Journal of Machine Learning Research 22 (2021) 1-36 Submitted 3/20; Published 5/21

Stochastic Proximal Methods for Non-Smooth Non-Convex Constrained Sparse Optimization*

Michael R. Metel MICHAELROS.METEL@RIKEN.JP
Center for Artificial Intelligence Project, RIKEN
Tokyo 103-0027, Japan

Akiko Takeda TAKEDA@MIST.I.U-TOKYO.AC.JP
Graduate School of Information Science and Technology
The University of Tokyo
Tokyo 103-0027, Japan
Center for Artificial Intelligence Project, RIKEN
Tokyo 103-0027, Japan

Editor: Tong Zhang

Abstract

This paper focuses on stochastic proximal gradient methods for optimizing a smooth non-convex loss function with a non-smooth non-convex regularizer and convex constraints. To the best of our knowledge we present the first non-asymptotic convergence bounds for this class of problem. We present two simple stochastic proximal gradient algorithms, for general stochastic and finite-sum optimization problems. In a numerical experiment we compare our algorithms with the current state-of-the-art deterministic algorithm and find our algorithms to exhibit superior convergence.

Keywords: stochastic optimization, non-convex optimization, non-smooth optimization, constrained optimization, sparse optimization

1. Introduction

In this paper we consider optimization problems of the form

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

where $f(w)$ has a Lipschitz continuous gradient and $h(w)$ is a proper closed convex function. The functions $f(w)$ and $g(w)$ can be non-convex, and $g(w)$ and $h(w)$ can be non-differentiable, but we assume that $g(w)$ and $h(w)$ have efficiently computable proximal operators. In addition, we assume that

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

is the expectation of a stochastic function $F(w,\xi)$, 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

*This paper is an extension of (Metel and Takeda, 2019), which was part of the 2019 ICML conference proceedings.

consider the finite-sum problem with

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

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

First-order stochastic methods for the case where the non-smooth non-convex function $g(w) = 0$ is an active research area. Non-asymptotic convergence bounds were first achieved in (Ghadimi et al., 2016). For finite-sum problems, Reddi et al. (2016) were the first to develop a proximal stochastic variance reduced gradient algorithm with an improved convergence complexity. The current state-of-the-art in terms of gradient call complexity (see Section 2) for the finite-sum problem is found in (Pham et al., 2019) using the ProxSARAH framework.

For the problem of solving (1) where $h(w) = 0$, the current body of research is limited given the non-convexity of $g(w)$. The first non-asymptotic convergence results for a non-smooth non-convex function $g(w)$ is found in (Xu et al., 2019b), where it is assumed that $f(w) = f^1(w) - f^2(w)$, where both $f^1(w)$ and $f^2(w)$ are convex, $f^1(w)$ is smooth, and $f^2(w)$ has a Hölder continuous gradient. The current best convergence complexity results are by Xu et al. (2019a), which are also state-of-the-art for the case of minimizing $f(w)$ as a general expectation (2) with $g(w) = 0$ and $h(w) \neq 0$. We are unaware of any existing non-asymptotic convergence results for the general problem setting of (1).

Given our generalized formulation, a wide range of applications can be dealt with. In many of which, a sparse solution is desirable as it avoids overfitting to sampled data, and simplifies the interpretation of the result and its implementation. Our motivation for $g(w)$ is to act as a non-smooth non-convex regularizer, such as SCAD (Fan and Li, 2001), MCP (Zhang et al., 2010), the log-sum penalty (Candes et al., 2008), or the capped l_1 norm, which are able to approximate the l_0 norm better than their convex or smooth counterparts. The function $h(w)$ allows us to include convex constraints to our problem through the use of an indicator function of the convex feasible region.

The function $f(w)$ is intended to model the objective of our optimization problem, such as a loss function in empirical risk minimization, an agent's utility function in portfolio optimization, or a statistical method performed on sample data, where non-convex smooth functions arise naturally. In order to show the importance of problem setting (1), three applications which fit our model's assumptions are investigated in Section 7, highlighting our proposed algorithms' usefulness in practice.

The subdifferential of sums of smooth and convex functions is well understood, as is the subdifferential of the sum of a smooth and a non-convex non-differentiable function. We have to contrast this with our problem setting, which is the sum of three functions where two are non-differentiable and two are non-convex. The subdifferential of this function does not have good calculus rules, and so it is not evident what form of convergence can be obtained. In order to analyze the convergence of algorithms for (1), we introduce a new convergence measure called the *subdifferential mapping*, for which we prove results bounding its distance to zero.

We present and analyze a mini-batch stochastic proximal algorithm (MBSPA) for general stochastic objectives of the form (2), and a variance reduced stochastic proximal algorithm (VRSPA) for finite-sum problems of the form (3). Convergence complexities for our algorithms, as well as the current best complexities for particular cases of our problem are summarized in Table 1. We are not aware of any other works proving non-asymptotic convergence bounds for our general problem setting. We implemented both algorithms and show superior convergence in practice compared to a state-of-the-art deterministic algorithm, the Successive Difference-of-Convex Approximation Method (SDCAM) (Liu et al., 2019).

Table 1: State-of-the-art convergence complexities for particular cases of (1). The convergence is measured in terms of the gradient mapping in (Pham et al., 2019), the subdifferential in (Xu et al., 2019a) and the subdifferential mapping in this paper. We note that the convergence complexities of SPGA for the non-finite-sum case are also state-of-art for when $h(w) \neq 0$ and $g(w) = 0$.

Algorithm	Reference	Finite -sum	$h(w) = 0$	$g(w) = 0$	Gradient Call Complexity	Proximal Operator Complexity
ProxSARAH	(Pham et al., 2019) Theorem 6	√	×	√	$O(n^{1/2}\epsilon^{-2})$	$O(n^{1/2}\epsilon^{-2})$
SPGA	(Xu et al., 2019a) Corollary 3	×	√	×	$O(\epsilon^{-3})$	$O(\epsilon^{-2})$
	(Xu et al., 2019a) Corollary 4	√	√	×	$O(n^{1/2}\epsilon^{-2} + n)$	$O(\epsilon^{-2})$
MBSPA	Corollary 10	×	×	×	$O(\epsilon^{-5})$	$O(\epsilon^{-3})$
VRSPA	Corollary 15	√	×	×	$O(n^{2/3}\epsilon^{-3})$	$O(\epsilon^{-3})$

2. Background

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

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

where $\hat{\partial}\Phi(w) := \{v : \liminf\limits_{x\to w, x\neq w} \frac{\Phi(x)-\Phi(w)-\langle v,x-w\rangle}{||x-w||_2} \geq 0\}$ and $w^k \xrightarrow{\Phi} w$ signifies the sequence $w^k \to w$ and $\Phi(w^k) \to \Phi(w)$. The limiting subdifferential is equal to the gradient and subdifferential when the function is continuously differentiable and proper convex, respectively. We also assume the proximal operators of $g(w)$ and $h(w)$ are nonempty for all w, and that they can be computed efficiently,

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

$$\text{prox}_{\gamma h}(w) := \underset{x\in\mathbb{R}^d}{\text{argmin}} \left\{ \frac{1}{2\gamma}||w-x||_2^2 + h(x) \right\}, \tag{6}$$

for $\lambda, \gamma > 0$. In particular, we denote an element of $\text{prox}_{\lambda g}(w)$ as

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

We note that $\text{prox}_{\gamma h}(w)$ maps to a singleton since $h(w)$ is proper, closed, and convex, see for example (Beck, 2017, Theorem 6.3).

We will measure algorithm complexity in terms of the number of gradient calls and proximal operations. 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. Subdifferential Mapping

For the analysis of non-asymptotic convergence bounds, the problem setting of (1) is of a more general form compared to objective functions previously studied. In order to prove bounds for our proposed algorithms, we introduce a new convergence measure, which we call the *subdifferential mapping*,

$$\mathcal{P}_\gamma(w,\mathcal{S}) := \left\{ \frac{1}{\gamma}\left(w - \text{prox}_{\gamma h}(w-\gamma s)\right) : s \in \mathcal{S} \right\},$$

where $\mathcal{S} \subseteq \mathbb{R}^d$ is the subdifferential of a function, which is a closed set wherever the function is finite (Rockafellar and Wets, 2009, Theorem 8.6). Defining

$$\mathcal{G}_\gamma(w) := \mathcal{P}_\gamma(w, \nabla f(w) + \partial g(w)),$$

we are interested in algorithm solutions $\bar{w}$, with accompanying $\bar{\gamma} > 0$, which satisfy

$$\mathbb{E}\left[\operatorname{dist}(0, \mathcal{G}_{\bar{\gamma}}(\bar{w}))\right] \leq \epsilon, \tag{8}$$

which we will call an ϵ-accurate solution. We will also use the notation $\mathcal{P}_{\gamma}(w, G)$, when $G \in \mathbb{R}^d$ is the gradient or a particular subgradient of a function in our analysis. $\mathcal{G}_{\gamma}(w)$ generalizes the gradient mapping $\mathcal{P}_{\gamma}(w, \nabla f(w))$ which has been used in the convergence criterion for proximal stochastic gradient methods for solving (1) with $g(w) = 0$, such as in (Ghadimi et al., 2016; Reddi et al., 2016; Li and Li, 2018). To motivate the measure of convergence (8), consider the case where

$$\operatorname{dist}(0, \mathcal{G}_{\bar{\gamma}}(\bar{w})) = 0.$$

This implies that there exists an element $s_g(\bar{w}) \in \partial g(\bar{w})$ such that

$$0 = \mathcal{P}_{\bar{\gamma}}(\bar{w}, \nabla f(\bar{w}) + s_g(\bar{w})), \tag{9}$$

and in particular

$$\bar{w} = \operatorname{prox}_{\bar{\gamma}h}(\bar{w} - \bar{\gamma}(\nabla f(\bar{w}) + s_g(\bar{w}))). \tag{10}$$

From the first order optimality condition of $\operatorname{prox}_{\bar{\gamma}h}(\bar{w} - \bar{\gamma}(\nabla f(\bar{w}) + s_g(\bar{w})))$ in (6),

$$0 \in -\mathcal{P}_{\bar{\gamma}}(\bar{w}, \nabla f(\bar{w}) + s_g(\bar{w})) + \nabla f(\bar{w}) + s_g(\bar{w}) + \partial h(\operatorname{prox}_{\bar{\gamma}h}(\bar{w} - \bar{\gamma}(\nabla f(\bar{w}) + s_g(\bar{w})))).$$

Applying (9) and (10),

$$0 \in \nabla f(\bar{w}) + \partial g(\bar{w}) + \partial h(\bar{w}).$$

We also note that in the case $g(w) = 0$, considering $\bar{w}^+ := \operatorname{prox}_{\bar{\gamma}h}(\bar{w} - \bar{\gamma}\nabla f(\bar{w}))$, it follows that (Drusvyatskiy and Paquette, 2019, Equation 4.1)

$$\operatorname{dist}(0, \partial\Phi(\bar{w}^+)) \leq (1 + L\bar{\gamma})||\mathcal{G}_{\bar{\gamma}}(\bar{w})||_2.$$

If (8) holds, then in expectation, $\bar{w}$ is a distance $\bar{\gamma}\epsilon$ away from a point which is an $(1+L\bar{\gamma})\epsilon$-stationary point.

4. Auxiliary Functions of $\Phi(w)$

The convergence analysis of our algorithms relies on a sequence of majorizing functions of

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

where $g(w)$ has been replaced by its Moreau envelope,

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

The Moreau envelope dates back to the 1960's, see for example (Rockafellar and Wets, 2009) for a thorough treatment of its mathematical properties and as well as its history. It has played a critical role in regularized optimization models, as well in the approximation of non-smooth non-convex functions. We recall in this section how the Moreau envelope can be decomposed into a difference of convex functions, which was first used in (Liu et al., 2019), as well as in (Xu et al., 2019b). For further theoretical developments and applications of the Moreau envelope for non-smooth non-convex sparse optimization, we refer readers to the papers (Laude et al., 2018; Liu and Yin, 2019; Shen et al., 2016) and the references therein.

Taking $x = w$, we note that

$$e_\lambda g(w) \leq g(w). \tag{11}$$

Expanding $\frac{1}{2\lambda}||w - x||_2^2$, the Moreau envelope can be written as

$$e_\lambda g(w) = \frac{1}{2\lambda}||w||_2^2 - D^\lambda(w), \tag{12}$$

where $D^\lambda(w) = \sup_{x\in\mathbb{R}^d}\left(\frac{1}{\lambda}w^T x - \frac{1}{2\lambda}||x||_2^2 - g(x)\right)$. 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 $D^\lambda(w)$. Given iteration w^k, a smooth majorizing function of $f(w) + e_\lambda g(w)$ can be written as

$$E^k_\lambda(w) := f(w) + U^k_\lambda(w), \tag{13}$$

where

$$U^k_\lambda(w) = \frac{1}{2\lambda}||w||_2^2 - \left(D^\lambda(w^k) + \frac{1}{\lambda}\zeta^\lambda(w^k)^\top(w - w^k)\right).$$

We will only need to evaluate the gradient of $E^k_\lambda(w)$, which is simply

$$\nabla E^k_\lambda(w) = \nabla f(w) + \frac{1}{\lambda}(w - \zeta^\lambda(w^k)). \tag{14}$$

The following property shows that $E^k_\lambda(w) + h(w)$ is a majorization of $\tilde{\Phi}_\lambda(w)$. The minimization a sequence of majorizations is a well established solution technique for difference of convex functions including DCA (Le Thi and Dinh, 2018) as well as CCCP (Sriperumbudur and Lanckriet, 2009).

Property 1 *The following holds for $E^k_\lambda(w)$:*

$$E^k_\lambda(w) + h(w) \geq \tilde{\Phi}_\lambda(w) \textit{ for all } w \in \mathbb{R}^d \tag{15}$$

$$E^k_\lambda(w^k) + h(w^k) = \tilde{\Phi}_\lambda(w^k) \tag{16}$$

$$E^k_\lambda(w) \textit{ is } L_\lambda := \left(L + \frac{1}{\lambda}\right) - \textit{smooth.} \tag{17}$$

Proof As what differs between $E^k_\lambda(w) + h(w)$ and $\tilde{\Phi}_\lambda(w)$ is only $U^k_\lambda(w)$ and $e_\lambda g(w)$, we will show that (15) and (16) hold between these two terms.

(15): As found in (Liu et al., 2019), for any $w, z \in \mathbb{R}^d$,

$$\begin{aligned}D^\lambda(w) - D^\lambda(z) =& \sup_{x\in\mathbb{R}^d}\left(\frac{1}{\lambda}w^\top x - \frac{1}{2\lambda}||x||_2^2 - g(x)\right) - \sup_{x\in\mathbb{R}^d}\left(\frac{1}{\lambda}z^\top x - \frac{1}{2\lambda}||x||_2^2 - g(x)\right)\\ \geq& \frac{1}{\lambda}w^\top\zeta^\lambda(z) - \frac{1}{2\lambda}||\zeta^\lambda(z)||_2^2 - g(\zeta^\lambda(z)) - \left(\frac{1}{\lambda}z^\top\zeta^\lambda(z) - \frac{1}{2\lambda}||\zeta^\lambda(z)||_2^2 - g(\zeta^\lambda(z))\right)\\ =& \frac{1}{\lambda}\zeta^\lambda(z)^\top(w-z).\end{aligned}$$

Setting $z = w^k$,

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

(16): $U_\lambda^k(w^k) = \frac{1}{2\lambda}||w^k||_2^2 - D^\lambda(w^k) = e_\lambda g(w^k)$ from (12).

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

■

5. Mini-Batch Stochastic Proximal Algorithm

The algorithm MBSPA presented in this section makes use of

$$\nabla A_{\lambda,M}^k(w,\xi^k) := \frac{1}{M}\sum_{j=1}^M \nabla F(w,\xi_j^k) + \frac{1}{\lambda}(w - \zeta^\lambda(w^k)), \tag{18}$$

which is a stochastic version of $\nabla E_\lambda^k(w)$, replacing $\nabla f(w)$ with an unbiased estimate using M samples ξ_j^k, $j = 1, ..., M$ in iteration k. The optimal values for parameters α and θ of MBSPA in terms of convergence complexity are given in Corollary 10.

5.1 Convergence analysis

The convergence analysis of MBSPA follows the technique of Ghadimi et al. (2016) adapted to our problem. We first define the following gradient mappings in iteration k,

$$\mathcal{G}_{\gamma,A}^k(w^k) := \mathcal{P}_\gamma(w^k, \nabla A_{\lambda,M}^k(w^k,\xi^k))$$

and

$$\mathcal{G}_{\gamma,E}^k(w^k) := \mathcal{P}_\gamma(w^k, \nabla E_\lambda^k(w^k)).$$

Algorithm 1 Mini-Batch Stochastic Proximal Algorithm (MBSPA)

```
**Input:** $w^1 \in \mathbb{R}^d$, $N \in \mathbb{Z}_{>0}$, $\alpha, \theta \in \mathbb{R}$
$M := \lceil N^\alpha \rceil$, $\lambda = \frac{1}{N^\theta}$
$L_\lambda = L + \frac{1}{\lambda}$, $\gamma = \frac{1}{L_\lambda}$
$R \sim$ uniform$\{1, ..., N\}$
**for** $k = 1, 2, ..., R-1$ **do**
  $\zeta^\lambda(w^k) \in \text{prox}_{\lambda g}(w^k)$
  Sample $\xi^k \sim P^M$
  Compute $\nabla A^k_{\lambda,M}(w^k, \xi^k)$ (18)
  $w^{k+1} = \text{prox}_{\gamma h}(w^k - \gamma \nabla A^k_{\lambda,M}(w^k, \xi^k))$
**end for**
**Output:** $\bar{w}^R \in \text{prox}_{\lambda g}(w^R)$
```

We also note that

$$\begin{aligned} w^{k+1} &= \text{prox}_{\gamma h}(w^k - \gamma \nabla A^k_{\lambda,M}(w^k, \xi^k)) \\ &= w^k - \gamma \left(\frac{1}{\gamma} \left(w^k - \text{prox}_{\gamma h}(w^k - \gamma \nabla A^k_{\lambda,M}(w^k, \xi^k)) \right) \right) \\ &= w^k - \gamma \mathcal{G}^k_{\gamma,A}(w^k). \end{aligned} \tag{19}$$

In order to offer some intuition for Algorithm 1, we analyze its convergence for minimizing $\Phi(w)$ when $f(w) = h(w) = 0$ and $g(w) = |w|$, and $f(w) = g(w) = 0$ and $h(w) = |w|$ for $w \in \mathbb{R}$ in the appendix.

The following lemma bounds $\mathbb{E}\left[||\mathcal{G}^R_{\gamma,E}(w^R)||_2^2 \right]$, which will be used to bound $\mathbb{E}\left[\text{dist}(0, \mathcal{G}_{\bar{\gamma}}(\bar{w}^R)) \right]$ in Theorem 9.

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

$$\mathbb{E}||\mathcal{G}^R_{\gamma,E}(w^R)||_2^2 \leq \frac{(L + N^\theta)}{N} \tilde{\Delta} + \frac{6}{\lceil N^\alpha \rceil} \sigma^2,$$

where $\tilde{\Delta} = 4(\tilde{\Phi}_\lambda(w^1) - \tilde{\Phi}_\lambda(w^*_\lambda))$ and w^*_λ is a global minimizer of $\tilde{\Phi}_\lambda(\cdot)$.

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

Property 3 *For any, including random $w \in \mathbb{R}^d$, it holds that*

$$\mathbb{E}||\nabla A^k_{\lambda,M}(w, \xi^k) - \nabla E^k_\lambda(w)||_2^2 \leq \frac{\sigma^2}{M}.$$

Proof From (14) and (18), $\nabla A^k_{\lambda,M}(w,\xi^k)-\nabla E^k_\lambda(w)=\frac{1}{M}\sum_{j=1}^M\nabla F(w,\xi^k_j)-\nabla f(w)$. Taking the expectation of its squared norm,

$$\begin{aligned}\mathbb{E}||\nabla A^k_{\lambda,M}(w,\xi^k)-\nabla E^k_\lambda(w)||_2^2 &= \mathbb{E}||\frac{1}{M}\sum_{j=1}^M(\nabla F(w,\xi^k_j)-\nabla f(w))||_2^2\\ &= \frac{1}{M^2}\mathbb{E}(\mathbb{E}[||\sum_{j=1}^M(\nabla F(w,\xi^k_j)-\nabla f(w))||_2^2|w])\\ &= \frac{1}{M^2}\mathbb{E}\sum_{i=1}^n\mathbb{E}[(\sum_{j=1}^M\nabla F(w,\xi^k_j)_i-\nabla f(w)_i)^2|w].\end{aligned}$$

For $j\neq l$, $\nabla F(w,\xi^k_j)_i-\nabla f(w)_i$ and $\nabla F(w,\xi^k_l)_i-\nabla f(w)_i$ are conditionally independent random variables with zero mean with respect to w. It follows that

$$\begin{aligned}&\mathbb{E}[(\nabla F(w,\xi^k_j)_i-\nabla f(w)_i)(\nabla F(w,\xi^k_l)_i-\nabla f(w)_i)|w]\\ =&\mathbb{E}[(\nabla F(w,\xi^k_j)_i-\nabla f(w)_i)|w]\mathbb{E}[(\nabla F(w,\xi^k_l)_i-\nabla f(w)_i)|w]\\ =&0,\end{aligned}$$

and

$$\begin{aligned}\frac{1}{M^2}\mathbb{E}\sum_{i=1}^n\mathbb{E}[(\sum_{j=1}^M\nabla F(w,\xi^k_j)_i-\nabla f(w)_i)^2|w] &= \frac{1}{M^2}\mathbb{E}\sum_{i=1}^n\mathbb{E}[\sum_{j=1}^M(\nabla F(w,\xi^k_j)_i-\nabla f(w)_i)^2|w]\\ &= \frac{1}{M^2}\mathbb{E}\sum_{i=1}^n\sum_{j=1}^M(\nabla F(w,\xi^k_j)_i-\nabla f(w)_i)^2\\ &= \frac{1}{M^2}\sum_{j=1}^M\mathbb{E}||\nabla F(w,\xi^k_j)-\nabla f(w)||_2^2\\ &\leq \frac{\sigma^2}{M}\end{aligned}$$

using (5). ■

The following property can be found in (Ghadimi et al., 2016). The notation we use is somewhat different, so we have included the proof in the appendix for clarity.

Property 4 (Ghadimi et al., 2016, Lemma 1) *Let $w,s\in\mathbb{R}^d$ and $\gamma>0$, then*

$$-\langle s,\mathcal{P}_\gamma(w,s)\rangle\leq\frac{1}{\gamma}\left(h(w)-h(\mathrm{prox}_{\gamma h}(w-\gamma s))\right)-||\mathcal{P}_\gamma(w,s)||_2^2.$$

Property 5 *For any $w\in\mathbb{R}^d$, it holds that*

$$|||\mathcal{G}^k_{\gamma,A}(w)-\mathcal{G}^k_{\gamma,E}(w)||_2\leq||\nabla A^k_{\lambda,M}(w,\xi^k)-\nabla E^k_\lambda(w)||_2.$$

Proof

$$\begin{aligned}
&||\mathcal{G}^k_{\gamma,A}(w) - \mathcal{G}^k_{\gamma,E}(w)||_2\\
=&||\frac{1}{\gamma}\left(w - \operatorname{prox}_{\gamma h}(w - \gamma\nabla A^k_{\lambda,M}(w,\xi^k))\right) - \frac{1}{\gamma}\left(w - \operatorname{prox}_{\gamma h}(w - \gamma\nabla E^k_\lambda(w))\right)||_2\\
=&\frac{1}{\gamma}||\operatorname{prox}_{\gamma h}(w - \gamma\nabla E^k_\lambda(w)) - \operatorname{prox}_{\gamma h}(w - \gamma\nabla A^k_{\lambda,M}(w,\xi^k))||_2\\
\leq&\frac{1}{\gamma}||w - \gamma\nabla E^k_\lambda(w) - w + \gamma\nabla A^k_{\lambda,M}(w,\xi^k)||_2\\
=&||\nabla A^k_{\lambda,M}(w,\xi^k)) - \nabla E^k_\lambda(w))||_2,
\end{aligned}$$

where the inequality holds due to the nonexpansivity of the proximal operator of proper closed convex functions (Beck, 2017, Theorem 6.42). ■

Proof of Lemma 2 Given the smoothness of $E^k_\lambda(w)$ as shown in Property 1,

$$\begin{aligned}
E^k_\lambda(w^{k+1}) &\leq E^k_\lambda(w^k) + \langle\nabla E^k_\lambda(w^k), w^{k+1} - w^k\rangle + \frac{L_\lambda}{2}||w^{k+1} - w^k||_2^2\\
&= E^k_\lambda(w^k) + \langle\nabla E^k_\lambda(w^k), -\gamma\mathcal{G}^k_{\gamma,A}(w^k)\rangle + \frac{L_\lambda}{2}||\gamma\mathcal{G}^k_{\gamma,A}(w^k)||_2^2,
\end{aligned}$$

from (19). Let $\delta_k := \nabla A^k_{\lambda,M}(w^k,\xi^k) - \nabla E^k_\lambda(w^k)$, then

$$E^k_\lambda(w^{k+1}) \leq E^k_\lambda(w^k) - \gamma\langle\nabla A^k_{\lambda,M}(w^k,\xi^k), \mathcal{G}^k_{\gamma,A}(w^k)\rangle + \gamma\langle\delta_k, \mathcal{G}^k_{\gamma,A}(w^k)\rangle + \frac{L_\lambda}{2}||\gamma\mathcal{G}^k_{\gamma,A}(w^k)||_2^2.$$

Using Property 4 with $w = w^k$ and $s = \nabla A^k_{\lambda,M}(w^k,\xi^k)$,

$$E^k_\lambda(w^{k+1}) \leq E^k_\lambda(w^k) + h(w^k) - h(w^{k+1}) - \gamma||\mathcal{G}^k_{\gamma,A}(w^k)||_2^2 + \gamma\langle\delta_k, \mathcal{G}^k_{\gamma,A}(w^k)\rangle + \frac{L_\lambda\gamma^2}{2}||\mathcal{G}^k_{\gamma,A}(w^k)||_2^2.$$

Applying (15) and (16),

$$\begin{aligned}
\tilde{\Phi}_\lambda(w^{k+1}) &\leq \tilde{\Phi}_\lambda(w^k) - \gamma||\mathcal{G}^k_{\gamma,A}(w^k)||_2^2 + \gamma\langle\delta_k, \mathcal{G}^k_{\gamma,A}(w^k)\rangle + \frac{L_\lambda\gamma^2}{2}||\mathcal{G}^k_{\gamma,A}(w^k)||_2^2\\
&= \tilde{\Phi}_\lambda(w^k) + \left(\frac{L_\lambda\gamma^2}{2} - \gamma\right)||\mathcal{G}^k_{\gamma,A}(w^k)||_2^2 + \gamma\langle\delta_k, \mathcal{G}^k_{\gamma,E}(w^k)\rangle + \gamma\langle\delta_k, \mathcal{G}^k_{\gamma,A}(w^k) - \mathcal{G}^k_{\gamma,E}(w^k)\rangle\\
&\leq \tilde{\Phi}_\lambda(w^k) + \left(\frac{L_\lambda\gamma^2}{2} - \gamma\right)||\mathcal{G}^k_{\gamma,A}(w^k)||_2^2 + \gamma\langle\delta_k, \mathcal{G}^k_{\gamma,E}(w^k)\rangle + \gamma||\delta_k||_2||\mathcal{G}^k_{\gamma,A}(w^k) - \mathcal{G}^k_{\gamma,E}(w^k)||_2\\
&\leq \tilde{\Phi}_\lambda(w^k) + \left(\frac{L_\lambda\gamma^2}{2} - \gamma\right)||\mathcal{G}^k_{\gamma,A}(w^k)||_2^2 + \gamma\langle\delta_k, \mathcal{G}^k_{\gamma,E}(w^k)\rangle + \gamma||\delta_k||_2^2,
\end{aligned}$$

where the last inequality uses Property 5. After N iterations,

$$\begin{aligned}
\left(\gamma - \frac{L_\lambda\gamma^2}{2}\right)\sum_{k=1}^N ||\mathcal{G}^k_{\gamma,A}(w^k)||_2^2 &\leq \tilde{\Phi}_\lambda(w^1) - \tilde{\Phi}_\lambda(w^{N+1}) + \gamma\sum_{k=1}^N\left(\langle\delta_k, \mathcal{G}^k_{\gamma,E}(w^k)\rangle + ||\delta_k||_2^2\right)\\
&\leq \tilde{\Phi}_\lambda(w^1) - \tilde{\Phi}_\lambda(w^*) + \gamma\sum_{k=1}^N\left(\langle\delta_k, \mathcal{G}^k_{\gamma,E}(w^k)\rangle + ||\delta_k||_2^2\right). \quad (20)
\end{aligned}$$

It follows from (4) that for ξ^k independent of w, $\mathbb{E}[\nabla A^k_{\lambda,M}(w,\xi^k)|w] = \nabla E^k_\lambda(w)$, and so $\mathbb{E}[\delta_k|w^k] = 0$. Taking the expectation of both sides of (20),

$$\begin{aligned}\left(\gamma - \frac{L_\lambda}{2}\gamma^2\right)\sum_{k=1}^N \mathbb{E}||\mathcal{G}^k_{\gamma,A}(w^k)||_2^2 \leq& \tilde{\Phi}_\lambda(w^1) - \tilde{\Phi}_\lambda(w^*) + \gamma\sum_{k=1}^N \mathbb{E}\left(\langle\delta_k, \mathcal{G}^k_{\gamma,E}(w^k)\rangle + ||\delta_k||_2^2\right)\\ =& \tilde{\Phi}_\lambda(w^1) - \tilde{\Phi}_\lambda(w^*) + \gamma\sum_{k=1}^N \left(\mathbb{E}(\mathbb{E}[\langle\delta_k, \mathcal{G}^k_{\gamma,E}(w^k)\rangle|w^k]) + \mathbb{E}||\delta_k||_2^2\right)\\ =& \tilde{\Phi}_\lambda(w^1) - \tilde{\Phi}_\lambda(w^*) + \gamma\sum_{k=1}^N \left(\mathbb{E}(\langle\mathbb{E}[\delta_k|w^k], \mathcal{G}^k_{\gamma,E}(w^k)\rangle) + \mathbb{E}||\delta_k||_2^2\right)\\ =& \tilde{\Phi}_\lambda(w^1) - \tilde{\Phi}_\lambda(w^*) + \gamma\sum_{k=1}^N \mathbb{E}||\delta_k||_2^2\\ \leq& \tilde{\Phi}_\lambda(w^1) - \tilde{\Phi}_\lambda(w^*) + \gamma\frac{N}{M}\sigma^2,\end{aligned}$$

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

$$\begin{aligned}\mathbb{E}||\mathcal{G}^R_{\gamma,A}(w^R)||_2^2 &= \frac{1}{N}\sum_{k=1}^N \mathbb{E}||\mathcal{G}^k_{\gamma,A}(w^k)||_2^2\\ &\leq \frac{1}{N\left(\gamma - \frac{L_\lambda}{2}\gamma^2\right)}\left(\tilde{\Phi}_\lambda(w^1) - \tilde{\Phi}_\lambda(w^*) + \gamma\frac{N}{M}\sigma^2\right)\\ &= \frac{2L_\lambda}{N}\left(\tilde{\Phi}_\lambda(w^1) - \tilde{\Phi}_\lambda(w^*)\right) + \frac{2}{M}\sigma^2,\end{aligned}$$

where the final equality holds since $\gamma = \frac{1}{L_\lambda}$.

$$\begin{aligned}\mathbb{E}||\mathcal{G}^R_{\gamma,E}(w^R)||_2^2 &= \mathbb{E}||\mathcal{G}^R_{\gamma,A}(w^R) + \mathcal{G}^R_{\gamma,E}(w^R) - \mathcal{G}^R_{\gamma,A}(w^R)||_2^2\\ &= \mathbb{E}\left(||\mathcal{G}^R_{\gamma,A}(w^R)||_2^2 + 2\langle\mathcal{G}^R_{\gamma,A}(w^R), \mathcal{G}^R_{\gamma,E}(w^R) - \mathcal{G}^R_{\gamma,A}(w^R)\rangle + ||\mathcal{G}^R_{\gamma,E}(w^R) - \mathcal{G}^R_{\gamma,A}(w^R)||_2^2\right)\\ &\leq 2\mathbb{E}||\mathcal{G}^R_{\gamma,A}(w^R)||_2^2 + 2\mathbb{E}||\mathcal{G}^R_{\gamma,E}(w^R) - \mathcal{G}^R_{\gamma,A}(w^R)||_2^2\\ &\leq \frac{4L_\lambda}{N}\left(\tilde{\Phi}_\lambda(w^1) - \tilde{\Phi}_\lambda(w^*)\right) + \frac{4}{M}\sigma^2 + 2\mathbb{E}||\nabla A^R_{\lambda,M}(w^R,\xi^R) - \nabla E^R_\lambda(w^R)||_2^2\\ &\leq \frac{4L_\lambda}{N}\left(\tilde{\Phi}_\lambda(w^1) - \tilde{\Phi}_\lambda(w^*)\right) + \frac{4}{M}\sigma^2 + \frac{2}{M}\sigma^2\\ &= \frac{4L_\lambda}{N}\left(\tilde{\Phi}_\lambda(w^1) - \tilde{\Phi}_\lambda(w^*)\right) + \frac{6}{M}\sigma^2\\ &= \frac{(L+N^\theta)}{N}\tilde{\Delta} + \frac{6}{\lceil N^\alpha\rceil}\sigma^2,\end{aligned}$$

where the first inequality uses Young's inequality on the middle term. ■

In order to prove the convergence of $\mathbb{E}\left[\text{dist}(0, \mathcal{G}_{\bar{\gamma}}(\bar{w}^R)\right]$, we require the following three properties. The next property can be found in (Beck, 2017). Our assumptions are slightly different, so we have included the proof in the appendix for clarity.

Property 6 (Beck, 2017, Theorem 10.9) *For $\gamma^1 \geq \gamma^2 > 0$ and any $w, s \in \mathbb{R}^d$,*

$$||\mathcal{P}_{\gamma^1}(w,s)||_2 \leq ||\mathcal{P}_{\gamma^2}(w,s)||_2.$$

Property 7 *Assume that $g(w)$ is Lipschitz continuous with parameter l and $\bar{\gamma} \geq \gamma$, then*

$$\mathrm{dist}(0, \mathcal{G}_{\bar{\gamma}}(\zeta^\lambda(w^k))) \leq ||\mathcal{G}^k_{\gamma,E}(w^k)||_2 + 2l\lambda\left(\frac{2}{\bar{\gamma}} + L\right).$$

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

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

It follows that

$$\begin{aligned}
&\mathrm{dist}(0, \mathcal{G}_{\bar{\gamma}}(\zeta^\lambda(w^k)))\\
\leq &||\mathcal{P}_{\bar{\gamma}}(\zeta^\lambda(w^k), \nabla f(\zeta^\lambda(w^k)) + \frac{1}{\lambda}(w^k - \zeta^\lambda(w^k)))||_2\\
= &||\mathcal{P}_{\bar{\gamma}}(\zeta^\lambda(w^k), \nabla f(\zeta^\lambda(w^k)) + \frac{1}{\lambda}(w^k - \zeta^\lambda(w^k))) + \mathcal{P}_{\bar{\gamma}}(w^k, \nabla E^k_\lambda(w^k)) - \mathcal{P}_{\bar{\gamma}}(w^k, \nabla E^k_\lambda(w^k))||_2\\
\leq &||\mathcal{P}_{\bar{\gamma}}(w^k, \nabla E^k_\lambda(w^k))||_2 + ||\mathcal{P}_{\bar{\gamma}}(\zeta^\lambda(w^k), \nabla f(\zeta^\lambda(w^k)) + \frac{1}{\lambda}(w^k - \zeta^\lambda(w^k))) - \mathcal{P}_{\bar{\gamma}}(w^k, \nabla E^k_\lambda(w^k))||_2.
\end{aligned}$$

Given that $||\mathcal{P}_{\bar{\gamma}}(w^k, \nabla E^k_\lambda(w^k))||_2 \leq ||\mathcal{P}_{\gamma}(w^k, \nabla E^k_\lambda(w^k))||_2 = ||\mathcal{G}^k_{\gamma,E}(w^k)||_2$ from Property 6,

$$\begin{aligned}
&\mathrm{dist}(0, \mathcal{G}_{\bar{\gamma}}(\zeta^\lambda(w^k)))\\
\leq &||\mathcal{G}^k_{\gamma,E}(w^k)||_2 + ||\mathcal{P}_{\bar{\gamma}}(\zeta^\lambda(w^k), \nabla f(\zeta^\lambda(w^k)) + \frac{1}{\lambda}(w^k - \zeta^\lambda(w^k))) - \mathcal{P}_{\bar{\gamma}}(w^k, \nabla E^k_\lambda(w^k))||_2.
\end{aligned} \tag{21}$$

Focusing on the second term,

$$\begin{aligned}
&||\mathcal{P}_{\bar{\gamma}}(\zeta^\lambda(w^k), \nabla f(\zeta^\lambda(w^k)) + \frac{1}{\lambda}(w^k - \zeta^\lambda(w^k))) - \mathcal{P}_{\bar{\gamma}}(w^k, \nabla E^k_\lambda(w^k))||_2\\
= &\Bigg\|\frac{1}{\bar{\gamma}}\left(\zeta^\lambda(w^k) - \mathrm{prox}_{\bar{\gamma}h}(\zeta^\lambda(w^k) - \bar{\gamma}(\nabla f(\zeta^\lambda(w^k)) + \frac{1}{\lambda}(w^k - \zeta^\lambda(w^k))))\right)\\
&-\frac{1}{\bar{\gamma}}\left(w^k - \mathrm{prox}_{\bar{\gamma}h}(w^k - \bar{\gamma}(\nabla f(w^k) + \frac{1}{\lambda}(w^k - \zeta^\lambda(w^k))))\right)\Bigg\|_2\\
\leq &\frac{1}{\bar{\gamma}}||\zeta^\lambda(w^k) - w^k||_2 + \frac{1}{\bar{\gamma}}||\,\mathrm{prox}_{\bar{\gamma}h}(w^k - \bar{\gamma}(\nabla f(w^k) + \frac{1}{\lambda}(w^k - \zeta^\lambda(w^k))))\\
&- \mathrm{prox}_{\bar{\gamma}h}(\zeta^\lambda(w^k) - \bar{\gamma}(\nabla f(\zeta^\lambda(w^k)) + \frac{1}{\lambda}(w^k - \zeta^\lambda(w^k))))||_2\\
\leq &\frac{1}{\bar{\gamma}}||\zeta^\lambda(w^k) - w^k||_2 + \frac{1}{\bar{\gamma}}||w^k - \bar{\gamma}(\nabla f(w^k) + \frac{1}{\lambda}(w^k - \zeta^\lambda(w^k)))\\
&- (\zeta^\lambda(w^k) - \bar{\gamma}(\nabla f(\zeta^\lambda(w^k)) + \frac{1}{\lambda}(w^k - \zeta^\lambda(w^k))))||_2\\
\leq &\frac{2}{\bar{\gamma}}||\zeta^\lambda(w^k) - w^k||_2 + ||\nabla f(\zeta^\lambda(w^k)) - \nabla f(w^k)||_2\\
\leq &\frac{2}{\bar{\gamma}}||\zeta^\lambda(w^k) - w^k||_2 + L||\zeta^\lambda(w^k) - w^k||_2,
\end{aligned} \tag{22}$$

where the second inequality follows from the nonexpansivity of the proximal operator. In order to bound $||\zeta^\lambda(w^k) - w^k||_2$, recall from (11) 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 of $g(w)$,

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

and dividing both sides by $\frac{1}{2\lambda}||w - \zeta^\lambda(w)||_2$,

$$||w - \zeta^\lambda(w)||_2 \leq 2l\lambda. \tag{23}$$

Using (21)-(23),

$$\begin{aligned} \mathrm{dist}(0, \mathcal{G}_{\bar{\gamma}}(\zeta^\lambda(w^k))) \leq& ||\mathcal{G}_{\gamma,E}^k(w^k)||_2 + \frac{2}{\bar{\gamma}}||\zeta^\lambda(w^k) - w^k||_2 + L||\zeta^\lambda(w^k) - w^k||_2 \\ \leq& ||\mathcal{G}_{\gamma,E}^k(w^k)||_2 + 2l\lambda\left(\frac{2}{\bar{\gamma}} + L\right). \end{aligned}$$

■

Property 8 *Let w^* be a global minimizer of $\Phi(\cdot)$ and let w^*_λ be a global minimizer of $\tilde{\Phi}_\lambda(\cdot)$. Assume that $g(w)$ is Lipschitz continuous with parameter l, then*

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

Proof

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

where the first inequality follows from (11). For any w,

$$-e_\lambda g(w) = -\frac{1}{2\lambda}||w - \zeta^\lambda(w)||_2^2 - g(\zeta^\lambda(w))$$

Adding $g(w)$ to both sides,

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

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

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

■

Theorem 9 *Assume that $g(w)$ is Lipschitz continuous with parameter l and $\bar{\gamma} = \frac{1}{N^\tau}$ for $\tau \leq \theta$. The output $\bar{w}^R$ of MBSPA satisfies*

$$\mathbb{E}\left[\text{dist}(0, \mathcal{G}_{\bar{\gamma}}(\bar{w}^R))\right] \leq \sqrt{\frac{(L+N^\theta)}{N}\left(\Delta + \frac{2l^2}{N^\theta}\right)} + \sqrt{\frac{6\sigma^2}{\lceil N^\alpha \rceil}} + \frac{2l}{N^\theta}\left(2N^\tau + L\right),$$

where $\Delta = 4(\Phi(w^1) - \Phi(w^*))$ and w^* is a global minimizer of $\Phi(\cdot)$.

Proof We first verify that $\bar{\gamma} = \frac{1}{N^\tau} \geq \frac{1}{N^\theta} \geq \frac{1}{L+N^\theta} = \gamma$. From Property 7, taking $\zeta^\lambda(w^R) = \bar{w}^R$,

$$\text{dist}(0, \mathcal{G}_{\bar{\gamma}}(\bar{w}^R)) \leq ||\mathcal{G}^R_{\gamma,E}(w^R)||_2 + 2l\lambda\left(\frac{2}{\bar{\gamma}} + L\right).$$

Taking its expectation,

$$\begin{aligned}\mathbb{E}\left[\text{dist}(0, \mathcal{G}_{\bar{\gamma}}(\bar{w}^R))\right] &\leq \mathbb{E}[||\mathcal{G}^R_{\gamma,E}(w^R)||_2] + 2l\lambda\left(\frac{2}{\bar{\gamma}} + L\right) \\ &\leq \sqrt{\mathbb{E}\left[||\mathcal{G}^R_{\gamma,E}(w^R)||_2^2\right]} + \frac{2l}{N^\theta}\left(2N^\tau + L\right) \\ &\leq \sqrt{\frac{(L+N^\theta)}{N}\tilde{\Delta}} + \sqrt{\frac{6\sigma^2}{\lceil N^\alpha \rceil}} + \frac{2l}{N^\theta}\left(2N^\tau + L\right),\end{aligned}$$

where the second inequality uses Jensen's inequality and the third inequality follows from Lemma 2. The result then follows using Property 8 as

$$\begin{aligned}\tilde{\Delta} = 4(\tilde{\Phi}_\lambda(w^1) - \tilde{\Phi}_\lambda(w^*_\lambda)) &\leq 4(\Phi(w^1) - \Phi(w^*)) + 2l^2\lambda \\ &= \Delta + \frac{2l^2}{N^\theta}.\end{aligned}$$

■

Having bounded the expected distance of $\mathcal{G}_{\bar{\gamma}}(\bar{w}^R)$ from the origin, we prove an ϵ-accurate point convergence complexity.

Corollary 10 *Assume that $g(w)$ is Lipschitz continuous with parameter l. To obtain an ϵ-accurate solution (8) using MBSPA, the gradient call complexity is $O(\epsilon^{-5})$ and the proximal operator complexity is $O(\epsilon^{-3})$ choosing $\theta = \frac{1}{3}$, $\alpha = \frac{2}{3}$, and $\tau = 0$.*

Proof From Theorem 9,

$$\begin{aligned}\mathbb{E}\left[\operatorname{dist}(0, \mathcal{G}_{\bar{\gamma}}(\bar{w}^R))\right] \leq & \sqrt{\frac{(L+N^\theta)}{N}\left(\Delta + \frac{2l^2}{N^\theta}\right)} + \sqrt{\frac{6\sigma^2}{\lceil N^\alpha \rceil}} + \frac{2l}{N^\theta}(2+L) \\ = & O(N^{0.5\theta-0.5}) + O(N^{-0.5\alpha}) + O(N^{-\theta}).\end{aligned}$$

Setting $\theta = \frac{1}{3}$ and $\alpha = \frac{2}{3}$,

$$\mathbb{E}\left[\operatorname{dist}(0, \mathcal{G}_{\bar{\gamma}}(\bar{w}^R))\right] \leq O(N^{-\frac{1}{3}}).$$

An ϵ-accurate solution will require less than $N = O(\epsilon^{-3})$ iterations. Two proximal operations are required per iteration, which establishes the proximal operator complexity of $O(\epsilon^{-3})$. The number of gradient calls per iteration is $\lceil N^\alpha \rceil = O(\epsilon^{-2})$. The number of gradient calls to get an ϵ-accurate solution is then bounded by

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

■

6. Variance Reduced Stochastic Proximal Algorithm for Finite-sum Problems

In this section we assume that $f(w)$ takes the form of (3). This is a common problem setting when optimizing over a collected dataset, such as when doing empirical risk minimization. The algorithm MBSPA applies for this setting as well, but in this section we analyze a variance reduced method, VRSPA, which in addition takes advantage of the finite-sum structure of $f(w)$ to acheive a better convergence complexity compared to MBSPA, see Table 1. The optimal values for parameters α and θ of VRSPA are given in Corollary 15.

6.1 Convergence analysis

We require the function $E_\lambda^{k,t}(w)$ in our convergence analysis, which is constructed in the same way as $E_\lambda^k(w)$ (13), but using w_t^k instead of w^k. This function possesses the same characteristics as found in Property 1. In addition, let

$$\mathcal{G}_{\gamma,E}^{k,t}(w_t^k) := \mathcal{P}_\gamma(w_t^k, \nabla E_\lambda^{k,t}(w_t^k)).$$

The convergence analysis follows the work of Li and Li (2018) adapted to our problem.

Lemma 11 *For an initial value $\tilde{w}_1$, $N \in \mathbb{Z}_{>0}$, and $\alpha, \theta \in \mathbb{R}$, VRSGA generates w_T^R satisfying the following bound.*

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

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

Algorithm 2 Variance reduced stochastic proximal algorithm (VRSPA)

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

In order to prove this result, we require the following properties. Property 12 can be found in (Li and Li, 2018). We have included the proof in our notation in the appendix.

Property 12 (Li and Li, 2018, Lemma 1) *Consider arbitrary $w, s, z \in \mathbb{R}^d$, and $w^+ = \text{prox}_{\gamma h}(w - \gamma s)$,*

$$E_\lambda^{k,t}(w^+) + h(w^+) \leq E_\lambda^{k,t}(z) + h(z) + \langle \nabla E_\lambda^{k,t}(w) - s, w^+ - z \rangle + \frac{L_\lambda}{2}||w^+ - w||_2^2 + \frac{L_\lambda}{2}||z - w||_2^2 - \frac{1}{\gamma}\langle w^+ - w, w^+ - z \rangle.$$

Property 13 *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

$$\begin{aligned} ||w - x||_2^2 &= ||w - z + z - x||_2^2 \\ &\leq (||w - z||_2 + ||z - x||_2)^2 \\ &= ||w - z||_2^2 + 2||w - z||_2||z - x||_2 + ||z - x||_2^2 \\ &\leq ||w - z||_2^2 + \left(\beta||w - z||_2^2 + \frac{1}{\beta}||z - x||_2^2\right) + ||z - x||_2^2 \\ &= (1 + \beta)||w - z||_2^2 + \left(1 + \frac{1}{\beta}\right)||z - x||_2^2, \end{aligned}$$

where the second inequality uses Young's inequality. ■

Proof of Lemma 11 Let $\hat{w}_{t+1}^k = \text{prox}_{\gamma h}(w_t^k - \gamma \nabla E_\lambda^{k,t}(w_t^k))$, with $w^+ = w_{t+1}^k$, $w = w_t^k$, $s = V_t^k$, and $z = \hat{w}_{t+1}^k$ in Property 12 to get the inequality

$$\begin{aligned} E_\lambda^{k,t}(w_{t+1}^k) + h(w_{t+1}^k) \leq & E_\lambda^{k,t}(\hat{w}_{t+1}^k) + h(\hat{w}_{t+1}^k) + \langle \nabla E_\lambda^{k,t}(w_t^k) - V_t^k, w_{t+1}^k - \hat{w}_{t+1}^k \rangle + \\ & \frac{L_\lambda}{2}||w_{t+1}^k - w_t^k||_2^2 + \frac{L_\lambda}{2}||\hat{w}_{t+1}^k - w_t^k||_2^2 - \frac{1}{\gamma}\langle w_{t+1}^k - w_t^k, w_{t+1}^k - \hat{w}_{t+1}^k \rangle. \end{aligned} \tag{24}$$

In addition, let $w^+ = \hat{w}_{t+1}^k$, $w = w_t^k$, $s = \nabla E_\lambda^{k,t}(w_t^k)$, and $z = w_t^k$ in Property 12 to get

$$\begin{aligned} E_\lambda^{k,t}(\hat{w}_{t+1}^k) + h(\hat{w}_{t+1}^k) \leq & E_\lambda^{k,t}(w_t^k) + h(w_t^k) + \langle \nabla E_\lambda^{k,t}(w_t^k) - \nabla E_\lambda^{k,t}(w_t^k), \hat{w}_{t+1}^k - w_{t+1}^k \rangle \\ & + \frac{L_\lambda}{2}||\hat{w}_{t+1}^k - w_t^k||_2^2 + \frac{L_\lambda}{2}||w_t^k - w_t^k||_2^2 - \frac{1}{\gamma}\langle \hat{w}_{t+1}^k - w_t^k, \hat{w}_{t+1}^k - w_t^k \rangle \\ = & E_\lambda^{k,t}(w_t^k) + h(w_t^k) + \left(\frac{L_\lambda}{2} - \frac{1}{\gamma}\right)||\hat{w}_{t+1}^k - w_t^k||_2^2. \end{aligned} \tag{25}$$

Adding (24) and (25),

$$\begin{aligned} E_\lambda^{k,t}(w_{t+1}^k) + h(w_{t+1}^k) \leq & E_\lambda^{k,t}(w_t^k) + h(w_t^k) + \langle \nabla E_\lambda^{k,t}(w_t^k) - V_t^k, w_{t+1}^k - \hat{w}_{t+1}^k \rangle + \frac{L_\lambda}{2}||w_{t+1}^k - w_t^k||_2^2 \\ & - \frac{1}{\gamma}\langle w_{t+1}^k - w_t^k, w_{t+1}^k - \hat{w}_{t+1}^k \rangle + \left(L_\lambda - \frac{1}{\gamma}\right)||\hat{w}_{t+1}^k - w_t^k||_2^2. \end{aligned}$$

From (15) and (16),

$$\begin{aligned} \tilde{\Phi}_\lambda(w_{t+1}^k) \leq & \tilde{\Phi}_\lambda(w_t^k) + \langle \nabla E_\lambda^{k,t}(w_t^k) - V_t^k, w_{t+1}^k - \hat{w}_{t+1}^k \rangle + \frac{L_\lambda}{2}||w_{t+1}^k - w_t^k||_2^2 \\ & - \frac{1}{\gamma}\langle w_{t+1}^k - w_t^k, w_{t+1}^k - \hat{w}_{t+1}^k \rangle + \left(L_\lambda - \frac{1}{\gamma}\right)||\hat{w}_{t+1}^k - w_t^k||_2^2. \end{aligned} \tag{26}$$

Plugging $\langle w_{t+1}^k - w_t^k, w_{t+1}^k - \hat{w}_{t+1}^k \rangle = \frac{1}{2}\left(||w_{t+1}^k - w_t^k||_2^2 + ||w_{t+1}^k - \hat{w}_{t+1}^k||_2^2 - ||\hat{w}_{t+1}^k - w_t^k||_2^2\right)$ into (26) and rearranging,

$$\begin{aligned} \tilde{\Phi}_\lambda(w_{t+1}^k) \leq & \tilde{\Phi}_\lambda(w_t^k) + \langle \nabla E_\lambda^{k,t}(w_t^k) - V_t^k, w_{t+1}^k - \hat{w}_{t+1}^k \rangle + \left(\frac{L_\lambda}{2} - \frac{1}{2\gamma}\right)||w_{t+1}^k - w_t^k||_2^2 \\ & - \frac{1}{2\gamma}||w_{t+1}^k - \hat{w}_{t+1}^k||_2^2 + \left(L_\lambda - \frac{1}{2\gamma}\right)||\hat{w}_{t+1}^k - w_t^k||_2^2. \end{aligned} \tag{27}$$

Focusing on the term $-\frac{1}{2\gamma}||w_{t+1}^k - \hat{w}_{t+1}^k||_2^2$, we apply Property 13 with $w = w_{t+1}^k$, $x = w_t^k$, and $z = \hat{w}_{t+1}^k$. After rearranging,

$$-(1+\beta)||w_{t+1}^k - \hat{w}_{t+1}^k||_2^2 \leq -||w_{t+1}^k - w_t^k||_2^2 + \left(1 + \frac{1}{\beta}\right)||\hat{w}_{t+1}^k - w_t^k||_2^2.$$

Dividing both sides by $2\gamma(1+\beta)$,

$$-\frac{1}{2\gamma}||w_{t+1}^k - \hat{w}_{t+1}^k||_2^2 \leq -\frac{1}{2\gamma(1+\beta)}||w_{t+1}^k - w_t^k||_2^2 + \frac{\left(1+\frac{1}{\beta}\right)}{2\gamma(1+\beta)}||\hat{w}_{t+1}^k - w_t^k||_2^2.$$

Choosing $\beta = 3$,

$$-\frac{1}{2\gamma}||w_{t+1}^k - \hat{w}_{t+1}^k||_2^2 \leq -\frac{1}{8\gamma}||w_{t+1}^k - w_t^k||_2^2 + \frac{1}{6\gamma}||\hat{w}_{t+1}^k - w_t^k||_2^2.$$

Using this inequality in (27),

$$\begin{aligned}
\tilde{\Phi}_\lambda(w_{t+1}^k) \leq & \tilde{\Phi}_\lambda(w_t^k) + \langle \nabla E_\lambda^{k,t}(w_t^k) - V_t^k, w_{t+1}^k - \hat{w}_{t+1}^k \rangle + \left(\frac{L_\lambda}{2} - \frac{1}{2\gamma}\right)||w_{t+1}^k - w_t^k||_2^2 \\
& - \frac{1}{8\gamma}||w_{t+1}^k - w_t^k||_2^2 + \frac{1}{6\gamma}||\hat{w}_{t+1}^k - w_t^k||_2^2 + \left(L_\lambda - \frac{1}{2\gamma}\right)||\hat{w}_{t+1}^k - w_t^k||_2^2 \\
= & \tilde{\Phi}_\lambda(w_t^k) + \langle \nabla E_\lambda^{k,t}(w_t^k) - V_t^k, w_{t+1}^k - \hat{w}_{t+1}^k \rangle + \left(\frac{L_\lambda}{2} - \frac{5}{8\gamma}\right)||w_{t+1}^k - w_t^k||_2^2 \\
& + \left(L_\lambda - \frac{1}{3\gamma}\right)||\hat{w}_{t+1}^k - w_t^k||_2^2 \\
\leq & \tilde{\Phi}_\lambda(w_t^k) + \gamma||\nabla E_\lambda^{k,t}(w_t^k) - V_t^k||_2^2 + \left(\frac{L_\lambda}{2} - \frac{5}{8\gamma}\right)||w_{t+1}^k - w_t^k||_2^2 \\
& + \left(L_\lambda - \frac{1}{3\gamma}\right)||\hat{w}_{t+1}^k - w_t^k||_2^2, \qquad (28)
\end{aligned}$$

where the last inequality holds since

$$\begin{aligned}
& \langle \nabla E_\lambda^{k,t}(w_t^k) - V_t^k, w_{t+1}^k - \hat{w}_{t+1}^k \rangle \\
\leq & ||\nabla E_\lambda^{k,t}(w_t^k) - V_t^k||_2||w_{t+1}^k - \hat{w}_{t+1}^k||_2 \\
= & ||\nabla E_\lambda^{k,t}(w_t^k) - V_t^k||_2||\operatorname{prox}_{\gamma h}(w_t^k - \gamma V_t^k) - \operatorname{prox}_{\gamma h}(w_t^k - \gamma \nabla E_\lambda^{k,t}(w_t^k))||_2 \\
\leq & \gamma||\nabla E_\lambda^{k,t}(w_t^k) - V_t^k||_2^2
\end{aligned}$$

using the Cauchy-Schwarz inequality and the nonexpansivity of the proximal operator of h. Taking the expectation of both sides of (28),

$$\begin{aligned}
\mathbb{E}\tilde{\Phi}_\lambda(w_{t+1}^k) \leq & \mathbb{E}\left[\tilde{\Phi}_\lambda(w_t^k) + \gamma||\nabla E_\lambda^{k,t}(w_t^k) - V_t^k||_2^2 + \left(\frac{L_\lambda}{2} - \frac{5}{8\gamma}\right)||w_{t+1}^k - w_t^k||_2^2\right. \\
& \left. + \left(L_\lambda - \frac{1}{3\gamma}\right)||\hat{w}_{t+1}^k - w_t^k||_2^2\right]. \qquad (29)
\end{aligned}$$

Focusing on $\mathbb{E}\left[||\nabla E_{\lambda}^{k,t}(w_t^k) - V_t^k||_2^2\right]$, from (14) and the definition of V_t^k found in Algorithm 2, $\nabla E_{\lambda}^{k,t}(w_t^k) - V_t^k = \nabla f(w_t^k) - (\frac{1}{b}\sum_{j\in I}\left(\nabla f_j(w_t^k) - \nabla f_j(\tilde{w}^k)\right) + G^k)$. After rearranging,

$$\begin{aligned}\mathbb{E}||\nabla E_{\lambda}^{k,t}(w_t^k) - V_t^k||_2^2 &= \mathbb{E}||\frac{1}{b}\sum_{j\in I}(\nabla f_j(\tilde{w}^k) - \nabla f_j(w_t^k) - G^k + \nabla f(w_t^k))||_2^2\\ &= \frac{1}{b^2}\mathbb{E}(\mathbb{E}[||\sum_{j\in I}(\nabla f_j(\tilde{w}^k) - \nabla f_j(w_t^k) - G^k + \nabla f(w_t^k))||_2^2|\tilde{w}^k, w_t^k])\\ &= \frac{1}{b^2}\sum_{j\in I}\mathbb{E}||\nabla f_j(\tilde{w}^k) - \nabla f_j(w_t^k) - G^k + \nabla f(w_t^k)||_2^2\\ &\leq \frac{1}{b^2}\sum_{j\in I}\mathbb{E}||\nabla f_j(\tilde{w}^k) - \nabla f_j(w_t^k)||_2^2\\ &\leq \frac{L^2}{b}\mathbb{E}||\tilde{w}^k - w_t^k||_2^2.\end{aligned}$$

The random variables $\nabla f_j(\tilde{w}^k) - \nabla f_j(w_t^k) - \left(G^k - \nabla f(w_t^k)\right)$ for $j \in I$ are conditionally independent with zero mean with respect to $\tilde{w}^k$ and w_t^k, and so the third equality holds using the same reasoning found in Property 3. The first inequality holds since $\mathbb{E}||x - \mathbb{E}[x]||_2^2 \leq \mathbb{E}||x||_2^2$ for any random variable x. Using this bound in (29),

$$\begin{aligned}\mathbb{E}\tilde{\Phi}_{\lambda}(w_{t+1}^k) \leq& \mathbb{E}\left[\tilde{\Phi}_{\lambda}(w_t^k) + \gamma\frac{L^2}{b}||\tilde{w}^k - w_t^k||_2^2 + \left(\frac{L_{\lambda}}{2} - \frac{5}{8\gamma}\right)||w_{t+1}^k - w_t^k||_2^2\right.\\ &\left. + \left(L_{\lambda} - \frac{1}{3\gamma}\right)||\hat{w}_{t+1}^k - w_t^k||_2^2\right]\\ \leq& \mathbb{E}\left[\tilde{\Phi}_{\lambda}(w_t^k) + \frac{L_{\lambda}}{6b}||\tilde{w}^k - w_t^k||_2^2 - \frac{13L_{\lambda}}{4}||w_{t+1}^k - w_t^k||_2^2 - L_{\lambda}||\hat{w}_{t+1}^k - w_t^k||_2^2\right]\\ =& \mathbb{E}\left[\tilde{\Phi}_{\lambda}(w_t^k) + \frac{L_{\lambda}}{6b}||\tilde{w}^k - w_t^k||_2^2 - \frac{13L_{\lambda}}{4}||w_{t+1}^k - w_t^k||_2^2 - \frac{1}{36L_{\lambda}}||\mathcal{G}_{\gamma E}^{k,t}(w_t^k)||_2^2\right],\end{aligned} \tag{30}$$

where the last two lines use $\gamma = \frac{1}{6L_{\lambda}}$. Focusing on $-\frac{13L_{\lambda}}{4}||w_{t+1}^k - w_t^k||_2^2$, we apply Property 13 with $w = w_{t+1}^k$, $x = \tilde{w}^k$, and $z = w_t^k$,

$$(1+\beta)||w_{t+1}^k - w_t^k||_2^2 \geq ||w_{t+1}^k - \tilde{w}^k||_2^2 - \left(1 + \frac{1}{\beta}\right)||w_t^k - \tilde{w}^k||_2^2$$

$$-\frac{13L_{\lambda}}{4}||w_{t+1}^k - w_t^k||_2^2 \leq -\frac{13L_{\lambda}}{4(1+\beta)}||w_{t+1}^k - \tilde{w}^k||_2^2 + \frac{13L_{\lambda}\left(1 + \frac{1}{\beta}\right)}{4(1+\beta)}||w_t^k - \tilde{w}^k||_2^2.$$

Setting $\beta = 2t - 1$,

$$-\frac{13L_{\lambda}}{4}||w_{t+1}^k - w_t^k||_2^2 \leq -\frac{13L_{\lambda}}{8t}||w_{t+1}^k - \tilde{w}^k||_2^2 + \frac{13L_{\lambda}}{8t-4}||w_t^k - \tilde{w}^k||_2^2.$$

Applying this bound in (30),

$$\mathbb{E}\tilde{\Phi}_\lambda(w^k_{t+1}) \leq \mathbb{E}\left[\tilde{\Phi}_\lambda(w^k_t) + \left(\frac{L_\lambda}{6b} + \frac{13L_\lambda}{8t-4}\right)||\tilde{w}^k - w^k_t||_2^2 - \frac{13L_\lambda}{8t}||w^k_{t+1} - \tilde{w}^k||_2^2 \right. \\ \left. - \frac{1}{36L_\lambda}||\mathcal{G}^{k,t}_{\gamma,E}(w^k_t)||_2^2\right].$$

Summing over t,

$$\mathbb{E}\tilde{\Phi}_\lambda(w^k_{m+1}) \leq \mathbb{E}\left[\tilde{\Phi}_\lambda(w^k_1) + \sum_{t=1}^{m}\left(\frac{L_\lambda}{6b} + \frac{13L_\lambda}{8t-4}\right)||\tilde{w}^k - w^k_t||_2^2 \right. \\ \left. - \sum_{t=1}^{m}\frac{13L_\lambda}{8t}||w^k_{t+1} - \tilde{w}^k||_2^2 - \frac{1}{36L_\lambda}\sum_{t=1}^{m}||\mathcal{G}^{k,t}_{\gamma,E}(w^k_t)||_2^2\right].$$

As $\tilde{w}^k = w^k_1$ and $||w^k_{m+1} - \tilde{w}^k||_2^2 \geq 0$,

$$\begin{aligned}
&\mathbb{E}\tilde{\Phi}_\lambda(w^k_{m+1}) \\
\leq&\mathbb{E}\left[\tilde{\Phi}_\lambda(w^k_1) + \sum_{t=2}^{m}\left(\frac{L_\lambda}{6b} + \frac{13L_\lambda}{8t-4}\right)||\tilde{w}^k - w^k_t||_2^2\right. \\
&\left. - \sum_{t=1}^{m-1}\frac{13L_\lambda}{8t}||w^k_{t+1} - \tilde{w}^k||_2^2 - \frac{1}{36L_\lambda}\sum_{t=1}^{m}||\mathcal{G}^{k,t}_{\gamma,E}(w^k_t)||_2^2\right] \\
=&\mathbb{E}\left[\tilde{\Phi}_\lambda(w^k_1) + \sum_{t=1}^{m-1}\left(\frac{L_\lambda}{6b} + \frac{13L_\lambda}{8t+4} - \frac{13L_\lambda}{8t}\right)||w^k_{t+1} - \tilde{w}^k||_2^2 - \frac{1}{36L_\lambda}\sum_{t=1}^{m}||\mathcal{G}^{k,t}_{\gamma,E}(w^k_t)||_2^2\right] \\
\leq&\mathbb{E}\left[\tilde{\Phi}_\lambda(w^k_1) + \sum_{t=1}^{m-1}\left(\frac{L_\lambda}{6b} - \frac{L_\lambda}{2t^2}\right)||w^k_{t+1} - \tilde{w}^k||_2^2 - \frac{1}{36L_\lambda}\sum_{t=1}^{m}||\mathcal{G}^{k,t}_{\gamma,E}(w^k_t)||_2^2\right] \\
\leq&\mathbb{E}\left[\tilde{\Phi}_\lambda(w^k_1) - \frac{1}{36L_\lambda}\sum_{t=1}^{m}||\mathcal{G}^{k,t}_{\gamma,E}(w^k_t)||_2^2\right],
\end{aligned}$$

where the last inequality holds since $6b = 6m^2 > 2(m-1)^2 \geq 2t^2$ for $t = 1, ..., m-1$. This summation can be equivalently written as

$$\mathbb{E}\tilde{\Phi}_\lambda(\tilde{w}^{k+1}) \leq \mathbb{E}\tilde{\Phi}_\lambda(\tilde{w}^k) - \mathbb{E}\left[\frac{1}{36L_\lambda}\sum_{t=1}^{m}||\mathcal{G}^{k,t}_{\gamma,E}(w^k_t)||_2^2\right].$$

Rearranging,

$$\mathbb{E}\left[\frac{1}{36L_\lambda}\sum_{t=1}^{m}||\mathcal{G}^{k,t}_{\gamma,E}(w^k_t)||_2^2\right] \leq \mathbb{E}\tilde{\Phi}_\lambda(\tilde{w}^k) - \mathbb{E}\tilde{\Phi}_\lambda(\tilde{w}^{k+1}),$$

summing over k,

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{36L_\lambda}\sum_{k=1}^{S}\sum_{t=1}^{m}||\mathcal{G}^{k,t}_{\gamma,E}(w^k_t)||_2^2\right] &\leq \tilde{\Phi}_\lambda(\tilde{w}^1) - \mathbb{E}\tilde{\Phi}_\lambda(\tilde{w}^{S+1}) \\
&\leq \tilde{\Phi}_\lambda(\tilde{w}^1) - \tilde{\Phi}_\lambda(w^*_\lambda),
\end{aligned}$$

and multiplying both sides by $\frac{36L_\lambda}{Sm}$,

$$\begin{aligned}\mathbb{E}\left[||\mathcal{G}^{R,T}_{\gamma,E}(w^R_T)||_2^2\right] &\leq \frac{36L_\lambda\left(\tilde{\Phi}_\lambda(\tilde{w}^1)-\tilde{\Phi}_\lambda(w^*_\lambda)\right)}{Sm}\\ &=\tilde{\Delta}\frac{L+(Sm)^\theta}{Sm}.\end{aligned}$$

■

We now prove the convergence of VRSPA in terms of the subdifferential mapping.

Theorem 14 *Assume that $g(w)$ is Lipschitz continuous with parameter l and $\bar{\gamma}=\frac{1}{N^\tau}$ for $\tau\leq\theta$. The output $\bar{w}^R_T$ of VRSPA satisfies the following inequality.*

$$\mathbb{E}\left[\mathrm{dist}(0,\mathcal{G}_{\bar{\gamma}}(\bar{w}^R_T))\right]\leq\sqrt{\frac{(L+(Sm)^\theta)\left(\Delta+18l^2(Sm)^{-\theta}\right)}{Sm}}+\frac{2l}{(Sm)^\theta}\left(2N^\tau+L\right),$$

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

Proof We check that $\bar{\gamma}=\frac{1}{N^\tau}\geq\frac{1}{(Sm)^\tau}\geq\frac{1}{(Sm)^\theta}\geq\frac{1}{6(L+(Sm)^\theta)}=\gamma$. From Property 7,

$$\mathrm{dist}(0,\mathcal{G}_{\bar{\gamma}}(\bar{w}^R_T))\leq||\mathcal{G}^{R,T}_{\gamma,E}(w^R_T)||_2+2l\lambda\left(\frac{2}{\bar{\gamma}}+L\right).$$

Taking its expectation,

$$\begin{aligned}\mathbb{E}\left[\mathrm{dist}(0,\mathcal{G}_{\bar{\gamma}}(\bar{w}^R_T)))\right]&\leq\mathbb{E}[||\mathcal{G}^{R,T}_{\gamma,E}(w^R_T)||_2]+2l\lambda\left(\frac{2}{\bar{\gamma}}+L\right)\\ &\leq\sqrt{\mathbb{E}\left[||\mathcal{G}^{R,T}_{\gamma,E}(w^R_T)||_2^2\right]}+\frac{2l}{(Sm)^\theta}\left(2N^\tau+L\right)\\ &\leq\sqrt{\frac{(L+(Sm)^\theta)\left(\Delta+18l^2(Sm)^{-\theta}\right)}{Sm}}+\frac{2l}{(Sm)^\theta}\left(2N^\tau+L\right),\end{aligned}$$

where the third inequality follows from Lemma 11 and Property 8,

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

■

Corollary 15 *Assume that $g(w)$ is Lipschitz continuous with parameter l. To obtain an ϵ-accurate solution* (8) *using VRSPA, the gradient call complexity is $O(n^{\frac{2}{3}}\epsilon^{-3})$ and the proximal operator complexity is $O(\epsilon^{-3})$ choosing $\alpha=\theta=\frac{1}{3}$, and $\tau=0$.*

Proof From Theorem 14,

$$\mathbb{E}\left[\operatorname{dist}(0, \mathcal{G}_{\bar{\gamma}}(\bar{w}_T^R))\right] \leq \sqrt{\frac{\left(L + (Sm)^{\frac{1}{3}}\right)\left(\Delta + 18l^2(Sm)^{-\frac{1}{3}}\right)}{Sm}} + \frac{2l}{(Sm)^{\frac{1}{3}}}(2+L)$$
$$= O((Sm)^{-\frac{1}{3}}).$$

An ϵ-accurate solution will require at most $Sm = O(\epsilon^{-3})$ iterations. Two proximal operations are required each iteration, giving a proximal operator complexity of $O(\epsilon^{-3})$. The number of gradient calls after Sm iterations is

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

■

7. Applications

We now present three examples of sparse constrained optimization problems which fit within our assumptions.

7.1 Sparse Binary Classification with Outlier Detection and Fairness Constraints

We are given training data $\{x, y\}$ where $x = \{x^1, x^2, ..., x^n\}$, $x^j \in \mathbb{R}^{d'}$ is the feature set, and $y = \{y^1, y^2, ..., y^n\}$, $y^j \in \{-1, 1\}$ is the label set. In the application of classifying people, there may be sensitive attributes such as race or sex. Even if a sensitive attribute $x_a \in x$ is removed from the feature set, our predictions may still be correlated to it, resulting in our model disproportionately treating a subset of the population unfairly. This is remedied by bounding the covariance between the sensitive attribute x_a and the model output as done in (Zafar et al., 2017),

$$\frac{1}{n}\left|\sum_{j=1}^{n}(x_a^j - \bar{x}_a)v^\top x_{-a}^j\right| \leq c,$$

where $\bar{x}_a$ is the mean of x_a, x_{-a}^j is the j^{th} feature vector with the sensitive attribute removed, $v^\top x_{-a}^j$ is our model output using decision variables $v \in \mathbb{R}^{d'}$, and $c > 0$ determines the maximum covariance tolerated.

We consider the smoothed 0-1 loss of Zhao et al. (2010) as our loss function,

$$\mathcal{L}(u) = \begin{cases} 0 & \text{if } u > 1 \\ \frac{1}{4}u^3 - \frac{3}{4}u + \frac{1}{2} & \text{if } -1 \leq u \leq 1 \\ 1 & \text{otherwise} . \end{cases} \tag{31}$$

We implement outlier detection by the mean-shift method, modifying our prediction to $v^\top x^j_{-a} + z^j$, using decision variables $z \in \mathbb{R}^n$ to reduce the loss incurred by outliers. It was shown in (She and Owen, 2011) that the l_1 norm is not effective as a penalizer of z when multiple outliers are present, which motivates the use of a non-convex regularizer. As all of the regularizers considered for $g(w)$ in Section 1 are separable, we are able to take $g(w) = g^1(v)+g^2(z)$, and are free to use different regularizers for v and z. The classification problem is then solved by the following minimization,

$$\begin{aligned}\min_{v,z}\ & \frac{1}{n}\sum_{j=1}^{n} \mathcal{L}(y^j(v^\top x^j_{-a} + z^j)) + g^1(v) + g^2(z)\\ \text{s.t.}\ & \frac{1}{n}\left|\sum_{j=1}^{n}(x^j_a - \bar{x}_a)v^\top x^j_{-a}\right| \le c.\end{aligned}$$

The feasible region of decision variables v can be rewritten as

$$C = \left\{ v : \begin{matrix} \hat{x}^\top v \le c \\ -\hat{x}^\top v \le c \end{matrix} \right\},$$

where $\hat{x} = \frac{1}{n}\sum_{j=1}^{n}(x^j_a - \bar{x}_a)x^j_{-a}$.

Property 16 *The projection of v onto C equals*

$$\text{proj}_C(v) = \begin{cases} v - \frac{\hat{x}^\top v - c}{||\hat{x}||_2^2}\hat{x} & \textit{if } \hat{x}^\top v > c \\ v - \frac{\hat{x}^\top v + c}{||\hat{x}||_2^2}\hat{x} & \textit{if } -\hat{x}^\top v > c \\ v & \textit{else.} \end{cases}$$

Proof The projection of v onto C can be found solving the following minimization problem,

$$\begin{aligned}\min_{\bar{v}} & \frac{1}{2}||\bar{v} - v||_2^2 \\ \text{s.t.}\quad & \hat{x}^\top \bar{v} \le c \\ & -\hat{x}^\top \bar{v} \le c,\end{aligned} \tag{32}$$

using its KKT conditions,

$$\begin{aligned}& \bar{v} = v + (\mu_2 - \mu_1)\hat{x} && \text{(Stationarity)}\\ & \hat{x}^T\bar{v} \le c, \quad -\hat{x}^T\bar{v} \le c && \text{(Primal feasibility)}\\ & \mu_1, \mu_2 \ge 0 && \text{(Dual feasibility)}\\ & \mu_1(\hat{x}^T\bar{v} - c) = 0, \quad \mu_2(-\hat{x}^T\bar{v} - c) = 0. && \text{(Complementary slackness)}\end{aligned}$$

We consider three scenarios for v, the first being when $\hat{x}^Tv > c$, where an optimal soluton is $\mu_1 = \frac{\hat{x}^Tv - c}{||\hat{x}||_2^2}$, $\mu_2 = 0$, and $\bar{v} = v - \mu_1\hat{x}$. The second scenario is when $-\hat{x}^Tv > c$, where we set $\mu_1 = 0$, $\mu_2 = -\frac{\hat{x}^Tv + c}{||\hat{x}||_2^2}$, and $\bar{v} = v + \mu_2\hat{x}$. Finally, when v is feasible in (32), we set $\mu_1 = \mu_2 = 0$, giving $\bar{v} = v$. ■

We take $g^1(v) := \sum_{i=1}^{d'} g_i^1(v_i)$ equal to MCP, where for $\kappa_1, \nu_1 > 0$,

$$\begin{aligned} g_i^1(v_i) &= \kappa_1 \int_0^{|v_i|} \max\left(0, 1 - \frac{u}{\nu_1 \kappa_1}\right) du \\ &= \begin{cases} \kappa_1 |v_i| - \frac{v_i^2}{2\nu_1} & \text{if } |v_i| \leq \nu_1 \kappa_1 \\ \nu_1 \kappa_1^2/2 & \text{if } |v_i| > \nu_1 \kappa_1. \end{cases} \end{aligned} \tag{33}$$

Property 17 *$g^1(v)$ is $\kappa_1\sqrt{d'}$-Lipschitz continuous.*

Proof Assume $v_i \geq 0$, over which $g_i^1(v_i)$ is differentiable for $v_i > 0$ with $\left|\frac{dg_i^1}{dv_i}(v_i)\right| \leq \kappa_1$. Using the mean value theorem, for $v_i' \geq 0$, $|g_i^1(v_i') - g_i^1(v_i)| \leq \kappa_1 |v_i' - v_i|$. Given the symmetry of $g_i^1(v_i)$, this bound holds for all $v_i', v_i \in \mathbb{R}$, and

$$\begin{aligned} |g^1(v') - g^1(v)| &= \left|\sum_{i=1}^{d'} (g_i^1(v_i') - g_i^1(v_i))\right| \\ &\leq \sum_{i=1}^{d'} |g_i^1(v_i') - g_i^1(v_i)| \\ &\leq \kappa_1 \sum_{i=1}^{d'} |v_i' - v_i| \\ &\leq \kappa_1 \sqrt{d'} ||v' - v||_2 \end{aligned}$$

■

As considered in (She and Owen, 2011), we set $g^2(z)$ equal to SCAD, which is also separable. For $\kappa_2 > 0$ and $\nu_2 > 2$,

$$\begin{aligned} g_i^2(z_i) &= \kappa_2 \int_0^{|z_i|} \min\left(1, \frac{\max(0, \nu_2\kappa_2 - u)}{(\nu_2 - 1)\kappa_2}\right) du \\ &= \begin{cases} \kappa_2 |z_i| & \text{if } |z_i| \leq \kappa_2 \\ \frac{-z_i^2 + 2\nu_2\kappa_2|z_i| - \kappa_2^2}{2(\nu_2 - 1)} & \text{if } \kappa_2 < |z_i| \leq \nu_2\kappa_2 \\ (\nu_2 + 1)\kappa_2^2/2 & \text{if } |z_i| > \nu_2\kappa_2. \end{cases} \end{aligned}$$

Similarly to MCP, SCAD is symmetric and $|\operatorname{dist}(0, \partial g_i^2(z_i))| \leq \kappa_2$. Using the same reasoning as in the proof of Property 17, we get the following property.

Property 18 *$g^2(z)$ is $\kappa_2\sqrt{n}$-Lipschitz continuous.*

For the closed form solutions of the proximal operators of MCP and SCAD see (Gong et al., 2013).

Property 19 *The function* $\frac{1}{n}\sum_{j=1}^{n}\mathcal{L}(y^j(v^\top x_{-a}^j + z^j))$*, with* $\mathcal{L}(\cdot)$ *as defined in* (31)*, is* $\frac{3}{2n}\sum_{j=1}^{n}||[(x_{-a}^j)^\top, 1]||_2^2$*-smooth in* v *and* z*.*

Proof Taking the second derivative, $|\mathcal{L}''(u)| \leq \frac{3}{2}$. Using the mean value theorem, $|\mathcal{L}'(u) - \mathcal{L}'(t)| \leq \frac{3}{2}|u-t|$. Composing $\mathcal{L}(u)$ with the affine function $y^j(v^\top x_{-a}^j + z^j)$, the resulting function is $\frac{3}{2}||[(x_{-a}^j)^\top, 1]||_2^2$-smooth (Shalev-Shwartz and Ben-David, 2014, Claim 12.9). We conclude that $\frac{1}{n}\sum_{j=1}^{n}\mathcal{L}(y^j(v^\top x_{-a}^j + z^j))$ is $\frac{3}{2n}\sum_{j=1}^{n}||[(x_{-a}^j)^\top, 1]||_2^2$-smooth. ■

For the following two applications, it is assumed $g(w)$ is taken as either MCP or SCAD for simplicity.

7.2 Sparse Portfolio Optimization using S-Shaped Utility with Loss Aversion

We assume there are d risky assets with stochastic returns r_i, $i = 1, ..., d$, and an investor desires to place a fraction w_i of their wealth into each asset i. Finding a sparse portfolio is desirable as trading fewer assets generally results in lower transaction costs. Motivated by prospect theory (Kahneman and Tversky, 1979), we assume the investor is risk adverse in gains (concave utility) and risk seeking in losses (convex utility). Our objective is to maximize the following exponential utility function

$$F(w,r) = \begin{cases} \frac{1-e^{-\psi^1(\sum_{i=1}^d w_i r_i)}}{\psi^1} & \text{if } \sum_{i=1}^d w_i r_i \geq 0 \\ \frac{e^{\psi^2(\sum_{i=1}^d w_i r_i)}-1}{\psi^2} & \text{otherwise,} \end{cases} \tag{34}$$

where $\psi^1, \psi^2 > 0$. This utility function has been considered in (Köbberling and Wakker, 2005; Pirvu and Schulze, 2012). Choosing $\psi^1 > \psi^2$ models loss aversion, where the investor has increased sensitivity to losses than to gains. The optimization problem is then

$$\begin{aligned} &\max \mathbb{E}_r[F(w,r)] + g(w) \\ &\text{s.t. } \sum_{i=1}^{d} w_i \leq 1 \\ &\qquad w_i \geq 0 \text{ for } i = 1, ..., d, \end{aligned} \tag{35}$$

where we assume there should be no short selling. Let

$$Q := \left\{ w : \begin{matrix} \sum_{i=1}^d w_i \leq 1 \\ w_i \geq 0 \text{ for } i = 1, ..., d \end{matrix} \right\},$$

and $I^N := \{i : w_i \leq 0\}$ and $I^P := \{i : w_i > 0\}$.

Property 20 *The projection of* w *onto* Q *can be computed in two steps: 1)* $\text{proj}_Q(w_i) = 0$ *for* $i \in I^N$*; 2) if* $\sum\limits_{i\in I^P} w_i > 1$*, project* $\{w_i : i \in I^P\}$ *onto the probability simplex.*

Proof Similar to Property 16, the projection of w onto Q is the following minimization problem,

$$\begin{aligned}
&\min_{\bar{w}} \frac{1}{2}||\bar{w} - w||_2^2 \\
&\text{s.t. } \sum_{i=1}^{d} \bar{w}_i \leq 1 \\
&\qquad \bar{w}_i \geq 0 \text{ for } i = 1, ..., d,
\end{aligned}$$

with KKT conditions,

$$\begin{aligned}
&\bar{w}_i = w_i - \mu_0 + \mu_i \text{ for } i = 1, ..., d && \text{(Stationarity)} \\
&\sum_{i=1}^{d} \bar{w}_i \leq 1, \quad \bar{w}_i \geq 0 \text{ for } i = 1, ..., d && \text{(Primal feasibility)} \\
&\mu_i \geq 0 \text{ for } i = 0, ..., d && \text{(Dual feasibility)} \\
&\mu_0 (\sum_{i=1}^{d} \bar{w}_i - 1) = 0, \quad \mu_i \bar{w}_i = 0 \text{ for } i = 1, ..., d. && \text{(Complementary slackness)}
\end{aligned}$$

First we confirm that if $w_i \leq 0$, then $\bar{w}_i = 0$, or else $0 < \bar{w}_i = w_i - \mu_0 + \mu_i \leq \mu_i$, which contradicts complementary slackness. Now if $\sum\limits_{i \in I^P} w_i \leq 1$, setting $\mu_i = -w_i$ for $i \in I^N$, $\mu_i = 0$ for $i \in I^P$, and $\mu_0 = 0$, gives the optimal solution of $\bar{w}_i = 0$ for $i \in I^N$ and $\bar{w}_i = w_i$ for $i \in I^P$. If instead $\sum\limits_{i \in I^P} w_i > 1$, $\sum\limits_{i \in I^P} \bar{w}_i = 1$, or else if $\sum\limits_{i \in I^P} \bar{w}_i < 1$, $\mu_0 = 0$ and there would exist a $j \in I^P$ with $w_j > \bar{w}_j = w_j + \mu_j$, which contradicts dual feasibility.

■

We note that projecting onto the probability simplex can be achieved using a simple non-iterative algorithm such as found in (Wang and Carreira-Perpinán, 2013). For the optimization problem (35), we assume that we have access to n historical observations of r, r^j for $j = 1, ..., n$, and take a distribution-free approach, optimizing directly over the observations,

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

Property 21 *The function* $f(w) = \frac{1}{n} \sum_{j=1}^{n} F(w, r^j)$, *where* $F(w, r^j)$ *is as defined in* (34) *is* $\frac{\max(\psi^1, \psi^2)}{n} \sum_{j=1}^{n} ||r^j||_2^2$*-smooth.*

Proof We first consider the univariate function

$$\hat{F}(u) = \begin{cases} \hat{F}_1(u) = \frac{1 - e^{-\psi^1 u}}{\psi^1} & \text{if } u \geq 0 \\ \hat{F}_2(u) = \frac{e^{\psi^2 u} - 1}{\psi^2} & \text{otherwise.} \end{cases}$$

The first and second derivatives of $\hat{F}_1(u)$ and $\hat{F}_2(u)$ are

$$\hat{F}_1'(u) = e^{-\psi^1 u},\ \hat{F}_1''(u) = -\psi^1 e^{-\psi^1 u}, \quad \hat{F}_2'(u) = e^{\psi^2 u},\ \text{and}\ \hat{F}_2''(u) = \psi^2 e^{\psi^2 u}.$$

We can see that $|\hat{F}_1''(u)| \leq \psi^1$ and $|\hat{F}_2''(u)| \leq \psi^2$ over their domains. Assume that $w, x \geq 0$ and $v, u \leq 0$. Using the mean value theorem,

$$\begin{aligned}|\hat{F}'(w) - \hat{F}'(x)| &= |\hat{F}_1'(w) - \hat{F}_1'(x)| \\ &\leq \psi^1 |w - x| \end{aligned} \tag{36}$$

$$\begin{aligned}|\hat{F}'(v) - \hat{F}'(u)| &= |\hat{F}_2'(v) - \hat{F}_2'(u)| \\ &\leq \psi^2 |v - u| \end{aligned} \tag{37}$$

Assuming $\psi^1 \geq \psi^2$, let $x = \frac{-\psi^2}{\psi^1} v$ with $\hat{F}_1'(x) = \hat{F}_2'(v)$, then

$$\begin{aligned}|\hat{F}'(w) - \hat{F}'(v)| &= |\hat{F}_1'(w) - \hat{F}_2'(v)| \\ &= |\hat{F}_1'(w) - \hat{F}_1'(x)| \\ &\leq \psi^1 |w - x| \\ &\leq \psi^1 |w - v|. \end{aligned} \tag{38}$$

Assuming now $\psi^2 \geq \psi^1$, let $u = \frac{-\psi^1}{\psi^2} w$ with $\hat{F}_2'(u) = \hat{F}_1'(w)$, then

$$\begin{aligned}|\hat{F}'(w) - \hat{F}'(v)| &= |\hat{F}_1'(w) - \hat{F}_2'(v)| \\ &= |\hat{F}_2'(u) - \hat{F}_2'(v)| \\ &\leq \psi^2 |u - v| \\ &\leq \psi^2 |w - v|, \end{aligned} \tag{39}$$

From (36)-(39), we conclude that $\hat{F}(u)$ is $\max(\psi^1, \psi^2)$-smooth. As shown in the proof of Property 19, since $F(w, r^j)$ is $\hat{F}(u)$ composed with the affine function $\sum_{i=1}^d w_i r_i^j$, it is $||r^j||_2^2 \max(\psi^1, \psi^2)$-smooth and $f(w)$ is $\frac{\max(\psi^1,\psi^2)}{n} \sum_{j=1}^n ||r^j||_2^2$-smooth. ■

7.3 Non-Negative Sparse Principal Component Analysis

Principal component analysis (PCA) finds a lower dimensional approximation of a dataset, with the non-negative sparse extension having applications in economics, bioinformatics and computer vision (Zass and Shashua, 2007). Given a data set $x \in \mathbb{R}^{d \times n}$, we find its first sparse non-negative principal component by solving

$$\begin{aligned}&\min\ -\frac{1}{2n}\sum_{j=1}^n (w^\top x_j)^2 + g(w) \\ &\text{s.t.}\ ||w||_2 \leq 1, w \geq 0.\end{aligned} \tag{40}$$

The projection onto $C = \{w : ||w||_2 \leq 1, w \geq 0\}$ has the explicit solution (Bauschke et al., 2018, Theorem 7.1)

$$\text{proj}_C(w) = \frac{\max(w, 0)}{\max(|| \max(w, 0)||_2, 1)}.$$

Property 22 *Given a dataset $x \in \mathbb{R}^{d\times n}$, the function $f(w) = -\frac{1}{2n}\sum_{j=1}^{n}(w^\top x_j)^2$ is $\frac{1}{n}||xx^\top||_2$-smooth, where $||\cdot||_2$ is the spectral norm.*

Proof The function $f(w)$ can be rewritten as

$$f(w) = -\frac{1}{2n}\sum_{j=1}^{n}(w^\top x_j)^2 = -\frac{1}{2n}w^\top\left(\sum_{j=1}^{n}x_jx_j^\top\right)w = -\frac{1}{2n}w^\top xx^\top w.$$

$$\begin{aligned}\left\|\nabla f(w) - \nabla f(w')\right\|_2 =& \left\| -\frac{1}{n}xx^\top w + \frac{1}{n}xx^\top w'\right\|_2 \\ \leq& \frac{1}{n}\left\|xx^\top\right\|_2\left\|w - w'\right\|_2.\end{aligned}$$

■

7.3.1 Experiment

We conducted experiments to observe our convergence analysis in practice and to compare our algorithms to SDCAM (Liu et al., 2019). We test on the problem of non-negative sparse PCA (40) on datasets MNIST (LeCun, 1998) and RCV1 (Lewis et al., 2004). The dimensions of MNIST are $n = 60,000$ and $d = 784$, and those of RCV1 are $n = 804,414$ and $d = 47,236$. 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. In Figure 1 we compare the performance of all algorithms. Given differences in convergence analysis and algorithm design, we plot the objective function versus wall-clock time as a general measure of algorithm performance. The values for α and θ established in Corollaries 10 and 15 were used to implement MBSPA and VRSPA. It was hypothesized that the inferior performance of VRSPA was due to its smaller stepsize, so VRSPA2 is VRSPA using the stepsize of MBSPA. All parameters of SDCAM were left unchanged as used in the available implementation[1]. The regularizer $g(w)$ was chosen as MCP with parameters chosen as $\kappa = \frac{1}{d}$ and $\nu = 1$. VRSPA requires that each $f_j(w)$ be L-smooth, so in our numerical experiments we took $L = \max_j \left\|x_jx_j^\top\right\|_2$. We chose the number of iterations of each algorithm so that they terminate at approximately the same time where at least half of the algorithms reached an observable level of convergence. Unlike for VRSPA, the number of samples M used to approximate $\nabla f(w)$ in MBSPA grows with the number of iterations. Given the rapid

[1] http://www.mypolyuweb.hk/~tkpong/Matrix_sparse_MP_codes/

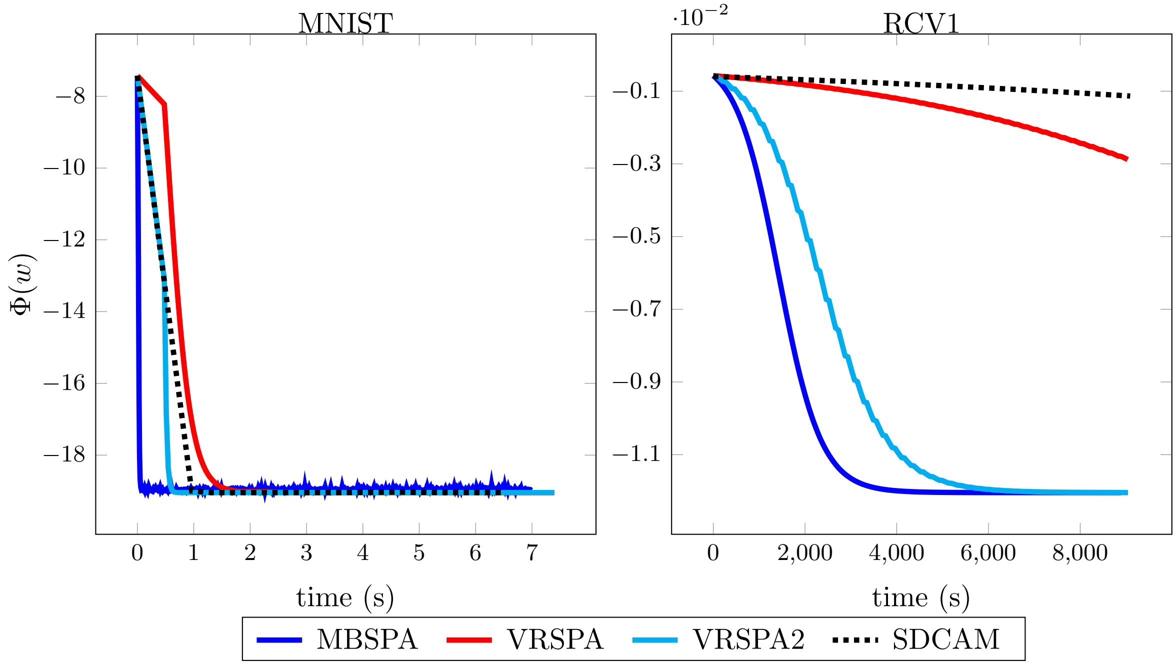

Figure 1: Comparison of algorithms of this paper and SDCAM (Liu et al., 2019) on datasets MNIST and RCV1.

convergence of all algorithms for the MNIST dataset, the number of samples used was only $M = 91$, which can be observed by the choppiness of the function value. We observe in general though that our algorithms were able to achieve a fast rate convergence in both experiments compared to SDCAM.

In Figure 2 we plot an upper bound of the subdifferential mapping of our algorithms. Given that the analysis of SDCAM did not use this measure of convergence we did not include it in the plots. Following Corollaries 10 and 15 we chose $\tau = 0$, resulting in $\bar{\gamma} = 1$. For each iteration k, we plot

$$||\hat{\mathcal{G}}_1(w^k)||_2 := ||w^k - \text{prox}_C(w^k - s^k)||_2 \tag{41}$$

where $s^k \in \nabla f(w^k) + \partial g^1(w^k)$, taking $\partial|0| = 0$ when computing a subgradient of $\partial g^1(w^k)$ (33). We can see that in general the convergence of the subdifferential mapping is not monotonic. For non-convex optimization, as we approach a local minimum, the subdifferential of the objective will not generally be monotonically decreasing. This allows the norm of the change in w to increase when applying a projected subgradient method while the function value is decreasing. We also observe that the convergence of each algorithm in terms of (41) is similar to their convergence in terms of their function values. This gives positive empirical evidence for the benefit of using the subdifferential mapping as a measure of convergence.

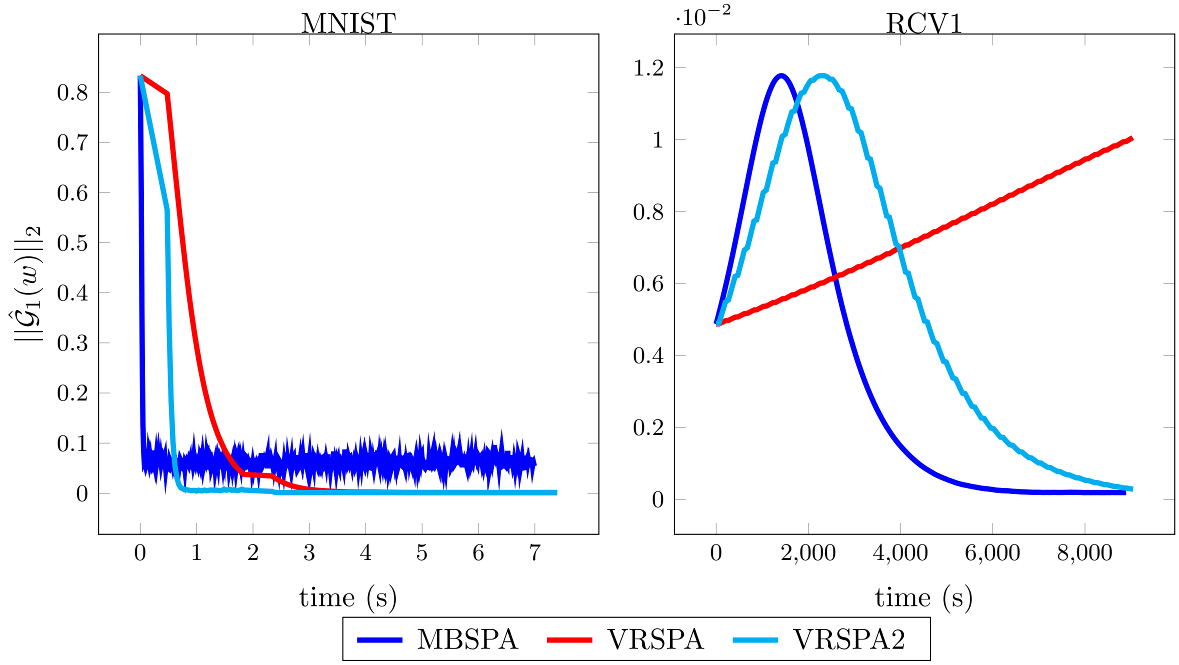

Figure 2: Bound on subdifferential mapping of algorithms of this paper on datasets MNIST and RCV1.

8. Conclusion

In this paper we considered minimizing a smooth non-convex loss function with a non-smooth non-convex regularizer with convex constraints. We proposed a new measure of convergence, the subdifferential mapping, and presented two stochastic proximal gradient algorithms. To the best of our knowledge, we have presented the first non-asymptotic convergence bounds for this class of objective function. In an empirical study we found our algorithms to converge faster than a state-of-the-art deterministic algorithm.

Acknowledgments

The research of the first author is supported in part by JSPS KAKENHI Grants No. 19H04069. The research of the second author is supported in part by JSPS KAKENHI Grants No. 17H01699 and 19H04069.

Appendix

Minimizing $|w|$ using Algorithm 1

Considering the case when $f(w) = g(w) = 0$ and $h(w) = |w|$, the proximal operator of $h(w)$ is

$$\begin{aligned}\operatorname{prox}_{\gamma h}(w) &= \max(0, |w| - \gamma)\operatorname{sgn}(w) \qquad \text{for } \gamma > 0\\ &= \begin{cases} w - \operatorname{sgn}(w)\gamma & \text{if } |w| - \gamma > 0\\ 0 & \text{if } |w| - \gamma \leq 0,\end{cases}\end{aligned}$$

and

$$\begin{aligned}\mathcal{G}_\gamma(w) &= \frac{1}{\gamma}\left(w - \operatorname{prox}_{\gamma h}(w)\right)\\ &= \begin{cases} \operatorname{sgn}(w) & \text{if } |w| - \gamma > 0\\ \frac{w}{\gamma} & \text{if } |w| - \gamma \leq 0.\end{cases}\end{aligned}$$

Examining Algorithm 1 in this setting,

$$\nabla A^k_{\lambda,M}(w, \xi^k) = 0,$$

and updates will be done as

$$w^{k+1} = \operatorname{prox}_{\gamma h}(w^k).$$

Each iteration $|w^k|$ is decreased by γ when $|w^k| - \gamma > 0$, and once $|w^k| \leq \gamma$, the algorithm will converge setting $w^{k+1} = 0$, so after $K := \lceil \frac{|w^1|}{\gamma} \rceil$ iterations $\mathcal{G}_\gamma(w^K) = 0$.

Considering the case when $f(w) = h(w) = 0$ and $g(w) = |w|$,

$$\begin{aligned}\operatorname{prox}_{\lambda g}(w) &:= \max(0, |w| - \lambda)\operatorname{sgn}(w) \text{ for } \lambda > 0\\ &= \begin{cases} w - \operatorname{sgn}(w)\lambda & \text{if } |w| - \lambda > 0\\ 0 & \text{if } |w| - \lambda \leq 0,\end{cases}\end{aligned}$$

and

$$\mathcal{G}_\gamma(w) = \left\{s : s \in \partial|w|\right\}.$$

Examining Algorithm 1,

$$\nabla A^k_{\lambda,M}(w, \xi^k) = \frac{1}{\lambda}\left(w - \operatorname{prox}_{\lambda g}(w)\right),$$

and updates will be done as

$$w^{k+1} = w^k - \frac{\gamma}{\lambda}\left(w^k - \operatorname{prox}_{\lambda g}(w^k)\right).$$

In this setting $L = 0$, so $\gamma = \lambda$ and

$$w^{k+1} = \operatorname{prox}_{\lambda g}(w^k),$$

and we recover the same convergence with $w^K = 0$, and $\operatorname{dist}(0, \mathcal{G}_\gamma(w^K)) = 0$.

Property 4 (Ghadimi et al., 2016, Lemma 1) *Let $w, s \in \mathbb{R}^d$ and $\gamma > 0$, then*

$$- \langle s, \mathcal{P}_\gamma(w,s) \rangle \leq \frac{1}{\gamma} \left(h(w) - h(\text{prox}_{\gamma h}(w - \gamma s)) \right) - ||\mathcal{P}_\gamma(w,s)||_2^2.$$

Proof By the optimality of $\text{prox}_{\gamma h}(w - \gamma s)$ in (6),

$$0 \in -\mathcal{P}_\gamma(w,s) + s + \partial h(\text{prox}_{\gamma h}(w - \gamma s)).$$

Taking $p \in \partial h(\text{prox}_{\gamma h}(w - \gamma s))$ such that $0 = -\mathcal{P}_\gamma(w,s) + s + p$, it follows that

$$\begin{aligned} 0 &= \langle -\mathcal{P}_\gamma(w,s) + s + p, \mathcal{P}_\gamma(w,s) \rangle \\ &= \langle s + p, \mathcal{P}_\gamma(w,s) \rangle - ||\mathcal{P}_\gamma(w,s)||_2^2 \\ &\leq \langle s, \mathcal{P}_\gamma(w,s) \rangle + \frac{1}{\gamma} \left(h(w) - h(\text{prox}_{\gamma h}(w - \gamma s)) \right) - ||\mathcal{P}_\gamma(w,s)||_2^2, \end{aligned}$$

where the inequality uses the convexity of h. ■

Property 6 (Beck, 2017, Theorem 10.9) *For $\gamma^1 \geq \gamma^2 > 0$ and any $w, s \in \mathbb{R}^d$,*

$$||\mathcal{P}_{\gamma^1}(w,s)||_2 \leq ||\mathcal{P}_{\gamma^2}(w,s)||_2.$$

Proof For an arbitrary $v \in \mathbb{R}^d$ and $\gamma > 0$, $\text{prox}_{\gamma h}(v)$ is the minimizer of $\frac{1}{2\gamma}||v - x||_2^2 + h(x)$ and so

$$\frac{1}{\gamma}(v - \text{prox}_{\gamma h}(v)) \in \partial h(\text{prox}_{\gamma h}(v)). \tag{42}$$

By the definition of a subgradient of a convex function, for any $y \in \mathbb{R}^d$,

$$h(y) - h(\text{prox}_{\gamma h}(v)) \geq \frac{1}{\gamma} \left\langle v - \text{prox}_{\gamma h}(v), y - \text{prox}_{\gamma h}(v) \right\rangle.$$

First let $\gamma = \gamma^1$, $v = w - \gamma^1 s$ and $y = \text{prox}_{\gamma^2 h}(w - \gamma^2 s)$,

$$\begin{aligned} &h(\text{prox}_{\gamma^2 h}(w - \gamma^2 s)) - h(\text{prox}_{\gamma^1 h}(w - \gamma^1 s)) \qquad (43) \\ \geq& \frac{1}{\gamma^1} \left\langle w - \gamma^1 s - \text{prox}_{\gamma^1 h}(w - \gamma^1 s), \text{prox}_{\gamma^2 h}(w - \gamma^2 s) - \text{prox}_{\gamma^1 h}(w - \gamma^1 s) \right\rangle \\ =& \left\langle \mathcal{P}_{\gamma^1}(w,s) - s, \gamma^1 \mathcal{P}_{\gamma^1}(w,s) - \gamma^2 \mathcal{P}_{\gamma^2}(w,s) \right\rangle. \end{aligned}$$

Exchanging γ^1 and γ^2, letting $\gamma = \gamma^2$, $v = w - \gamma^2 s$ and $y = \text{prox}_{\gamma^1 h}(w - \gamma^1 s)$,

$$h(\text{prox}_{\gamma^1 h}(w - \gamma^1 s)) - h(\text{prox}_{\gamma^2 h}(w - \gamma^2 s)) \geq \left\langle \mathcal{P}_{\gamma^2}(w,s) - s, \gamma^2 \mathcal{P}_{\gamma^2}(w,s) - \gamma^1 \mathcal{P}_{\gamma^1}(w,s) \right\rangle. \tag{44}$$

Adding inequalities (43) and (44),

$$0 \geq \left\langle \mathcal{P}_{\gamma^1}(w,s) - \mathcal{P}_{\gamma^2}(w,s), \gamma^1 \mathcal{P}_{\gamma^1}(w,s) - \gamma^2 \mathcal{P}_{\gamma^2}(w,s) \right\rangle. \tag{45}$$

Expanding and rearranging (45),

$$\begin{aligned}\gamma^1||\mathcal{P}_{\gamma^1}(w,s)||_2^2+\gamma^2||\mathcal{P}_{\gamma^2}(w,s)||_2^2 &\leq (\gamma^1+\gamma^2)\left\langle \mathcal{P}_{\gamma^1}(w,s),\mathcal{P}_{\gamma^2}(w,s)\right\rangle \\ &\leq (\gamma^1+\gamma^2)||\mathcal{P}_{\gamma^1}(w,s)||_2||\mathcal{P}_{\gamma^2}(w,s)||_2,\end{aligned} \tag{46}$$

using the Cauchy-Schwarz inequality. Assume $||\mathcal{P}_{\gamma^1}(w,s)||_2>0$, otherwise the property trivially holds, and set $t=||\mathcal{P}_{\gamma^2}(w,s)||_2/||\mathcal{P}_{\gamma^1}(w,s)||_2$. Inequality (46) can now be written as

$$\gamma^1+\gamma^2t^2-(\gamma^1+\gamma^2)t\leq 0.$$

The roots of the left hand side function occur at $t=1$ and $t=\frac{\gamma^1}{\gamma^2}$, so for the inequality to hold,

$$1\leq t\leq\frac{\gamma^1}{\gamma^2},$$

which includes the desired inequality,

$$||\mathcal{P}_{\gamma^1}(w,s)||_2\leq||\mathcal{P}_{\gamma^2}(w,s)||_2.$$

■

Property 12 (Li and Li, 2018, Lemma 1) *Consider arbitrary $w,s,z\in\mathbb{R}^d$, and $w^+=\operatorname{prox}_{\gamma h}(w-\gamma s)$,*

$$\begin{aligned}E_\lambda^{k,t}(w^+)+h(w^+) \leq & E_\lambda^{k,t}(z)+h(z)+\langle\nabla E_\lambda^{k,t}(w)-s,w^+-z\rangle+\frac{L_\lambda}{2}||w^+-w||_2^2+\frac{L_\lambda}{2}||z-w||_2^2 \\ & -\frac{1}{\gamma}\langle w^+-w,w^+-z\rangle.\end{aligned}$$

Proof As was done in the proof of Property 4, let us take $p\in\partial h(\operatorname{prox}_{\gamma h}(w-\gamma s))$ such that $0=-\mathcal{P}_\gamma(w,s)+s+p=\frac{1}{\gamma}(w^+-w)+s+p$. It follows by the convexity of $h(\cdot)$ that

$$\begin{aligned}h(w^+)&\leq h(z)+\langle p,w^+-z\rangle \\ &=h(z)-\left\langle\frac{1}{\gamma}(w^+-w)+s,w^+-z\right\rangle.\end{aligned} \tag{47}$$

Adding (47) with the following two inequalities, which come from the smoothness of $E_\lambda^{k,t}(w)$ and $-E_\lambda^{k,t}(w)$, see Property 1, proves the result.

$$\begin{aligned}E_\lambda^{k,t}(w^+)&\leq E_\lambda^{k,t}(w)+\langle\nabla E_\lambda^{k,t}(w),w^+-w\rangle+\frac{L_\lambda}{2}||w^+-w||_2^2 \\ -E_\lambda^{k,t}(z)&\leq -E_\lambda^{k,t}(w)+\langle-\nabla E_\lambda^{k,t}(w),z-w\rangle+\frac{L_\lambda}{2}||z-w||_2^2\end{aligned}$$

■

References

Heinz H Bauschke, Minh N Bui, and Xianfu Wang. Projecting onto the Intersection of a Cone and a Sphere. *SIAM Journal on Optimization*, 28(3):2158–2188, 2018.

Amir Beck. *First-Order Methods in Optimization.* SIAM, 2017.

Emmanuel J Candes, Michael B Wakin, and Stephen P Boyd. Enhancing Sparsity by Reweighted l_1 Minimization. *Journal of Fourier Analysis and Applications*, 14(5-6): 877–905, 2008.

Dmitriy Drusvyatskiy and Courtney Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, 178(1):503–558, 2019.

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, 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.

Daniel Kahneman and Amos Tversky. Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2):263–292, 1979.

Veronika Köbberling and Peter P Wakker. An index of loss aversion. *Journal of Economic Theory*, 122(1):119–131, 2005.

Emanuel Laude, Tao Wu, and Daniel Cremers. A Nonconvex Proximal Splitting Algorithm under Moreau-Yosida Regularization. In *International Conference on Artificial Intelligence and Statistics*, pages 491–499. PMLR, 2018.

Hoai An Le Thi and Tao Pham Dinh. DC programming and DCA: thirty years of developments. *Mathematical Programming*, 169(1):5–68, 2018.

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

David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. RCV1: A New Benchmark Collection for Text Categorization Research. *Journal of Machine Learning Research*, 5 (Apr):361–397, 2004.

Zhize Li and Jian Li. A Simple Proximal Stochastic Gradient Method for Nonsmooth Nonconvex Optimization. In *Advances in Neural Information Processing Systems*, pages 5564–5574, 2018.

Tianxiang Liu, Ting Kei Pong, and Akiko Takeda. A successive difference-of-convex approximation method for a class of nonconvex nonsmooth optimization problems. *Mathematical Programming*, 176(1-2):339–367, 2019.

Yanli Liu and Wotao Yin. An Envelope for Davis–Yin Splitting and Strict Saddle-Point Avoidance. *Journal of Optimization Theory and Applications*, 181(2):567–587, 2019.

Michael R Metel and Akiko Takeda. Simple Stochastic Gradient Methods for Non-Smooth Non-Convex Regularized Optimization. In *International Conference on Machine Learning*, pages 4537–4545, 2019.

Nhan H Pham, Lam M Nguyen, Dzung T Phan, and Quoc Tran-Dinh. ProxSARAH: An Efficient Algorithmic Framework for Stochastic Composite Nonconvex Optimization. *arXiv preprint arXiv:1902.05679*, 2019.

Traian A Pirvu and Klaas Schulze. Multi-stock portfolio optimization under prospect theory. *Mathematics and Financial Economics*, 6(4):337–362, 2012.

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.

R Tyrrell Rockafellar and Roger J-B Wets. *Variational Analysis*, volume 317. Springer Science & Business Media, 2009.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms.* Cambridge University Press, 2014.

Yiyuan She and Art B Owen. Outlier detection using nonconvex penalized regression. *Journal of the American Statistical Association*, 106(494):626–639, 2011.

Lixin Shen, Yuesheng Xu, and Xueying Zeng. Wavelet inpainting with the ℓ_0 sparse regularization. *Applied and Computational Harmonic Analysis*, 41(1):26–53, 2016.

Bharath K Sriperumbudur and Gert RG Lanckriet. On the Convergence of the Concave-Convex Procedure. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, pages 1759–1767. Curran Associates Inc., 2009.

Weiran Wang and Miguel A Carreira-Perpinán. Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application. *arXiv preprint arXiv:1309.1541*, 2013.

Yi Xu, Rong Jin, and Tianbao Yang. Stochastic Proximal Gradient Methods for Non-smooth Non-Convex Regularized Problems. *arXiv preprint arXiv:1902.07672*, 2019a.

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. In *International Conference on Machine Learning*, pages 6942–6951, 2019b.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness Constraints: Mechanisms for Fair Classification. In *Artificial Intelligence and Statistics*, pages 962–970, 2017.

Ron Zass and Amnon Shashua. Nonnegative Sparse PCA. In *Advances in neural information processing systems*, pages 1561–1568, 2007.

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

Lei Zhao, Musa Mammadov, and John Yearwood. From Convex to Nonconvex: A Loss Function Analysis for Binary Classification. In *2010 IEEE International Conference on Data Mining Workshops*, pages 1281–1288. IEEE, 2010.