Generalized Majorization-Minimization for Non-Convex Optimization

Hu Zhang[1] , **Pan Zhou**[2] , **Yi Yang**[1,3*] and **Jiashi Feng**[2]
[1]School of Computer Science, University of Technology Sydney, Australia
[2]Dept. ECE, National University of Singapore, Singapore
[3]Baidu Research
Hu.Zhang-1@student.uts.edu.au, pzhou@u.nus.edu, Yi.Yang@uts.edu.au, elefjia@nus.edu.sg

Abstract

Majorization-Minimization (MM) algorithms optimize an objective function by iteratively minimizing its majorizing surrogate and offer attractively fast convergence rate for convex problems. However, their convergence behaviors for non-convex problems remain unclear. In this paper, we propose a novel MM surrogate function from strictly upper bounding the objective to bounding the objective in expectation. With this generalized surrogate conception, we develop a new optimization algorithm, termed SPI-MM, that leverages the recent proposed SPIDER for more efficient non-convex optimization. We prove that for finite-sum problems, the SPI-MM algorithm converges to an stationary point within deterministic and lower stochastic gradient complexity. To our best knowledge, this work gives the first non-asymptotic convergence analysis for MM-alike algorithms in general non-convex optimization. Extensive empirical studies on non-convex logistic regression and sparse PCA demonstrate the advantageous efficiency of the proposed algorithm and validate our theoretical results.

1 Introduction

In this paper, let us consider the following finite-sum optimization problem:

$$\min_{\boldsymbol{\theta}} \left[\Phi(\boldsymbol{\theta}) \triangleq \frac{1}{n} \sum_{i=1}^{n} f_i(\boldsymbol{\theta}) + h(\boldsymbol{\theta}) \right], \tag{1}$$

where each component $f_i : \mathbb{R}^p \to \mathbb{R}$ is a continuous, non-convex but smooth function, associated with one sample. The second term $h(\boldsymbol{\theta})$ could be non-smooth and non-convex. n is the number of samples. Such a formulation encapsulates many statistical learning tasks, *e.g.* principle component analysis [Feng *et al.*, 2013], regression [Draper and Smith, 2014] and training neural networks [LeCun *et al.*, 2015].

Majorization-Minimization (MM) [Lange *et al.*, 2000] is an optimization framework for designing well-behaved optimization algorithms for non-convex functions. MM algorithms solve problem (1) via two steps. The first step is to construct a proper surrogate g that upper bounds the objective function $\Phi(\boldsymbol{\theta})$ tightly, *i.e.*, $g(\boldsymbol{\theta}) \geq \Phi(\boldsymbol{\theta})$. The second step is to optimize the surrogate whose optimum is much easier to obtain. Since the surrogate constructed at the current estimator is majorant to the objective function, each minimization step over the surrogate will decrease the objective function monotonically. MM is attractive in practice as one can decompose the original complex problem to a series of much simpler sub-problems that are easier and faster to optimize.

Previous works show that the convergence rate of MM algorithms is nearly optimal when the objective is convex [Mairal, 2013b]. Specifically, they are shown to converge at a rate of $O(1/\sqrt{t})$ in a finite-sum setting and $O(1/t)$ in a stochastic setting for strongly convex objective functions. However, when the component f_i in (1) is non-convex, an exact global optimum is unreachable and the theoretical convergence guarantee is hard to obtain by nature. Though previous studies have provided analysis of the convergence for asymptotic stationary points [Borwein and Lewis, 2010], their results are rather limited, and little work has revealed the specific convergence rate.

In this work, we aim to conquer this challenge and give concrete convergence rate analysis of MM-alike algorithms for solving non-convex problems. Inspired by the recently proposed SPIDER (Stochastic Path Integrated Differential EstimatoR) [Fang *et al.*, 2018] method, we propose a generalized surrogate which fully exploits the historical gradient information and develop a new MM algorithm, called SPI-MM. The proposed SPI-MM algorithm significantly relaxes the requirement on the surrogate from classic MM algorithms. Instead of tightly bounding the objective for all the samples, it only requires the surrogate to bound the objective *in expectation*. The SPI-MM is general and can instantiate existing MM methods. Figure 1 illustrates such generalization over the surrogate construction by SPI-MM. The red line is the classic surrogate which strictly upper bounds the objective function at the solution $\boldsymbol{\theta}_t$. Our proposed surrogate can lie in the space between two dotted lines and the requirement is much milder.

In particular, when we adopt a first-order generalized surrogate, we prove that the proposed SPI-MM algorithm decreases the objective value in expectation. In addition, we prove that the SPI-MM algorithm terminates af-

*Part of this work was done when Yi Yang was visiting Baidu Research during his Professional Experience Program.

ter $\mathcal{O}(\frac{\sqrt{n}}{\epsilon^2} + n)$ gradient computations when searching for an ϵ-approximation first-order stationary point $\boldsymbol{\theta}$ with $\mathbb{E}\|\nabla\Phi(\boldsymbol{\theta})\| \leqslant \epsilon$. These results are obtained by our developed novel proof techniques that we will detail in the method section, on top of the techniques from SPIDER.

The main contributions of our paper are as follows:

- We propose a generalized surrogate for non-convex optimization problems. Instead of requiring the surrogate to exactly upper bound the objective, we only require that the surrogate upper bound the objective *in expectation*. With such generalization, we put forward a new MM algorithm, named SPI-MM.
- We give the first non-asymptotic convergence rate and *IFO* complexity result for MM-alike algorithms, by non-trivially extending the techniques introduced in SPIDER. For the first time, we prove that MM algorithms can achieve $\mathcal{O}(\frac{\sqrt{n}}{\epsilon^2} + n)$ in gradient computation complexity, for non-convex optimization.
- We conduct extensive experiments on *logistic regression with a non-convex regularizer* and *sparse PCA* to show the superior effectiveness of our proposed SPI-MM algorithm over well-established baseline algorithms including MM, MISO, MISO1 and SMM.

2 Related Work

The Majorization-Minimization (MM) framework was first proposed in [Lange *et al.*, 2000]. It generalizes methods like EM by "transferring" the optimization to a sequence of surrogate functions which upper bound the original objective function. [Mairal, 2013a] proposed a stochastic majorization-minimization scheme, extending the MM principle to large-scale or possibly infinite datasets. [Mairal, 2015] proposed an incremental MM algorithm where a single function is obtained in each iteration, based on which the approximate surrogate function is updated. [Parizi *et al.*, 2015] provided a general MM scheme that is less sensitive to initialization in Concave-Convex Procedure (CCCP) problems. [Xu *et al.*, 2016] presented a relaxed version of the MM algorithm to solve the robust matrix factorization (RMF) problems, which only requires a locally majorant surrogate. More recently, [Bietti and Mairal, 2017] proposed variance reduced MISO for data augmentation problems.

Regarding convergence analysis of MM algorithms, [Vaida, 2005] revealed the global convergence of EM algorithms, extended the result of EM to that of MM algorithms under some conditions. However, they could only solve cases where the objective function is differentiable and convex, whereas we consider a non-convex problem here. [Mairal, 2013b] only studied the asymptotic stationary point conditions with first-order surrogate functions for non-convex problems, although this work gave concrete convergence results for convex and strongly-convex problems. [Kang *et al.*, 2015] established sublinear convergence results for non-covex problems by applying the theory of the Kurdyka-Lojasiewicz inequality. They only considered the regularizer $h(\boldsymbol{\theta})$ to be non-convex rather than a general non-convex problem. [Xu *et al.*, 2016] gave a statement on asymptotic stationary point conditions with the relaxed MM algorithm.

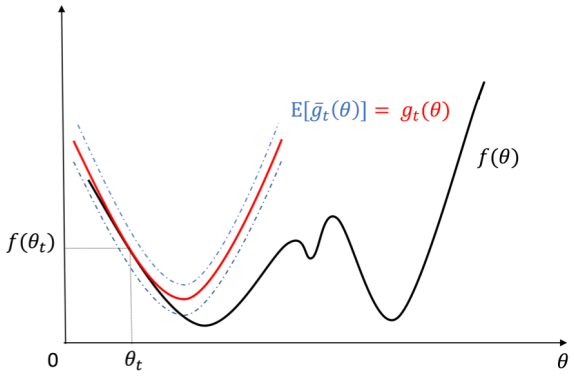

Figure 1: Illustration of classical MM and SPI-MM. A globally majorant surrogate $g_t(\boldsymbol{\theta})$ in classic MM algorithms is shown in red; our proposed surrogate $\bar{g}_t(\boldsymbol{\theta})$ possibly lies in the region between two dotted lines.

SGD as well as its variants [Nesterov, 1998; Cappé and Moulines, 2009; Kingma and Ba, 2014] is another line of approaches for solving problem (1). Based on SGD, various variance reduction methods have been proposed, like SAGA [Defazio *et al.*, 2014], SDCA[Shalev-Shwartz and Zhang, 2013], SVRG [Johnson and Zhang, 2013] and SARAH [Nguyen *et al.*, 2017]. SPIDER was proposed in [Fang *et al.*, 2018], which has been proved to be the most efficient algorithm for non-convex problems so far w.r.t. gradient complexity. It achieved $\mathcal{O}(\frac{\sqrt{n}}{\epsilon^2} + n)$ *IFO* complexity, which outperforms the results in previous methods, *e.g.* $\mathcal{O}(\frac{n^{\frac{2}{3}}}{\epsilon^2} + n)$ in SVRG.

Despite the considerable previous research, the convergence analysis of existing MM algorithms for non-convex problems is still not limited. When one considers a first-order surrogate, either a specific convergence result is absent or the gradient complexity has never been considered in MM algorithms. We apply the idea of SPIDER in extended MM definition and propose a generalized SPI-MM algorithm. We also give a concrete convergence rate and a clear *IFO* complexity analysis.

3 Proposed Algorithm

In this section, we first introduce useful assumptions and definitions in Section 3.1. Then we revisit the classic MM algorithms by examining some surrogate examples in details in Section 3.2. After that, we introduce our proposed SPI-MM algorithm in Section 3.3 and provide convergence analysis results in Section 3.4 and 3.5.

3.1 Preliminaries

We consider problem (1) with the objective function $\Phi(\boldsymbol{\theta})$ satisfying the following usual assumptions.

Assumption 1. *The objective function in problem* (1) *satisfies*

1. $\Phi(\boldsymbol{\theta})$ *is bounded below, i.e.,* $\Phi^* = \inf_{\boldsymbol{\theta}\in\mathbb{R}^p} \Phi(\boldsymbol{\theta}) > -\infty$;

Algorithm 1 Classic MM Algorithm

Input: initial estimator $\boldsymbol{\theta}_0$, iteration number K;
1: **for** $t = 0, ..., T-1$ **do**
2: Compute a surrogate function g_t of $\Phi(\boldsymbol{\theta})$ near $\boldsymbol{\theta}_t$,
3: Update the solution: $\boldsymbol{\theta}_t \in \arg\min_{\boldsymbol{\theta}} g_t(\boldsymbol{\theta}; \boldsymbol{\theta}_t)$;
4: **end for**
Output: $\boldsymbol{\theta}_T$.

2. *For every i = 1,...,n, the gradient ∇f_i is L-Lipschitz continuous, i.e.,*

$$\left\| \nabla f_i(\boldsymbol{\theta}) - \nabla f_i(\boldsymbol{\theta}') \right\| \leq L \left\| \boldsymbol{\theta} - \boldsymbol{\theta}' \right\|,$$

and f_i is also said to be L-smooth.

MM algorithms heavily rely on constructing a proper surrogate function for the objective $\Phi(\boldsymbol{\theta})$. Typically, the surrogate function should have the following properties.

Definition 1 (Classic surrogate)**.** *A function $g : \mathbb{R}^p \to \mathbb{R}$ is a first-order surrogate function of Φ in problem (1) near $\boldsymbol{\theta}_t$ when*

1. *$g(\boldsymbol{\theta}; \boldsymbol{\theta}_t)$ is a global majorant surrogate if the general condition $g(\boldsymbol{\theta}; \boldsymbol{\theta}_t) \geqslant \Phi(\boldsymbol{\theta})$ holds for all $\boldsymbol{\theta}$;*
2. *$g(\boldsymbol{\theta}_t; \boldsymbol{\theta}_t) = \Phi(\boldsymbol{\theta}_t)$; $\nabla g(\boldsymbol{\theta}_t; \boldsymbol{\theta}_t) = \nabla \Phi(\boldsymbol{\theta}_t)$ when $\Phi(\boldsymbol{\theta})$ is smooth.*

The incremental first-order oracle (IFO) is usually used to measure the computation complexity of evaluating stochastic optimization algorithms[Agarwal and Bottou, 2014; Wang *et al.*, 2018].

Definition 2 (IFO complexity)**.** *For $\Phi(\boldsymbol{\theta})$ in problem* (1)*, an IFO means selecting an index i and a datum $\boldsymbol{x}_i$ and returning the pair $(\Phi(\boldsymbol{\theta}; \boldsymbol{x}_i), \nabla \Phi(\boldsymbol{\theta}; \boldsymbol{x}_i))$.*

3.2 The MM Algorithms

We here briefly review the classic MM algorithm [Lange *et al.*, 2000; Razaviyayn *et al.*, 2016]. In each iteration, it requires the surrogate function g to tightly upper bound the objective at estimator obtained in last iteration, *i.e.*, $\boldsymbol{\theta}_t$, as explained in Definition 1.

One popular choice for the surrogate g is the proximal gradient surrogates. Given the gradient $\nabla f(\boldsymbol{\theta}_t)$ and another sufficiently large parameter L, the surrogate of $\Phi(\boldsymbol{\theta})$ is constructed as follows:

$$g_t(\boldsymbol{\theta}) = f(\boldsymbol{\theta}_t) + \nabla f(\boldsymbol{\theta}_t)(\boldsymbol{\theta} - \boldsymbol{\theta}_t) + \frac{L}{2} \left\| \boldsymbol{\theta} - \boldsymbol{\theta}_t \right\|^2 + h(\boldsymbol{\theta}). \quad (2)$$

Surrogates like the above one are strongly convex thus are much easier to optimize. MM sets the new $\boldsymbol{\theta}_t$ from

$$\boldsymbol{\theta}_{t+1} = \arg\min g_t(\boldsymbol{\theta})$$

as the starting point for the next iteration.

A surrogate constructed as above satisfies the strict majorization condition in Definition 1 and will result in non-increasing loss function on the sequential estimate $\{\boldsymbol{\theta}_0, ..., \boldsymbol{\theta}_t\}$. Namely, $\Phi(\boldsymbol{\theta}_{t+1}) \leq g(\boldsymbol{\theta}_{t+1}) \leq g(\boldsymbol{\theta}_t) = \Phi(\boldsymbol{\theta}_t)$. With such a property, previous works derive convergence rate for convex and strongly convex problems. Several previous algorithms like SMM [Mairal, 2013b] and MISO [Mairal, 2015] adopt different combinations of surrogates constructed in different iterations, but it is always required that the final surrogates at the current iteration upper bound the original objective function. A classic procedure of MM algorithms is summarized in Algorithm 1.

The globally majorant requirement can conveniently facilitate the convergence proof for convex problems. As above mentioned, one can directly establish the relation between the values of an objective function in two adjacent iterations. However, under this strict requirement, it is hard to fully utilize the historical gradient information though the construction of the surrogate relies on gradients in the past iterations. Moreover, the strictly selected surrogates may lead to a very small step size under a non-convex setting, resulting in a slower convergence rate in practice. In this work, we substantially relax such a requirement and surprisingly obtain convergence guarantees by leveraging recent non-convex optimization techniques.

Algorithm 2 SPI-MM Algorithm

Input: Initial $\boldsymbol{\theta}_0 \in \Theta$, iteration number T, iteration interval p, mini-batch size $\mathcal{S}_2$. Choose a surrogate g_0^i of f_i near $\boldsymbol{\theta}_0$ for all i;
1: **for** $t = 0, ..., T-1$ **do**
2: **if** mod(t, p)==0 **then**
3: Draw all samples and choose some surrogates g_i for all $i \in$ [n], $\bar{g}_t(\boldsymbol{\theta}; \boldsymbol{\theta}_t) = \frac{1}{|\mathcal{S}_1|} \sum_{i \in \mathcal{S}_1} g_i$,
4: **else**
5: Randomly draw mini-batch $\mathcal{S}_2$ and choose base surrogate g_i for all $i \in \mathcal{S}_2$,
$g_{\mathcal{S}_2^t}(\boldsymbol{\theta}; \boldsymbol{\theta}_t) = \frac{1}{|\mathcal{S}_2^t|} \sum_{i \in \mathcal{S}_2^t} g_i$,
6: $\bar{g}_t(\boldsymbol{\theta}; \boldsymbol{\theta}_t) = g_{\mathcal{S}_2^t}(\boldsymbol{\theta}; \boldsymbol{\theta}_t) + (-\nabla g_{\mathcal{S}_2^t}(\boldsymbol{\theta}_{t-1}; \boldsymbol{\theta}_{t-1}) + \mathcal{V}_{t-1})^\top (\boldsymbol{\theta} - \boldsymbol{\theta}_t) + \frac{\mu}{2} \left\| \boldsymbol{\theta} - \boldsymbol{\theta}_t \right\|^2$;
7: **end if**
8: Update the solution: $\boldsymbol{\theta}_{t+1} \in \arg\min_{\boldsymbol{\theta}} \bar{g}_t(\boldsymbol{\theta}; \boldsymbol{\theta}_t)$;
9: **end for**
Output: $\boldsymbol{\theta}_\xi$ that is uniformly chosen at random from $\{\boldsymbol{\theta}_t\}_{t=0}^{T-1}$.

3.3 The SPI-MM Algorithm

We first describe our generalized MM idea and then elaborate on the surrogate construction process. In our generalized MM definition, we measure progress per iteration over the objective function using expectation. It allows us to relax the surrogate constraint in classic MM.

Generalized Surrogate

Under our generalized MM framework, we only require the surrogate to satisfy below conditions.

Definition 3 (Generalized surrogate)**.** *A function g_{gen} is said to be a surrogate function in generalized MM if*

1. *$g_{gen}(\boldsymbol{\theta}_t; \boldsymbol{\theta}_t) = \Phi(\boldsymbol{\theta}_t)$,*

2. $\mathbb{E}\nabla g_{gen}(\boldsymbol{\theta}_t;\boldsymbol{\theta}_t) = \nabla\Phi(\boldsymbol{\theta}_t)$, *if* $\Phi(\boldsymbol{\theta})$ *is smooth,*

3. $\mathbb{E}g_{gen}(\boldsymbol{\theta};\boldsymbol{\theta}_t) \geq \Phi(\boldsymbol{\theta})$ *for all* $\boldsymbol{\theta}$.

The conditions above are closely related to those in classic MM. For the 2nd condition in Definition 3, we expect the value of a new surrogate to equal the objective function at $\boldsymbol{\theta}_{t-1}$. The expectation of the new surrogate's gradient at $\boldsymbol{\theta}_{t-1}$ is also required to equal that of the objective function. However, instead of ensuring the surrogate to be globally majorant, we only require its expectation to be globally majorant. When we construct the surrogate on a single datum $\boldsymbol{x}_i$, we also just require $\mathbb{E}g^i_{gen}(\boldsymbol{\theta};\boldsymbol{\theta}_t) \geq \Phi^i(\boldsymbol{\theta})$ for all $\boldsymbol{\theta}$. This relaxed condition ensures that our surrogate be valid but of sufficient flexibility in exploring the solution space. Such milder constraints in generalized MM do not directly imply $\Phi(\boldsymbol{\theta}_t) \leq \Phi(\boldsymbol{\theta}_{t-1})$, but they ensure $\forall t, \mathbb{E}\Phi(\boldsymbol{\theta}_t) \leq \Phi(\boldsymbol{\theta}_0)$. This implies guarantees for the objective function value to be non-increased in expectation.

Note that the above surrogate is valid in the classic sense. The first condition requires that the surrogate concentrate around a classic surrogate. However, with slight uncertainty, the choice over the surrogates becomes more flexible. For example, we can make slight compromise in accuracy to obtain a larger step size, giving faster convergence in practice. Also, by generalizing the surrogate, we are able to use the techniques for proving the objective decrease in expectation. This is key for obtaining our following convergence guarantees.

With the generalized surrogate definition, we propose a method to construct the surrogate and develop the SPI-MM algorithm for solving problem (1). Overall, this method considers the combination of past and current gradient information, and an arbitrary valid surrogate obtained in the classic MM setting. Some surrogates proposed in [Mairal, 2013a] and [Xu *et al.*, 2016] also apply here. We show that the constructed surrogate is not necessarily restricted to the traditional MM setting.

SPI-MM Algorithm

Here, we propose a concrete algorithm, named SPI-MM that employs a specific generalized surrogate construction.

Our proposed SPI-MM algorithm fully leverages the gradient of past surrogates as follows. Suppose each epoch includes p iterations and totally we need T iterations. At the first step t_0 in each epoch, we use all the samples n to construct a strict surrogate as in classic MM. For each datum, we choose a surrogate g_i of Φ_i near $\boldsymbol{\theta}_0$. When $h(\boldsymbol{\theta})$ in $\Phi(\boldsymbol{\theta})$ is non-smooth, we choose a surrogate g_i of f_i near $\boldsymbol{\theta}_0$. We define $|\mathcal{S}_1| = n$. Then we have

$$\bar{g}_0(\boldsymbol{\theta};\boldsymbol{\theta}_0) = \frac{1}{|\mathcal{S}_1|}\sum_{i\in\mathcal{S}_1} g_i. \tag{3}$$

We minimize Eqn.(3) to get $\boldsymbol{\theta}_1$. For the next step, we only sample $\mathcal{S}_2$ samples. Relying on $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1$ respectively, we construct a base surrogate $g_{\mathcal{S}_2}(\boldsymbol{\theta};\boldsymbol{\theta}_1)$ of $\Phi_i \in \mathcal{S}_2$ near $\boldsymbol{\theta}_1$ and an auxiliary surrogate $g_{\mathcal{S}_2}(\boldsymbol{\theta};\boldsymbol{\theta}_0)$ near $\boldsymbol{\theta}_0$. These two surrogates here at least satisfy the 2nd condition in Definition 1.

We use the base surrogate $g_{\mathcal{S}_2}(\boldsymbol{\theta};\boldsymbol{\theta}_1)$ as the first term of the new surrogate in this step. We compute the gradient of the auxiliary surrogate $g_{\mathcal{S}_2}(\boldsymbol{\theta};\boldsymbol{\theta}_0)$ and $\bar{g}_0(\boldsymbol{\theta};\boldsymbol{\theta}_0)$ at $\boldsymbol{\theta}_0$, *i.e.* $\nabla g_{\mathcal{S}_2}(\boldsymbol{\theta}_0;\boldsymbol{\theta}_0), \nabla\bar{g}_0(\boldsymbol{\theta}_0;\boldsymbol{\theta}_0)$ to construct a linear term $(-\nabla g_{\mathcal{S}_2}(\boldsymbol{\theta}_0;\boldsymbol{\theta}_0) + \nabla\bar{g}_0(\boldsymbol{\theta}_0;\boldsymbol{\theta}_0))^\top(\boldsymbol{\theta}-\boldsymbol{\theta}_1)$, and take it as the second term in the new surrogate. We add a second-order term $\frac{\mu}{2}\|\boldsymbol{\theta}-\boldsymbol{\theta}_1\|^2$ to complement for the relaxing in the base surrogate. We minimize the sum of these three terms to get $\boldsymbol{\theta}_2$. The following optimization iterations repeat the above surrogate construction, summarized as

$$\begin{aligned}\bar{g}_t(\boldsymbol{\theta};\boldsymbol{\theta}_t) = g_{\mathcal{S}_2^t}(\boldsymbol{\theta};\boldsymbol{\theta}_t) + \{&-\nabla g_{\mathcal{S}_2^t}(\boldsymbol{\theta}_{t-1};\boldsymbol{\theta}_{t-1}) \\ &+ \sum_{i=1}^{t-1}[\nabla g_{\mathcal{S}_2^i}(\boldsymbol{\theta}_i;\boldsymbol{\theta}_i) - \nabla g_{\mathcal{S}_2^i}(\boldsymbol{\theta}_{i-1};\boldsymbol{\theta}_{i-1})] \\ &+ \nabla\bar{g}_0(\boldsymbol{\theta}_0,\boldsymbol{\theta}_0)\}^\top(\boldsymbol{\theta}-\boldsymbol{\theta}_t) + \frac{\mu}{2}\|\boldsymbol{\theta}-\boldsymbol{\theta}_t\|^2.\end{aligned} \tag{4}$$

We rewrite the gradient part in the linear term as $-\nabla g_{\mathcal{S}_2^t}(\boldsymbol{\theta}_{t-1};\boldsymbol{\theta}_{t-1}) + \mathcal{V}_{t-1}$ for simplicity:

$$\mathcal{V}_{t-1} \triangleq \sum_{i=1}^{t-1}[\nabla g_{\mathcal{S}_2^i}(\boldsymbol{\theta}_i;\boldsymbol{\theta}_i) - \nabla g_{\mathcal{S}_2^i}(\boldsymbol{\theta}_{i-1};\boldsymbol{\theta}_{i-1})] + \nabla\bar{g}_0(\boldsymbol{\theta}_0,\boldsymbol{\theta}_0).$$

Samples $\{\mathcal{S}_2^{t-1}, \mathcal{S}_2^{t-2}, ..., \mathcal{S}_2^1\}$ in different iterations are equal in number. The idea for constructing this surrogate is borrowed from SPIDER [Fang *et al.*, 2018], where current gradient is estimated by utilizing the gradient in the past to reduce the variance, leading to smaller sampling numbers.

We demonstrate below that the surrogate in SPI-MM satisfies conditions in our Definition 3. For the first condition, we have

$$\bar{g}_t(\boldsymbol{\theta}_t;\boldsymbol{\theta}_t) = g_{\mathcal{S}_2^t}(\boldsymbol{\theta}_t;\boldsymbol{\theta}_t) \tag{5}$$

from Eqn. 4. On the other side,

$$g_{\mathcal{S}_2^t}(\boldsymbol{\theta}_t;\boldsymbol{\theta}_t) = \Phi(\boldsymbol{\theta}_t) \tag{6}$$

as we require the base surrogate $g_{\mathcal{S}_2^t}(\boldsymbol{\theta};\boldsymbol{\theta}_t)$ to satisfy the 2nd condition in Definition 1, leading to $\bar{g}_t(\boldsymbol{\theta}_t;\boldsymbol{\theta}_t) = \Phi(\boldsymbol{\theta}_t)$ verifying the first condition.

For $\nabla\bar{g}_t(\boldsymbol{\theta};\boldsymbol{\theta}_t)$ at $\boldsymbol{\theta}_t$, the gradient is computed as

$$\nabla\bar{g}_t(\boldsymbol{\theta}_t;\boldsymbol{\theta}_t) = \nabla g_{\mathcal{S}_2^t}(\boldsymbol{\theta}_t;\boldsymbol{\theta}_t) - \nabla g_{\mathcal{S}_2^t}(\boldsymbol{\theta}_{t-1};\boldsymbol{\theta}_{t-1}) + \mathcal{V}_{t-1}. \tag{7}$$

It is easy to obtain the expectation of $-\nabla g_{\mathcal{S}_2^t}(\boldsymbol{\theta}_{t-1};\boldsymbol{\theta}_{t-1}) + \mathcal{V}_{t-1}$ is zero by iteratively unfolding this term and i.i.d. sampling condition. Thus, we have $\mathbb{E}\nabla\bar{g}_t(\boldsymbol{\theta}_t;\boldsymbol{\theta}_t) = \mathbb{E}\nabla g_{\mathcal{S}_2^t}(\boldsymbol{\theta}_t;\boldsymbol{\theta}_t) = \nabla\Phi(\boldsymbol{\theta}_t)$, which satisfies the 2nd condition.

Finally, for the 3rd condition in Definition 3, we have $\mathbb{E}\bar{g}_t(\boldsymbol{\theta},\boldsymbol{\theta}_t) = \mathbb{E}g_{\mathcal{S}_2^t}(\boldsymbol{\theta}-\boldsymbol{\theta}_t) + \frac{\mu}{2}\|\boldsymbol{\theta}-\boldsymbol{\theta}_t\|^2$. Suppose the objective function $\Phi(\boldsymbol{\theta})$ is L-smooth, $\nabla^2\mathbb{E}g_{\mathcal{S}_2^t}(\boldsymbol{\theta}-\boldsymbol{\theta}_t) \succeq \mu_f$. Then we just need to tune μ large enough to make $\bar{\mu} = \mu+\mu_f$ larger than L to satisfy the 3rd condition in Definition 3. We argue that μ_f here is reasonable, since the base surrogate $g_{\mathcal{S}_2^t}(\boldsymbol{\theta};\boldsymbol{\theta}_t)$ is selected by ourselves. For simplicity, we can even directly select a strongly convex function here.

Our way of constructing the new surrogate accords with the basic idea in SPIDER. Based on the base surrogate, we use the gradient of past surrogates to construct a linear term, which is able to reduce the variance of the base surrogate. Such a composition also facilitates the convergence analysis, which is explained in next subsection. The overall algorithm is summarized in Algorithm 2.

3.4 Convergence Guarantees

We here analyze the convergence properties of the proposed scheme under the generalized MM setting. We focus on the non-convex problem and searching for a stationary point $\mathbb{E}\|\nabla\Phi(\boldsymbol{\theta})\| \leq \epsilon$. We first show our loss function is decreasing every epoch in expectation and then give the concrete convergence results and corresponding *IFO* complexity.

Lemma 1. *Suppose Assumption 3.1 holds, and a sequence $\{\theta_{n_t p}\}$ is produced by Algorithm 2 after every* p *iterations. The base surrogate $g_t(\boldsymbol{\theta};\boldsymbol{\theta}_t)$ is L_f-smooth, $\alpha = \frac{1}{2\mu} - \frac{L_f}{2\mu\bar{\mu}^2} - \frac{L}{2\bar{\mu}^2} - \frac{L^2 p}{2\bar{\mu}^2\mu|\mathcal{S}_2|}$, $\mathcal{V}_i = \nabla\bar{g}_i(\boldsymbol{\theta}_i;\boldsymbol{\theta}_i)$. Then the objective function $\Phi(\boldsymbol{\theta})$ after every p iterations is guaranteed to decrease in expectation:*

$$\mathbb{E}\Phi(\boldsymbol{\theta}_{n_t p}) - \mathbb{E}\Phi(\boldsymbol{\theta}_{(n_t-1)p}) \leq -\sum_{i=(n_t-1)p}^{n_t p-1} \alpha\mathbb{E}\left\|\mathcal{V}_i\right\|^2. \quad (8)$$

By Lemma 1, if each epoch contains p iterations, we guarantee that the objective function be almost certain to be decreasing in expectation. It is also very easy to derive $\mathbb{E}\Phi(\boldsymbol{\theta}_T) - \Phi(\boldsymbol{\theta}_0) \leq -\sum_{i=0}^{T-1}\alpha\mathbb{E}\left\|\mathcal{V}_i\right\|^2$, meaning the objective function value is driven to be shrinking. Theorem 1 gives the final convergence rate and IFO complexity by applying Lemma 1. It shows that the objective function converges to a stationary point at a rate of $\mathcal{O}(1/\sqrt{t})$ and the total gradient computation is $\mathcal{O}(\sqrt{n}\epsilon^{-2}+n)$.

Theorem 1. *Suppose Assumptions 3.1 holds and apply SPI-MM in Algorithm 2. Let $p = \sqrt{n}$, $\mathcal{S}_2 = \sqrt{n}$ and μ be large enough. Then we have final output satisfying $\mathbb{E}\|\nabla\Phi(\boldsymbol{\theta}_\xi)\| \leq \epsilon$ as long as the total number of iterations T satisfies*

$$T \geq \mathcal{O}\left(\frac{\Phi(\boldsymbol{\theta}_0) - \Phi^*}{\epsilon^2}\right). \quad (9)$$

And the total resulting IFO complexity is $\mathcal{O}(\sqrt{n}\epsilon^{-2}+n)$.

3.5 Proof Roadmap

Due to space limit, we omit details of the proof and provide a proof roadmap here to illustrate the basic idea.

We aim to bound the iteration steps and gradient computations for attaining the first-order stationary point $\mathbb{E}\|\nabla\Phi(\boldsymbol{\theta}_\xi)\| \leq \varepsilon$ in non-convex problems. To this end, we leverage the structure in SPI-MM and rewrite $\mathbb{E}\|\nabla\Phi(\boldsymbol{\theta}_\xi)\| \leq \varepsilon$ as follows:

$$\begin{aligned}\mathbb{E}\|\nabla\Phi(\boldsymbol{\theta}_\xi)\|^2 &= \mathbb{E}\|\nabla\Phi(\boldsymbol{\theta}_\xi) - \mathcal{V}_\xi + \mathcal{V}_\xi\|^2 \\ &\leq 2\mathbb{E}\|\nabla\Phi(\boldsymbol{\theta}_\xi) - \mathcal{V}_\xi\|^2 + 2\mathbb{E}\|\mathcal{V}_\xi\|^2.\end{aligned} \quad (10)$$

Then we bound the above two last terms by establishing following two lemmas respectively.

Lemma 2. Under Assumption 1, let $n_t = [t/p]$ such that $(n_t-1)p \leq t \leq n_t p - 1$, $(n_t-1)p$ is the beginning of epoch n_t. Then the estimator $\mathcal{V}_k$ satisfies

$$\begin{aligned}\mathbb{E}\|\mathcal{V}_t - \nabla\Phi(\boldsymbol{\theta}_t)\|^2 &\leq \sum_{i=(n_t-1)p}^{t} \frac{L^2}{|\mathcal{S}_2|}\|\boldsymbol{\theta}_{i+1} - \boldsymbol{\theta}_i\|^2 \\ &\leq \sum_{i=(n_t-1)p}^{t} \frac{L^2}{|\mathcal{S}_2|\bar{\mu}^2}\mathbb{E}\|\mathcal{V}_i\|^2.\end{aligned} \quad (11)$$

The above lemma bounds the first term. Now we proceed to bound the second term $\mathbb{E}\|\mathcal{V}_t\|$ as follows. We have

$$\begin{aligned}\boldsymbol{\theta}_{t+1} &= \arg\min_{\boldsymbol{\theta}} \bar{g}_t(\boldsymbol{\theta};\boldsymbol{\theta}_t) \\ &= \arg\min_{\boldsymbol{\theta}} \{g_{\mathcal{S}_2^t}(\boldsymbol{\theta};\boldsymbol{\theta}_t) + \frac{\mu}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_t\|^2 + \\ &\quad (-\nabla g_{\mathcal{S}_2^t}(\boldsymbol{\theta}_{t-1};\boldsymbol{\theta}_{t-1}) + \mathcal{V}_{t-1})^\top(\boldsymbol{\theta} - \boldsymbol{\theta}_t)\}, \\ \boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t &= -\frac{1}{\mu}(\mathcal{V}_t + \nabla g_{\mathcal{S}_2^t}(\boldsymbol{\theta}_{t+1};\boldsymbol{\theta}_t) - \nabla g_{\mathcal{S}_2^t}(\boldsymbol{\theta}_t;\boldsymbol{\theta}_t)).\end{aligned}$$

By substituting $\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t$ into the following formulation

$$\Phi(\boldsymbol{\theta}_{t+1}) \leq \Phi(\boldsymbol{\theta}_t) + \langle\nabla\Phi(\boldsymbol{\theta}_t), \boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t\rangle + \frac{L}{2}\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t\|^2.$$

we get Lemma 3.

Lemma 3. Under Assumption 1, our new surrogate is $\bar{\mu}$-strongly convex and the base surrogate is L_f-smooth. If the parameters $\mu, \bar{\mu}, L_f, p$ and $\mathcal{S}_2$ are chosen satisfying

$$\alpha \triangleq \frac{1}{2\mu} - \frac{L_f}{2\mu\bar{\mu}^2} - \frac{L}{2\bar{\mu}^2} - \frac{L^2 p}{2\bar{\mu}^2\mu|\mathcal{S}_2|} > 0,$$

we have

$$\mathbb{E}\|\mathcal{V}_\xi\|^2 = \frac{1}{T}\sum_{i=1}^{T-1}\mathbb{E}\|\mathcal{V}_i\|^2 \leq \frac{\Phi(\boldsymbol{\theta}_0) - \Phi^*}{T\alpha}. \quad (12)$$

By applying Lemma 2 and Lemma 3, we can prove $\mathbb{E}\|\nabla\Phi(\boldsymbol{\theta}_\xi)\|^2 \leq \frac{2}{T\alpha}\left(1 + \frac{L^2 p}{\bar{\mu}^2|\mathcal{S}_2|}\right)(\Phi(\boldsymbol{\theta}_0) - \Phi^*)$, and with proper parameters selected we get the convergence rate and IFO complexity results in Theorem 2.

4 Experiments

We conduct two groups of experiments on non-convex problems to evaluate our proposed results. The first group is to optimize a logistic regression loss with a non-convex regularizer[Gasso *et al.*, 2009]. Specifically, we optimize the following problem:

$$\Phi(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N}\log(1 + e^{-y_i \boldsymbol{x}_i^\top \boldsymbol{\theta}}) + \frac{\lambda}{2}\sum_{j=1}^{p}\log(|\theta[j]| + \varepsilon),$$

where $\theta[j]$ is the j-th element in $\boldsymbol{\theta}$. The function $h(\boldsymbol{\theta})$ here is not differentiable. We write it as a composition of $h(u) = \log(u+\varepsilon)$ and $u = |\theta[j]|$. We construct the following surrogate for $h(\boldsymbol{\theta})$ through linear approximation:

$$\frac{\lambda}{2}\sum_{j=1}^{p}\log(|\theta_{t-1}[j]| + \varepsilon) + \frac{\lambda}{2}\sum_{j=1}^{p}\frac{|\theta[j]| - |\theta_{t-1}[j]|}{|\theta_{t-1}[j]| + \varepsilon}.$$

By choosing a second-order approximation of the first logistic term $f_{\mathcal{S}_2^{t-1}}(\boldsymbol{\theta}_{t-1}) + \nabla f_{\mathcal{S}_2^{t-1}}(\boldsymbol{\theta}_{t-1})^\top(\boldsymbol{\theta} - \boldsymbol{\theta}_{t-1}) + \frac{L_f}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_{t-1}\|^2$, we can establish the linear term $(-\nabla f_{\mathcal{S}_2^{t-1}}(\boldsymbol{\theta}_{t-2}) + \sum_{i=1}^{t-2}[\nabla f_{\mathcal{S}_2^i}(\boldsymbol{\theta}_i) - \nabla f_{\mathcal{S}_2^i}(\boldsymbol{\theta}_{i-1})] + \nabla f(\boldsymbol{\theta}_0))(\boldsymbol{\theta} - \boldsymbol{\theta}_{t-1})$ and the second-order term $\frac{\mu}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_{t-1}\|^2$

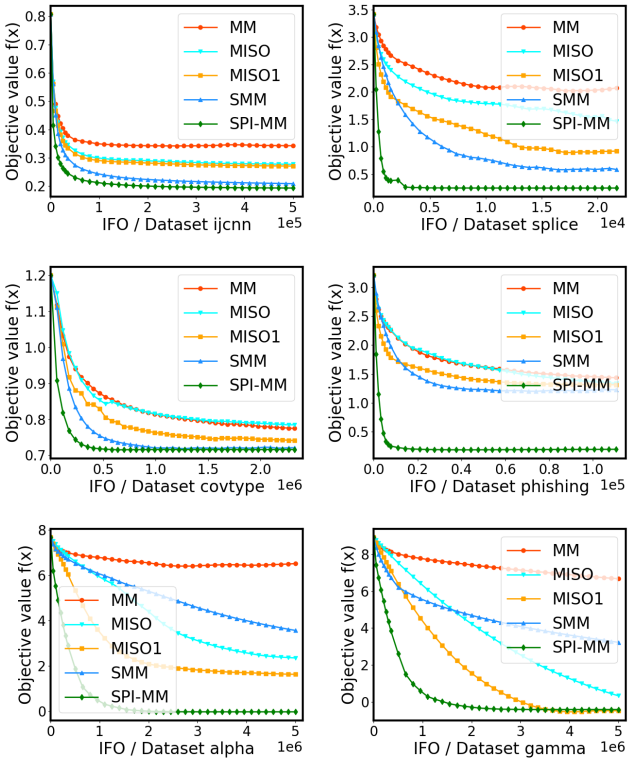

Figure 2: Comparison of algorithms on *non-convex logistic regression* problem.

according to SPI-MM. We then have the final surrogate function with omitted constants:

$$\bar{g}_t(\boldsymbol{\theta}) = (\sum_{i=1}^{t-1} [\nabla f_{\mathcal{S}_2^i}(\boldsymbol{\theta}_i) - \nabla f_{\mathcal{S}_2^i}(\boldsymbol{\theta}_{i-1})])^\top (\boldsymbol{\theta} - \boldsymbol{\theta}_{t-1}) + \frac{\lambda}{2} \sum_{j=1}^{p} \frac{|\theta[j]| - |\theta_{t-1}[j]|}{|\theta_{t-1}[j] + \varepsilon} + \frac{\mu + L_f}{2} \|\boldsymbol{\theta} - \boldsymbol{\theta}_{t-1}\|^2 .$$

We choose four algorithms as baselines, *i.e.* classic MM in Algorithm 1, MISO [Mairal, 2015], MISO1 [Mairal, 2015] and SMM [Mairal, 2013b], for performance comparison with ours on four datasets including *ijcnn, splice, covtype, phishing*[1][Chang and Lin, 2011] and two larger datasets including *alpha and gamma*[2]. MISO1 is a modification of vanilla MISO by tuning the Lipschitz parameter of the surrogate on 5% of the samples.

Figure 2 shows the curves of different algorithms on IFO vs. objective function value. One can observe that our proposed SPI-MM algorithm outperforms all compared algorithms on all the six datasets. It is obvious that SPI-MM has sharper convergence behavior w.r.t. the IFO complexity. SPI-MM also gives a lower objective function value. In contrast, the compared algorithms are more likely to get stuck in a poorer local minimum.

[1]https://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/

[2]ftp://largescale.ml.tu-berlin.de/largescale/

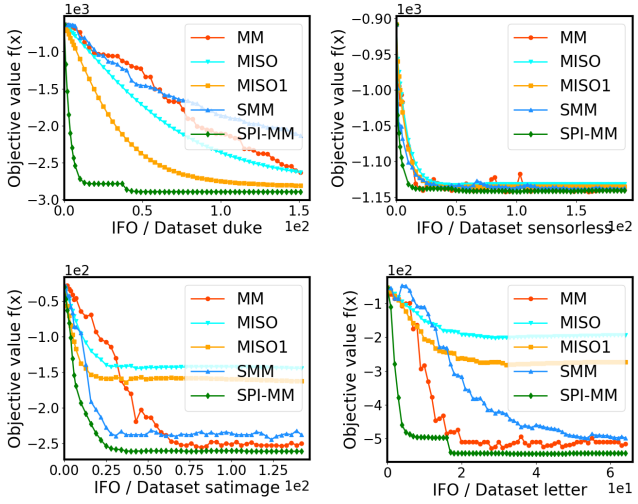

Figure 3: Comparison of selected algorithms on *sparse-PCA* problem.

In the second group of experiments, we compare our SPI-MM algorithm with the same four baselines as above on four datasets including *duke, satimage, sensorless* and *letter* [1], on the sparse-PCA problem that is formulated as below [Zou *et al.*, 2006]:

$$\arg\min_{\boldsymbol{\theta}} \left\{ -\boldsymbol{\theta}^\top \boldsymbol{\Sigma} \boldsymbol{\theta} + \lambda \|\boldsymbol{\theta}\|^2 \right\}, \text{ with } \boldsymbol{\Sigma} \triangleq X^\top X.$$

where $\boldsymbol{\Sigma} \triangleq X^\top X, X \in \mathbb{R}^{n \times p}$, n is the number of samples and p is the dimension. Our algorithm SPI-MM exhibits faster convergence rate w.r.t. *IFO* as shown in Figure 3. SPI-MM decreases the objective function value much faster than others. We observe that the loss curves of SMM and MM are less stable here. Our assumed explanation is that they lack past gradient information as supervision.

Both groups of experiments clearly demonstrate the advantage of the proposed SPI-MM over well-developed baseline algorithms.

5 Conclusions

In this work, we propose a generalized concept of surrogate that is core to MM algorithms, which requires milder conditions for bounding the objective functions. This generalized conception facilitates the convergence analysis for non-convex problems. We then develop a specific algorithm SPI-MM based on the generalized surrogate and prove its non-asymptotic convergence rate and *IFO* complexity. Numerical results confirm the computational superiority of SPI-MM over other popular algorithms.

Acknowledgments

Hu Zhang (No. 201706340188) is partially supported by the Chinese Scholarship Council. The work of Jiashi Feng was partially supported by NUS IDS R-263-000-C67-646, ECRA R-263-000-C87-133 and MOE Tier-II R-263-000-D17-112.

References

[Agarwal and Bottou, 2014] Alekh Agarwal and Leon Bottou. A lower bound for the optimization of finite sums. *arXiv preprint arXiv:1410.0723*, 2014.

[Bietti and Mairal, 2017] Alberto Bietti and Julien Mairal. Stochastic optimization with variance reduction for infinite datasets with finite sum structure. In *Advances in Neural Information Processing Systems*, pages 1623–1633, 2017.

[Borwein and Lewis, 2010] Jonathan Borwein and Adrian S Lewis. *Convex analysis and nonlinear optimization: theory and examples*. Springer Science & Business Media, 2010.

[Cappé and Moulines, 2009] Olivier Cappé and Eric Moulines. On-line expectation–maximization algorithm for latent data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):593–613, 2009.

[Chang and Lin, 2011] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[Defazio *et al.*, 2014] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in neural information processing systems*, pages 1646–1654, 2014.

[Draper and Smith, 2014] Norman R Draper and Harry Smith. *Applied regression analysis*, volume 326. John Wiley & Sons, 2014.

[Fang *et al.*, 2018] Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, pages 687–697, 2018.

[Feng *et al.*, 2013] Jiashi Feng, Huan Xu, and Shuicheng Yan. Online robust pca via stochastic optimization. In *Advances in Neural Information Processing Systems*, pages 404–412, 2013.

[Gasso *et al.*, 2009] Gilles Gasso, Alain Rakotomamonjy, and Stéphane Canu. Recovering sparse signals with a certain family of nonconvex penalties and dc programming. *IEEE Transactions on Signal Processing*, 57(12):4686–4698, 2009.

[Johnson and Zhang, 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.

[Kang *et al.*, 2015] Yangyang Kang, Zhihua Zhang, and Wu-Jun Li. On the global convergence of majorization minimization algorithms for nonconvex optimization problems. *arXiv preprint arXiv:1504.07791*, 2015.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Lange *et al.*, 2000] Kenneth Lange, David R Hunter, and Ilsoon Yang. Optimization transfer using surrogate objective functions. *Journal of computational and graphical statistics*, 9(1):1–20, 2000.

[LeCun *et al.*, 2015] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[Mairal, 2013a] Julien Mairal. Optimization with first-order surrogate functions. In *International Conference on Machine Learning*, pages 783–791, 2013.

[Mairal, 2013b] Julien Mairal. Stochastic majorization-minimization algorithms for large-scale optimization. In *Advances in Neural Information Processing Systems*, pages 2283–2291, 2013.

[Mairal, 2015] Julien Mairal. Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM Journal on Optimization*, 25(2):829–855, 2015.

[Nesterov, 1998] Yurii Nesterov. Introductory lectures on convex programming volume i: Basic course. *Lecture notes*, 1998.

[Nguyen *et al.*, 2017] Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2613–2621. JMLR. org, 2017.

[Parizi *et al.*, 2015] Sobhan Naderi Parizi, Kun He, Stan Sclaroff, and Pedro Felzenszwalb. Generalized majorization-minimization. *arXiv preprint arXiv:1506.07613*, 2015.

[Razaviyayn *et al.*, 2016] Meisam Razaviyayn, Maziar Sanjabi, and Zhi-Quan Luo. A stochastic successive minimization method for nonsmooth nonconvex optimization with applications to transceiver design in wireless communication networks. *Mathematical Programming*, 157(2):515–545, 2016.

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

[Vaida, 2005] Florin Vaida. Parameter convergence for em and mm algorithms. *Statistica Sinica*, 15(3):831, 2005.

[Wang *et al.*, 2018] Zhe Wang, Kaiyi Ji, Yi Zhou, Yingbin Liang, and Vahid Tarokh. Spiderboost: A class of faster variance-reduced algorithms for nonconvex optimization. *arXiv preprint arXiv:1810.10690*, 2018.

[Xu *et al.*, 2016] Chen Xu, Zhouchen Lin, Zhenyu Zhao, and Hongbin Zha. Relaxed majorization-minimization for non-smooth and non-convex optimization. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[Zou *et al.*, 2006] Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.