with ProxSVRG and ProxSAGA to get a set of specific parameters. Whenever most of the analysis may be derived regarding those algorithms. The reason at the back choice associated with parameters of equation(3).We got the following convergent results for ProxSVRG andProxSAGA.

Theorem 1.

Assume m≤ n+1 in algorithms 1.let T1 is multiple of m_1 and $\eta=\rho/L2$ were $\rho<1/2$ and fulfills the condition : $\frac{4\rho^2 m1^2}{m}$+ ρ≤1 where m is the mini-batch size.

In that case for output x_b of Algorithm 2,We get :

$$E\,[\|G_\eta(y_b)\|^2]\leq\frac{2L2(F1^0)-F1(y^*))}{\rho(1-2\,\rho)T1},$$

Optimal solution y^*in equation (1)

Proof of Convergence Analysis: We are defining the particular gradient iterate

$$\bar{y}_{t+1}^{s+1}= \mathrm{prox}_{\eta g}(y_t^{s+1}-\eta\nabla f(y_t^{s+1})), \tag{8}$$

That is simply to our analysis and is no means computed. Consider. Using Lemma 2 to equation (8)(With $y1=\bar{y}_t^{s+1}$,$z=y_t^{s+1}$,and $d=\nabla f(y_t^{s+1})$), and Using expectations we obtained the particular bounded. $E[F1(\bar{y}_{t+1}^{s+1})]\leq E[F1(y_t^{s+1})+[\frac{L2}{2}-\frac{1}{2\eta}]\|\bar{y}_{t+1}^{s+1}-y_t^{s+1}\|^2-\frac{1}{2\eta}\|\bar{y}_{t+1}^{s+1}-y_t^{s+1}\|^2]$. (9)

Remember the iteration of Algorithm 1to find out using followingan update :

$y_{t+1}^{s+1}= \mathrm{prox}_{\eta g}(y_t^{s+1}-\eta(v1_t^{s+1}))$, where $v1$ is a random vector(10)

$v1_t^{s+1}=\frac{1}{m}\sum j_t\epsilon I_t\,(\nabla fj_t(y_t^{s+1})\,\text{-}\nabla fj_t\,(\bar{y}^s))+\,g1^{s+1}$ (see Algorithms1). using Lemma 2 to update (10) (with $y1=y_{t+1}^{s+1}$,$z=\bar{y}_{t+1}^{s+1}$,and $d=v1_t^{s+1}$) and getting expectations we obtain.

$$E[F1(y_{t+1}^{s+1})]\leq E[F1(\bar{y}_t^{s+1})+(y_{t+1}^{s+1}-\bar{y}_{t+1}^{s+1},\nabla f(y_t^{s+1})-v1_t^{s+1})+\left[\frac{L2}{2}+\frac{1}{2\eta}\right]||\bar{y}_{t+1}^{s+1}-y_t^{s+1}||^2$$

$$+[\frac{L2}{2}-\frac{1}{2\eta}]\|y_{t+1}^{s+1}-y_t^{s+1}\|^2-\frac{1}{2\eta}\|y_{t+1}^{s+1}-\bar{y}_{t+1}^{s+1}\|^2],\text{where}v1\text{is the random vector.} \tag{11}$$

Adding equation (9) and (11), we will get

$$E[F1(y_{t+1}^{s+1})]\leq E[F1(y_t^{s+1})+[g-\frac{1}{2\eta}]||\bar{y}_{t+1}^{s+1}-y_t^{s+1}||^2+[\frac{g}{2}-\frac{1}{2\eta}]\|y_{t+1}^{s+1}-y_t^{s+1}\|^2-\|y_{t+1}^{s+1}-\|^2+(y_{t+1}^{s+1}-\bar{y}_{t+1}^{s+1},\nabla f(y_t^{s+1})-v1_t^{s+1})] \tag{12}$$

where $T_2=(y_{t+1}^{s+1}-\bar{y}_{t+1}^{s+1},\nabla f(y_t^{s+1})-v1_t^{s+1})$

We could bound the T_2term in the following :

$$E[T_2]\leq\frac{1}{4\eta}E\|y_{t+1}^{s+1}-\bar{y}_{t+1}^{s+1}\|^2+\frac{\eta}{4}E||\nabla f(y_t^{s+1})-v1_t^{s+1}||^2\leq\frac{1}{4\eta}E\|y_{t+1}^{s+1}-\bar{y}_{t+1}^{s+1}\|^2+\frac{\eta g^2}{2m}E||y_t^{s+1}-\bar{y}^s||^2]$$

The initial inequality from young's inequality and Cauchy-Schwarz, even though the second inequality is a consequence of Lemma 3.Substitutig the equation (12) in T2,we can see.

$$E[(y_{t+1}^{s+1})]\leq E[(F1(y_t^{s+1})]+[g-\frac{1}{4\eta}]||\bar{y}_{t+1}^{s+1}-y_t^{s+1}||^2+[\frac{g}{4}-\frac{1}{4\eta}]||y_{t+1}^{s+1}-y_t^{s+1}||^2+\frac{\eta g^2}{2m}E||y_t^{s+1}-\bar{y}^s||^2]. \tag{13}$$

To analyze equation (13) further, we discovered which we use the following Lyapunov function:

R_t^{s+1}:=$E[F1\,(y_t^{s+1})]+c1_t||y_t^{s+1}-\bar{y}^s||^2]$.

for certain entity $c1_b$=0 and $c1_t$=$c1_{t+1}$ (1+λ)+$\frac{ŋg^2}{2m}$.Further, for the remainder of analysis bounded set

λ=1/b. We will then certain bounded set R_{t+1}^{s+1}as follow

$$R_{t+1}^{s+1} = E[F1\,(y_{t+1}^{s+1})+c1_{t+1}||\,y_{t+1}^{s+1}-y_t^{s+1}+y_t^{s+1}-\bar{y}^s||^{\,2}\,]$$
$$=E[F1\,(y_{t+1}^{s+1})+c1_{t+1}(||y_{t+1}^{s+1}-y_t^{s+1}||^2+||y_t^{s+1}-\bar{y}^s||^2+2(y_{t+1}^{s+1}-y_t^{s+1}+y_t^{s+1}-\bar{y}^s))]$$

$$\leq E[(F1(y_{t+1}^{s+1})+c1_{t+1}\,(1+λ)||y_{t+1}^{s+1}-y_t^{s+1}||^2+c1_{t+1}\,(1+λ)||y_t^{s+1}-\bar{y}^s||^2]$$

$$\leq E[(F1(y_t^{s+1})+[g\text{-}\frac{1}{4ŋ}]||\bar{y}_{t+1}^{s+1}-y_t^{s+1}||^2\ +[c1_{t+1}\ (1+\frac{1}{λ})\ +\frac{g}{4}-\frac{1}{4ŋ}]\,||y_{t+1}^{s+1}-y_t^{s+1}||^2\ +[c1_{t+1}\ (1+λ)$$
$$+\frac{ŋg^2}{2m}\,]||y_t^{s+1}-\bar{y}^s||^2] \qquad (14)$$

$$\leq E[(F1(y_t^{s+1})+[g\text{-}\frac{1}{4ŋ}]||\bar{y}_{t+1}^{s+1}-y_t^{s+1}||^2+[c1_{t+1}\,(1+λ)+\frac{ŋg^2}{2m}]||y_t^{s+1}-\bar{y}^s||^2]$$
$$=R_t^{s+1}+[g\text{-}\frac{1}{4ŋ}]E||\bar{y}_{t+1}^{s+1}-y_t^{s+1}||^2. \qquad (15)$$

The initial inequality follows Young's inequality and Cauchy-Schwarz. The other inequality is to the bounded set of equation (13),although the final equality is to a description of Lyapunov function R_t^{s+1}.The next inequality holds the pattern of that values $c1_t$ fulfills the bounded set:

$$c1_{t+1}\,(1+\frac{1}{λ})+\frac{g}{4}\leq\frac{1}{4ŋ}. \qquad (16)$$

to verify equation (16), the initial notice that $c1_{\mathrm{m1}}$=0 and $c1_{t+1}$ (1+λ)+$\frac{ŋg^2}{2m}$. Recursion of parameter t, we will obtain

$$c1_{t=}\frac{ŋg^2}{2m}\frac{(1+λ)^{b-t}-1}{λ}=\frac{ρg\mathrm{m1}}{2m}((1+\frac{1}{\mathrm{m1}})^{\mathrm{m1}-t}\ \text{-}1)\leq\frac{ρg\mathrm{m1}}{2m}\,(e_1\text{-}1)\leq\frac{ρg\mathrm{m1}}{m}$$

in which the initial equality is to credit to the meaning associated withŋ and λ. Follow the initial inequality follow that$\lim_{h\to\infty}\left(1+\frac{1}{h}\right)^h$=$e_1$ and $\left(1+\frac{1}{h}\right)^h$ is an increasing function where h>0(e_1 denoted the Euler's number).

$$c1_{t+1}\,(1+\frac{1}{λ})+\frac{g}{4}\leq\frac{ρg\mathrm{m1}}{m}(1+\mathrm{m1})+\frac{g}{4}\leq\frac{ρg\mathrm{m1}^2}{m}+\frac{g}{4}\leq\frac{g}{2ρ}=\frac{1}{2ŋ},$$

where use the 2nd inequalitym1≥1.Use of 3rd inequality that is following condition follow.

$$\frac{4ρ^2\mathrm{m1}^2}{m}+ρ\leq1.$$

where equation (16) of an inequality follows. Right now, equation (15) adding all of the iterations in epoch and in that case microscopic sums, we have

$$R_b^{s+1}\leq R_1^{s+1}+\sum_{t=0}^{b-1}[g-\frac{1}{4ŋ}]E||\bar{y}_{t+1}^{s+1}-y_t^{s+1}||^2. \qquad (17)$$

The definition of $\bar{y}_t^{s+1}$ and since$c1_{\mathrm{m1}}$=0, it follows that R_{m1}^{s+1} =E[F($\bar{y}_{\mathrm{m1}}^{s+1}$)]=E[F($\bar{y}^{s+1}$)].Moreover,$R_1^{s+1}$=E[F1($\bar{y}_1^{s+1}$)]=E[F1($\bar{y}^s$)].In fact that y_1^{s+1}=$\bar{y}^s$. Accordingly, the above equation (17) of inequality used to, we get

$$E[F(\bar{y}^{s+1})]\leq E[F1(\bar{y}^s)]+\sum_{t=0}^{\mathrm{m1}-1}[g-\frac{1}{4ŋ}]E||\bar{y}_{t+1}^{s+1}-y_t^{s+1}||^2. \qquad (18)$$

Adding of equation(18) all of the iterations in epochs and rearranging all terms, we get bounded set

$$\sum_{s=1}^{s1}\sum_{t=1}^{m1}[\frac{1}{4ŋ}-g]\text{E}||\bar{y}_{t+1}^{s+1}-y_t^{s+1}||^2\leq \text{F1}(\text{y}^0)\text{-E}[\text{F}(\bar{y}^s)]\leq\text{F1}(\text{y}^0)\text{-E}[\text{F1}(\text{y*})],$$
(19)

The optimality of y* in 2nd inequality as follow

$$\text{Ǥ}_{ŋ}(y_t^{s+1})=\frac{1}{ŋ}[y_t^{s+1}-\text{prox}_{ŋh}(y_t^{s+1}-ŋ\nabla \text{f}(y_t^{s+1}))]=\frac{1}{ŋ}[y_t^{s+1}-\bar{y}_{t+1}^{s+1}].$$

Using the relation of an equation in (19) we get

$$\sum_{s=1}^{s1}\sum_{t=1}^{m1}[\frac{1}{4ŋ}-g]\,ŋ^2\text{E}||\,\text{Ǥ}ŋ(y_t^{s+1})||^2\leq\text{F1}(\text{y}^0)\text{-E}[\text{F1}(\text{y*})] \qquad (20)$$

Right now associated with y_b from Algorithm 1 and simplifying we get the desired results

7.3 . Convergence Analysis of theorem1

Theorem1**:** theorem1 show proxSVRG for constant MBGD of size m=1.this result is strongly more opposing to proxSGD where convergence with MBGD certainly not know. Hence, the results obtained by theorem1 is weaker than that of proxSGD. This point is highlighted by theorem1 to the given corollary.

Corollary 1. To obtaining ϵ-accurate solution to obtain with m=1 and parameters from Theorem 1,the IFOO and PO(PL) Complexity of Algorithm 1 happen to be ɵ(n/ϵ) and $1/\epsilon$)correspondingly.

Corollary1. Our study of corollary1 follows Algorithm1 of each inner iteration has IFOO complexity of ɵ(1⁄ϵ) since m1=n, we assume that n is an integer .This IFOO complexity includes the IFOO call for calculating average SGD at the end of every epoch. Also, each internal iteration invokes the PO, where the PO complexity is ɵ(n/ϵ). it is similar to those with MBGD proxSVRG and proxSGD ,this is because n IFOO calls in proxSGD is correspond to the PO call whereas one PO call corresponds to one IFOO call in proxSVRG.

Theorem 2. Suppose m=$n^{2/3}$ in Algorithm Let ŋ=1/(3L2),m1=[n1/3] and T1 is a multiple of m1.Then for the output y_b associated with Algorithm 1,we have got :

$$\text{E }[||\text{Ǥ}_{ŋ}(\text{y}_b)||^2]\leq\frac{18\text{L}2(\text{F}1^0)-\text{F}1(\text{y}^*))}{\text{T}1},$$

where y^* isan optimal solution of this equation (1).

Proof:

Corollary 2. Let m=$n^{2/3}$ and set of parameters in Theorem2.Certainly ,to obtain **ϵ**-proper solution, the IFOO and PO(PL) complexity associated with Algorithm are **Ɵ**(kn+($kn^{2/3}/\epsilon$)) and ɵ($1/\epsilon$) respectively. From the Theorem2, it can be seen that the total number of inner iterations in all epochs of Algorithm 1 to obtain **ϵ**-proper solution is ɵ($1/\epsilon$) inner iteration of all epochs of Algorithm 2 involves the call to PO(PL),we have a PO proper complexity ɵ($1/\epsilon$).Hence, Since m=$n^{2/3}$ IFOO call is produced at each inner iteration, we obtain total complexity of ($n^{2/3}/\epsilon$).

```
**Algorithm 1:** NC ProxSvrg($y^1, T1, m1, m, ŋ$)Input : $\bar{y}^1$=$\bar{y}^1_{m1}$=$y^1 \epsilon R^d$,epoch length m1,step size $ŋ > 1$,s1=$^{T1}/_{m1}$
for s=0 to s1-1 do
$y_1^{s+1}$ =$y_{m1}^s$
$\mathbf{g}^{s+1}$=$\frac{1}{n}\sum_{j=2}^{n+1} \nabla f_j\left(\bar{y}^s\right)$
for t=0 to m1-1 do
**consistently** randomly chosen $I_t \subset 1,2,3,.,.,.,.,.n+1\}$ such that $|I_t|$=m
$v1_t^{s+1}$=$\frac{1}{m}\sum j_t \epsilon I_t (\nabla f j_t (y_t^{s+1}) - \nabla f j_t (\bar{y}^s)) + g1^{s+1}$
$y_{t+1}^{s+1}$= $\mathrm{prox}_{ŋh}(y_t^{s+1} - ŋ v1_t^{s+1})$
end for
$\bar{y}^{s+1}$=$y_{m1}^{s+1}$
end for
Output: Iterate $y_b$ pick random from$\{\{y_{t+1}^{s+1}\}\}_{t=0}^{m1-1}\}_{s=0}^{s1-1}$.
```

7.4 . Non-Convex Proximal SAGA

In the previous section, we studied ProxSVRG for finding a solution for (1). and we found that ProxSVRG requires full SGD calculation per each epoch and it's not a fully incremental algorithm. There is an alternative algorithm to ProxSVRG developed in [4] and based on the work of [4] we develop ProxSAGA, a non-convex variant of SAGA. In algorithm 2 we showed pseudo-code for ProxSAGA. ProxSAGA is the main difference between algorithm 1 and algorithm 2, which avoids calculation of full SGD per each epoch. It maintains an average SGD vector $g1^{t+1}$ for each iteration. We need to store the SGD $\{\nabla f_j(\alpha_j^t)\}_{j=1}^n$, (which in general can cost O(nd) in storage but in some cases, we can reduce it to O(n)). ProxSAGA will give results better than ProxSVRG and its implementation is also easy. ProxSAGA in algorithm 2 is minutely different from an alternative algorithm which is mentioned in [4].Especially, where two seats I_t ,J_t are sampled at each iteration when uses in mini-batches as gradient one in [4].This is chiefly applicable to the case of theoretical analysis. it has been proved that non-convex proxSVRG and proxSAGA shape similar guarantees convex case. Particular, our first result for proxSAGA for proxSVRG in theorem1 is similar.

```
Algorithm 2:NC ProxSAGA($y^1, T1, m, ŋ$)
Input:$y^1 \epsilon R^d$,$\alpha 1_i^1$=$y^1$ for i∈⌈n⌋,step size $ŋ > 0$
$g^1$=$\frac{1}{n}\sum_{j=2}^{n+1} \nabla f_j\left(\alpha 1_j^1\right)$
for t=0 to T1-1 do
consistently randomly chosen sets $I_t$,$J1_t$ from ⌈n⌋ so that $|I_t|$=$|J1_t|$=m
$v1^t$=$\frac{1}{m}\sum j_t \epsilon I_t (\nabla f j_t (y_t^{s+1}) - \nabla f j_t (\bar{y}^s)) + g1^s$
$y^{t+1}$= $\mathrm{prox}_{ŋh}(y^t - ŋ v1^t)$
$\alpha 1_j^{t+1}$=$y^t$ for k∈$J1_t$ and $\alpha 1_k^{t+1}$ =$\alpha 1_k^t$ for k∈$J1_t$
```

$g^{t+1=} g^t - \frac{1}{n} \sum k_t \epsilon J1_t \, (\nabla fk_{t(}\alpha 1^t_{k_t}) \text{-} \nabla fk_{t(}\alpha 1^{t+1}_{k_t}))$

end for

output: Iterate y_b pick random from$\{y^t\}_{t=1}^{T1}$

8. Experiments

In this section, we are presenting our results. We have studied the problem of non-negative PCA(Principle Component Analysis). More importantly, for a set of samples, we solve the following optimization problems. $\{z_j\}^n{}_{j=1,} \frac{1}{4} y^\top (\sum_{j=1}^{n} z_j z_j{}^\top) y$. (7)

Generally, the optimization problem is NP-hard i.e (non-deterministic polynomial-time hard). Standard PCA with this particular form can be formulated as (1) with $f_j(y) = -(y^\top z_j)^2$ for all j ϵ[n] and $h(y)=I_c(y)$ where C is the convex set $\{y \ \epsilon R^d | \ ||y|| <= 1, y>=0\}$. In our experiments, there is comparison between proxSGD with non-convex figure2.For proxSGD the choice of step is important have been taken as

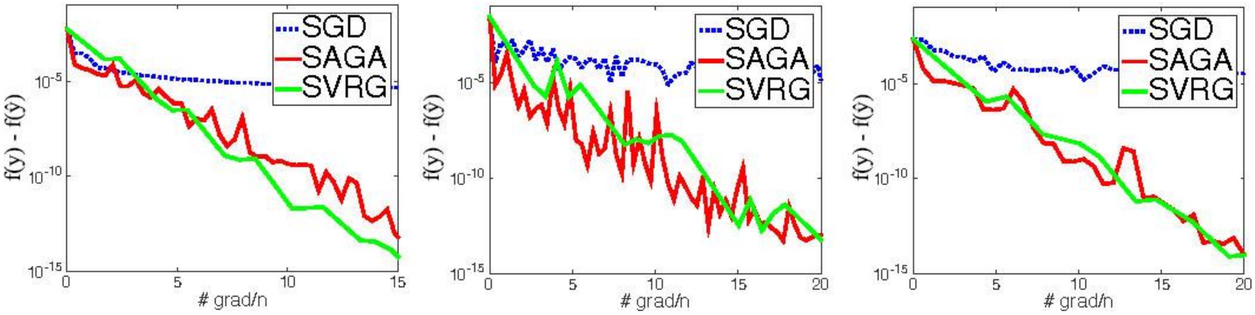

Figure1:Analysis of Non-negative Principal Component .

The analytical performance of components ProxSGD, ProxSVRG, and ProxSAGA on 'rec'(left) ,'mnist'(left-right-center) and (right) datasets have been Graphically Presented. Hence the Y-axis is the function of sub-optimality i.e , $f(Y)-f(Y^*)$,in this analysis, represents the best solution which is obtained by running gradients descent for analyzing the longest time duration with multiple restarts.

$\eta_t = \eta_0 \left(1 + \eta' \left[\frac{t}{n}\right]\right)^{-1}$ where $\eta_0, \eta' > 0$, using the popular t-inverse step size. ProxSVRG and proxSAGA are motivated and based on theoretical analysis. The step sizes for each method are chosen based on the best performance on the objective value. We fixed step size η'=0 to ProxSGD and use epoch length as m1=n. All experiments are done in LIBSVM with normalized samples ($\|z_j\|$=1 for all j ϵ[n]) which are taken from the standard machine learning datasets. initialized by each of the methods all running ProxSGDfor n iterations to serve two purposes. One is typically beneficial for variance and covariance reduction techniques by providing the best initial point. The other one is for calculating the initial mean gradient g^1. Having a mini-batch size m=1 is demonstrated the performance of algorithms with fixed mini-batches in our experiments. We reported the sub-optimality in the objective function for standard machine learning datasets i.e f $(y_t^{s+1}) - f(\hat{y})$ (for ProxSVRG) and f $(y^t) - f(\hat{y})$ (for ProxSAGA) where ŷ is the solution calculated by running proximal gradient descent for multiple random initializations and a huge number of iterations. We compared IFOO complexity divided by n for all algorithms and this includes the cost for the full gradient at the end of each epoch for ProxSVRG.The performances of ProxSGD, ProxSVRG, and ProxSAGA on the NN-Principle Component Analysis problem are shown in Figure 2 the objective value for ProxSVRG and ProxSAGA is much greater compared to ProxSGD. We found significant gain for all datasets and the selection of step size was faster for ProxSVRG and ProxSAGA than that for ProxSGD. For this particular task, we didn't find any significant difference in performances of ProxSAGA and ProxSVRG.

9. Concluding Remarks:

In this paper, the authors have proposed an advanced use of Non-convex, Non-smooth finite sum problems. Experimentally, the variance reduction techniques have been used from better and fast results. By employing this technique we can correctly prove that we can design the methods that one can comparatively perform better then ProxSGD and Proximal gradient descent. The algorithms of stochastic gradient descent (SGD) and its variance are used for solving non-convex problems, particularly deep leering. Thus, the theoretical convergence results of the proposed algorithms of Non-convex, Non-smooth optimization problems have been provided in the paper. The practical outcome approach ofthis paper shows that the proximal stochastic gradient to a stationary point with constant mini-batches has probably convergence. This provable fact adds a milestone in our knowledge of Non-smooth, Non-convex problems. In addition, this paper aims to address many questions and bridge the gap between theory and practical. The authors proposed an advanced and fast stochastic method for a broad family of Non-smooth, Non-convex problems. The key features of these problems include:-

(i) any suitable stochastic convex optimization algorithms example SGD, when employed for minimizing regularized convex problems, can return an averaged solution at each stage.**(ii)**likewise, an averaged solution is returned as the final solution.

1 Taking future works into account, it may be suggested that researchers may consider developing more variants of the meta algorithms. These proposed meta algorithms conclude stage-wise RMS Prop, stage-wise AMS grad, etc .Furthermore, this model may also be largely considered for the empirical studies of the image net data set.

References

[1] Agarwal, Alekh, and Leon Bottou. "A lower bound for the optimization of finite sums." In International conference on machine learning,(2015), pp. 78-86. PMLR.
[2] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Convex optimization with sparsity-inducing norms. In S. Sra, S. Nowozin, and S. J. Wright, editors, Optimization for Machine Learning. MIT Press, (2011).
[3] Léon Bottou. Stochastic gradient learning in neural networks. Proceedings of Neuro-Nımes,(1991) 91(8).
[4] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In NIPS 27,(2014), pages 1646–1654.
[5] Masao Fukushima and Hisashi Mine. A generalized proximal point algorithm for certain nonconvex minimization problems. International Journal of Systems Science,(1981), 12(8):989–1000.
[6] Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. SIAM Journal on Optimization,(2013), 23(4):2341–2368.
[7] Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. Mathematical Programming, 155(1-2):267–305, (2014),December.
[8] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In NIPS 26,(2013), pages 315–323.
[9] Hamed Karimi and Mark Schmidt. Linear convergence of proximal-gradient methods under the polyak-lojasiewicz condition. In NIPS Workshop, (2015).
[10] G. Lan and Y. Zhou. An optimal randomized incremental gradient method. arXiv:1507.02000, (2015).
[11] Xingguo Li, Tuo Zhao, Raman Arora, Han Liu, and Jarvis Haupt. Stochastic variance reduced optimization for nonconvex sparse learning. In ICML, (2016). arXiv:1605.02711.
[12] Ji Liu and Stephen J. Wright. Asynchronous stochastic coordinate descent: Parallelism and convergence properties. SIAM Journal on Optimization, 25(1):351–376,(2015).
[13] Hisashi Mine and Masao Fukushima. A minimization method for the sum of a convex function and a continuously differentiable function. Journal of Optimization Theory and Applications,(1981), 33(1):9–23.
[14] J. J. Moreau. Fonctions convexes duales et points proximaux dans un espace hilbertien. C. R. Acad. Sci. Paris SÃlr. A Math., (1962),255:2897–2899.
[15] Arkadi Nemirovski and D Yudin. Problem Complexity and Method Efficiency in Optimization. John Wiley and Sons, (1983).
[16] Yu Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. SIAM Journal on Optimization, (2012),22(2):341–362.
[17] Yurii Nesterov. Introductory Lectures On Convex Optimization: A Basic Course. Springer, (2003).
[18] N. Parikh and S. Boyd. Proximal algorithms. Foundations and Trends in Optimization, 1(3)(2014),:127–239.
11
[19] BT Poljak and Ya Z Tsypkin. Pseudogradient adaptation and training algorithms. Automation and Remote Control,(1973), 34:45–67.
[20] B.T. Polyak. Gradient methods for the minimisation of functionals. USSR Computational Mathematics and Mathematical Physics, 3(4):864–878,(1963), January .
[21] Sashank Reddi, Ahmed Hefny, Suvrit Sra, Barnabas Poczos, and Alex J Smola. On variance reduction in stochastic gradient descent and its asynchronous variants. In NIPS 28,(2015), pages 2629–2637.

[22] Sashank J. Reddi, Ahmed Hefny, Suvrit Sra, Barnabás Póczos, and Alexander J. Smola. Stochastic variance reduction for nonconvex optimization. CoRR, abs/1603.06160, (2016).

[23] Sashank J. Reddi, Suvrit Sra, Barnabás Póczos, and Alexander J. Smola. Fast incremental method for nonconvex optimization. CoRR, abs/1603.06159, (2016).

[24] H. Robbins and S. Monro. A stochastic approximation method. Annals of Mathematical Statistics, 22☹1951),400–407.

[25] R Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. SIAM journal on control and optimization, (1976),14(5):877–898.

[26] Mark W. Schmidt, Nicolas Le Roux, and Francis R. Bach. Minimizing Finite Sums with the Stochastic Average Gradient. arXiv:1309.2388, (2013).

[27] Shai Shalev-Shwartz. SDCA without duality. CoRR, abs/1502.06177, (2015).

[28] Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss. The Journal of Machine Learning Research, 14(1):567–599, (2013).

[29] Ohad Shamir. A stochastic PCA and SVD algorithm with an exponential convergence rate. arXiv:1409.2848, (2014).

[30] Ohad Shamir. Fast stochastic algorithms for SVD and PCA: Convergence properties and convexity. arXiv:1507.08788, (2015).

[31] Suvrit Sra. Scalable nonconvex inexact proximal splitting. In NIPS, pages(2012), 530–538.

[32] Lin Xiao and Tong Zhang. A proximal stochastic gradient method with progressive variance reduction. SIAM Journal on Optimization,(2014), 24(4):2057–207.

[33] Zeyuan Allen Zhu and Elad Hazan. Variance reduction for faster non-convex optimization. CoRR, abs/(2016),1603.05643.