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

Michael R. Metel\(^1\) and Akiko Takeda\(^2\)

\(^1\)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; RIKEN Center for Advanced Intelligence Project, Tokyo, Japan

January 25, 2019

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 very 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 demonstrate our algorithms’ better performance in practice for empirical risk minimization on well known datasets.

Keywords: optimization, regularization, stochastic gradient, non-convex, non-smooth, variance reduction

1 Introduction

In this work we consider regularized optimization problems of the form

\[
\min_{w \in \mathbb{R}^d} h(w) := f(w) + g(w),
\]

(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

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

(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 $\xi$, $\xi_j$ for $j = 1, \ldots, n$:

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

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 and Li, 2001) and minimax concave penalty (MCP) (Zhang et al., 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 (Chapelle et al., 2009; Wu and Liu, 2007). Smooth non-convex loss functions exhibiting these beneficial qualities include the sigmoid loss, Lorenz loss (Barbu et al., 2017), and Savage loss (Masnadi-Shirazi and 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 and Lan, 2016; 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 and Zhang (2013). The current state of the art for the finite-sum problem seems to be the work of Li and Li (2018) where one can also find a table of the convergence complexities of competing algorithms.

In the pursuit of solving (11) 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 and 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 and Suzuki (2017), considering 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 $\epsilon$-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 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, 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 our algorithms to outperform in practice, finding better solutions in significantly less time.

- We highlight that our methods do not require the use of a DC algorithm, which require solving a sequence of approximating functions. This results in a more direct solution approach with simpler algorithms to implement, which require fewer iterations and significantly less computation time to obtain a desired solution.

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

| Algorithm   | Reference                  | Finite-sum | Complexity                  |
|-------------|----------------------------|------------|-----------------------------|
| SSDC-SPG    | Xu et al. (2018, Theorem 7 a.) | ×          | $O(\epsilon^{-8})$         |
| SSDC-SVRG   | Xu et al. (2018, Theorem 7 c.) | √          | $O(n\epsilon^{-4})$         |
| MBSGA       | Corollary 7                | ×          | $O(\epsilon^{-5})$         |
| VRSGA       | Corollary 12               | √          | $O(n^{2/3}\epsilon^{-3})$   |

2 Preliminaries

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

$$||\nabla f(w) - \nabla f(x)||_2 \leq L||w - x||_2,$$  \hspace{1cm} (4)

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

$$\mathbb{E}[\nabla F(w, \xi^k)] = \nabla f(w),$$  \hspace{1cm} (5)

and for some constant $\sigma$,

$$\mathbb{E}||\nabla F(w, \xi^k) - \nabla f(w)||_2^2 \leq \sigma^2.$$  \hspace{1cm} (6)
Let $\partial h(w)$ denote the Clarke generalized gradient (Clarke, 2013, Ch. 10) of our objective, which is always a nonempty convex compact set for locally Lipschitz functions, and coincides with the gradient and subdifferential when the function is continuously differentiable and convex, respectively. With the generalized directional derivative,

$$h^\circ(w,v) := \lim_{z \to w, t \downarrow 0} \inf \frac{h(z + tv) - h(z)}{t},$$

the Clarke generalized gradient is the set

$$\partial h(w) := \{ \phi : h^\circ(w,v) \geq \phi^T v, \forall v \in \text{dom}(h) \}. $$

The Clarke generalized gradient has previously been used to analyze non-smooth, non-convex stochastic optimization algorithms, see for example (Majewski et al., 2018). We also assume the proximal operator of $g(w)$ is nonempty for all $w$, and can be computed in closed form,

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

and in particular, let us denote an element as

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

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

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

We will measure algorithm complexity in terms of the number of gradient calls. For any $w$, this is either computing $\nabla F(w, \xi^k)$ given a sample $\xi^k$, 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 two stages. We first prove the convergence of the following function, $\tilde{h}_{\lambda}(w)$, and then bound the convergence of $h(w)$ in terms of $\tilde{h}_{\lambda}(w)$ and $\lambda$. Proving the attainment of an $\epsilon$-stationary solution of $\tilde{h}_{\lambda}(w)$ relies on a smooth majorant function, which we will describe in this section.

$$\tilde{h}_{\lambda}(w) := f(w) + e_{\lambda g}(w), \quad (9)$$

where

$$e_{\lambda g}(w) := \inf_{x \in \mathbb{R}^d} \left\{ \frac{1}{2\lambda} ||w - x||^2 + g(x) \right\} \quad (10)$$

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

$$e_{\lambda g}(w) \leq g(w). \quad (11)$$
The Moreau envelope can be written as a DC function,

\[ e_{\lambda}g(w) = \frac{1}{2\lambda}||w||^2 - D^{\lambda}(w), \]  

(12)

where \( D^{\lambda}(w) = \sup_{x \in \mathbb{R}^d} \left( \frac{1}{\lambda} w^T x - \frac{1}{2\lambda}||x||^2 - g(x) \right) \).

We note that as the supremum of a set of affine functions, \( D^{\lambda}(w) \) is convex, and we see from (7), \( \zeta^{\lambda}(w) \) attains the supremum of \( D^{\lambda}(w) \). We can write down a smooth majorant of \( \tilde{h}_{\lambda}(w) \) as

\[ E^k_{\lambda}(w) := f(w) + U^k_{\lambda}(w) \]

in iteration \( k \), where

\[ U^k_{\lambda}(w) = \frac{1}{2\lambda}||w||^2 - (D^{\lambda}(w^k) + \frac{1}{\lambda}\zeta^{\lambda}(w^k)^T(w - w^k)). \]

The gradient of \( E^k_{\lambda}(w) \) is

\[ \nabla E^k_{\lambda}(w) = \nabla f(w) + \frac{1}{\lambda}(w - \zeta^{\lambda}(w^k)). \]  

(13)

**Property 1.** The following holds for \( E^k_{\lambda}(w) \).

\[ E^k_{\lambda}(w) \geq \tilde{h}_{\lambda}(w) \text{ for all } w \in \mathbb{R}^d \]  

(14)

\[ E^k_{\lambda}(w^k) = \tilde{h}_{\lambda}(w^k) \]  

(15)

\[ \nabla E^k_{\lambda}(w^k) \in \partial \tilde{h}_{\lambda}(w^k) \]  

(16)

\[ E^k_{\lambda}(w) \text{ is } L_{E_{\lambda}} := \left( L + \frac{1}{\lambda} \right) \text{ smooth.} \]  

(17)

**Proof.** Given that both functions contain \( f(w) \), it is sufficient to show that (14)-(16) hold between the second terms \( U^k_{\lambda}(w) \) and \( e_{\lambda}g(w) \).

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

\[ D^{\lambda}(w) - D^{\lambda}(z) = \sup_{x \in \mathbb{R}^d} \left( \frac{1}{\lambda} w^T x - \frac{1}{2\lambda}||x||^2 - g(x) \right) - \sup_{x \in \mathbb{R}^d} \left( \frac{1}{\lambda} z^T x - \frac{1}{2\lambda}||x||^2 - g(x) \right) \]

\[ \geq \frac{1}{\lambda} w^T \zeta^{\lambda}(z) - \frac{1}{2\lambda}||\zeta^{\lambda}(z)||^2 - g(\zeta^{\lambda}(z)) \]

\[ - \left( \frac{1}{\lambda} z^T \zeta^{\lambda}(z) - \frac{1}{2\lambda}||\zeta^{\lambda}(z)||^2 - g(\zeta^{\lambda}(z)) \right) \]

\[ = \frac{1}{\lambda} \zeta^{\lambda}(z)(w - z). \]  

(18)

Setting \( z = w^k \),

\[ e_{\lambda}g(w) = \frac{1}{2\lambda}||w||^2 - D^{\lambda}(w) \leq \frac{1}{2\lambda}||w||^2 - (D^{\lambda}(w^k) + \frac{1}{\lambda}\zeta^{\lambda}(w^k)^T(w - w^k)) = U^k_{\lambda}(w), \]
which proves (14). We also note that by the convexity of $D^\lambda(w)$ and (18),

$$\frac{1}{\lambda} \zeta^{\lambda}(w) \in \partial D^\lambda(w),$$  \hspace{1cm} (19)

(15): $U^k_\lambda(w^k) = \frac{1}{2\lambda}||w^k||^2 - D^\lambda(w^k) = e_\lambda g^k(w)$ from (12).

(16): As the supremum of lower semi-continuous functions, $D^\lambda(w)$ is lower semi-continuous, and given its convexity, it is locally Lipschitz continuous \cite[Theorem 5.17]{clarke2013}. Its Clarke generalized gradient exists, and

$$\partial e_\lambda g(w) = \partial \left( \frac{1}{2\lambda}||w||^2 - D^\lambda(w) \right)$$

$$= \frac{w}{\lambda} + \partial (-D^\lambda(w))$$

$$= \frac{w}{\lambda} - \partial D^\lambda(w)$$

$$\ni \frac{1}{\lambda}(w - \zeta(w)).$$

The second equality uses the fact that the Clarke generalized gradient of a continuously differentiable function and a locally Lipschitz function is equal to the sum of their Clarke generalized gradients, which can be found in \cite[Exercise 10.16]{clarke2013}. The third equality follows from the homogeneity of the Clarke generalized gradient \cite[Proposition 10.11]{clarke2013}. It holds that

$$\nabla U^k_\lambda(w^k) = \frac{1}{\lambda} w^k - \frac{1}{\lambda} \zeta^\lambda(w^k) \in \partial e_\lambda g(w^k).$$

(17): 

$$||\nabla E^k_\lambda(w) - \nabla E^k_\lambda(w')||_2 = ||\nabla f(w) + \frac{1}{\lambda} \left( w - \zeta^\lambda(w^k) \right) - \left( \nabla f(w') + \frac{1}{\lambda} \left( w' - \zeta^\lambda(w^k) \right) \right)||_2$$

$$\leq (L + \frac{1}{\lambda})||w - w'||_2.$$ 

We note that the Moreau envelope of a convex function is also $\frac{1}{\lambda}$-smooth \cite[Theorem 6.60]{beck2017}), 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 \cite{ghadimi2013} adapted to our problem. As $\nabla E^k_\lambda(w^R) \in \partial \bar{h}_\lambda(w^R)$ from (16), the following lemma bounds $\mathbb{E} \left[ \text{dist}(0, \partial \bar{h}_\lambda(w^R)) \right]$, with which we will ultimately bound $\mathbb{E} \left[ \text{dist}(0, \partial \bar{h}(\bar{w}^R)) \right]$ in Theorem 6.
Proof. From the definition of Lemma 2. For an initial value \( w^1 \in \mathbb{R}^d, N \in \mathbb{Z}_{>0}, \alpha \geq 0, \theta \in (0, 0.5 + \alpha), \) 
\[ \lambda = \frac{1}{N^\alpha} \in \mathbb{Z}_{N>0} \] 
\[ L_{\lambda} = L + \frac{1}{\lambda} \] 
\[ \gamma = \min \left\{ \frac{1}{L_{\lambda}}, \frac{1}{\sigma \sqrt{N}} \right\} \] 
\[ R \sim \text{uniform}\{1, \ldots, N\} \] 
for \( k = 1, 2, \ldots, R - 1 \) do 
\[ \zeta_{\lambda}(w^k) \in \text{prox}_{\lambda g}(w^k) \] 
Sample \( \xi_k \sim P^M \) 
\[ \nabla A_{\lambda M}^k(w^k, \xi_k) = \frac{1}{M} \sum_{j=1}^{M} \nabla F(w, \xi_j^k) + \frac{1}{\lambda}(w^k - \zeta_{\lambda}(w^k)) \] 
\[ w^{k+1} = w^k - \gamma \nabla A_{\lambda M}^k(w^k, \xi_k) \] 
end for 
Output: \( \bar{w}^R \in \text{prox}_{\lambda g}(w^R) \)

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

\[ \mathbb{E}\left[ \| \nabla E_{\lambda}^R(w^R) \|^2 \right] \leq \tilde{D} \left( L + N^\theta \right) + \frac{\sigma}{\sqrt{N}} \left( \tilde{D} + \frac{L + N^\theta}{|N^\alpha|} \right), \]

where \( \tilde{D} = 2(\tilde{h}_\lambda(w_1) - \tilde{h}_\lambda(w^*_\lambda)) \) and \( w^*_\lambda \) is a global minimizer of \( \tilde{h}_\lambda(\cdot). \)

In order to prove this result, we require the following property.

Property 3.

\[ \mathbb{E}\left[ \| \nabla A_{\lambda M}^k(w^k, y^k) - \nabla E_{\lambda}^k(w^k) \|^2 \right] \leq \frac{\sigma^2}{M} \]

Proof. From the definition of \( \nabla A_{\lambda M}^k(w^k, y^k) \) found in Algorithm 1 and (13), \( \nabla A_{\lambda M}^k(w^k, y^k) - \nabla E_{\lambda}^k(w^k) = \frac{1}{M} \sum_{j=1}^{M} \nabla F(w, y_j^k) - \nabla f(w). \) Taking the expectation of its squared norm,

\[ \mathbb{E}\left[ \| \nabla A_{\lambda M}^k(w^k, y^k) - \nabla E_{\lambda}^k(w^k) \|^2 \right] = \mathbb{E}\left[ \| \frac{1}{M} \sum_{j=1}^{M} (\nabla F(w, y_j^k) - \nabla f(w)) \|^2 \right] = \frac{1}{M^2} \mathbb{E} \sum_{i=1}^{n} \left( \sum_{j=1}^{M} \nabla F(w, y_j^k)_i - \nabla f(w)_i \right)^2. \]

For \( j \neq l, \nabla F(w, y_j^k)_i - \nabla f(w)_i \) and \( \nabla F(w, y_l^k)_i - \nabla f(w)_i \) are independent random variables with zero mean. It follows that

\[ \mathbb{E}[(\nabla F(w, y_j^k)_i - \nabla f(w)_i)(\nabla F(w, y_l^k)_i - \nabla f(w)_i)] = \mathbb{E}[(\nabla F(w, y_j^k)_i - \nabla f(w)_i)]\mathbb{E}[(\nabla F(w, y_l^k)_i - \nabla f(w)_i)] = 0, \]
and
\[
\frac{1}{M^2} \mathbb{E} \sum_{i=1}^{n} \left( \sum_{j=1}^{M} \nabla F(w, y_j^k) - \nabla f(w) \right)^2 = \frac{1}{M^2} \mathbb{E} \sum_{i=1}^{n} \sum_{j=1}^{M} (\nabla F(w, y_j^k) - \nabla f(w))^2
\]
\[
= \frac{1}{M^2} \sum_{j=1}^{M} \mathbb{E} \| \nabla F(w, y_j^k) - \nabla f(w) \|_2^2 \leq \frac{\sigma^2}{M}
\]

using (6).

Proof of Lemma: Given the smoothness of \( E_\lambda^k(w) \) as shown in Property 11,
\[
E_\lambda^k(w^{k+1}) \leq E_\lambda^k(w^k) + \langle \nabla E_\lambda^k(w^k), w^{k+1} - w^k \rangle + \frac{L_{E\lambda}}{2} \| w^{k+1} - w^k \|^2
\]
\[
= E_\lambda^k(w^k) + \langle \nabla E_\lambda^k(w^k), -\gamma \nabla A_{\lambda M}^k(w^k, \xi^k) \rangle + \frac{L_{E\lambda}}{2} \| -\gamma \nabla A_{\lambda M}^k(w^k, \xi^k) \|^2.
\]

Using (14) and (15),
\[
\tilde{h}(w^{k+1}) \leq \tilde{h}(w^k) - \gamma \langle \nabla E_\lambda^k(w^k), \nabla A_{\lambda M}^k(w^k, \xi^k) \rangle + \frac{L_{E\lambda}}{2} \gamma^2 \| \nabla A_{\lambda M}^k(w^k, \xi^k) \|^2.
\]

Setting \( \delta_k = \nabla A_{\lambda M}^k(w^k, \xi^k) - \nabla E_\lambda^k(w^k) \),
\[
\tilde{h}(w^{k+1}) \leq \tilde{h}(w^k) - \gamma \left( \| \nabla E_\lambda^k(w^k) \|_2^2 + \langle \nabla E_\lambda^k(w^k), \delta_k \rangle \right)
\]
\[
+ \frac{L_{E\lambda}}{2} \gamma^2 \left( \| \nabla E_\lambda^k(w^k) \|_2^2 + 2 \langle \nabla E_\lambda^k(w^k), \delta_k \rangle + \| \delta_k \|^2 \right)
\]
\[
= \tilde{h}(w^k) + \left( \frac{L_{E\lambda}}{2} \gamma^2 - \gamma \right) \| \nabla E_\lambda^k(w^k) \|_2^2 + \left( L_{E\lambda} \gamma^2 - \gamma \right) \langle \nabla E_\lambda^k(w^k), \delta_k \rangle + \frac{L_{E\lambda}}{2} \gamma^2 \| \delta_k \|_2^2,
\]
as
\[
\langle \nabla E_\lambda^k(w^k), \nabla A_{\lambda M}^k(w^k, \xi^k) \rangle = \| \nabla E_\lambda^k(w^k) \|_2^2 + \langle \nabla E_\lambda^k(w^k), \delta_k \rangle
\]
and
\[
\| \nabla A_{\lambda M}^k(w^k, \xi^k) \|^2 = \| \nabla E_\lambda^k(w^k) \|^2 + 2 \langle \nabla E_\lambda^k(w^k), \delta_k \rangle + \| \delta_k \|^2.
\]

After \( N \) iterations,
\[
\left( \gamma - \frac{L_{E\lambda}}{2} \gamma^2 \right) \sum_{k=1}^{N} \| \nabla E_\lambda^k(w^k) \|_2^2
\]
\[
\leq \tilde{h}(w^1) - \tilde{h}(w^{N+1}) + \left( L_{E\lambda} \gamma^2 - \gamma \right) \sum_{k=1}^{N} \langle \nabla E_\lambda^k(w^k), \delta_k \rangle + \frac{L_{E\lambda}}{2} \gamma^2 \sum_{k=1}^{N} \| \delta_k \|^2
\]
\[
\leq \tilde{h}_\lambda(w^1) - \tilde{h}_\lambda(w^*) + \left( L_{E\lambda} \gamma^2 - \gamma \right) \sum_{k=1}^{N} \langle \nabla E_\lambda^k(w^k), \delta_k \rangle + \frac{L_{E\lambda}}{2} \gamma^2 \sum_{k=1}^{N} \| \delta_k \|^2.
\]
It follows from (5) that for $w$ independent of $\xi_k$, $E \nabla A^k_{\lambda M}(w, \xi_k) = \nabla E^k_{\lambda}(w)$, and so $E[\delta_k] = 0$. Taking the expectation of both sides,

$$
\left( \gamma - \frac{L_{E\lambda}}{2}\gamma^2 \right) \sum_{k=1}^{N} \mathbb{E}\|\nabla E^k_{\lambda}(w^k)\|^2 \leq \tilde{h}(w^1) - \tilde{h}(w^\ast) + \frac{L_{E\lambda}}{2}\gamma^2 \sum_{k=1}^{N} \mathbb{E}\|\delta_k\|^2
$$

$$
\leq \tilde{h}(w^1) - \tilde{h}(w^\ast) + \frac{L_{E\lambda}}{2}\gamma^2 \frac{N}{M}\sigma^2,
$$

where the second inequality uses Property 3. As we choose $R$ uniformly over $\{1, ..., N\}$,

$$
\mathbb{E}\|\nabla E^R_{\lambda}(w^R)\|^2 = \frac{1}{N} \sum_{k=1}^{N} \mathbb{E}\|\nabla E^k_{\lambda}(w^k)\|^2
$$

$$
\leq \frac{1}{N \left( \gamma - \frac{L_{E\lambda}}{2}\gamma^2 \right)} \left( \tilde{h}(w^1) - \tilde{h}(w^\ast) + \frac{L_{E\lambda}}{2}\gamma^2 \frac{N}{M}\sigma^2 \right).
$$

Since $\gamma \leq \frac{1}{L_{E\lambda}}$, $\gamma - \frac{L_{E\lambda}}{2}\gamma^2 \geq \frac{1}{2}\gamma$, and

$$
\frac{1}{N \left( \gamma - \frac{L_{E\lambda}}{2}\gamma^2 \right)} \left( \frac{\tilde{D}}{2} + \frac{L_{E\lambda}}{2}\gamma^2 \frac{N}{M}\sigma^2 \right) \leq \frac{1}{N\gamma} \left( \tilde{D} + L_{E\lambda}\gamma^2 \frac{N}{M}\sigma^2 \right)
$$

$$
= \frac{\tilde{D}}{N\gamma} + \frac{L_{E\lambda}\gamma^2}{M}\sigma^2
$$

$$
\leq \frac{\tilde{D}}{N} \max \left\{ L_{E\lambda}, \sigma\sqrt{N} \right\} + \frac{L_{E\lambda}\sigma}{M\sqrt{N}}
$$

$$
\leq \frac{\tilde{D}L_{E\lambda}}{N} + \frac{\sigma}{\sqrt{N}} \left( \tilde{D} + \frac{L_{E\lambda}}{M} \right).
$$

In order to prove the convergence of $\mathbb{E} [\text{dist}(0, \partial h(\bar{w}^R))]$, we will require the following two properties.

**Property 4.** Assume that $g(w)$ is Lipschitz continuous with parameter $l$.

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

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

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

and

$$
\nabla f(\zeta^\lambda(w^k)) + \frac{1}{\lambda}(w^k - \zeta^\lambda(w^k)) \in \partial h(\zeta^\lambda(w^k)).
$$
It follows that
\[
\text{dist}(0, \partial h(\zeta^\lambda(w^k))) \leq \|\nabla f(\zeta^\lambda(w^k)) + \frac{1}{\lambda}(w^k - \zeta^\lambda(w^k))\|_2
\]
\[
= \|\nabla f(w^k) - \nabla f(w^k) + \nabla f(\zeta^\lambda(w^k)) + \frac{1}{\lambda}(w^k - \zeta^\lambda(w^k))\|_2
\]
\[
\leq \|\nabla f(w^k) + \frac{1}{\lambda}(w^k - \zeta^\lambda(w^k))\|_2 + \|\nabla f(\zeta^\lambda(w^k)) - \nabla f(w^k)\|_2
\]
\[
\leq \|\nabla E_\lambda(w^k)\|_2 + L\|w^k - \zeta^\lambda(w^k)\|_2.
\]
In order to bound \(\|w^k - \zeta^\lambda(w^k)\|_2\), recall from (11) that
\[
g(w) \geq e_\lambda g(w)
\]
\[
= \frac{1}{2\lambda}\|w - \zeta^\lambda(w)\|_2^2 + g(\zeta^\lambda(w)).
\]
Rearranging and using the Lipschitz continuity,
\[
\frac{1}{2\lambda}\|w - \zeta^\lambda(w)\|_2^2 \leq g(w) - g(\zeta^\lambda(w))
\]
\[
\leq l\|w - \zeta^\lambda(w)\|_2
\]
\[
\|w - \zeta^\lambda(w)\|_2 \leq 2l\lambda.
\]

\[\square\]

**Property 5.** Let \(w^*\) be a global minimizer of \(h(\cdot)\) and let \(w_\lambda^*\) 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.**
\[
\tilde{h}_\lambda(w) - \tilde{h}_\lambda(w_\lambda^*) - (h(w) - h(w^*)) = e_\lambda g(w) - f(w_\lambda^*) - e_\lambda g(w_\lambda^*) - (g(w) - f(w^*) - g(w^*))
\]
\[
\leq - f(w_\lambda^*) - e_\lambda g(w_\lambda^*) + f(w^*) + g(w^*)
\]
\[
\leq - f(w_\lambda^*) - e_\lambda g(w_\lambda^*) + f(w_\lambda^*) + g(w_\lambda^*)
\]
\[
eq g(w_\lambda^*) - e_\lambda g(w_\lambda^*),
\]
where the first inequality follows from (11). For any \(w\), by the definition of the Moreau envelope,
\[
e_\lambda g(w) = \frac{1}{2\lambda}\|w - \zeta^\lambda(w)\|_2^2 + g(\zeta^\lambda(w))
\]
\[
g(w) - e_\lambda g(w) = g(w) - g(\zeta^\lambda(w)) - \frac{1}{2\lambda}\|w - \zeta^\lambda(w)\|_2^2
\]
\[
\leq l\|w - \zeta^\lambda(w)\|_2 - \frac{1}{2\lambda}\|w - \zeta^\lambda(w)\|_2^2.
\]
The right-hand side is maximized when \(\|w - \zeta^\lambda(w)\|_2 = l\lambda\), giving the desired result,
\[
g(w) - e_\lambda g(w) \leq \frac{l^2\lambda}{2}.
\]

\[\square\]
We note that (20) cannot be improved under the further assumption that $g(w)$ is convex, which can be found in (Beck, 2017, Theorem 10.51).

**Theorem 6.** Assume that $g(w)$ is Lipschitz continuous with parameter $l$. The output $\bar{w}^R$ of MBSGA satisfies

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

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

**Proof.** From Property 4, considering $\zeta(\bar{w}^R) = \bar{w}^R$,

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

Taking its expectation,

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

$$
\leq \sqrt{\mathbb{E}[||\nabla E(\bar{w}^R)||^2_2]} \text{+} \frac{2lL}{N^\theta}
$$

$$
\leq \sqrt{\tilde{D}(L + N^\theta)} \frac{N}{N} \text{+} \sqrt{\frac{\sigma}{\sqrt{N}} \left( \tilde{D} + \frac{L + N^\theta}{N^\theta} \right) + \frac{2lL}{N^\theta}},
$$

where the second inequality follows from Jensen’s inequality and the third inequality follows from Lemma 2. The result then follows using Property 5 as

$$
\tilde{D} = 2(\tilde{h}(w^1) - \tilde{h}(w^*_\lambda)) \leq 2(h(w^1) - h(w^*)) + l^2 \lambda
$$

$$
= D + \frac{l^2 N^\theta}{N^\theta}.
$$

Now that we have bounded the distance of $\partial h(w)$ from the origin, we prove an $\epsilon$-stationary point convergence complexity for the single sample case, $M = 1$, and for an optimal size of $M$.

**Corollary 7.** Assume that $g(w)$ is Lipschitz continuous with parameter $l$. To obtain an $\epsilon$-stationary solution (8) using MBSGA, the optimal gradient call complexity with $\alpha = 0$ is $O(\epsilon^{-6})$ choosing $\theta = \frac{1}{6}$. The optimal mini-batch complexity is $O(\epsilon^{-5})$, and occurs by choosing $\alpha = 0.25$ and $\theta = 0.25$.

**Proof.** From Theorem 6,

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

$$
= O(N^{0.5\theta-0.5}) + O(N^{-0.25} + N^{0.5\theta-0.5\alpha-0.25}) + O(N^{-\theta}).
$$
Setting $\theta = \alpha = 0.25$,
\[ \mathbb{E} \left[ \text{dist}(0, \partial h(\tilde{w}^R)) \right] \leq O(N^{-0.25}). \]

An $\epsilon$-stationary solution will require less than $N = O(\epsilon^{-4})$ iterations. The number of gradient calls per iteration is $[N^\alpha] = O(\epsilon^{-1})$. The number of gradient calls to get an $\epsilon$-stationary solution is then
\[ N[N^\alpha] = O(\epsilon^{-5}). \]

When $\alpha = 0$,
\[ \mathbb{E} \left[ \text{dist}(0, \partial h(\tilde{w}^R)) \right] = O(N^{0.5\theta - 0.25} + N^{-\theta}). \]

Choosing $\theta = \frac{1}{6}$, the number of gradient calls to get an $\epsilon$-stationary solution is $O(\epsilon^{-6})$.

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 $E^k_{t\lambda}(w)$, which is constructed in the same manner as $E^k_\lambda(w)$, using $w^k_t$ instead of $w^k$. This function possesses the same characteristics as found in Property II. The convergence analysis follows very closely to the work of Li and Li (2018) adapted to our problem.

**Lemma 8.** For an initial value $\tilde{w}_1$ and $N \in \mathbb{Z}_{>0}$, VRSGA generates $w^R_T$ satisfying the following bounds.
\[ \mathbb{E} \left[ ||\nabla E^R_{t\lambda}(w^R_T)||^2 \right] \leq \frac{\tilde{D}L_{\lambda}E}{Sm} \]
\[ \leq \frac{\tilde{D}L_{\lambda}E}{N}, \]
where $\tilde{D} = 36(\tilde{h}_\lambda(w^1) - \tilde{h}_\lambda(w_\lambda^*))$ and $w_\lambda^*$ is a global minimizer of $\tilde{h}_\lambda(\cdot)$.

In order to prove this result, we require the following lemmas.

**Lemma 9.** Consider arbitrary $w, V, z \in \mathbb{R}^d$, $\gamma \in \mathbb{R}$, and $w^+ = w - \gamma V$,
\[ E^k_{t\lambda}(w^+) \leq E^k_{t\lambda}(z) + \langle \nabla E^k_{t\lambda}(w) - V, w^+ - z \rangle + \frac{L_{\lambda}E}{2} ||w^+ - w||^2 + \frac{L_{\lambda}E}{2} ||z - w||^2 \]
\[ - \frac{1}{\gamma} \langle w^+ - w, w^+ - z \rangle. \]
Algorithm 2 Variance reduced stochastic gradient algorithm (VRSGA)

\textbf{Input:} \tilde{w}^1 \in \mathbb{R}^d, N \in \mathbb{Z}_{>0}
\quad m = \lceil n \frac{M}{M} \rceil, \quad b = \lceil n \frac{M}{M} \rceil
\quad S = \lceil \frac{N}{m} \rceil, \quad \lambda = (Sm)^{-\frac{1}{2}}
\quad L_{E\lambda} = L + \frac{1}{\lambda}, \quad \gamma = \frac{1}{6L_{E\lambda}}
\quad R \sim \text{uniform}\{1, \ldots, S\}

\textbf{for} k = 1, 2, \ldots, R \textbf{do}
\quad w_1^k = \tilde{w}^k
\quad G^k = \nabla f(\tilde{w}^k)
\quad \textbf{for} t = 1, 2, \ldots, m \textbf{do}
\quad \quad \zeta^k_\lambda(w_t^k) \in \text{prox}_{\lambda g}(w_t^k)
\quad \quad I \sim \text{uniform}\{1, \ldots, n\}^b
\quad \quad 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^k_\lambda(w_t^k))
\quad \quad w_{t+1}^k = w_t^k - \gamma V_t^k
\quad \textbf{end for}
\quad \tilde{w}^{k+1} = w_{m+1}^k
\textbf{end for}

T \sim \text{uniform}\{1, \ldots, m\}
\textbf{Output:} \tilde{w}_T^R \in \text{prox}_{\lambda g}(w_T^R)

\textbf{Proof.} Adding the following three inequalities proves the result, where the first two inequalities come from the smoothness of \(E_{t\lambda}^k(w)\) and \(-E_{t\lambda}^k(w)\), see Property 1, and the third inequality comes from the fact that, by definition, \(V + \frac{1}{\gamma} (w^+ - w) = 0\).

\[ E_{t\lambda}^k(w^+) \leq E_{t\lambda}^k(w) + \langle \nabla E_{t\lambda}^k(w), w^+ - w \rangle + \frac{L_{E\lambda}}{2} ||w^+ - w||^2 \]
\[ -E_{t\lambda}^k(z) \leq -E_{t\lambda}^k(w) + \langle -\nabla E_{t\lambda}^k(w), z - w \rangle + \frac{L_{E\lambda}}{2} ||z - w||^2 \]
\[ 0 = -\langle V + \frac{1}{\gamma} (w^+ - w), w^+ - z \rangle \]

\[ \sum_{t=1}^{m} \sum_{j \in I} \nabla f_j(w_t^k) - \nabla f_j(\tilde{w}^k) = \sum_{t=1}^{m} \sum_{j \in I} \nabla f_j(\tilde{w}^k) - \nabla f_j(w_t^k) \]

Lemma 10. For vectors \(w, x, z\), and \(\beta > 0\),

\[ ||w - x||_2^2 \leq (1 + \beta) ||w - z||_2^2 + \left(1 + \frac{1}{\beta}\right) ||z - x||_2^2. \]
Proof.
\[ ||w - x||^2 = ||w - z + z - x||^2 \]
\[ \leq (||w - z||^2 + ||z - x||^2)^2 \]
\[ = ||w - z||^2 + 2||w - z||^2||z - x||^2 + ||z - x||^2 \]
\[ \leq ||w - z||^2 + \left( \beta||w - z||^2 + \frac{1}{\beta}||z - x||^2 \right) + ||z - x||^2 \]
\[ = (1 + \beta)||w - z||^2 + \left( 1 + \frac{1}{\beta} \right) ||z - x||^2, \]
where the second inequality uses Young’s inequality.

\[ \square \]

Proof of Lemma 9 Let \( \bar{w}_{t+1}^k = w_t^k - \gamma \nabla E_{\lambda t}^k(w_t^k) \), with \( w^+ = w_{t+1}^k \), \( w = w_t^k \), \( V = V_t^k \), and \( z = \bar{w}_{t+1}^k \) in Lemma 9 to get the inequality

\[ E_{\lambda t}^k(w_{t+1}^k) \leq E_{\lambda t}^k(\bar{w}_{t+1}^k) + \langle \nabla E_{\lambda t}^k(w_t^k) - V_t^k, w_{t+1}^k - \bar{w}_{t+1}^k \rangle + \frac{L_{E\lambda}}{2} ||w_{t+1}^k - w_t^k||^2 \]
\[ + \frac{L_{E\lambda}}{2} ||w_{t+1}^k - w_t^k||^2 - \frac{1}{\gamma} \langle w_{t+1}^k - w_t^k, \bar{w}_{t+1}^k - w_t^k \rangle \]
\[ = E_{\lambda t}^k(w_t^k) + \left( \frac{L_{E\lambda}}{2} - \frac{1}{\gamma} \right) ||w_{t+1}^k - w_t^k||^2. \]

Adding (21) and (22),

\[ E_{\lambda t}^k(w_{t+1}^k) \leq E_{\lambda t}^k(w_t^k) + \langle \nabla E_{\lambda t}^k(w_t^k) - V_t^k, w_{t+1}^k - w_t^k \rangle + \frac{L_{E\lambda}}{2} ||w_{t+1}^k - w_t^k||^2 \]
\[ - \frac{1}{\gamma} \langle w_{t+1}^k - w_t^k, w_{t+1}^k - \bar{w}_{t+1}^k \rangle + \left( L_{E\lambda} - \frac{1}{\gamma} \right) ||w_{t+1}^k - w_t^k||^2. \]

Plugging \( \langle w_{t+1}^k - w_t^k, w_{t+1}^k - \bar{w}_{t+1}^k \rangle = \frac{1}{2} (||w_{t+1}^k - w_t^k||^2 + ||w_{t+1}^k - \bar{w}_{t+1}^k||^2 - ||\bar{w}_{t+1}^k - w_t^k||^2) \) into (23) and rearranging,

\[ E_{\lambda t}^k(w_{t+1}^k) \leq E_{\lambda t}^k(w_t^k) + \langle \nabla E_{\lambda t}^k(w_t^k) - V_t^k, w_{t+1}^k - \bar{w}_{t+1}^k \rangle + \left( \frac{L_{E\lambda}}{2} - \frac{1}{2\gamma} \right) ||w_{t+1}^k - w_t^k||^2 \]
\[ - \frac{1}{2\gamma} ||w_{t+1}^k - \bar{w}_{t+1}^k||^2 + \left( L_{E\lambda} - \frac{1}{2\gamma} \right) ||\bar{w}_{t+1}^k - w_t^k||^2. \]

(24)
Focusing on the term $-\frac{1}{\gamma}||w^k_{t+1} - \bar{w}^k_{t+1}||_2^2$, we apply Lemma 10 with $w = w^k_{t+1}$, $x = w^k_t$, and $z = \bar{w}^k_{t+1}$. Rearranging,

$$-(1 + \beta)||w^k_{t+1} - \bar{w}^k_{t+1}||_2^2 \leq -||w^k_{t+1} - w^k_t||_2^2 + \left(1 + \frac{1}{\beta}\right)||\bar{w}^k_{t+1} - w^k_t||_2^2$$

$$-\frac{1}{2\gamma}||w^k_{t+1} - \bar{w}^k_{t+1}||_2^2 \leq -\frac{1}{(1 + \beta)2\gamma}||w^k_{t+1} - w^k_t||_2^2 + \frac{1}{(1 + \beta)2\gamma}||\bar{w}^k_{t+1} - w^k_t||_2^2.$$

Choosing $\beta = 3$,

$$-\frac{1}{2\gamma}||w^k_{t+1} - \bar{w}^k_{t+1}||_2^2 \leq -\frac{1}{8\gamma}||w^k_{t+1} - w^k_t||_2^2 + \frac{1}{6\gamma}||\bar{w}^k_{t+1} - w^k_t||_2^2.$$

Using this inequality in (24),

$$E^k_{t+1}(w^k_{t+1}) \leq E^k_{t\lambda}(w^k_t) + \langle \nabla E^k_{t\lambda}(w^k_t) - V^k_t, w^k_{t+1} - \bar{w}^k_{t+1} \rangle + \left(\frac{L_{E\lambda}}{2} - \frac{1}{2\gamma}\right)||w^k_{t+1} - w^k_t||_2^2$$

$$-\frac{1}{8\gamma}||w^k_{t+1} - w^k_t||_2^2 + \frac{1}{6\gamma}||\bar{w}^k_{t+1} - w^k_t||_2^2 + \left(\frac{L_{E\lambda}}{2} - \frac{1}{2\gamma}\right)||\bar{w}^k_{t+1} - w^k_t||_2^2$$

$$= E^k_{t\lambda}(w^k_t) + \langle \nabla E^k_{t\lambda}(w^k_t) - V^k_t, w^k_{t+1} - \bar{w}^k_{t+1} \rangle + \left(\frac{L_{E\lambda}}{2} - \frac{5}{8\gamma}\right)||w^k_{t+1} - w^k_t||_2^2$$

$$+ \left(\frac{L_{E\lambda}}{2} - \frac{5}{8\gamma}\right)||\bar{w}^k_{t+1} - w^k_t||_2^2,$$

where the last equality holds since $w^k_{t+1} - \bar{w}^k_{t+1} = \gamma(\nabla E^k_{t\lambda}(w^k_t) - V^k_t)$. Using (14) and (15) and taking the expectation of both sides,

$$\mathbb{E}\tilde{h}_\lambda(w^k_{t+1}) \leq \mathbb{E}\left[\tilde{h}_\lambda(w^k_t) + \gamma||\nabla E^k_{t\lambda}(w^k_t) - V^k_t||_2^2 + \left(\frac{L_{E\lambda}}{2} - \frac{5}{8\gamma}\right)||w^k_{t+1} - w^k_t||_2^2$$

$$+ \left(\frac{L_{E\lambda}}{2} - \frac{5}{8\gamma}\right)||\bar{w}^k_{t+1} - w^k_t||_2^2\right].$$

(25)

Focusing on $\mathbb{E}||\nabla E^k_{t\lambda}(w^k_t) - V^k_t||_2^2$, from (13) and the definition of $V^k_t$ found in Algorithm 2, $\nabla E^k_{t\lambda}(w^k_t) - V^k_t = \nabla f(w) - (\frac{1}{b} \sum_{j \in I} (\nabla f_j(w^k_t) - \nabla f_j(\bar{w}^k_t)) + G^k)$. Rearranging, and taking
the expectation of its squared norm,

\[ \mathbb{E}\|\nabla E_{\lambda}^k(w^k_t) - V^k_t\|^2 = \mathbb{E}\left|\frac{1}{b} \sum_{j \in I} (\nabla f_j(\tilde{w}^k) - \nabla f_j(w^k_t)) - (G^k - \nabla f(w^k_t)) \right|^2 \]

\[ = \frac{1}{b^2} \mathbb{E} \sum_{j \in I} \|\nabla f_j(\tilde{w}^k) - \nabla f_j(w^k_t) - (G^k - \nabla f(w^k_t))\|^2 \]

\[ \leq \frac{1}{b^2} \mathbb{E} \sum_{j \in I} \|\nabla f_j(\tilde{w}^k) - \nabla f_j(w^k_t)\|^2 \]

\[ \leq \frac{L_E^2}{b} \mathbb{E}\|\tilde{w}^k - w^k_t\|^2. \]

As the squared norm of a sum of independent random variables with zero mean, the second equality holds using the same reasoning as found in Property 3 and the first inequality holds since \( \mathbb{E}\|x - \mathbb{E}[x]\|^2 \leq \mathbb{E}\|x\|^2 \) for any random variable \( x \). Using this bound in (25),

\[ \mathbb{E} \tilde{h}_\lambda(w^k_{t+1}) \leq \mathbb{E} \left[ \tilde{h}_\lambda(w^k_t) + \frac{L_E^2}{b} \|\tilde{w}^k - w^k_t\|^2 + \left( \frac{L_E}{2} - \frac{5}{8\gamma} \right) \|w^k_{t+1} - w^k_t\|^2 \right. \]

\[ \left. + \left( L_E - \frac{1}{3\gamma} \right) \|\bar{w}^k_{t+1} - w^k_t\|^2 \right] \]

\[ = \mathbb{E} \left[ \tilde{h}_\lambda(w^k_t) + \frac{L_E}{6b} \|\tilde{w}^k - w^k_t\|^2 - \frac{13L_E}{4} \|w^k_{t+1} - w^k_t\|^2 - L_E \|\bar{w}^k_{t+1} - w^k_t\|^2 \right] \]

\[ = \mathbb{E} \left[ \tilde{h}_\lambda(w^k_t) + \frac{L_E}{6b} \|\tilde{w}^k - w^k_t\|^2 - \frac{13L_E}{4} \|w^k_{t+1} - w^k_t\|^2 - \frac{1}{36L_E} \|\nabla E_{\lambda}^k(w^k_t)\|^2 \right], \quad (26) \]

where both equalities use the fact that \( \gamma = \frac{1}{6L_E} \). Focusing on \( -\frac{13L_E}{4} \|w^k_{t+1} - w^k_t\|^2 \), we apply Lemma 10 with \( w = w^k_{t+1}, x = \tilde{w}^k, \) and \( z = w^k_t \),

\[ (1 + \beta) \|w^k_{t+1} - w^k_t\|^2 \geq \|w^k_{t+1} - \tilde{w}^k\|^2 - \left( 1 + \frac{1}{\beta} \right) \|w^k_t - \tilde{w}^k\|^2 \]

\[ \geq \frac{13L_E}{4} \|w^k_{t+1} - w^k_t\|^2 \leq \frac{13L_E}{4(1 + \beta)} \|w^k_{t+1} - \tilde{w}^k\|^2 + \frac{13L_E}{4(1 + \beta)} \|w^k_t - \tilde{w}^k\|^2. \]

Setting \( \beta = 2t - 1, \)

\[ -\frac{13L_E}{4} \|w^k_{t+1} - w^k_t\|^2 \leq -\frac{13L_E}{8t} \|w^k_{t+1} - \tilde{w}^k\|^2 + \frac{13L_E}{8t - 4} \|w^k_t - \tilde{w}^k\|^2. \]

Applying this bound in (26),

\[ \mathbb{E} \tilde{h}_\lambda(w^k_{t+1}) \leq \mathbb{E} \left[ \tilde{h}_\lambda(w^k_t) + \left( \frac{L_E}{6b} + \frac{13L_E}{8t - 4} \right) \|\tilde{w}^k - w^k_t\|^2 - \frac{13L_E}{8t} \|w^k_{t+1} - \tilde{w}^k\|^2 \right. \]

\[ \left. - \frac{1}{36L_E} \|\nabla E_{\lambda}^k(w^k_t)\|^2 \right]. \quad (27) \]
Summing over $t$, 
\[
\mathbb{E} \tilde{h}_\lambda(w_{m+1}^k) \leq \mathbb{E} \left[ \tilde{h}_\lambda(w_1^k) + \sum_{t=1}^{m} \left( \frac{L_{E\lambda}}{6b} + \frac{13L_{E\lambda}}{8t} \right) \| \tilde{w}^k - w_t^k \|^2 \right. \\
- \sum_{t=1}^{m} \frac{13L_{E\lambda}}{8t} \| w_{t+1}^k - \tilde{w}^k \|^2 - \frac{1}{36L_{E\lambda}} \sum_{t=1}^{m} \| \nabla E_{t\lambda}(w_t^k) \|^2 \right].
\]
Considering that $\tilde{w}^k = w_1^k$ and $\| w_{m+1}^k - \tilde{w}^k \|^2 \geq 0$, 
\[
\mathbb{E} \tilde{h}_\lambda(w_{m+1}^k) \leq \mathbb{E} \left[ \tilde{h}_\lambda(w_1^k) + \sum_{t=1}^{m-1} \left( \frac{L_{E\lambda}}{6b} + \frac{13L_{E\lambda}}{8t} \right) \| w_t^k - \tilde{w}^k \|^2 \right. \\
- \sum_{t=1}^{m-1} \frac{13L_{E\lambda}}{8t} \| w_{t+1}^k - \tilde{w}^k \|^2 - \frac{1}{36L_{E\lambda}} \sum_{t=1}^{m} \| \nabla E_{t\lambda}(w_t^k) \|^2 \right] \\
= \mathbb{E} \left[ \tilde{h}_\lambda(w_1^k) + \sum_{t=1}^{m-1} \left( \frac{L_{E\lambda}}{6b} + \frac{13L_{E\lambda}}{8t} - \frac{13L_{E\lambda}}{8t+4} \right) \| w_t^k - \tilde{w}^k \|^2 \right. \\
- \frac{1}{36L_{E\lambda}} \sum_{t=1}^{m} \| \nabla E_{t\lambda}(w_t^k) \|^2 \right] \\
\leq \mathbb{E} \left[ \tilde{h}_\lambda(w_1^k) - \frac{1}{36L_{E\lambda}} \sum_{t=1}^{m} \| \nabla E_{t\lambda}(w_t^k) \|^2 \right],
\]
where the last inequality holds since $6b = 6m^2 > 2(m-1)^2 \geq 2t^2$ for $t = 1, \ldots, m-1$. This summation can be equivalently written as
\[
\mathbb{E} \tilde{h}_\lambda(\tilde{w}^{k+1}) \leq \mathbb{E} \tilde{h}_\lambda(\tilde{w}^k) - \mathbb{E} \left[ \frac{1}{36L_{E\lambda}} \sum_{t=1}^{m} \| \nabla E_{t\lambda}(w_t^k) \|^2 \right]
\]
\[
\mathbb{E} \left[ \frac{1}{36L_{E\lambda}} \sum_{t=1}^{m} \| \nabla E_{t\lambda}(w_t^k) \|^2 \right] \leq \mathbb{E} \tilde{h}_\lambda(\tilde{w}^k) - \mathbb{E} \tilde{h}_\lambda(\tilde{w}^{k+1})
\]
\[
\mathbb{E} \left[ \frac{1}{36L_{E\lambda}} \sum_{k=1}^{S} \sum_{t=1}^{m} \| \nabla E_{t\lambda}(w_t^k) \|^2 \right] \leq \tilde{h}_\lambda(\tilde{w}^1) - \mathbb{E} \tilde{h}_\lambda(\tilde{w}^{S+1})
\]
\[
\leq \tilde{h}_\lambda(\tilde{w}^1) - \tilde{h}_\lambda(w_\lambda^{*})
\]
\[
\mathbb{E} \left[ \| \nabla E_{T\lambda}(w_T^R) \|^2 \right] \leq \frac{36L_{E\lambda} (\tilde{h}_\lambda(\tilde{w}^1) - \tilde{h}_\lambda(w_\lambda^{*}))}{Sm} \leq \frac{36L_{E\lambda} (\tilde{h}_\lambda(\tilde{w}^1) - \tilde{h}_\lambda(w_\lambda^{*}))}{N}.
\]

\[\square\]
Theorem 11. Assume that $g(w)$ is Lipschitz continuous with parameter $l$. The output $\bar{w}_T^R$ of VRSGA satisfies

$$\mathbb{E} \left[ \| \text{dist}(0, \partial h(\bar{w}_T^R)) \|_2 \right] \leq \sqrt{\frac{\left( L + (Sm)^{\frac{1}{2}} \right) \left( D + 18l^2(Sm)^{\frac{1}{2}} \right)}{Sm}} + \frac{2lL}{(Sm)^{\frac{1}{3}}}$$

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

Proof. The proof follows exactly what was done to prove Theorem 6. From Property 4,

$$\text{dist}(0, \partial h(\bar{w}_T^R)) \leq \| \nabla \mathbb{E} \lambda (w_T^R) \|_2^{\frac{1}{2}} + 2l\lambda L.$$ 

Taking its expectation,

$$\mathbb{E} \left[ \| \text{dist}(0, \partial h(\bar{w}_T^R)) \|_2 \right] \leq \mathbb{E} \left[ \| \nabla \mathbb{E} \lambda (w_T^R) \|_2 \right]^{\frac{1}{2}} + 2l\lambda L \leq \sqrt{\mathbb{E} \left[ \| \nabla \mathbb{E} \lambda (w_T^R) \|_2 \right]}^{\frac{1}{2}} + \frac{2lL}{(Sm)^{\frac{1}{3}}}.$$

where the third inequality follows from Lemma 8. The fourth inequality holds since $L_{E\lambda} = L + (Sm)^{\frac{1}{2}}$, and using Property 5,

$$\hat{D} = 36(h_\lambda(w^1) - h_\lambda(w^*_\lambda)) \leq 36(h(w^1) - h(w^*)) + 18l^2\lambda$$

$$= D + \frac{18l^2}{(Sm)^{\frac{1}{2}}}$$

$\square$

Corollary 12. Assume that $g(w)$ is Lipschitz continuous with parameter $l$. To obtain an $\epsilon$-stationary solution using VRSGA, the gradient call complexity is $O(n^{\frac{2}{3}}\epsilon^{-3})$.

Proof. From Theorem 11

$$\mathbb{E} \left[ \| \text{dist}(0, \partial h(\bar{w}_T^R)) \|_2 \right] \leq \sqrt{\frac{\left( L + (Sm)^{\frac{1}{2}} \right) \left( D + 18l^2(Sm)^{\frac{1}{2}} \right)}{Sm}} + \frac{2lL}{(Sm)^{\frac{1}{3}}}$$

An $\epsilon$-stationary solution will require at most $Sm = O(\epsilon^{-3})$ iterations. The number of gradient calls after $Sm$ iterations is

$$Sn + Sm b = Sm \left[ \frac{n}{n^{\frac{1}{3}}} \right] + Sm \left[ n^{\frac{2}{3}} \right] = O(n^{\frac{2}{3}}\epsilon^{-3}).$$

$\square$
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 13.** The log-sum penalty is $\sqrt{n}^{\kappa/\nu}$-Lipschitz continuous.

**Proof.** As $g_i(w_i)$ is symmetric, assume $w_i \geq 0$ over which $g_i(w_i)$ is concave,

$$g(z) - g(w) \leq \nabla g(w)^T (z - w)$$

$$||g(z) - g(w)||_2 \leq ||\nabla g(w)||_2 ||z - w||_2.$$

Since $0 < \nabla g_i(w) = \frac{\kappa}{\nu + w_i} \leq \frac{\kappa}{\nu},$

$$||\nabla g(w)||_2 \leq \sqrt{n}^{\kappa/\nu}.$$


Smooth non-convex loss functions, which are known to be robust to outliers, include the sigmoid loss, Lorenz loss (Barbu et al., 2017), Savage loss (Masnadi-Shirazi and 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 14.** 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

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

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

We can see that \(v = 1\) simultaneously maximizes the first component and minimizes the second component of \(\hat{L}''(v)\), and so we conclude that
\[
||\hat{L}''(v)||_2 \leq ||\hat{L}''(1)||_2 = 2.
\]

Using the mean value theorem, for any \(v\) and \(u\),
\[
||\hat{L}'(v) - \hat{L}'(u)||_2 \leq ||\hat{L}''(1)||_2 ||(v - u)||_2.
\]

We now show that \(\mathcal{L}(v)\) is also 2-smooth. For \(v > 1\), \(\mathcal{L}'(v) = \hat{L}'(1) = 0\). Taking \(v > 1\) and \(u \leq 1\),
\[
||\mathcal{L}'(v) - \mathcal{L}'(u)||_2 = ||\hat{L}'(1) - \hat{L}'(u)||_2 \\
\leq 2||1 - u||_2 \\
\leq 2||v - u||_2.
\]

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

\[
\square
\]

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

**Property 15.** The Lorenz loss function is DC-decomposable,
\[
\mathcal{L}(v) = \mathcal{L}^1(v) - \mathcal{L}^2(v),
\]
where \(\mathcal{L}^1(v) = \frac{1}{4}v^2 + \mathcal{L}(v)\) and \(\mathcal{L}^2(v) = -\frac{1}{4}v^2\).

**Proof.** Taking the third derivative of \(\hat{L}(v)\) (28),
\[
\hat{L}'''(v) = \frac{-4(v - 1)}{(1 + (v - 1)^2)^2} \left( 3 - \frac{4(v - 1)^2}{1 + (v - 1)^2} \right),
\]
and already knowing that the maximum of \(\hat{L}''(v)\) lies at \(v = 1\) from the proof of Property 14, the minimum of \(\mathcal{L}''(v)\) lies at \(v' = 1 \pm \sqrt{3}\), with \(\hat{L}''(v') = -\frac{1}{4}\). Since \(\mathcal{L}''(v) \geq -\frac{1}{4}\), we write the DC decomposition of \(\mathcal{L}(v)\) as
\[
\mathcal{L}(v) = \mathcal{L}^1(v) - \mathcal{L}^2(v),
\]
where \(\mathcal{L}^1(v) = \frac{1}{4}v^2 + \mathcal{L}(v)\) and \(\mathcal{L}^2(v) = -\frac{1}{4}v^2\).  

\[
\square
\]
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 (Dheeru and Karra Taniskidou, 2017) and MNIST (LeCun, 1998), as used in Allen-Zhu and Hazan (2016; Li and Li, 2018; Reddi et al., 2016). 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 an iMac with a 3.4 GHz quad core Intel Core i5 processor and 8GB of RAM. We compare performance in terms of gradient calls, and for the variance reduced algorithms, we amortize the gradient calls used to calculate \( G_k \) over each iteration as done in (Li and Li, 2018). We tested over different experiment lengths, implementing each algorithm so that they will do at least \( e = 15 \) effective passes for a9a and \( e = 9 \) effective passes for MNIST. We compare performance in terms of the log of the objective value.

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 \( R = N \). In both MBSGA and VRSGA, \( N \) was chosen as small as possible while ensuring that the number of gradient calls was at least \( en \). The regularizer’s parameters were chosen as \( \kappa = \frac{1}{d} \) and \( \nu = 1 \). The only undetermined parameter of our algorithms is the upper bound \( \sigma \) used in MBSGA. This parameter was estimated by doing 50 iterations of MBSGA with step size \( \gamma = \frac{1}{L} \), using a different random seed than was used for the experiments, and computing the sample estimate \( \hat{\sigma} \) each iteration with the \( M \) samples used in the algorithm. An estimate of \( \sigma \) was then taken as \( \hat{\sigma} = \max_{k} \hat{\sigma}^k \).

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 by a paper that their work is based on. In the remainder of this paragraph we describe all chosen parameter values using the notation found in (Xu et al., 2018). The algorithm SSDC-SPG calls a stochastic proximal gradient (SPG) algorithm \( K \) times. For the \( k \)th iteration, the number of iterations of SPG equals \( T_k = 4k \). Each iteration of SPG uses one gradient call. We used the minimum \( K \) which ensured at least \( en \) gradient calls were used. The convex majorant parameter \( \gamma = 3L \), and the step size \( \eta_t = 1/(L(t + 1)) \). The Moreau envelope parameter \( \mu = \epsilon \), where \( K = O(1/\epsilon^4) \), is the only non-explicitly given parameter, which we set to \( \mu = 1/(K^{1/4}) \). SSDC-SVRG calls a stochastic variance reduced gradient (SVRG) algorithm \( K \) times. We set the inner loop length \( T_k = \max(2, 200L/\gamma) \), and the outer loop length \( S_k = \lceil \log_2(k) \rceil \). The step size \( \eta_k = 0.05/L \). Two parameters are not explicitly given, similar to in SSDC-SPG, we set \( \mu = 1/(K^{1/4}) \). For these parameter settings, there seems to be no restriction on \( \gamma \). Their SVRG algorithm is based off of the work of Xiao and Zhang (2014), where empirical testing of different sizes of \( T_k \) was done for a binary classification problem. The best performance was found with a choice of \( T_k = 2n \), from which we were able to determine \( \gamma \). Given \( \gamma \), we were then able to solve for \( K \), ensuring at least \( en \) gradient calls were used.
In Figures 1 and 2 we compare the performance of our algorithms with those of (Xu et al., 2018) on the datasets a9a and MNIST, respectively, where the y-axis is the logarithm of the objective value and the x-axis is the effective passes over the dataset. We observe that our algorithms are able to achieve better convergence using the same number of gradient calls in both experiments. We provide further experimental results in Table 2, which show the number of iterations and the amount of time in seconds which were required to do the described experiments. We conclude that our algorithms were able to obtain better solutions, compared to their counterparts of (Xu et al., 2018), while being at least one order of magnitude faster in terms of computation time. We also note that in contradiction to our theoretical results and those of (Xu et al., 2018), MBSGA and SSDC-SPG outperformed VRSGA and SSDC-SVRG. In this application, using a larger number of iterations with less accurate gradient estimates seems to perform better.
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, with our algorithms exhibiting superior convergence complexities 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 stochastic variance reduced algorithm. In
an empirical study we found our algorithms to obtain better solutions in significantly less
time. We attribute the superior performance of our algorithms to their more direct and
simple approach, avoiding the use of a DC algorithm, which requires optimizing not one,
but a sequence of approximating functions using a stochastic gradient algorithm. 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.

References

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

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

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

Emmanuel J Candes, Michael B Wakin, and Stephen P Boyd. Enhancing sparsity by
rewighted $l_1$ minimization. *Journal of Fourier analysis and applications*, 14(5-6):877–
905, 2008.

Olivier Chapelle, Chuong B Do, Choon H Teo, Quoc V Le, and Alex J Smola. Tighter
bounds for structured estimation. In *Advances in neural information processing systems*,
pages 281–288, 2009.

Francis Clarke. *Functional analysis, calculus of variations and optimal control*, volume 264.
Springer Science & Business Media, 2013.

Dua Dheeru and Efi Karra Taniskidou. UCI machine learning repository, 2017. URL
http://archive.ics.uci.edu/ml

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

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

Saeed Ghadimi and Guanghui Lan. Accelerated gradient methods for nonconvex nonlinear
and stochastic programming. *Mathematical Programming*, 156(1-2):59–99, 2016.

Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Mini-batch stochastic approximation
methods for nonconvex stochastic composite optimization. *Mathematical Programming*,
155(1-2):267–305, 2016.
Pinghua Gong, Changshui Zhang, Zhaosong Lu, Jianhua Huang, and Jieping Ye. A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. In *International Conference on Machine Learning*, pages 37–45, 2013.

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

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

Yann LeCun. The MNIST database of handwritten digits. [http://yann.lecun.com/exdb/mnist/], 1998.

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

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

Szymon Majewski, Błażej Miasojedow, and Eric Moulines. Analysis of nonsmooth stochastic approximation: the differential inclusion approach. *arXiv preprint arXiv:1805.01916*, 2018.

Hamed Masnadi-Shirazi and Nuno Vasconcelos. On the design of loss functions for classification: theory, robustness to outliers, and savageboost. In *Advances in neural information processing systems*, pages 1049–1056, 2009.

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

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

Sashank J Reddi, Suvrit Sra, Barnabás Póczos, and Alexander J Smola. Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization. In *Advances in Neural Information Processing Systems*, pages 1145–1153, 2016.

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

Xiaotong Shen, George C Tseng, Xuegong Zhang, and Wing Hung Wong. On ψ-learning. *Journal of the American Statistical Association*, 98(463):724–734, 2003.

Yichao Wu and Yufeng Liu. Robust truncated hinge loss support vector machines. *Journal of the American Statistical Association*, 102(479):974–983, 2007.
Lin Xiao and Tong Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

Yi Xu, Qi Qi, Qihang Lin, Rong Jin, and Tianbao Yang. Stochastic optimization for DC functions and non-smooth non-convex regularizers with non-asymptotic convergence. *arXiv preprint arXiv:1811.11829*, 2018.

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