Simple Stochastic Gradient Methods for Non-Smooth Non-Convex Regularized Optimization

Michael R. Metel [1] **Akiko Takeda** [1 2]

Abstract

Our work focuses on stochastic gradient methods for optimizing a smooth non-convex loss function with a non-smooth non-convex regularizer. Research on this class of problem is quite limited, and until recently no non-asymptotic convergence results have been reported. We present two simple stochastic gradient algorithms, for finite-sum and general stochastic optimization problems, which have superior convergence complexities compared to the current state-of-the-art. We also compare our algorithms' performance in practice for empirical risk minimization.

1. Introduction

In this work we consider regularized optimization problems of the form

$$\min_{w \in \mathbb{R}^d} h(w) := f(w) + g(w), \tag{1}$$

where $f(w)$ has a Lipschitz continuous gradient and $g(w)$ has a proximal operator that can be efficiently computed. In addition, we assume that

$$f(w) := \mathbb{E}_\xi[F(w, \xi)], \tag{2}$$

where $\xi \in \mathbb{R}^p$ is a random vector following a probability distribution P from which i.i.d. samples can be generated. We will also consider what is known as the finite-sum problem, where the expectation of $F(w, \xi)$ is taken over an empirical distribution function created by taking n samples of ξ, ξ_j for $j = 1, ..., n$:

$$f(w) := \frac{1}{n} \sum_{j=1}^{n} f_j(w), \tag{3}$$

where $f_j(w) = F(w, \xi_j)$ and has a Lipschitz continuous gradient.

Our motivation for studying this problem is empirical risk minimization in machine learning. The purpose of $g(w)$, as a regularizer, is to induce a sparse solution when minimizing $f(w)$. Non-convex regularizers have been shown to outperform their convex counterparts with reduced bias in parameter estimation, including smoothly clipped absolute deviation (SCAD) (Fan & Li, 2001) and minimax concave penalty (MCP) (Zhang, 2010), as well as possess enhanced sparse signal recovery, such as the log-sum penalty (Candes et al., 2008). In addition, improved generalization accuracy has been found using non-convex instead of convex loss functions (Shen et al., 2003), with better robustness to outliers and noisy sample data (Wu & Liu, 2007; Chapelle et al., 2009). Smooth non-convex loss functions exhibiting these beneficial qualities include the sigmoid loss, Lorenz loss (Barbu et al., 2017), and Savage loss (Masnadi-Shirazi & Vasconcelos, 2009).

The literature concerning first-order stochastic methods for regularized optimization is vast, so we restrict our attention to algorithms achieving non-asymptotic rates of convergence for a non-convex function $f(w)$. Stochastic gradient methods for the case of a convex regularizer has been an active research area where algorithms with non-asymptotic convergence results were first achieved in (Ghadimi et al., 2016). For finite-sum problems, Reddi et al. (2016) were the first to develop a proximal algorithm using the stochastic variance reduced gradient approach of Johnson & Zhang (2013). The current state-of-the-art for the finite-sum problem seems to be the work of Li & Li (2018) where one can also find a table of the convergence complexities of competing algorithms.

In the pursuit of solving (1) where neither function $f(w)$ nor $g(w)$ are convex, the current body of research is quite limited. A generalization of (Ghadimi et al., 2016) with $g(w)$ being quasi-convex can be found in (Kawashima & Fujisawa, 2018), where the same convergence complexity is achieved. The only other work for non-convex regularizers to our knowledge is that of Xu et al. (2018), which recently improved upon the stochastic difference of convex (DC) algorithm of Nitanda & Suzuki (2017), considering

[1]RIKEN Center for Advanced Intelligence Project, Tokyo, Japan [2]Department of Creative Informatics, Graduate School of Information Science and Technology, the University of Tokyo, Tokyo, Japan. Correspondence to: Michael R. Metel <michaelros.metel@riken.jp>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

an objective of the form $c^1(w) - c^2(w) + g(w)$ where $c^1(w) := \mathbb{E}_\xi[C^1(w,\xi)]$ and $c^2(w) := \mathbb{E}_\varsigma[C^2(w,\varsigma)]$ are convex functions. It is assumed that $c^1(w)$ has a Lipschitz continuous gradient and $c^2(w)$ has a Hölder continuous gradient, and the proximal mapping of $g(w)$ can be efficiently computed. In their algorithms, a sequence of subproblems must be solved with increasing accuracy using a first-order stochastic algorithm, where convergence to a nearly ϵ-critical point in a finite number of iterations is proved. The best convergence complexities in their work are achieved when it is assumed that $g(w)$ is Lipschitz continuous and $c^2(w)$ has a Lipschitz continuous gradient, which we will assume when discussing their work.

We now summarize the two main contributions of this paper:

- Two algorithms are presented, a mini-batch stochastic gradient algorithm for general stochastic objectives of the form (2), and a variance reduced stochastic gradient algorithm for finite-sum problems of the form (3). We are aware of only one other work, (Xu et al., 2018), which has proven non-asymptotic convergence for the class of problem we focus on in this paper. We attain superior convergence results under both objective assumptions, which are summarized in Table 1. The complexities are in terms of the number of gradient calls and proximal operations, see Section 2.
- No numerical experiments were conducted in (Xu et al., 2018). We implemented all algorithms for an application in empirical risk minimization and found the simplest algorithm to implement also performed the best in practice.

Remark: In a subsequent revision uploaded after submission of this work, Xu et al. (2019) present improved complexity results, as well as numerical experiments. The first row of Table 1 would be $O(\epsilon^{-5})$ and $O(\epsilon^{-5})$, and the second row would be $\tilde{O}(n\epsilon^{-3})$ and $\tilde{O}(\epsilon^{-3})$ following the latest version of their work.

2. Preliminaries

We assume that $f(w)$ has a Lipschitz continuous gradient with parameter L,

$$||\nabla f(w) - \nabla f(x)||_2 \leq L||w - x||_2,$$

which we will denote as being an L-smooth function. In the finite-sum case, we assume that each $f_j(w)$ is also L-smooth. Given a sample $\xi^k \sim P$, generated in iteration k of an algorithm, we assume we can generate an unbiased stochastic gradient $\nabla F(w,\xi^k)$ such that

$$\mathbb{E}[\nabla F(w,\xi^k)] = \nabla f(w), \tag{4}$$

and for some constant σ,

$$\mathbb{E}||\nabla F(w,\xi^k) - \nabla f(w)||_2^2 \leq \sigma^2. \tag{5}$$

Let $\partial h(w)$ denote the limiting subdifferential of our objective, defined as

$$\partial h(w) := \{v : \exists w^k \xrightarrow{h} w, v^k \in \hat{\partial} h(w^k) \text{ with } v^k \to v\},$$

where $\hat{\partial} h(w) := \{v : \liminf_{x \to w, x \neq w} \frac{h(x) - h(w) - \langle v, x-w\rangle}{||x-w||_2} \geq 0\}$, and $w^k \xrightarrow{h} w$ signifies $w^k \to w$ with $h(w^k) \to h(w)$. The limiting subdifferential coincides with the gradient and subdifferential when the function is continuously differentiable and proper convex, respectively. We make use of the property that

$$\partial h(w) = \nabla f(w) + \partial g(w), \tag{6}$$

for finite $g(w)$ (Rockafellar & Wets, 2009, Exercise 8.8 (c)). We also assume the proximal operator of $g(w)$ is nonempty for all $w \in \mathbb{R}^d$ and $\lambda > 0$, and can be efficiently computed,

$$\text{prox}_{\lambda g}(w) := \underset{x \in \mathbb{R}^d}{\text{argmin}} \left\{ \frac{1}{2\lambda}||w - x||_2^2 + g(x) \right\}.$$

In particular, let us denote an element as

$$\zeta^\lambda(w) \in \text{prox}_{\lambda g}(w). \tag{7}$$

We are interested in the convergence complexity of finding an ϵ-stationary solution, such that for an algorithm solution $\overline{w}$,

$$\mathbb{E}\left[\text{dist}(0, \partial h(\overline{w}))\right] \leq \epsilon. \tag{8}$$

We will measure algorithm complexity in terms of the number of gradient calls and proximal operations. For any w, a gradient call is either computing $\nabla F(w,\xi^k)$ given a sample ξ^k, or in the finite-sum case, returning $\nabla f_j(w)$ for a given j.

3. Auxiliary functions of $h(w)$

Our convergence results rely on bounding the gradient of a sequence of majorant functions of the auxiliary function

$$\tilde{h}_\lambda(w) := f(w) + e_\lambda g(w)$$

in expectation, where

$$e_\lambda g(w) := \inf_{x \in \mathbb{R}^d} \left\{ \frac{1}{2\lambda}||w - x||_2^2 + g(x) \right\}$$

is the Moreau envelope of $g(w)$. By considering $x = w$, we observe that

$$e_\lambda g(w) \leq g(w). \tag{9}$$

Table 1. Comparison of convergence complexities obtained in (Xu et al., 2018) and this paper.

Algorithm	Reference	Finite-sum Assumption	Gradient Call Complexity	Proximal Operator Complexity
SSDC-SPG	Theorem 7 a, Xu et al. (2018)	$\times$	$O(\epsilon^{-8})$	$O(\epsilon^{-8})$
SSDC-SVRG	Theorem 7 c, Xu et al. (2018)	$\surd$	$O(n\epsilon^{-4})$	$O(\epsilon^{-4})$
MBSGA	Corollary 6	$\times$	$O(\epsilon^{-5})$	$O(\epsilon^{-4})$
VRSGA	Corollary 9	$\surd$	$O(n^{2/3}\epsilon^{-3})$	$O(\epsilon^{-3})$

The Moreau envelope can be written as a DC function,

$$e_\lambda g(w) = \frac{1}{2\lambda}||w||_2^2 - D^\lambda(w), \tag{10}$$

where $D^\lambda(w) = \sup_{x\in\mathbb{R}^d}\left(\frac{1}{\lambda}w^T x - \frac{1}{2\lambda}||x||_2^2 - g(x)\right)$. We note that as the supremum of a set of affine functions, $D^\lambda(w)$ is convex, and we see from (7) that $\zeta^\lambda(w)$ attains the supremum of $D^\lambda(w)$. We can write down a smooth majorant of $\tilde{h}_\lambda(w)$ as

$$E_\lambda^k(w) := f(w) + U_\lambda^k(w)$$

in iteration k, where
$U_\lambda^k(w) = \frac{1}{2\lambda}||w||_2^2 - \left(D^\lambda(w^k) + \frac{1}{\lambda}\zeta^\lambda(w^k)^T(w-w^k)\right)$.
The gradient of $E_\lambda^k(w)$ is

$$\nabla E_\lambda^k(w) = \nabla f(w) + \frac{1}{\lambda}(w - \zeta^\lambda(w^k)). \tag{11}$$

Property 1. *The following holds for $E_\lambda^k(w)$.*

$$E_\lambda^k(w) \geq \tilde{h}_\lambda(w) \text{ for all } w \in \mathbb{R}^d \tag{12}$$
$$E_\lambda^k(w^k) = \tilde{h}_\lambda(w^k) \tag{13}$$
$$E_\lambda^k(w) \text{ is } L_{E\lambda} := \left(L + \frac{1}{\lambda}\right) - \text{smooth.} \tag{14}$$

Proof. Given that both functions contain $f(w)$, it is sufficient to show that (12) and (13) hold between the second terms $U_\lambda^k(w)$ and $e_\lambda g(w)$.

(12): As found in (Liu et al., 2017), for any $w, z \in \mathbb{R}^d$,

$$\begin{aligned}
&D^\lambda(w) - D^\lambda(z)\\
=&\sup_{x\in\mathbb{R}^d}\left(\frac{1}{\lambda}w^T x - \frac{1}{2\lambda}||x||^2 - g(x)\right)\\
&-\sup_{x\in\mathbb{R}^d}\left(\frac{1}{\lambda}z^T x - \frac{1}{2\lambda}||x||^2 - g(x)\right)\\
\geq&\frac{1}{\lambda}w^T\zeta^\lambda(z) - \frac{1}{2\lambda}||\zeta^\lambda(z)||^2 - g(\zeta^\lambda(z))\\
&-\left(\frac{1}{\lambda}z^T\zeta^\lambda(z) - \frac{1}{2\lambda}||\zeta^\lambda(z)||^2 - g(\zeta^\lambda(z))\right)\\
=&\frac{1}{\lambda}\zeta^\lambda(z)(w-z).
\end{aligned}$$

Setting $z = w^k$,

$$\begin{aligned}
e_\lambda g(w) &= \frac{1}{2\lambda}||w||^2 - D^\lambda(w)\\
&\leq \frac{1}{2\lambda}||w||^2 - (D^\lambda(w^k) + \frac{1}{\lambda}\zeta^\lambda(w^k)^T(w-w^k))\\
&= U_\lambda^k(w).
\end{aligned}$$

(13): $U_\lambda^k(w^k) = \frac{1}{2\lambda}||w^k||_2^2 - D^\lambda(w^k) = e_\lambda g(w^k)$ from (10).

(14):

$$\begin{aligned}
&\left\|\nabla E_\lambda^k(w) - \nabla E_\lambda^k(w')\right\|_2\\
=&\|\nabla f(w) + \frac{1}{\lambda}\left(w - \zeta^\lambda(w^k)\right)\\
&-\left(\nabla f(w') + \frac{1}{\lambda}\left(w' - \zeta^\lambda(w^k)\right)\right)\|_2\\
\leq&(L + \frac{1}{\lambda})\|w - w'\|_2.
\end{aligned}$$

□

We note that the Moreau envelope of a convex function is also $\frac{1}{\lambda}$-smooth (Beck, 2017, Theorem 6.60), so there is no increase in the smoothness parameter for non-convex functions by taking a first-order approximation of the Moreau envelope.

4. Mini-batch stochastic gradient algorithm

4.1. Convergence analysis

The convergence analysis of MBSGA follows the technique of Ghadimi & Lan (2013) adapted to our problem. The following lemma bounds $\mathbb{E}||\nabla E_\lambda^R(w^R)||_2^2$, with which we will ultimately bound $\mathbb{E}\left[\text{dist}(0, \partial h(\bar{w}^R)\right]$ in Theorem 5.

Lemma 2. *For an initial value $w_1 \in \mathbb{R}^d$, $N \in \mathbb{Z}_{>0}$, and $\alpha, \theta \in \mathbb{R}$, MBSGA generates w^R satisfying the following bound.*

$$\mathbb{E}||\nabla E_\lambda^R(w^R)||_2^2 \leq \frac{\tilde{\Delta}}{N}(L + N^\theta) + \frac{\sigma}{\sqrt{N}}\left(\tilde{\Delta} + \frac{L + N^\theta}{\lceil N^\alpha \rceil}\right),$$

Algorithm 1 Mini-batch stochastic gradient algorithm (MBSGA)

Input: $w^1 \in \mathbb{R}^d$, $N \in \mathbb{Z}_{>0}$, $\alpha, \theta \in \mathbb{R}$
$M := \lceil N^\alpha \rceil$, $\lambda = \frac{1}{N^\theta}$
$L_{E\lambda} = L + \frac{1}{\lambda}$
$\gamma = \min\left\{\frac{1}{L_{E\lambda}}, \frac{1}{\sigma\sqrt{N}}\right\}$
$R \sim$ uniform$\{1, ..., N\}$
for $k = 1, 2, ..., R-1$ **do**
  $\zeta^\lambda(w^k) \in \text{prox}_{\lambda g}(w^k)$
  Sample $\xi^k \sim P^M$
  $\nabla A^k_{\lambda M}(w^k, \xi^k) = \frac{1}{M}\sum_{j=1}^M \nabla F(w^k, \xi^k_j) + \frac{1}{\lambda}(w^k - \zeta^\lambda(w^k))$
  $w^{k+1} = w^k - \gamma \nabla A^k_{\lambda M}(w^k, \xi^k)$
end for
Output: $\bar{w}^R \in \text{prox}_{\lambda g}(w^R)$

where $\tilde{\Delta} = 2(\tilde{h}_\lambda(w^1) - \tilde{h}_\lambda(w^*_\lambda))$ *and* w^*_λ *is a global minimizer of* $\tilde{h}_\lambda(\cdot)$.

Due to a lack of space, the proof of Lemma 2 can be found in Section 1 of the supplementary material. In order to prove the convergence of $\mathbb{E}\left[\text{dist}(0, \partial h(\bar{w}^R)\right]$, we will require the following two properties.

Property 3. *Assume that* $g(w)$ *is Lipschitz continuous with parameter* l*,*

$$\text{dist}(0, \partial h(\zeta^\lambda(w^k))) \leq ||\nabla E^k_\lambda(w^k)||_2 + 2l\lambda L.$$

Proof. Given that $\zeta^\lambda(w)$ is a minimizer of $\frac{1}{2\lambda}||w - x||_2^2 + g(x)$ from (7),

$$\frac{1}{\lambda}(w - \zeta^\lambda(w)) \in \partial g(\zeta^\lambda(w))$$

and

$$\nabla f(\zeta^\lambda(w^k)) + \frac{1}{\lambda}(w^k - \zeta^\lambda(w^k)) \in \partial h(\zeta^\lambda(w^k))$$

using (6). It follows that

$$\begin{aligned}
&\text{dist}(0, \partial h(\zeta^\lambda(w^k))) \\
\leq &||\nabla f(\zeta^\lambda(w^k)) + \frac{1}{\lambda}(w^k - \zeta^\lambda(w^k))||_2 \\
= &||\nabla f(w^k) - \nabla f(w^k) + \nabla f(\zeta^\lambda(w^k)) \\
&+ \frac{1}{\lambda}(w^k - \zeta^\lambda(w^k))||_2 \\
\leq &||\nabla f(w^k) + \frac{1}{\lambda}(w^k - \zeta^\lambda(w^k))||_2 \\
&+ ||\nabla f(\zeta^\lambda(w^k)) - \nabla f(w^k)||_2 \\
\leq &||\nabla E^k_\lambda(w^k)||_2 + L||w^k - \zeta^\lambda(w^k)||_2.
\end{aligned}$$

In order to bound $||w^k - \zeta^\lambda(w^k)||_2$, recall from (9) that

$$\begin{aligned}
g(w) &\geq e_\lambda g(w) \\
&= \frac{1}{2\lambda}||w - \zeta^\lambda(w)||_2^2 + g(\zeta^\lambda(w)).
\end{aligned}$$

Rearranging and using the Lipschitz continuity assumption,

$$\begin{aligned}
\frac{1}{2\lambda}||w - \zeta^\lambda(w)||_2^2 &\leq g(w) - g(\zeta^\lambda(w)) \\
&\leq l||w - \zeta^\lambda(w)||_2 \\
||w - \zeta^\lambda(w)||_2 &\leq 2l\lambda.
\end{aligned}$$

□

Property 4. *Let* w^* *be a global minimizer of* $h(\cdot)$ *and let* w^*_λ *be a global minimizer of* $\tilde{h}_\lambda(\cdot)$*. Assume that* $g(w)$ *is Lipschitz continuous with parameter* l*, then*

$$\tilde{h}_\lambda(w) - \tilde{h}_\lambda(w^*_\lambda) \leq h(w) - h(w^*) + \frac{l^2\lambda}{2}.$$

Proof.

$$\begin{aligned}
&\tilde{h}_\lambda(w) - \tilde{h}_\lambda(w^*_\lambda) - (h(w) - h(w^*)) \\
=&e_\lambda g(w) - f(w^*_\lambda) - e_\lambda g(w^*_\lambda) \\
&- (g(w) - f(w^*) - g(w^*)) \\
\leq &- f(w^*_\lambda) - e_\lambda g(w^*_\lambda) + f(w^*) + g(w^*) \\
\leq &- f(w^*_\lambda) - e_\lambda g(w^*_\lambda) + f(w^*_\lambda) + g(w^*_\lambda) \\
=&g(w^*_\lambda) - e_\lambda g(w^*_\lambda),
\end{aligned}$$

where the first inequality follows from (9). For any w, by the definition of the Moreau envelope,

$$\begin{aligned}
e_\lambda g(w) &= \frac{1}{2\lambda}||w - \zeta^\lambda(w)||_2^2 + g(\zeta^\lambda(w)) \\
g(w) - e_\lambda g(w) &= g(w) - g(\zeta^\lambda(w)) - \frac{1}{2\lambda}||w - \zeta^\lambda(w)||_2^2 \\
&\leq l||w - \zeta^\lambda(w)||_2 - \frac{1}{2\lambda}||w - \zeta^\lambda(w)||_2^2.
\end{aligned}$$

The right-hand side is maximized when $||w - \zeta^\lambda(w)||_2 = l\lambda$, giving the desired result,

$$g(w) - e_\lambda g(w) \leq \frac{l^2\lambda}{2}. \tag{15}$$

□

We note that (15) cannot be improved under the further assumption that $g(w)$ is convex, which can be found in (Beck, 2017, Theorem 10.51).

Theorem 5. *Assume that $g(w)$ is Lipschitz continuous with parameter l. The output $\bar{w}^R$ of MBSGA satisfies*

$$\mathbb{E}\left[\text{dist}(0, \partial h(\bar{w}^R))\right] \leq \sqrt{\frac{(\Delta + l^2 N^{-\theta})(L + N^\theta)}{N}} + \sqrt{\frac{\sigma}{\sqrt{N}}\left(\Delta + \frac{l^2}{N^\theta} + \frac{L + N^\theta}{\lceil N^\alpha \rceil}\right)} + \frac{2lL}{N^\theta},$$

where $\Delta = 2(h(w^1) - h(w^*))$ and w^* is a global minimizer of $h(\cdot)$.

Proof. From Property 3, choosing $\zeta^\lambda(w^R) = \bar{w}^R$,

$$\text{dist}(0, \partial h(\bar{w}^R)) \leq ||\nabla E_\lambda^R(w^R)||_2 + 2l\lambda L.$$

Taking its expectation,

$$\begin{aligned}
&\mathbb{E}\left[\text{dist}(0, \partial h(\bar{w}^R))\right] \\
\leq &\mathbb{E}[||\nabla E_\lambda^R(w^R)||_2] + 2l\lambda L \\
\leq &\sqrt{\mathbb{E}\left[||\nabla E_\lambda^R(w^R)||_2^2\right]} + \frac{2lL}{N^\theta} \\
\leq &\sqrt{\frac{\tilde{\Delta}(L + N^\theta)}{N}} + \sqrt{\frac{\sigma}{\sqrt{N}}\left(\tilde{\Delta} + \frac{L + N^\theta}{\lceil N^\alpha \rceil}\right)} + \frac{2lL}{N^\theta},
\end{aligned}$$

where the second inequality follows from Jensen's inequality and the third inequality uses Lemma 2. The result then follows using Property 4 as

$$\begin{aligned}
\tilde{\Delta} = 2(\tilde{h}_\lambda(w^1) - \tilde{h}_\lambda(w_\lambda^*)) &\leq 2(h(w^1) - h(w^*)) + l^2\lambda \\
&= \Delta + \frac{l^2}{N^\theta}
\end{aligned}$$

□

Now that we have bounded the expected distance of $\partial h(\bar{w}^R)$ from the origin, we prove an ϵ-stationary point convergence complexity.

Corollary 6. *Assume that $g(w)$ is Lipschitz continuous with parameter l. To obtain an ϵ-stationary solution* (8) *using MBSGA, the gradient call complexity is $O(\epsilon^{-5})$ and the proximal operator complexity is $O(\epsilon^{-4})$ when $\alpha = \theta = 0.25$.*

Proof. From Theorem 5,

$$\begin{aligned}
&\mathbb{E}\left[\text{dist}(0, \partial h(\bar{w}^R))\right] \\
\leq &\sqrt{\frac{(\Delta + l^2 N^{-\theta})(L + N^\theta)}{N}} \\
&+ \sqrt{\frac{\sigma}{\sqrt{N}}\left(\Delta + \frac{l^2}{N^\theta} + \frac{L + N^\theta}{\lceil N^\alpha \rceil}\right)} + \frac{2lL}{N^\theta} \\
= &O(N^{0.5\theta - 0.5}) + O(N^{-0.25} + N^{0.5\theta - 0.5\alpha - 0.25}) \\
&+ O(N^{-\theta}).
\end{aligned}$$

Algorithm 2 Variance reduced stochastic gradient algorithm (VRSGA)

Input: $\tilde{w}^1 \in \mathbb{R}^d$, $N \in \mathbb{Z}_{>0}$, $\alpha, \theta \in \mathbb{R}$
$m = \lceil n^\alpha \rceil$, $b = m^2$
$S = \lceil \frac{N}{m} \rceil$, $\lambda = (Sm)^{-\theta}$
$L_{E\lambda} = L + \frac{1}{\lambda}$, $\gamma = \frac{1}{6L_{E\lambda}}$
$R \sim$ uniform$\{1, ..., S\}$
for $k = 1, 2, ..., R$ **do**
  $w_1^k = \tilde{w}^k$
  $G^k = \nabla f(\tilde{w}^k)$
  for $t = 1, 2, ..., m$ **do**
    $\zeta^\lambda(w_t^k) \in \text{prox}_{\lambda g}(w_t^k)$
    $I \sim$ uniform$\{1, ..., n\}^b$
    $V_t^k = \frac{1}{b}\sum_{j \in I}\left(\nabla f_j(w_t^k) - \nabla f_j(\tilde{w}^k)\right) + G^k + \frac{1}{\lambda}(w_t^k - \zeta^\lambda(w_t^k))$
    $w_{t+1}^k = w_t^k - \gamma V_t^k$
  end for
  $\tilde{w}^{k+1} = w_{m+1}^k$
end for
$T \sim$ uniform$\{1, ..., m\}$
Output: $\bar{w}_T^R \in \text{prox}_{\lambda g}(w_T^R)$

Setting $\theta = \alpha = 0.25$,

$$\mathbb{E}\left[\text{dist}(0, \partial h(\bar{w}^R))\right] \leq O(N^{-0.25}).$$

An ϵ-stationary solution will require less than $N = O(\epsilon^{-4})$ iterations. One proximal operation is done per iteration, which establishes the proximal operator complexity of $O(\epsilon^{-4})$. The number of gradient calls per iteration is $\lceil N^\alpha \rceil = O(\epsilon^{-1})$. The number of gradient calls to get an ϵ-stationary solution is then less than

$$N\lceil N^\alpha \rceil = O(\epsilon^{-5}).$$

□

5. Variance reduced method for finite-sum problems

In this section we assume that

$$f(w) = \frac{1}{n}\sum_{j=1}^n f_j(w),$$

where each $f_j(w)$ is L-smooth.

5.1. Convergence analysis

In our convergence analysis, we make use of the function $E_{t\lambda}^k(w)$, which is constructed in the same manner as $E_\lambda^k(w)$, using w_t^k instead of w^k. This function possesses the same characteristics as found in Property 1. The convergence analysis follows closely to the work of Li & Li (2018) adapted to our problem.

Lemma 7. *For an initial value $\tilde{w}_1 \in \mathbb{R}^d$, $N \in \mathbb{Z}_{>0}$, and $\alpha, \theta \in \mathbb{R}$, VRSGA generates w_T^R satisfying the following bound.*

$$\mathbb{E}\left[||\nabla E_{T\lambda}^R(w_T^R)||_2^2\right] \leq \tilde{\Delta}\frac{L+(Sm)^\theta}{Sm},$$

where $\tilde{\Delta} = 36(\tilde{h}_\lambda(\tilde{w}^1) - \tilde{h}_\lambda(w_\lambda^*))$ and w_λ^* is a global minimizer of $\tilde{h}_\lambda(\cdot)$.

The proof of Lemma 7 can be found in Section 2 of the supplementary material.

Theorem 8. *Assume that $g(w)$ is Lipschitz continuous with parameter l. The output $\bar{w}_T^R$ of VRSGA satisfies*

$$\begin{aligned}&\mathbb{E}\left[||\operatorname{dist}(0, \partial h(\bar{w}_T^R))||_2\right]\\ \leq&\sqrt{\frac{(L+(Sm)^\theta)\left(\Delta + 18l^2(Sm)^{-\theta}\right)}{Sm}} + \frac{2lL}{(Sm)^\theta},\end{aligned}$$

where $\Delta = 36(h(w^1) - h(w^*))$ and w^* is a global minimizer of $h(\cdot)$.

Proof. The proof follows what was done to prove Theorem 5. From Property 3,

$$\operatorname{dist}(0, \partial h(\bar{w}_T^R)) \leq ||\nabla E_{T\lambda}^R(w_T^R)||_2 + 2l\lambda L.$$

Taking its expectation,

$$\begin{aligned}&\mathbb{E}\left[||\operatorname{dist}(0, \partial h(\bar{w}_T^R))||_2\right]\\ \leq&\mathbb{E}[||\nabla E_{T\lambda}^R(w_T^R)||_2] + 2l\lambda L\\ \leq&\sqrt{\mathbb{E}\left[||\nabla E_{T\lambda}^R(w_T^R)||_2^2\right]} + \frac{2lL}{(Sm)^\theta}\\ \leq&\sqrt{\frac{(L+(Sm)^\theta)\,\tilde{\Delta}}{Sm}} + \frac{2lL}{(Sm)^\theta}\\ \leq&\sqrt{\frac{(L+(Sm)^\theta)\left(\Delta + 18l^2(Sm)^\theta\right)}{Sm}} + \frac{2lL}{(Sm)^\theta},\end{aligned}$$

where the third inequality follows from Lemma 7. The fourth inequality holds using Property 4,

$$\begin{aligned}\tilde{\Delta} = 36(\tilde{h}_\lambda(w^1) - \tilde{h}_\lambda(w_\lambda^*)) &\leq 36(h(w^1) - h(w^*)) + 18l^2\lambda\\ &= \Delta + \frac{18l^2}{(Sm)^\theta}.\end{aligned}$$

□

Corollary 9. *Assume that $g(w)$ is Lipschitz continuous with parameter l. To obtain an ϵ-stationary solution* (8) *using VRSGA, the gradient call complexity is $O(n^{\frac{2}{3}}\epsilon^{-3})$ and the proximal operator complexity is $O(\epsilon^{-3})$ choosing $\alpha = \theta = \frac{1}{3}$.*

Proof. From Theorem 8 with $\theta = \frac{1}{3}$,

$$\begin{aligned}&\mathbb{E}\left[||\operatorname{dist}(0, \partial h(\bar{w}_T^R))||_2\right]\\ \leq&\sqrt{\frac{\left(L+(Sm)^{\frac{1}{3}}\right)\left(\Delta + 18l^2(Sm)^{\frac{-1}{3}}\right)}{Sm}} + \frac{2lL}{(Sm)^{\frac{1}{3}}}\\ =&\, O((Sm)^{-\frac{1}{3}})\end{aligned}$$

An ϵ-stationary solution will require at most $Sm = O(\epsilon^{-3})$ iterations, which establishes the proximal operator complexity. The number of gradient calls after Sm iterations, taking $\alpha = \frac{1}{3}$ is

$$Sn + Smb = Sm\frac{n}{\lceil n^{\frac{1}{3}}\rceil} + Sm\lceil n^{\frac{1}{3}}\rceil^2 = O(n^{\frac{2}{3}}\epsilon^{-3}).$$

□

6. Application

In this section we consider the application of binary classification for a particular choice of loss function and regularizer, which will be used in our numerical experiments. Non-convex Lipschitz continuous regularizers which have proximal operators with closed form solutions include the log-sum penalty, SCAD, MCP, and the capped l_1-norm. For their closed form solutions, see (Gong et al., 2013). All of these functions are separable, $g(w) := \sum_{i=1}^d g_i(w_i)$. For $\kappa, \nu > 0$, the log-sum penalty is

$$g_i(w_i) = \kappa \log(1 + |w_i|/\nu).$$

Property 10. *The log-sum penalty is $\frac{\kappa}{\nu}\sqrt{d}$-Lipschitz continuous.*

Proof. Assume $w_i \geq 0$ over which $g_i(w_i)$ is differentiable and $|\frac{dg_i}{dw_i}(w_i)| \leq \frac{\kappa}{\nu}$. Using the mean value theorem with $z_i \geq 0$, $|g_i(z_i) - g_i(w_i)| \leq \frac{\kappa}{\nu}|z_i - w_i|$. Given the symmetry of $g_i(w_i)$, this bound holds for general w_i and z_i. It then follows that for any w and z,

$$\begin{aligned}|g(z) - g(w)| &= \left|\sum_{i=1}^d (g_i(z_i) - g_i(w_i))\right|\\ &\leq \sum_{i=1}^d |g_i(z_i) - g_i(w_i)|\\ &\leq \frac{\kappa}{\nu}\sum_{i=1}^d |z_i - w_i|\\ &\leq \frac{\kappa}{\nu}\sqrt{d}||z - w||_2\end{aligned}$$

□

Smooth non-convex loss functions, which are known to be robust to outliers, include the sigmoid loss, $\frac{1}{1+e^v}$, Lorenz loss (Barbu et al., 2017), Savage loss (Masnadi-Shirazi & Vasconcelos, 2009), and the tangent loss (Masnadi-Shirazi et al., 2010). We will consider the Lorenz loss,

$$\mathcal{L}(v) = \begin{cases} 0 & \text{if } v > 1 \\ \log(1 + (v-1)^2) & \text{otherwise} \end{cases}$$

for $v \in \mathbb{R}$, which is differentiable everywhere (Barbu et al., 2017). For the problem setting of binary classification, we have a set of training data $\{x, y\}$ where $y = \{y^1, y^2, ..., y^n\}$, $y^j \in \{-1, 1\}$, is the label set, and $x = \{x^1, x^2, ..., x^n\}$, $x^j \in \mathbb{R}^d$, is the feature set. Our loss function is then

$$f(w) = \frac{1}{n}\sum_{j=1}^{n} f_j(w),$$

where

$$f_j(w) = \mathcal{L}(y^j w^T x^j).$$

Property 11. *Using the Lorenz loss function, $f(w)$ is $\frac{2}{n}\sum_{j=1}^{n} ||x^j||_2^2$-smooth.*

Proof. We first consider the function

$$\hat{\mathcal{L}}(v) = \log(1 + (v-1)^2).$$

Its first and second derivatives are

$$\hat{\mathcal{L}}'(v) = \frac{2(v-1)}{1 + (v-1)^2}$$

and

$$\hat{\mathcal{L}}''(v) = \frac{2}{1 + (v-1)^2} - \left(\frac{2(v-1)}{1 + (v-1)^2}\right)^2.$$

We can see that $v = 1$ maximizes $\hat{\mathcal{L}}''(v)$, with $\hat{\mathcal{L}}''(1) = 2$. Examining the third derivative,

$$\hat{\mathcal{L}}'''(v) = \frac{-4(v-1)}{(1 + (v-1)^2)^2}\left(3 - \frac{4(v-1)^2}{1 + (v-1)^2}\right),$$

$v = 1 \pm \sqrt{3}$ minimizes $\hat{\mathcal{L}}''(v)$ with $\hat{\mathcal{L}}''(1 \pm \sqrt{3}) = -0.25$, so we conclude that

$$|\hat{\mathcal{L}}''(v)| \leq |\hat{\mathcal{L}}''(1)| = 2.$$

Using the mean value theorem, for any v and u,

$$|\hat{\mathcal{L}}'(v) - \hat{\mathcal{L}}'(u)| \leq 2|v - u|.$$

We now show that $\mathcal{L}(v)$ is also 2-smooth. For $v > 1$, $\mathcal{L}'(v) = \hat{\mathcal{L}}'(1) = 0$. Taking $v > 1$ and $u \leq 1$,

$$\begin{aligned} |\mathcal{L}'(v) - \mathcal{L}'(u)| &= |\hat{\mathcal{L}}'(1) - \hat{\mathcal{L}}'(u)| \\ &\leq 2|1 - u| \\ &\leq 2|v - u|. \end{aligned}$$

An L-smooth function composed with the linear function, $y^j w^T x^j$, is $L||y^j x^j||_2^2$-smooth (Shalev-Shwartz & Ben-David, 2014, Claim 12.9), so $f_j(w)$ is $2||x^j||_2^2$-smooth and the result follows. □

We also note that the Lorenz loss function is DC-decomposable, which is required to implement the algorithms of (Xu et al., 2018).

Property 12. *The Lorenz loss function is DC-decomposable,*

$$\mathcal{L}(v) = \mathcal{L}^1(v) - \mathcal{L}^2(v),$$

where $\mathcal{L}^1(v) = \frac{1}{8}v^2 + \mathcal{L}(v)$ and $\mathcal{L}^2(v) = \frac{1}{8}v^2$.

Proof. Since $\mathcal{L}''(v) \geq -\frac{1}{4}$, from the proof of Property 11, we write the DC decomposition of $\mathcal{L}(v)$ as $\mathcal{L}(v) = \mathcal{L}^1(v) - \mathcal{L}^2(v)$, where $\mathcal{L}^1(v) = \frac{1}{8}v^2 + \mathcal{L}(v)$ and $\mathcal{L}^2(v) = \frac{1}{8}v^2$. □

7. Numerical experiments

We conducted experiments comparing our algorithms to those of (Xu et al., 2018) for the problem of binary classification as described in Section 6, on datasets a9a (Fan, 2018) and MNIST (LeCun, 1998), as used in (Reddi et al., 2016; Allen-Zhu & Hazan, 2016; Li & Li, 2018). For the MNIST dataset, our objective was to learn class 1. The dimensions of a9a are $n = 32,561$ and $d = 123$, and those of MNIST are $n = 60,000$ and $d = 784$. All experiments were conducted using MATLAB 2017b on a Mac Pro with a 2.7 GHz 12-core Intel Xeon E5 processor and 64GB of RAM. We compare performance in terms of the log of the objective function and wall-clock time.

All algorithms' convergence rates rely on outputting a random iteration. In order to fairly compare algorithms we ignore this step, e.g. for MBSGA, we set $R = N$. The algorithms were initially run taking $e = 15$ effective passes over the data for a9a and $e = 9$ for MNIST. These values were adjusted so that all algorithms ended at approximately the same time. The regularizer's parameters were chosen as $\kappa = \frac{1}{d}$ and $\nu = 1$. All parameter values used in MBSGA and VRSGA were obtained from the theoretical convergence results, except for the upper bound σ (5) used in MBSGA. This parameter was estimated by doing 50 iterations of MBSGA with step size $\gamma = \frac{1}{L_{E\lambda}}$, using a different random seed than was used for the experiments, and computing the sample estimate $\hat{\sigma}^k$ each iteration with the M samples used in the algorithm. An estimate of σ was then taken as $\hat{\sigma} = \max_k \hat{\sigma}^k$.

The proof of convergence of algorithms VRSGA and SSDC-SVRG rely on the assumption that each $f_j(w)$ is L-smooth, so for these instances $L = 2\max_j ||x^j||_2^2$. For algorithms MBSGA and VRSGA, the final proximal operation at the

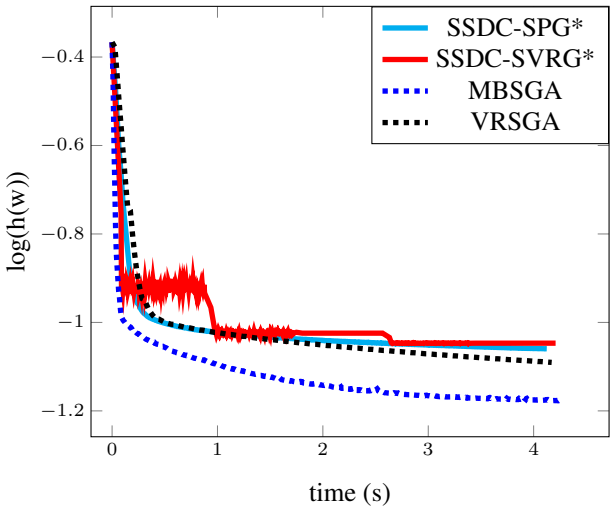

Figure 1. Comparison of algorithms of this paper and (Xu et al., 2018) (marked with *) using the a9a dataset

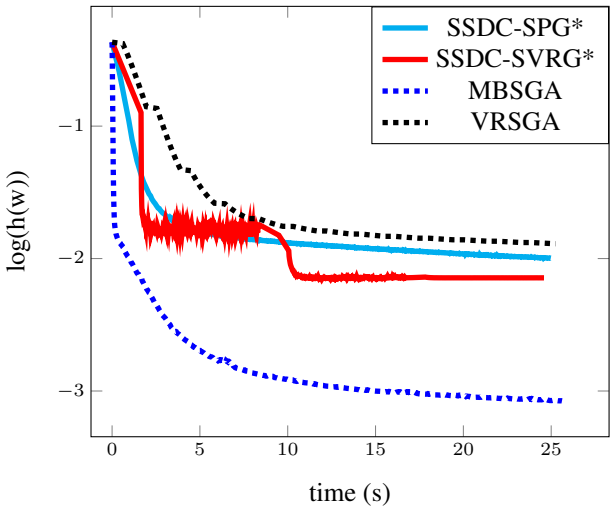

Figure 2. Comparison of algorithms of this paper and (Xu et al., 2018) (marked with *) using the MNIST dataset

output was omitted and can be considered as simply a means of proving the non-asymptotic convergence of the algorithms.

No experiments were done in (Xu et al., 2018), so we implemented their algorithms following the parameter values found in their theoretical results and remarks, and recommended in (Xiao & Zhang, 2014), from which their work is partially based on. Full details of their algorithms' implementation can be found in Section 3 of the supplementary material.

Figures 1 and 2 show the results of the experiments. We observe that MBSGA outperformed all other algorithms. MBSGA is also the simplest algorithm to implement, making it an appealing choice for use in practice. It appears all other algorithms would require further parameter tuning in order for them to possibly perform comparably.

8. Conclusion and future research

We have presented two simple stochastic gradient algorithms for optimizing a smooth non-convex loss function with a non-smooth non-convex regularizer. Our work improves upon the only other known non-asymptotic convergence results of Xu et al. (2018) for this class of problem. Superior convergence complexities were shown for the case of a general stochastic loss function using a mini-batch stochastic gradient algorithm, and for the case of a finite-sum loss function using a variance reduced stochastic gradient algorithm. In an empirical study we found that the simplest algorithm to implement was also the best performing, making it the most appealing algorithm considered for this problem setting. Future research using the techniques developed in this work could consider additional regularizers in the objective to induce desirable properties of the solution in addition to sparsity.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers 15K00031 and 19H04069, and supported by JST CREST Grant Numbers JPMJCR15K5 and JPMJCR14D2.

References

Allen-Zhu, Z. and Hazan, E. Variance reduction for faster non-convex optimization. In *International Conference on Machine Learning*, pp. 699–707, 2016.

Barbu, A., She, Y., Ding, L., and Gramajo, G. Feature selection with annealing for computer vision and big data learning. *IEEE transactions on pattern analysis and machine intelligence*, 39(2):272–286, 2017.

Beck, A. First-order methods in optimization, volume 25 of MOS-SIAM Series on Optimization. *Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA*, 2017.

Candes, E. J., Wakin, M. B., and Boyd, S. P. Enhancing sparsity by reweighted l_1 minimization. *Journal of Fourier analysis and applications*, 14(5-6):877–905, 2008.

Chapelle, O., Do, C. B., Teo, C. H., Le, Q. V., and Smola, A. J. Tighter bounds for structured estimation. In

Advances in neural information processing systems, pp. 281–288, 2009.

Fan, J. and Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456): 1348–1360, 2001.

Fan, R.-E. a9a dataset. *https://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/,* Access date: December 23, 2018.

Ghadimi, S. and Lan, G. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Ghadimi, S., Lan, G., and Zhang, H. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1-2):267–305, 2016.

Gong, P., Zhang, C., Lu, Z., Huang, J., and Ye, J. A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. In *International Conference on Machine Learning*, pp. 37–45, 2013.

Johnson, R. and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pp. 315–323, 2013.

Kawashima, T. and Fujisawa, H. Stochastic Gradient Descent for Stochastic Doubly-Nonconvex Composite Optimization. *arXiv preprint arXiv:1805.07960*, 2018.

LeCun, Y. The MNIST database of handwritten digits. *http://yann.lecun.com/exdb/mnist/*, 1998.

Li, Z. and Li, J. A Simple Proximal Stochastic Gradient Method for Nonsmooth Nonconvex Optimization. *arXiv preprint arXiv:1802.04477*, 2018.

Liu, T., Pong, T. K., and Takeda, A. A successive difference-of-convex approximation method for a class of nonconvex nonsmooth optimization problems. *arXiv preprint arXiv:1710.05778*, 2017.

Masnadi-Shirazi, H. and Vasconcelos, N. On the design of loss functions for classification: theory, robustness to outliers, and savageboost. In *Advances in Neural Information Processing Systems*, pp. 1049–1056, 2009.

Masnadi-Shirazi, H., Mahadevan, V., and Vasconcelos, N. On the design of robust classifiers for computer vision. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 779–786. IEEE, 2010.

Nitanda, A. and Suzuki, T. Stochastic difference of convex algorithm and its application to training deep boltzmann machines. In *Artificial Intelligence and Statistics*, pp. 470–478, 2017.

Reddi, S. J., Sra, S., Póczos, B., and Smola, A. J. Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization. In *Advances in Neural Information Processing Systems*, pp. 1145–1153, 2016.

Rockafellar, R. T. and Wets, R. J.-B. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Shen, X., Tseng, G. C., Zhang, X., and Wong, W. H. On ψ-learning. *Journal of the American Statistical Association*, 98(463):724–734, 2003.

Wu, Y. and Liu, Y. Robust truncated hinge loss support vector machines. *Journal of the American Statistical Association*, 102(479):974–983, 2007.

Xiao, L. and Zhang, T. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

Xu, Y., Qi, Q., Lin, Q., Jin, R., and Yang, T. Stochastic optimization for DC functions and non-smooth non-convex regularizers with non-asymptotic convergence. *arXiv preprint arXiv:1811.11829v1,* Access date: December 3, 2018.

Xu, Y., Qi, Q., Lin, Q., Jin, R., and Yang, T. Stochastic optimization for dc functions and non-smooth non-convex regularizers with non-asymptotic convergence. *arXiv preprint arXiv:1811.11829v2*, 2019.

Zhang, C.-H. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38 (2):894–942, 2010.