Incremental Majorization-Minimization Optimization with Application to Large-Scale Machine Learning

Julien Mairal

Incremental Majorization-Minimization Optimization with Application to Large-Scale Machine Learning*

Julien Mairal†

February 18, 2014

Abstract

Majorization-minimization algorithms consist of successively minimizing a sequence of upper bounds of the objective function. These upper bounds are tight at the current estimate, and each iteration monotonically drives the objective function downhill. Such a simple principle is widely applicable and has been very popular in various scientific fields, especially in signal processing and statistics. In this paper, we propose an incremental majorization-minimization scheme for minimizing a large sum of continuous functions, a problem of utmost importance in machine learning. We present convergence guarantees for non-convex and convex optimization when the upper bounds approximate the objective up to a smooth error; we call such upper bounds "first-order surrogate functions". More precisely, we study asymptotic stationary point guarantees for non-convex problems, and for convex ones, we provide convergence rates for the expected objective function value. We apply our scheme to composite optimization and obtain a new incremental proximal gradient algorithm with linear convergence rate for strongly convex functions. In our experiments, we show that our method is competitive with the state of the art for solving large-scale machine learning problems such as logistic regression, and we demonstrate its usefulness for sparse estimation with non-convex penalties.

1 Introduction

The principle of successively minimizing upper bounds of the objective function is often called *majorization-minimization* [31] or *successive upper-bound minimization* [43]. Each upper bound is locally tight at the current estimate, and each minimization step decreases the value of the objective function. Even though this principle does not provide any theoretical guarantee about the quality of the returned solution, it has been very popular and widely used because of its simplicity. Many existing approaches can indeed be interpreted from the majorization-minimization point of view. For instance, this is the case of gradient-based or proximal methods [3, 13, 24, 40, 48], expectation-maximization (EM) algorithms in statistics [16, 37], difference-of-convex (DC) programming [26], boosting [12, 15], some variational Bayes techniques used in machine learning [47], and the mean-shift algorithm for finding modes of a distribution [21]. Majorizing surrogates have also been used successfully in the signal processing literature about sparse optimization [10, 14, 22], linear inverse problems in image processing [1, 19], and matrix factorization [33, 36].

In this paper, we are interested in making the majorization-minimization principle scalable for minimizing a large sum of functions:

$$\min_{\theta \in \Theta} \left[f(\theta) \triangleq \frac{1}{T} \sum_{t=1}^{T} f^t(\theta) \right], \tag{1}$$

*This work was partially supported by the Gargantua project (program Mastodons - CNRS) and the LabEx PERSYVAL (ANR-11-LABX-0025). A short version of this work was presented at the International Conference of Machine Learning (ICML) in 2013 [34].

†LEAR Project-Team, INRIA Grenoble Rhône-Alpes, 655, avenue de l'Europe, 38330 Montbonnot, France. (julien.mairal@inria.fr).

where the functions $f^t : \mathbb{R}^p \to \mathbb{R}$ are continuous, and Θ is a convex subset of $\mathbb{R}^p$. When f is non-convex, exactly solving (1) is intractable in general, and when f is also non-smooth, finding a stationary point of (1) can be difficult. The problem above when T is large can be motivated by machine learning applications, where θ represents some model parameters and each function f^t measures the adequacy of the parameters θ to an observed data point indexed by t. In this context, minimizing f amounts to finding parameters θ that explain well some observed data. In the last few years, stochastic optimization techniques have become very popular in machine learning for their empirical ability to deal with a large number T of training points [8, 18, 45, 49]. Even though these methods have inherent sublinear convergence rates for convex and strongly convex problems [30, 38], they typically have a cheap computational cost per iteration, enabling them to efficiently find an approximate solution. Recently, incremental algorithms have also been proposed for minimizing finite sums of functions [5, 44, 45]. At the price of a higher memory cost than stochastic algorithms, these incremental methods enjoy faster convergence rates, while also having a cheap per-iteration computational cost.

Our paper follows this line of work: in order to exploit the particular structure of problem (1), we propose an incremental scheme whose cost per iteration is independent of T, as soon as the upper bounds of the objective are appropriately chosen. We call the resulting scheme "MISO" (*Minimization by Incremental Surrogate Optimization*). We present convergence results when the upper bounds are chosen among the class of "first-order surrogate functions", which approximate the objective function up to a smooth error—that is, differentiable with a Lipschitz continuous gradient. For non-convex problems, we obtain almost sure convergence and asymptotic stationary point guarantees. In addition, when assuming the surrogates to be strongly convex, we provide convergence rates for the expected value of the objective function. Remarkably, the convergence rate of MISO is linear for minimizing strongly convex composite objective functions, a property shared with two other incremental algorithms for smooth and composite convex optimization: the *stochastic average gradient* method (SAG) of Schmidt, Le Roux and Bach [44], and the *stochastic dual coordinate ascent* method (SDCA) of Shalev-Schwartz and Zhang [45]. Our scheme MISO is inspired in part by these two works, but yields different update rules than SAG or SDCA, and is also appropriate for non-convex optimization problems.

In the experimental section of this paper, we show that MISO can be useful for solving large-scale machine learning problems, and that it matches or outperforms cutting-edge solvers for large-scale logistic regression [3, 44]. Then, we show that our approach provides an effective incremental DC programming algorithm, which we apply to sparse estimation problems with nonconvex penalties [10].

The paper is organized as follows: Section 2 introduces the majorization-minimization principle with first-order surrogate functions. Section 3 is devoted to our incremental scheme MISO. Section 4 presents some numerical experiments, and Section 5 concludes the paper. Some basic definitions are given in Appendix A.

2 Majorization-minimization with first-order surrogate functions

In this section, we present the generic majorization-minimization scheme for minimizing a function f without exploiting the structure of f—that is, without using the fact that f is a sum of functions. We describe the procedure in Algorithm 1 and illustrate its principle in Figure 1. At iteration n, the estimate θ_n is obtained by minimizing a surrogate function g_n of f. When g_n uniformly upper-bounds f and when $g_n(\theta_{n-1}) = f(\theta_{n-1})$, it is clear that the objective function value monotonically decreases.

Algorithm 1 Basic majorization-minimization scheme.

input $\theta_0 \in \Theta$ (initial estimate); N (number of iterations).

1: **for** $n = 1, \ldots, N$ **do**

2: Compute a surrogate function g_n of f near θ_{n-1};

3: Minimize the surrogate and update the solution: $\theta_n \in \arg\min_{\theta \in \Theta} g_n(\theta)$.

4: **end for**

output θ_N (final estimate);

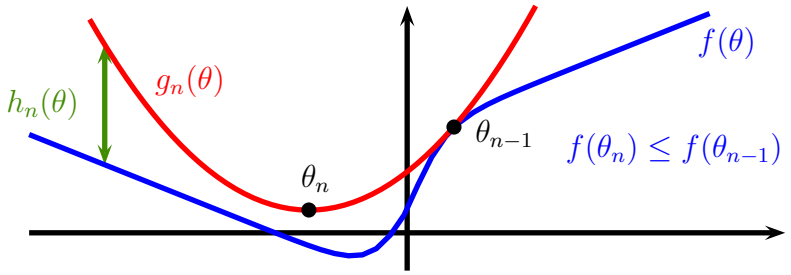

Figure 1: Illustration of the basic majorization-minimization principle. We compute a surrogate g_n of f near the current estimate θ_{n-1}. The new estimate θ_n is a minimizer of g_n. The function $h_n = g_n - f$ is the approximation error that is made when replacing f by g_n.

For this approach to be effective, we intuitively need functions g_n that are easy to minimize and that approximate well the objective f. In our paper, we measure the quality of the approximation through the smoothness of the error function $h_n \triangleq g_n - f$, which is a key quantity arising in the convergence analysis. More precisely, we require h_n to be L-smooth for some constant $L > 0$, as defined below:

Definition 2.1 (L-smooth functions)**.** *A function $f : \mathbb{R}^p \to \mathbb{R}$ is called L-smooth when it is differentiable and its gradient ∇f is L-Lipschitz continuous.*

With this definition in hand, we now introduce the class of "first-order surrogate functions", which will be shown to have good enough properties for analyzing the convergence of Algorithm 1 and the variants that we propose.

Definition 2.2 (First-order surrogate functions)**.** *A function $g : \mathbb{R}^p \to \mathbb{R}$ is a first-order surrogate function of f near κ in Θ when*

(i) *$g(\theta') \geq f(\theta')$ for all minimizers θ' of g over Θ. When the more general condition $g \geq f$ holds, we say that g is a* majorizing *surrogate;*

(ii) *the approximation error $h \triangleq g - f$ is L-smooth, $h(\kappa) = 0$, and $\nabla h(\kappa) = 0$.*

We denote by $\mathcal{S}_L(f, \kappa)$ the set of first-order surrogate functions and by $\mathcal{S}_{L,\rho}(f, \kappa) \subset \mathcal{S}_L(f, \kappa)$ the subset of ρ-strongly convex surrogates.

First-order surrogates are interesting because their approximation error—the difference between the surrogate and the objective—can be easily controlled. This is formally stated in the next lemma, which is a building block of our analysis:

Lemma 2.3 (Basic properties of first-order surrogate functions)**.** *Let g be a surrogate function in $\mathcal{S}_L(f, \kappa)$ for some κ in Θ. Define the approximation error $h \triangleq g - f$, and let θ' be a minimizer of g over Θ. Then, for all θ in Θ,*

- $|h(\theta)| \leq \frac{L}{2}\|\theta - \kappa\|_2^2$;
- $f(\theta') \leq f(\theta) + \frac{L}{2}\|\theta - \kappa\|_2^2$.

Assume that g is ρ-strongly convex, i.e., g is in $\mathcal{S}_{L,\rho}(f, \kappa)$. Then, for all θ in Θ,

- $f(\theta') + \frac{\rho}{2}\|\theta' - \theta\|_2^2 \leq f(\theta) + \frac{L}{2}\|\theta - \kappa\|_2^2$.

For the sake of conciseness, the proofs of all lemmas and propositions in this paper are relegated to Appendix B, and basic definitions are presented in Appendix A.

2.1 Non-convex convergence analysis

For general non-convex problems, proving convergence to a global (or local) minimum is impossible in general, and classical analysis studies instead asymptotic stationary point conditions (see, e.g., [4]). To do so, we make the following mild assumption when f is non-convex:

(A) f is bounded below and for all θ, θ' in Θ, the directional derivative $\nabla f(\theta, \theta' - \theta)$ of f at θ in the direction $\theta' - \theta$ exists.

The definitions of directional derivatives and stationary points are provided in Appendix A. A necessary first-order condition for θ to be a local minimum of f is to have $\nabla f(\theta, \theta' - \theta) \geq 0$ for all θ' in Θ (see, e.g., [7]). In other words, there is no feasible descent direction $\theta' - \theta$ and θ is a stationary point. Thus, we consider the following condition for assessing the quality of a sequence $(\theta_n)_{n \geq 0}$ for non-convex problems:

Definition 2.4 (Asymptotic stationary point)**.** *Under assumption* (A)*, a sequence $(\theta_n)_{n \geq 0}$ satisfies the asymptotic stationary point condition if*

$$\liminf_{n \to +\infty} \inf_{\theta \in \Theta} \frac{\nabla f(\theta_n, \theta - \theta_n)}{\|\theta - \theta_n\|_2} \geq 0. \tag{2}$$

Note that if f is differentiable on $\mathbb{R}^p$ and $\Theta = \mathbb{R}^p$, $\nabla f(\theta_n, \theta - \theta_n) = \nabla f(\theta_n)^\top (\theta - \theta_n)$, and the condition (2) implies that the sequence $(\nabla f(\theta_n))_{n \geq 0}$ converges to 0*.*

As noted above, we recover the classical definition of critical points for the smooth unconstrained case. We now give a first convergence result about Algorithm 1.

Proposition 2.5 (Non-convex analysis for Algorithm 1)**.** *Assume that* (A) *holds and that the surrogates g_n from Algorithm 1 are in $\mathcal{S}_L(f, \theta_{n-1})$ and are either majorizing f or strongly convex. Then, $(f(\theta_n))_{n \geq 0}$ monotonically decreases, and $(\theta_n)_{n \geq 0}$ satisfies the asymptotic stationary point condition.*

This proposition provides convergence guarantees for a large class of existing algorithms, including cases where f is non-smooth. In the next proposition, we relax some of the assumptions and obtain similar guarantees when only part of the objective function is approximated by a first-order surrogate.

Proposition 2.6 (Non-convex analysis for Algorithm 1 - partial surrogate)**.** *Assume that* (A) *holds and that the cost function f can be written as $f = f' \circ e$, where $e : \mathbb{R}^p \to \mathbb{R}^d$ is a C-Lipschitz function for some $C > 0$, $f' : \mathbb{R}^d \to \mathbb{R}$, and $\circ$ is the composition operator. In other words, $f(\theta) = f'(e(\theta))$ for all θ in $\mathbb{R}^p$. Assume that the function g_n in Algorithm 1 is defined as $g_n \triangleq g'_n \circ e$, where g'_n is a majorizing surrogate in $\mathcal{S}_L(f', e(\theta_{n-1}))$. Then the conclusions of Proposition 2.5 hold.*

In this proposition, g_n is a partial first-order surrogate of $f = f' \circ e$, where the part e is Lipschitz continuous. This extension of Proposition 2.5 is useful since it provides convergence results for classical approaches that will be described later in Section 2.3. Note that convergence results for non-convex problems are by nature weak, and our non-convex analysis does not provide any convergence rate. This is not the case when f is convex, as shown in the next section.

2.2 Convex analysis

The next proposition is based on a proof technique from Nesterov [40], originally designed for the proximal gradient method. We indeed obtain the same convergence rates as in [40].

Proposition 2.7 (Convex analysis for $\mathcal{S}_L(f, \kappa)$)**.** *Assume that f is convex, bounded below, and that there exists a constant $R > 0$ such that*

$$\|\theta - \theta^\star\|_2 \leq R \quad \textit{for all } \theta \in \Theta \;\; \textit{s.t.} \;\; f(\theta) \leq f(\theta_0), \tag{3}$$

where $\theta^\star$ is a minimizer of f on Θ. When the functions g_n in Algorithm 1 are in $\mathcal{S}_L(f, \theta_{n-1})$, we have for all $n \geq 1$,

$$f(\theta_n) - f^\star \leq \frac{2LR^2}{n+2},$$

where $f^\star \triangleq f(\theta^\star)$. Assume now that f is μ-strongly convex. Regardless of condition (3), we have for all $n \geq 1$,

$$f(\theta_n) - f^\star \leq \beta^n (f(\theta_0) - f^\star),$$

where $\beta \triangleq \frac{L}{\mu}$ if $\mu > 2L$ or $\beta \triangleq \left(1 - \frac{\mu}{4L}\right)$ otherwise.

The result of Proposition 2.7 is interesting because it does not make any strong assumption about the surrogate functions, except the ones from Definition 2.2. The next proposition shows that slightly better rates can be obtained with additional strong convexity assumptions.

Proposition 2.8 (Convex analysis for $\mathcal{S}_{L,\rho}(f, \kappa)$)**.** *Assume that f is convex, bounded below, and let $\theta^\star$ be a minimizer of f on Θ. When the surrogates g_n of Algorithm 1 are in $\mathcal{S}_{L,\rho}(f, \theta_{n-1})$ with $\rho \geq L$, we have for all $n \geq 1$,*

$$f(\theta_n) - f^\star \leq \frac{L\|\theta_0 - \theta^\star\|_2^2}{2n},$$

where $f^\star \triangleq f(\theta^\star)$. When f is μ-strongly convex, we have for all $n \geq 1$,

$$f(\theta_n) - f^\star \leq \left(\frac{L}{\rho + \mu}\right)^{n-1} \frac{L\|\theta_0 - \theta^\star\|_2^2}{2}.$$

Even though the constants obtained in the rates of Proposition 2.8 are slightly better than the ones of Proposition 2.7, the condition g_n in $\mathcal{S}_{L,\rho}(f, \kappa)$ with $\rho \geq L$ is much stronger than the simple assumption that g_n is in $\mathcal{S}_L(f, \kappa)$. It can indeed be shown that f is necessarily $(\rho - L)$-strongly convex if $\rho > L$, and convex if $\rho = L$. In the next section, we present some examples where such a condition holds.

2.3 Examples of first-order surrogate functions

In this section, we present practical first-order surrogate functions and different links between Algorithm 1 and existing approaches described in the literature.

2.3.1 Lipschitz gradient surrogates

When f is L-smooth, the following function is a majorizing surrogate in $\mathcal{S}_{2L,L}(f, \kappa)$:

$$g : \theta \mapsto f(\kappa) + \nabla f(\kappa)^\top (\theta - \kappa) + \frac{L}{2}\|\theta - \kappa\|_2^2.$$

Moreover, when f is convex, g is in $\mathcal{S}_{L,L}(f, \kappa)$, and when f is μ-strongly convex, g is in $\mathcal{S}_{L-\mu,L}(f, \kappa)$. This statement can be shown by using Lemmas A.5 and A.7 from the appendix. We remark that minimizing g amounts to performing a classical gradient descent step: $\theta' \leftarrow \kappa - \frac{1}{L}\nabla f(\kappa)$.

2.3.2 Proximal gradient surrogates

Let us now consider a composite optimization problem, meaning that f splits into two parts $f = f_1 + f_2$, where f_1 is L-smooth. Then, f admits the following majorizing surrogate in $\mathcal{S}_{2L}(f, \kappa)$, or in $\mathcal{S}_{2L,L}(f, \kappa)$ when f_2 is convex:

$$g : \theta \mapsto f_1(\kappa) + \nabla f_1(\kappa)^\top (\theta - \kappa) + \frac{L}{2}\|\theta - \kappa\|_2^2 + f_2(\theta).$$

The approximation error $g - f$ is indeed the same as in Section 2.3.1 and thus,

- when f_1 is convex, g is in $\mathcal{S}_L(f,\kappa)$. If f_2 is also convex, g is in $\mathcal{S}_{L,L}(f,\kappa)$;
- when f_1 is μ-strongly convex, g is in $\mathcal{S}_{L-\mu}(f,\kappa)$. If f_2 is also convex, g is in $\mathcal{S}_{L-\mu,L}(f,\kappa)$.

Minimizing g amounts to performing one step of the proximal gradient algorithm [3, 40, 48]. It is indeed easy to show that the minimum θ' of g—assuming it is unique—can be equivalently obtained as follows:

$$\theta' = \operatorname*{arg\,min}_{\theta\in\Theta} \left[\frac{1}{2} \left\| \theta - \left(\kappa - \frac{1}{L} \nabla f_1(\kappa) \right) \right\|_2^2 + \frac{1}{L} f_2(\theta) \right],$$

which is often written under the form $\theta' = \mathrm{Prox}_{f_2/L}[\kappa - (1/L)\nabla f_1(\kappa)]$, where "Prox" is called the "proximal operator". In some cases, the proximal operator can be computed efficiently in closed form, for example when f_2 is the ℓ_1-norm; it yields the iterative soft-thresholding algorithm for sparse estimation [14]. For a review of proximal operators and their computations, we refer the reader to [2].

2.3.3 Linearizing concave functions and DC programming

Assume that $f = f_1 + f_2$, where f_2 is concave and L-smooth. Then, the following function g is a majorizing surrogate in $\mathcal{S}_L(f,\kappa)$:

$$g : \theta \mapsto f_1(\theta) + f_2(\kappa) + \nabla f_2(\kappa)^\top (\theta - \kappa).$$

Such a surrogate appears in DC (difference of convex) programming (see [26]). When f_1 is convex, f can indeed be interpreted as the difference of two convex functions. It is also used in sparse estimation for dealing with some non-convex sparsity-inducing penalties [2]. For example, consider a cost function of the form $\theta \mapsto f_1(\theta) + \lambda \sum_{j=1}^p \log(|\theta[j]| + \varepsilon)$, where $\theta[j]$ is the j-th entry in θ. Even though the functions $\theta \mapsto \log(|\theta[j]| + \varepsilon)$ are not differentiable, they can be written as the composition of a concave smooth function $u \mapsto \log(u + \varepsilon)$ on $\mathbb{R}^+$, and a Lipschitz function $\theta \mapsto |\theta[j]|$. By upper-bounding the logarithm function by its linear approximation, it is then possible to use Proposition 2.6 to justify using the following partial surrogate:

$$g : \theta \mapsto f_1(\theta) + \lambda \sum_{j=1}^p \log(|\kappa[j]| + \varepsilon) + \lambda \sum_{j=1}^p \frac{|\theta[j]| - |\kappa[j]|}{|\kappa[j]| + \varepsilon}, \tag{4}$$

and minimizing g amounts to performing one step of the reweighted-ℓ_1 algorithm of Candès, Wakin and Boyd [10]. Similarly, other penalty functions are adapted to this framework. For instance, the logarithm can be replaced by any smooth concave non-decreasing function, or group-sparsity penalties [46, 50] can be used, such as $\theta \mapsto \sum_{g\in\mathcal{G}} \log(\|\theta_g\|_2 + \varepsilon)$, where $\mathcal{G}$ is a partition of $\{1,\dots,p\}$ and θ_g records the entries of θ corresponding to the set g. Proposition 2.6 indeed applies to this setting.

2.3.4 Variational surrogates

Let us now consider a real-valued function f defined on $\mathbb{R}^{p_1} \times \mathbb{R}^{p_2}$. Let $\Theta_1 \subseteq \mathbb{R}^{p_1}$ and $\Theta_2 \subseteq \mathbb{R}^{p_2}$ be two convex sets. Minimizing f over $\Theta_1 \times \Theta_2$ is equivalent to minimizing the function $\tilde{f}$ over Θ_1 defined as $\tilde{f}(\theta_1) \triangleq\!\!\mapsto \min_{\theta_2\in\Theta_2} f(\theta_1,\theta_2)$. Assume now that

- $\theta_2 \mapsto f(\theta_1,\theta_2)$ is μ-strongly convex for all θ_1 in $\mathbb{R}^{p_1}$;
- $\theta_1 \mapsto f(\theta_1,\theta_2)$ is differentiable for all θ_2;
- $(\theta_1,\theta_2) \mapsto \nabla_1 f(\theta_1,\theta_2)$ is L'-Lipschitz with respect to θ_1 and L-Lipschitz with respect to θ_2.[1]

Let us fix κ_1 in Θ_1. Then, the following function is a majorizing surrogate in $\mathcal{S}_{L'}(\tilde{f},\kappa)$:

$$g : \theta_1 \mapsto f(\theta_1, \kappa_2^\star) \text{ with } \kappa_2^\star \triangleq \operatorname*{arg\,min}_{\theta_2\in\Theta_2} \tilde{f}(\kappa_1,\theta_2),$$

[1] The notation ∇_1 denotes the gradient with respect to θ_1.

with $L'' = 2L' + L^2/\mu$. We can indeed apply Lemma A.8, which ensures that $\tilde{f}$ is differentiable with $\nabla\tilde{f}(\theta_1) = \nabla_1 f(\theta_1, \theta_2^\star)$ and $\theta_2^\star \triangleq \arg\min f(\theta_1, \theta_2)$ for all θ_1. Moreover, g is L'-smooth and $\tilde{f}$ is $L' + L^2/\mu$-smooth according to Lemma A.8, and thus $h \triangleq g - \tilde{f}$ is L''-smooth.[2]

When the surrogate g is used, Algorithm 1 corresponds to a block-coordinate descent algorithm with two blocks. Variational surrogates might also be useful for problems of a single variable θ_1. Let us give a concrete example from a regression problem with a Huber loss function $H : \mathbb{R} \mapsto \mathbb{R}$, defined for all u in $\mathbb{R}$ as

$$H(u) \triangleq \begin{cases} \frac{u^2}{2\delta} + \frac{\delta}{2} & \text{if } |u| \leq \delta \\ |u| & \text{otherwise} \end{cases},$$

where δ is a positive constant.[3] The Huber loss can be seen as a smoothed version of the ℓ_1-norm when δ is small, or simply a robust variant of the squared loss $u \mapsto \frac{1}{2}u^2$ that asymptotically grows linearly. Then, it is easy to show that the Huber loss admits the following variational representation:

$$H(u) = \frac{1}{2} \min_{w \geq \delta} \left[\frac{u^2}{w} + w \right].$$

Consider now a regression problem with m training data points represented by vectors $\mathbf{x}_i$ in $\mathbb{R}^p$, associated to real numbers y_i, for $i = 1, \ldots, m$. The robust regression problem with the Huber loss can be formulated as the minimization over $\mathbb{R}^p$ of

$$\tilde{f} : \theta_1 \mapsto \sum_{i=1}^m H(y_i - \mathbf{x}_i^\top \theta_1) = \min_{\theta_2 \in \mathbb{R}^m : \theta_2 \geq \delta} \left[f(\theta_1, \theta_2) \triangleq \frac{1}{2} \sum_{i=1}^m \frac{(y_i - \mathbf{x}_i^\top \theta_1)^2}{\theta_2[i]} + \theta_2[i] \right],$$

where θ_1 is the parameter vector of a linear model. The conditions described at the beginning of this section can be shown to be satisfied with a Lipschitz constant proportional to $(1/\delta)$; the resulting algorithm is the iterative reweighted least-square method, which appears both in the literature about robust statistics [31], and about sparse estimation where the Huber loss is used to approximate the ℓ_1-norm [2].

2.3.5 Jensen surrogates

Jensen's inequality also provides a natural mechanism to obtain surrogates for convex functions. Following the presentation of Lange, Hunger and Yang [31], we consider a convex function $f : \mathbb{R} \mapsto \mathbb{R}$, a vector $\mathbf{x}$ in $\mathbb{R}^p$, and define $\tilde{f} : \mathbb{R}^p \to \mathbb{R}$ as $\tilde{f}(\theta) \triangleq f(\mathbf{x}^\top \theta)$ for all θ. Let $\mathbf{w}$ be a weight vector in $\mathbb{R}^p_+$ such that $\|\mathbf{w}\|_1 = 1$ and $\mathbf{w}_i \neq 0$ whenever $\mathbf{x}_i \neq 0$. Then, we define for any κ in $\mathbb{R}^p$:

$$g : \theta \mapsto \sum_{i=1}^p \mathbf{w}_i f\left(\frac{\mathbf{x}_i}{\mathbf{w}_i}(\theta_i - \kappa_i) + \mathbf{x}^\top \kappa \right),$$

When f is L-smooth, and when $\mathbf{w}_i \triangleq |\mathbf{x}_i|^\nu / \|\mathbf{x}\|_\nu^\nu$, g is in $\mathcal{S}_{L'}(\tilde{f}, \kappa)$ with

- $L' = L\|\mathbf{x}\|_\infty^2 \|\mathbf{x}\|_0$ for $\nu = 0$;
- $L' = L\|\mathbf{x}\|_\infty \|\mathbf{x}\|_1$ for $\nu = 1$;
- $L' = L\|\mathbf{x}\|_2^2$ for $\nu = 2$.

As far as we know, the convergence rates we provide when using such surrogates are new. We also note that Jensen surrogates have been successfully used in machine learning. For instance, Della Pietra [15] interpret boosting procedures under this point of view through the concept of *auxiliary functions*.

[2] Note that tighter estimates of the constant L'' can be obtained in specific cases, as noted in [34].

[3] To simplify the notation, we present a shifted version of the traditional Huber loss, which usually satisfies $H(0) = 0$.

2.3.6 Quadratic surrogates

When f is twice differentiable and admits a matrix $\mathbf{H}$ such that $\mathbf{H} - \nabla^2 f$ is always positive definite, the following function is a first-order majorizing surrogate:

$$g : \theta \mapsto f(\kappa) + \nabla f(\kappa)^\top (\theta - \kappa) + \frac{1}{2}(\theta - \kappa)^\top \mathbf{H}(\theta - \kappa).$$

The Lipschitz constant of $\nabla(g - f)$ is the largest eigenvalue of $\mathbf{H} - \nabla^2 f(\theta)$ over Θ. Such surrogates appear frequently in the statistics and machine learning literature [6, 27, 29]. The goal is to to model the global curvature of the objective function during each iteration, without resorting to the Newton method. Even though quadratic surrogates do not necessarily lead to better theoretical convergence rates than simpler Lipschitz gradient surrogates, they can be quite effective in practice [27].

3 An incremental majorization-minimization algorithm: MISO

In this section, we introduce an incremental scheme that exploits the structure (1) of f as a large sum of T components. The most popular method for dealing with such a problem when f is smooth and $\Theta = \mathbb{R}^p$ is probably the *stochastic gradient descent* algorithm (SGD) and its variants (see [38]). It consists of drawing at iteration n an index $\hat{t}_n$ and updating the solution as $\theta_n \leftarrow \theta_{n-1} - \eta_n \nabla f^{\hat{t}_n}(\theta_{n-1})$, where the scalar η_n is a step size. Another popular algorithm is the *stochastic mirror descent* algorithm (see [28]) for general non-smooth convex problems, a setting we do not consider in this paper since non-smooth functions do not always admit practical first-order surrogates.

Recently, linear convergence rates for strongly convex functions f^t have been obtained in [44] and [45] by using randomized incremental algorithms whose cost per iteration is independent of T. The method SAG [44] for smooth unconstrained convex optimization is a randomized variant of the incremental gradient descent algorithm of Blatt, Hero and Gauchman [5], where an estimate of the gradient ∇f is incrementally updated at each iteration. The method SDCA [45] for strongly convex composite optimization is a dual coordinate ascent algorithm that performs incremental updates in the primal (1). Unlike SGD, both SAG and SDCA require storing information about past iterates, which is a key for obtaining fast convergence rates.

In a different context, incremental EM algorithms have been proposed by Neal and Hinton [37], where upper-bounds of a non-convex negative log-likelihood function are incrementally updated. By using similar ideas, we introduce the scheme MISO in Algorithm 2. At every iteration, a single function is observed, and an approximate surrogate of f is updated. Note that in the same line of work, Ahn et al. [1] have proposed a block-coordinate descent majorization-minimization algorithm, which corresponds to MISO when the variational surrogates of Section 2.3.4 are used.

Algorithm 2 Incremental scheme MISO.

input $\theta_0 \in \Theta$ (initial estimate); N (number of iterations).

1: Initialization: choose some surrogates g_0^t of f^t near θ_0 for all t;
2: **for** $n = 1, \ldots, N$ **do**
3: Randomly pick up one index $\hat{t}_n$ and choose a surrogate $g_n^{\hat{t}_n}$ of $f^{\hat{t}_n}$ near θ_{n-1}; set $g_n^t \triangleq g_{n-1}^t$ for all $t \neq \hat{t}_n$.
4: Update the solution: $\theta_n \in \arg\min_{\theta \in \Theta} \frac{1}{T} \sum_{t=1}^T g_n^t(\theta)$.
5: **end for**

output θ_N (final estimate);

In the next section, we study the convergence properties of the scheme MISO.

3.1 Convergence analysis

As in Section 2, we successively study non-convex and convex optimization problems. We start with the non-convex case, and make the following assumption:

(B) f is bounded below and for all θ, θ' in Θ and all t, the directional derivative $\nabla f^t(\theta, \theta' - \theta)$ of f^t at θ in the direction $\theta' - \theta$ exists.

Then, we obtain a first convergence result.

Proposition 3.1 (Non-convex analysis)**.** *Assume that* (B) *holds and that the surrogates $g_n^{\hat{t}_n}$ from Algorithm 2 are majorizing $f^{\hat{t}_n}$ and are in $\mathcal{S}_L(f^{\hat{t}_n}, \theta_{n-1})$. Then, the conclusions of Proposition 2.5 hold with probability one.*

We also give the counterpart of Proposition 2.6 for Algorithm 2.

Proposition 3.2 (Non-convex analysis - partial surrogates)**.** *Assume that* (B) *is satisfied and that the functions f^t can be written as $f^t = f'^t \circ e^t$, where the functions e^t are C-Lipschitz continuous for some $C > 0$. Assume also that the functions $g_n^{\hat{t}_n}$ in Algorithm 2 can be written as $g_n^{\hat{t}_n} = g_n'^{\hat{t}_n} \circ e^{\hat{t}_n}$, where $g_n'^{\hat{t}_n}$ is majorizing $f'^{\hat{t}_n}$ and is in $\mathcal{S}_L(f'^{\hat{t}_n}, e^{\hat{t}_n}(\theta_{n-1}))$. Then, the conclusions of Proposition 3.1 hold.*

The next lemma provides convergence rates for the convex case, under the assumption that the surrogate functions are ρ-strongly convex with $\rho \geq L$. The result notably applies to the proximal gradient surrogates of Section 2.3.2.

Proposition 3.3 (Convex analysis for strongly convex surrogate functions)**.** *Assume that f is convex and bounded below, and let $\theta^\star$ be a minimizer of f on Θ. Define $f^\star \triangleq \min_{\theta \in \Theta} f(\theta)$ and $\delta \triangleq \frac{1}{T}$. When the surrogates g_n^t in Algorithm 2 are majorizing f^t and are in $\mathcal{S}_{L,\rho}(f^t, \theta_{n-1})$ with $\rho \geq L$, we have for all $n \geq 1$,*

$$\mathbb{E}[f(\bar{\theta}_n) - f^\star] \leq \frac{L\|\theta^\star - \theta_0\|_2^2}{2\delta n}, \tag{5}$$

where $\bar{\theta}_n \triangleq \frac{1}{n}\sum_{i=1}^n \theta_i$ is the average of the iterates. Assume now that f is μ-strongly convex. For all $n \geq 1$,

$$\mathbb{E}[f(\theta_n) - f^\star] \leq \left((1-\delta) + \delta\frac{L}{\rho+\mu}\right)^{n-1} \frac{L\|\theta^\star - \theta_0\|_2^2}{2}. \tag{6}$$

The convergence rate of the previous proposition in the convex case suggests that the incremental scheme and the batch one of Section 2 have the same overall complexity, assuming that each iteration of the batch algorithm is T times the one of MISO. For strongly convex functions f^t, we obtain linear convergence rates, a property shared by SAG or SDCA; it is thus natural to make a more precise comparison with these other incremental approaches, which we present in the next two sections.

3.2 MISO for smooth unconstrained optimization

In this section, we assume that the optimization domain is unbounded—that is, $\Theta = \mathbb{R}^p$, and that the functions f^t are L-smooth. When using the Lipschitz gradient surrogates of Section 2.3.1, MISO amounts to iteratively using the following update rule:

$$\theta_n \leftarrow \frac{1}{T}\sum_{t=1}^T \kappa_{n-1}^t - \frac{1}{LT}\sum_{t=1}^T \nabla f^t(\kappa_{n-1}^t), \tag{7}$$

where the vectors κ_{n-1} are recursively defined for $n \geq 2$ as $\kappa_{n-1}^{\hat{t}_n} = \theta_{n-1}$ and $\kappa_{n-1}^t = \kappa_{n-2}^t$ for $t \neq \hat{t}_n$, with $\kappa_0^t = \theta_0$ for all t. It is then easy to see that the complexity of updating θ_n is independent of T, by storing

the vectors $\mathbf{z}_n^t = \kappa_{n-1}^t - (1/L)\nabla f^t(\kappa_{n-1}^t)$ and performing the update $\theta_n = \theta_{n-1} + (1/T)(\mathbf{z}_n^t - \mathbf{z}_{n-1}^t)$. In comparison, the approach SAG yields a different, but related, update rule:

$$\theta_n \leftarrow \theta_{n-1} - \frac{\alpha}{T}\sum_{t=1}^{T} \nabla f^t(\kappa_{n-1}^t), \tag{8}$$

where the value $\alpha = 1/(16L)$ is suggested in [44]. Even though the rules (7) and (8) seem to be similar to each other at first sight, they behave differently in practice and do not have the same theoretical properties. For non-convex problems, MISO is guaranteed to converge, which is not the case for SAG. For convex problems, both methods have a convergence rate of the same nature—that is, $O(T/n)$. For μ-strongly-convex problems, however, the convergence rate of SAG reported in [44] is substantially better than ours. Whereas the expected objective of SAG decreases with the rate $O(\rho^n)$ with $\rho_{\text{SAG}} = 1 - \min(\mu/(16L), 1/(8T))$, ours decreases with $\rho_{\text{MISO}} = 1 - 2\mu/(T(L+\mu))$, which is larger than ρ_{SAG} unless the problem is very well conditioned.

By maximizing the convex dual of (1) when the functions f^t are μ-strongly convex, the approach SDCA yields another update rule that resembles (7) and (8), and offers similar convergence rates as SAG. As part of the procedure, SDCA involves large primal gradient steps $\theta_{n-1} - (1/\mu)\nabla f^{\hat{t}_n}(\theta_{n-1})$ for updating the dual variables. It is thus appealing to study whether such large gradient steps can be used in (7) in the strongly convex case, regardless of the majorization-minimization principle. In other words, we want to study the use of the following surrogates within MISO:

$$g_n^t : \theta \mapsto f^t(\kappa_{n-1}^t) + \nabla f^t(\kappa_{n-1}^t)^\top(\theta - \kappa_{n-1}^t) + \frac{\mu}{2}\|\theta - \kappa_{n-1}^t\|_2^2, \tag{9}$$

which are lower bounds of the functions f^t instead of upper bounds. Then, minimizing $(1/T)\sum_{t=1}^T g_n^t$ amounts to performing the update (7) when replacing L by μ. The resulting algorithm is slightly different than SDCA, but resembles it. As shown in the next proposition, the method achieves a fast convergence rate when T is large.

Proposition 3.4 (MISO for strongly-convex unconstrained smooth problems)**.** *Assume that the functions f^t are μ-strongly convex, L-smooth, and bounded below. Let $\theta^\star$ be a minimizer of f on Θ. Assume that $T \geq 2L/\mu$. When the functions g_n^t of Eq. (9) are used in Algorithm 2, we have for all $n \geq 1$,*

$$\mathbb{E}[f(\theta_n) - f^\star] \leq \left(1 - \frac{1}{3T}\right)^n \frac{2T}{\mu}\|\nabla f(\theta_0)\|_2^2. \tag{10}$$

When the functions f^t are lower-bounded by the function $\theta \mapsto (\mu/2)\|\theta\|_2^2$, we can use the initialization $\theta_0 = 0$ and $g_0^t : \theta \mapsto (\mu/2)\|\theta\|_2^2$ for all t. Then, the constant $(2T/\mu)\|\nabla f(\theta_0)\|_2^2$ in (10) can be replaced by $Tf^\star$.

The proof technique is inspired in part by the one of SDCA [45]; the quantity $\sum_{t=1}^T g_n^t(\theta_n)$ is indeed a lower bound of $f^\star$, and plays a similar role as the dual value in SDCA. We remark that the convergence rate (10) improves significantly upon the original one (6), and is similar to the one of SAG when T is larger than $2L/\mu$.[4] However, Proposition 3.4 only applies to strongly convex problems. In other cases, the more conservative rule (7) should be preferred in theory, even though we present heuristics in Section 3.4 that suggest using larger step sizes than $1/L$ in practice.

3.3 MISO for composite optimization

When f can be written as $f = (1/T)\sum_{t=1}^T f_1^t + f_2$, where the functions f_1^t are L-smooth, we can use the proximal gradient surrogate presented in Section 2.3.2; it yields the following rule:

$$\theta_n \in \operatorname*{arg\,min}_{\theta\in\Theta} \frac{1}{2}\left\|\theta - \left(\frac{1}{T}\sum_{t=1}^T \kappa_{n-1}^t - \frac{1}{LT}\sum_{t=1}^T \nabla f_1^t(\kappa_{n-1}^t)\right)\right\|_2^2 + \frac{\lambda}{L} f_2(\theta), \tag{11}$$

[4] Note that a similar assumption appears in the first analysis of SAG published in [32] before its refinement in [44].

where the vectors κ_{n-1}^t are defined as in Section 3.2. This update is related to SDCA, as well as to stochastic methods for composite convex optimization such as the regularized dual averaging algorithm of Xiao [49]. As in the previous section, we obtain guarantees for non-convex optimization, but our linear convergence rate for strongly convex problems is not as fast as the one of SDCA. Even though we do not have a similar result as Proposition 3.4 for the composite setting, we have observed that using a smaller value for L than the theoretical one could work well in practice. We detail such an empirical strategy in the next section.

3.4 Practical implementation and heuristics

We have found the following strategies to improve the practical performance of MISO.

Initialization A first question is how to initialize the surrogates g_0^t in practice. Even though we have suggested the functions g_0^t to be in $\mathcal{S}_L(f^t, \theta_0)$ in Algorithm 2, our analysis weakly relies on this assumption. In fact, most of our results hold when choosing surrogates computed at points κ_0^t that are not necessarily equal to θ_0; at most only constants from the convergence rates would be affected by such a change. An effective empirical strategy is inspired by the second part of Proposition 3.4: we first define functions $g_0^t : \theta \mapsto (L/2)\|\theta - \theta_0\|_2^2$, and perform T iterations of MISO without randomization, selecting the function f^t at iteration t, such that each surrogate is updated exactly once. Then, we use these updated surrogates for initializing the regular randomized scheme.

Warm restart and continuation When available, warm restart can be used for initializing the surrogates. Assume that we are interested in minimizing a composite function $(1/T)\sum_{t=1}^T f_1^t(\theta) + \lambda f_2(\theta)$, which is parameterized by a scalar λ, and that we want to obtain a minimizer for several parameter values $\lambda_1 < \lambda_2 < ... < \lambda_M$. We first solve the problem for $\lambda = \lambda_M$, and then use the surrogates obtained at the end of the optimization for initializing the algorithm when addressing the problem with $\lambda = \lambda_{M-1}$. We proceed similarly going from larger to smaller values of λ. We have empirically observed that the warm restart strategy could be extremely efficient in practice, and would deserve further study in a future work.

Heuristics for selecting step sizes Choosing proximal gradient surrogates g^t requires choosing some Lipschitz constant L (or a strong convexity parameter μ for Proposition 3.4), which leads to a specific step size in (11). However, finding an appropriate step size can be difficult in practice for several reasons: (i) in some cases, these parameters are unknown; (ii) even though a global Lipschitz constant might be available, a local Lipschitz constant could be more effective; (iii) the convergence rates of Proposition 3.3 can be obtained by choosing a smaller value for L than the "true" Lipschitz constant, as long as the inequality $\mathbb{E}[f(\theta_n)] \leq \mathbb{E}[\bar{g}_n(\theta_n)]$ is always satisfied, where $\bar{g}_n \triangleq (1/T)\sum_{t=1}^T g_n^t$. This motivates the following heuristics:

- MISO1 first perform one pass over $\eta = 5\%$ of the data to select a constant L' of the form $2^{-k}L$ yielding the smallest objective on the data subset.
- MISO2 proceed as in MISO1, but choose a more aggressive strategy $L = L'\eta$; during the optimization, compute the quantities a_n^t and b_n^t defined as $a_n^t = a_{n-1}^t$, $b_n^t = b_{n-1}^t$ if $t \neq \hat{t}_n$, and otherwise $a_n^{\hat{t}_n} = f^{\hat{t}_n}(\theta_{n-1})$, $b_n^{\hat{t}_n} = g_L^{\hat{t}_n}(\theta_{n-1})$, where we have parameterized the surrogates g^t by L. Every T iterations, compare the sums $A_n = \sum_{t=1}^T a_n^t$ and $B_n = \sum_{t=1}^T b_n^t$. If $A_n \leq B_n$, do nothing; otherwise, increase the value of L until this inequality is satisfied.

The heuristic MISO2 is more aggressive than MISO1 since it starts with a smaller value for L. After every iteration, this value is possibly increased such that on average, the surrogates "behave" as majorizing functions. Even though this heuristic does not come with any theoretical guarantee, it was found to perform slightly better than MISO1 for strongly-convex problems.

Using a different parameter L_t for every function f_t Even though our analysis was conducted with a global parameter L for simplicity, it is easy to extend the analysis when the parameter L is adjusted individually for every surrogate. This is useful when the functions f_t are heterogeneous.

Table 1: Description of datasets used in our experiments.

name	T	p	storage	density	size (GB)
covtype	581 012	54	dense	1	0.23
alpha	500 000	500	dense	1	1.86
ocr	2 500 000	1 155	dense	1	21.5
real-sim	72 309	20 958	sparse	0.0024	0.056
rcv1	781 265	47 152	sparse	0.0016	0.89
webspam	250 000	16 091 143	sparse	0.0002	13.90

Parallelization with mini-batches The complexity of MISO is often dominated by the cost of updating the surrogates $g_n^{\hat{t}_n}$, which typically requires computing the gradient of a function. A simple extension is to update several surrogates at the same time, when parallel computing facilities are available.

4 Experimental validation

In this section, we evaluate MISO on large-scale machine learning problems. Our implementation is coded in C++ interfaced with Matlab and is freely available in the open-source software package SPAMS [36].[5] All experiments were conducted on a single core of a 2GHz Intel CPU with 64GB of RAM.

Datasets We use six publicly available datasets, which consist of pairs $(y_t, \mathbf{x}_t)_{t=1}^T$, where the y_t's are labels in $\{-1, +1\}$ and the $\mathbf{x}_t$'s are vectors in $\mathbb{R}^p$ representing data points. The datasets are described in Table 1. alpha, rcv1, ocr, and webspam are obtained from the 2008 Pascal large-scale learning challenge.[6] covtype and real-sim are obtained from the LIBSVM website.[7] The datasets are pre-processed as follows: all dense datasets are standardized to have zero-mean and unit variance for every feature. The sparse datasets are normalized such that each $\mathbf{x}_t$ has unit ℓ_2-norm.

4.1 ℓ_2-logistic regression

We consider the ℓ_2-regularized logistic regression problem, which can be formulated as follows:

$$\min_{\theta \in \mathbb{R}^p} \frac{1}{T} \sum_{t=1}^{T} \ell(y_t, \mathbf{x}_t^\top \theta) + \frac{\lambda}{2} \|\theta\|_2^2, \tag{12}$$

where $\ell(u, \hat{u}) = \log(1 + e^{-u\hat{u}})$ for all $(u, \hat{u})$. Following [44], we evaluate different methods with the parameter $\lambda = 1/T$, which is argued to be of the same order of magnitude as the smallest value that would be used in practice for machine learning experiments. The algorithms included in the comparison are:

SGD-h the stochastic gradient descent algorithm with a heuristic for choosing the step-size similar to MISO2, and inspired by Leon Bottou's sgd toolbox for machine learning.[8] A step-size of the form $\rho/\sqrt{n + n_0}$ is automatically adjusted when performing one pass on $\eta = 5\%$ of the training data. We obtain consistent results with the performance of SGD reported by Schmidt et al. [44] when the step-size is chosen from hindsight. Based on their findings, we do not include in our figures other variants of SGD, e.g., [17, 23, 25, 49].

FISTA the accelerated gradient method proposed by Beck and Teboulle [3] with a line-search for automatically adjusting the Lipschitz constant.

[5] `http://spams-devel.gforge.inria.fr/`.
[6] `http://largescale.ml.tu-berlin.de`.
[7] `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`.
[8] available here: `http://leon.bottou.org/projects/sgd`.

SDCA the algorithm of Shalev-Schwartz and Zhang [51], efficiently implemented in the language C by Mark Schmidt.[9]

SAG a fast implementation in C provided by the first author of [44]. We use the step-size $1/L$ since it performed similarly as their heuristic line search.

MISO0 the majorization-minimization algorithm MISO, using the trivial upper bound $L^t = 0.25\|\mathbf{x}_t\|_2^2$ on the Lipschitz constant for example t.

MISO2 the majorization-minimization heuristic MISO2 described in Section 3.4.

MISOμ the update rule corresponding to Proposition 3.4, which is using lower bounds of the objective instead of upper bounds.

For sparse datasets, the update rules MISO0 and MISO2 are not practical since they require storing a dense vector of size p for every example. Thus, we use mini-batches of size $\lfloor 1/d \rfloor$, where d is the density of the dataset; the resulting algorithms, which we denote by MISO0-mb and MISO2-mb, have a storage cost equal to the one of the dataset. On the other hand, the update rule MISOμ applied to the λ-strongly convex functions $f^t : \theta \mapsto \ell(y_t, \mathbf{x}_t^\top \theta) + \frac{\lambda}{2}\|\theta\|_2^2$ admits a simple and computationally cheap form:

$$\theta_n \leftarrow \theta_{n-1} - \frac{1}{T\lambda}\left(\ell'(y_{\hat{t}_n}, \mathbf{x}_{\hat{t}_n}^\top \theta_{n-1}) - \ell'(y_{\hat{t}_n}, \mathbf{x}_{\hat{t}_n}^\top \kappa_{n-1}^{\hat{t}_n})\right)\mathbf{x}_{\hat{t}_n}, \tag{13}$$

where ℓ' denotes the derivative of ℓ with respect to its second argument. Assuming that the dataset fits into memory, the only extra quantities to store are the scalars $\ell'(y_{\hat{t}_n}, \mathbf{x}_{\hat{t}_n}^\top \kappa_{n-1}^{\hat{t}_n})$, and the resulting memory cost is simply $O(T)$.

We report our evaluation for the above methods on Figures 2 and 3, where we plot the relative duality gap defined as $(f(\theta_n) - g^\star)/g^\star$, where $g^\star$ is the best value of the Fenchel dual that we have obtained during our experiments. The conclusions of our study are the following:

- SAG, SDCA, MISOμ perform similarly for a given number of passes over the data. MISO2 performs as well for dense datasets, but becomes slower for sparse datasets due to the use of minibatches.
- MISOμ is always the fastest in terms of CPU time, due to the extreme simplicity of the update rule (13). However, we have observed that MISOμ could diverge for significantly lower values of λ, as predicted by the condition $T \geq 2L/\mu$ in Proposition 3.4, unlike SAG, SDCA, and MISO2.
- MISO0 does not perform as well; in fact its performance appears to be the same as ISTA [3] without line-search (not reported in the figures).
- SGD-h performs well at the beginning of the procedure, but is not competitive compared to incremental approaches after a few passes over the data.

Note that an evaluation of a preliminary version of MISO2 is presented in [34] for the ℓ_1-regularized logistic regression problem, where the objective function is not strongly convex. Our experimental findings showed that MISO2 was competitive with state-of-the-art solvers based on active-set and coordinate descent algorithms [20].

4.2 Non-convex sparse estimation

The majorization-minimization principle is appealing for non-convex and non-smooth optimization, where only few algorithms apply. Here, we address a sparse estimation problem presented in Section 2.3.3:

$$\min_{\theta \in \mathbb{R}^p} \frac{1}{T}\sum_{t=1}^{T} \frac{1}{2}(y_t - \mathbf{x}_t^\top \theta)^2 + \lambda \sum_{j=1}^{p} \log(|\theta[j]| + \varepsilon), \tag{14}$$

[9] available here: `http://www.di.ens.fr/~mschmidt/Software/SAG.html`.

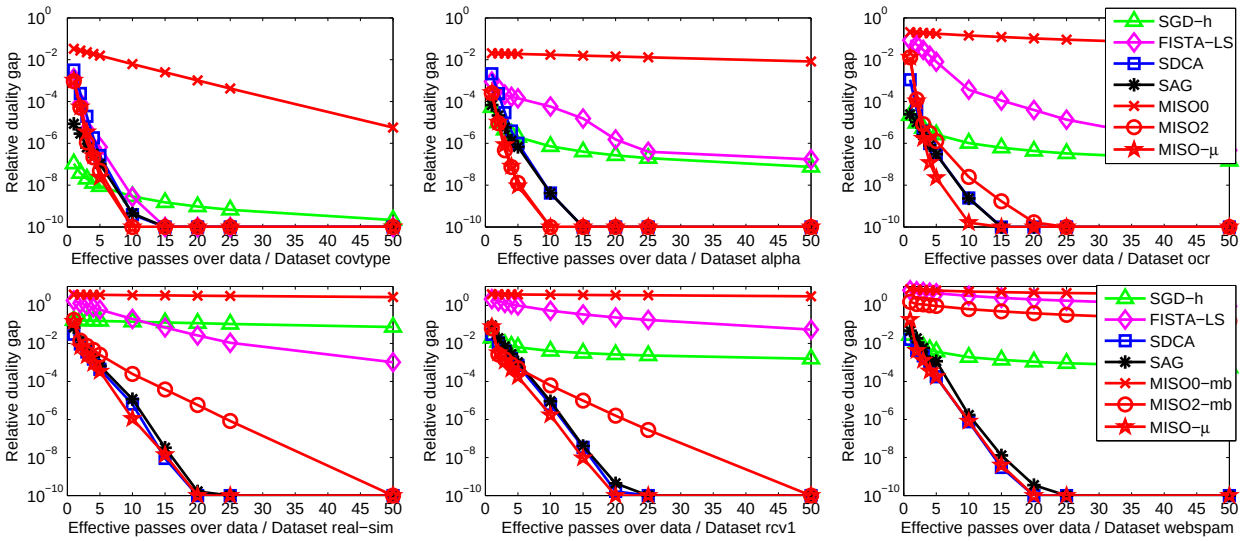

Figure 2: Relative duality gap obtained for the logistic regression experiment for different numbers of passes over the data.

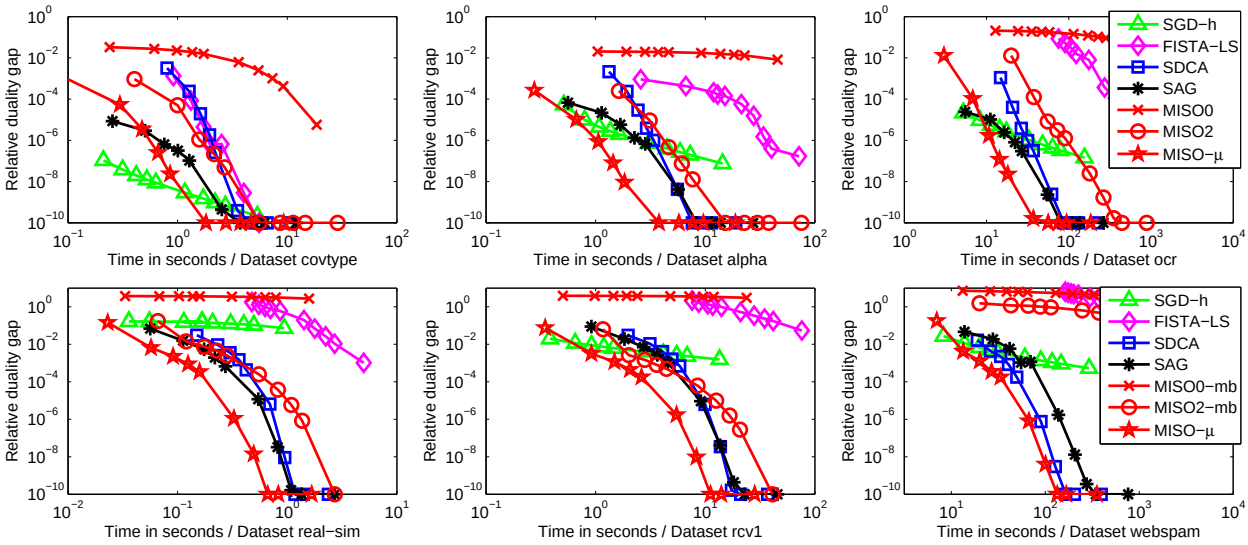

Figure 3: Relative duality gap obtained for logistic regression with respect to the CPU time.

where the scalars y_t and the vectors $\mathbf{x}_t$ are the same as in the previous section, and ε is set to 0.01. The model parameter λ controls the sparsity of the solution. Even though (14) is non-convex and non-smooth, stationary points can be obtained in various ways. In this section, we consider majorization-minimization approaches where the penalty function $\theta \mapsto \sum_{j=1}^{p} \log(|\theta[j]| + \varepsilon)$ is upper-bounded as in Eq. (4), whereas the functions $\theta \mapsto (1/2)(y_t - \mathbf{x}_t^\top \theta)^2$ are upper-bounded by the Lipschitz gradient surrogates of Section 2.3.1. We compare four approaches for finding a sparse approximate solution of (14):

MM Algorithm 1 with the trivial Lipschitz constant $L = (1/T) \sum_{t=1}^{T} 0.25 \|\mathbf{x}_t\|_2^2$;

MM-LS Algorithm 1 with the line-search scheme of ISTA [3] for adjusting L;

MISO0 as defined in Section 4.1;

MISO1 the scheme MISO with the heuristic line-search presented in Section 3.4.

We choose a parameter λ for each dataset, such that the solution with the lowest objective function obtained by any of the tested method has approximately a sparsity of 10 for datasets covtype and alpha, 100 for ocr and real-sim, and 1 000 for rcv1 and webspam. The methods were initialized with the point $\theta_0 = (\|\mathbf{y}\|_2 / \|\mathbf{X}^\top \mathbf{y}\|_2) \mathbf{X}^\top \mathbf{y}$; indeed, the initialization $\theta_0 = 0$ that was a natural choice in Section 4.1 appears to be often a bad stationary point of problem (14) and thus an inappropriate intial point. We report the objective function values for different passes over the data in Figure 4, and the sparsity of the solution in Figure 5. Our conclusions are the following:

- methods with line searches did significantly better than those without, showing that adjusting the constant L is important for these datasets;
- MISO1 did asymptotically better than MM-LS for five of the datasets after 50 epochs and slightly worse for real-sim; in general, MISO1 seems to converge substantially faster than other approaches, both in terms of objective function and in terms of the support of the solution.

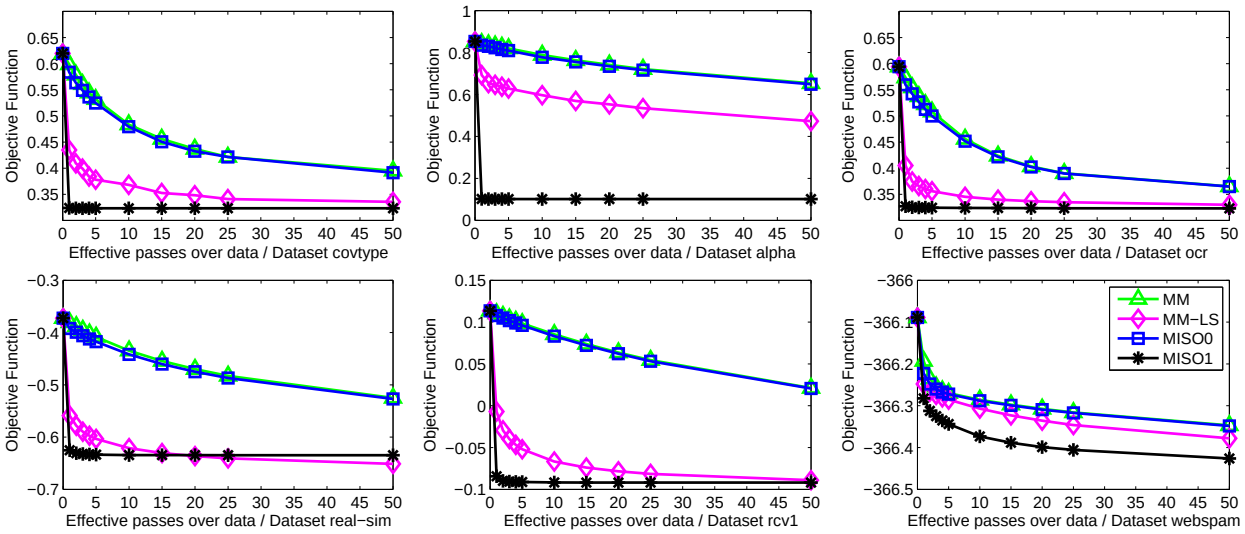

Figure 4: Objective function during the sparse estimation experiment.

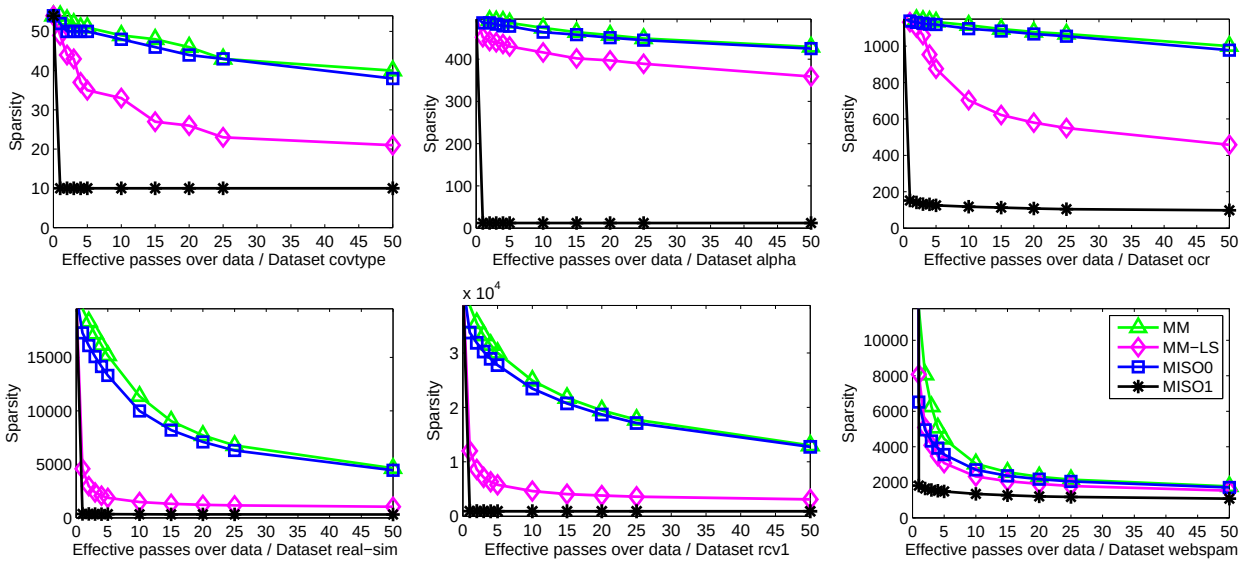

Figure 5: Sparsity of the solution during the sparse estimation experiment.

5 Conclusion

In this paper, we have presented new algorithms based on the majorization-minimization principle for minimizing a large sum of functions. The main asset of our approach is probably its applicability to a large class of non-convex problems, including non-smooth ones, where we obtain convergence and asymptotic stationary point guarantees. For convex problems, we also propose new incremental rules for composite optimization, which are competitive with state-of-the-art solvers in the context of large-scale machine learning problems such as logistic regression.

We note that other majorization-minimization algorithms have recently been analyzed, such as block coordinate variants in [34, 42] and stochastic ones in [11, 35, 43]. In particular, we have proposed in [35] a stochastic majorization-minimization algorithm that does not require to store information about past iterates, when the objective function is an expectation. Since the first version of our work was published in [35], MISO has also been extended by other authors in [52] using the alternating direction method of multipliers framework.

Acknowledgments

The author would like to thank Zaid Harchaoui, Francis Bach, Simon Lacoste-Julien, Mark Schmidt, and Martin Jaggi for fruitful discussions.

A Basic definitions and useful mathematical results

The following basic definitions can be found in classical textbooks, e.g, [4, 7, 9, 41].

Definition A.1 (Directional derivative)**.** *Let us consider a function* $f : \mathbb{R}^p \to \mathbb{R}$ *and* θ, θ' *be in* $\mathbb{R}^p$*. When it exists, the following limit is called the directional derivative of* f *at* θ *in the direction* $\theta' - \theta$*:* $\nabla f(\theta, \theta' - \theta) \triangleq \lim_{t \to 0^+} (f(\theta + t(\theta' - \theta)) - f(\theta))/t$*. When* f *is differentiable at* θ*, directional derivatives exist in every direction, and* $\nabla f(\theta, \theta' - \theta) = \nabla f(\theta)^\top (\theta' - \theta)$*.*

Definition A.2 (Stationary point)**.** *Let us consider a function $f : \Theta \subseteq \mathbb{R}^p \to \mathbb{R}$, where Θ is a convex set, such that f admits a directional derivative $\nabla f(\theta, \theta' - \theta)$ for all θ, θ' in Θ. We say that θ in Θ is a stationary point if for all θ' in Θ, $\nabla f(\theta, \theta' - \theta) \geq 0$.*

Definition A.3 (Lipschitz continuity)**.** *A function $f : \mathbb{R}^p \to \mathbb{R}$ is L-Lipschitz continuous for some $L > 0$ when for all θ, θ' in $\mathbb{R}^p$, $|f(\theta') - f(\theta)| \leq L\|\theta - \theta'\|_2$.*

Definition A.4 (Strong convexity)**.** *Let Θ be a convex set. A function $f : \Theta \subseteq \mathbb{R}^p \to \mathbb{R}$ is called μ-strongly convex when there exists a constant $\mu > 0$ such that for all θ' in Θ, the function $\theta \mapsto f(\theta) - \frac{\mu}{2}\|\theta - \theta'\|_2^2$ is convex.*

We now present a few lemmas that we use in the paper. The first one is classical and its proof can be found in Lemma 1.2.3 of [39].

Lemma A.5 (Quadratic upper bound for L-smooth functions)**.** *Let $f : \mathbb{R}^p \to \mathbb{R}$ be L-smooth. Then, for all θ, θ' in $\mathbb{R}^p$,*

$$|f(\theta') - f(\theta) - \nabla f(\theta)^\top(\theta' - \theta)| \leq \frac{L}{2}\|\theta - \theta'\|_2^2. \tag{15}$$

Lemma A.6 (Second-order growth property)**.** *Let $f : \mathbb{R}^p \to \mathbb{R}$ be μ-strongly convex, and $\Theta \subseteq \mathbb{R}^p$ be a convex set. Let $\theta^\star$ be the minimizer of f on Θ. Then, the following condition holds for all θ in Θ:*

$$f(\theta) \geq f(\theta^\star) + \frac{\mu}{2}\|\theta - \theta^\star\|_2^2.$$

Proof. Let us define the function $g : \theta \mapsto f(\theta) - \frac{\mu}{2}\|\theta - \theta^\star\|_2^2$. We show that $\theta^\star$ is a minimizer of the convex function g by checking first-order optimality conditions based on directional derivatives. For all θ in Θ, $\nabla g(\theta^\star, \theta - \theta^\star) = \nabla f(\theta^\star, \theta - \theta^\star) \geq 0$, where $\nabla f(\theta^\star, \theta - \theta^\star)$ is non-negative because $\theta^\star$ is a stationary point of f on Θ. Thus, the point $\theta^\star$ is also a stationary point of the function g on Θ, and is a minimizer of g on Θ since g is convex (see Proposition 2.1.2 of [7]). □

The next two lemmas are useful for characterizing first-order surrogate functions. Their proofs can be found in the appendix of [34].

Lemma A.7 (Regularity of residual functions)**.** *Let $f, g : \mathbb{R}^p \to \mathbb{R}$ be two functions. Define $h \triangleq g - f$. Then, if g is ρ-strongly convex and f is L-smooth, with $\rho \geq L$, h is $(\rho - L)$-strongly convex; if g and f are convex and L-smooth, h is also L-smooth; if g and f are μ-strongly convex and L-smooth, h is $(L - \mu)$-smooth.*

Lemma A.8 (Regularity of optimal value functions)**.** *Let $f : \mathbb{R}^{p_1} \times \Theta_2 \to \mathbb{R}$ be a function of two variables where $\Theta_2 \subseteq \mathbb{R}^{p_2}$ is a convex set. Assume that*

- *$\theta_1 \mapsto f(\theta_1, \theta_2)$ is differentiable for all θ_2 in Θ_2;*
- *$\theta_2 \mapsto \nabla_1 f(\theta_1, \theta_2)$ is L-Lipschitz continuous for all θ_1 in $\mathbb{R}^{p_1}$;*
- *$\theta_2 \mapsto f(\theta_1, \theta_2)$ is μ-strongly convex for all θ_1 in $\mathbb{R}^{p_1}$.*

Also define $\tilde{f}(\theta_1) \triangleq \min_{\theta_2 \in \Theta_2} f(\theta_1, \theta_2)$. Then, $\tilde{f}$ is differentiable and $\nabla \tilde{f}(\theta_1) = \nabla_1 f(\theta_1, \theta_2^\star)$, where $\theta_2^\star \triangleq \arg\min_{\theta_2 \in \Theta_2} f(\theta_1, \theta_2)$. Moreover, if $\theta_1 \mapsto \nabla_1 f(\theta_1, \theta_2)$ is L'-Lipschitz continuous for all θ_1 in $\mathbb{R}^{p_1}$, the gradient $\nabla \tilde{f}$ is $(L' + L^2/\mu)$-Lipschitz.

B Proofs of the main lemmas and propositions

We present in this section the proofs of the different lemmas and propositions in the paper.

B.1 Proof of theorem 2.3

Proof. The first inequality is a direct application of Lemma A.5 applied to the function h at the point κ when noticing that $h(\kappa) = 0$ and $\nabla h(\kappa) = 0$. Then, for all θ in Θ, we have $f(\theta') \leq g(\theta') \leq g(\theta) = f(\theta) + h(\theta)$, and we obtain the second inequality from the first one. When g is ρ-strongly convex, we use the second-order growth property of g presented in Lemma A.6, and obtain

$$f(\theta') + \frac{\rho}{2}\|\theta' - \theta\|_2^2 \leq g(\theta') + \frac{\rho}{2}\|\theta - \theta'\|_2^2 \leq g(\theta) = f(\theta) + h(\theta),$$

and the third inequality follows from the first one. □

B.2 Proof of theorem 2.5

Proof. The fact that $(f(\theta_n))_{n\geq 0}$ is non-increasing and convergent because bounded below is clear: for all $n \geq 1$,

$$f(\theta_n) \leq g_n(\theta_n) \leq g_n(\theta_{n-1}) = f(\theta_{n-1}),$$

where the first inequality and the last equality are obtained from Definition 2.2. The second inequality comes from the definition of θ_n. Denote by $f^\star$ the limit of the sequence $(f(\theta_n))_{n\geq 1}$ and by $h_n \triangleq g_n - f$ the approximation error function at iteration n. The functions h_n are L-smooth and the quantities $h_n(\theta_n)$ are non-negative. Then, $h_n(\theta_n) \leq f(\theta_{n-1}) - f(\theta_n)$, and by summation,

$$\sum_{n=1}^{\infty} h_n(\theta_n) \leq f(\theta_0) - f^\star,$$

and the non-negative sequence $(h_n(\theta_n))_{n\geq 0}$ necessarily converges to zero. Then, we have two possibilities (according to the assumptions made in the proposition).

- If the functions g_n are majorizing surrogates, plugging $\theta' = \theta_n - \frac{1}{L}\nabla h_n(\theta_n)$ in Lemma A.5 yields

$$h_n(\theta') \leq h_n(\theta_n) - \frac{1}{2L}\|\nabla h_n(\theta_n)\|_2^2,$$

and therefore, by using the fact that $h_n(\theta') \geq 0$ because $g_n \geq f$,

$$\|\nabla h_n(\theta_n)\|_2^2 \leq 2L(h_n(\theta_n) - h_n(\theta')) \leq 2Lh_n(\theta_n) \underset{n\to+\infty}{\longrightarrow} 0.$$

- If instead the functions g_n are ρ-strongly convex, we can use Lemma 2.3:

$$\frac{\rho}{2}\|\theta_n - \theta_{n-1}\|_2^2 \leq f(\theta_{n-1}) - f(\theta_n).$$

Summing over n yields that $\|\theta_n - \theta_{n-1}\|_2^2$ converges to zero, and

$$\|\nabla h_n(\theta_n)\|_2 = \|\nabla h_n(\theta_n) - \nabla h_n(\theta_{n-1})\|_2 \leq L\|\theta_n - \theta_{n-1}\|_2 \underset{n\to+\infty}{\longrightarrow} 0,$$

since $\nabla h_n(\theta_{n-1}) = 0$ according to Definition 2.2.

We now compute the directional derivative of f at a point θ_n and a direction $\theta - \theta_n$, where $n \geq 1$ and θ is in Θ:

$$\nabla f(\theta_n, \theta - \theta_n) = \nabla g_n(\theta_n, \theta - \theta_n) - \nabla h_n(\theta_n)^\top(\theta - \theta_n).$$

Note that θ_n minimizes g_n on Θ and therefore $\nabla g_n(\theta_n, \theta - \theta_n) \geq 0$. Therefore,

$$\nabla f(\theta_n, \theta - \theta_n) \geq -\|\nabla h_n(\theta_n)\|_2\|\theta - \theta_n\|_2,$$

where we use Cauchy-Schwarz's inequality. Then, by minimizing over θ and taking the infimum limit, we obtain

$$\liminf_{n\to+\infty} \inf_{\theta\in\Theta} \frac{\nabla f(\theta_n, \theta - \theta_n)}{\|\theta - \theta_n\|_2} \geq -\lim_{n\to+\infty} \|\nabla h_n(\theta_n)\|_2 = 0.$$

□

B.3 Proof of theorem 2.6

Proof. The proof follows the same steps as the one of Proposition 2.5. It is easy to show that $(f(\theta_n))_{n\geq 0}$ monotonically decreases and that $h_n(\theta_n) \triangleq g_n(\theta_n) - f(\theta_n)$ converges to zero when n grows to infinity. Note that h_n can be written as $h_n = h'_n \circ e$, where $h'_n \triangleq g'_n - f'$ is L-smooth. Proceeding as in the proof of Proposition 2.5, we can show that $\|\nabla h'_n(e(\theta_n))\|_2$ converges to zero.

Let us now fix $n \geq 1$ and consider δ such that $\theta_n + \delta$ is in Θ. We have

$$h_n(\theta_n + \delta) = h'_n(e(\theta_n + \delta)) = h'_n(e(\theta_n) + \|\delta\|_2 \mathbf{z}),$$

where $\mathbf{z}$ is a vector whose ℓ_2-norm is bounded by a universal constant $C > 0$ because the function e is Lipschitz continuous. Since h'_n is L-smooth, we also have

$$h_n(\theta_n + \delta) = h'_n(e(\theta_n) + \|\delta\|_2 \mathbf{z}) = h_n(\theta_n) + \|\delta\|_2 \nabla h'_n(e(\theta_n))^\top \mathbf{z} + O(\|\delta\|_2^2).$$

Plugging this simple relation with $\delta = t(\theta - \theta_n)$ for $0 < t < 1$ for some θ in Θ, into the definition of the directional derivative $\nabla h_n(\theta_n, \theta - \theta_n)$, we obtain the relation

$$|\nabla h_n(\theta_n, \theta - \theta_n)| \leq C \|\nabla h'_n(e(\theta_n))\|_2 \|\theta - \theta_n\|_2,$$

and since $\nabla f(\theta_n, \theta - \theta_n) = \nabla g_n(\theta_n, \theta - \theta_n) - \nabla h_n(\theta_n, \theta - \theta_n)$, and $\nabla g_n(\theta_n, \theta - \theta_n) \geq 0$,

$$\liminf_{n \to +\infty} \inf_{\theta \in \Theta} \frac{\nabla f(\theta_n, \theta - \theta_n)}{\|\theta - \theta_n\|_2} \geq -C \lim_{n \to +\infty} \|\nabla h'_n(e(\theta_n))\|_2 = 0.$$

□

B.4 Proof of theorem 2.7

Proof.

Non-strongly convex case:

Let us define $h_n \triangleq g_n - f$ the approximation error function at iteration $n \geq 1$. From Lemma 2.3, we have

$$f(\theta_n) \leq \min_{\theta \in \Theta} \left[f(\theta) + \frac{L}{2} \|\theta - \theta_{n-1}\|_2^2 \right].$$

Then, following a similar proof technique as Nesterov in [40], we have

$$\begin{aligned} f(\theta_n) &\leq \min_{\alpha \in [0,1]} \left[f(\alpha \theta^\star + (1-\alpha)\theta_{n-1}) + \frac{L\alpha^2}{2} \|\theta^\star - \theta_{n-1}\|_2^2 \right] \\ &\leq \min_{\alpha \in [0,1]} \left[\alpha f(\theta^\star) + (1-\alpha) f(\theta_{n-1}) + \frac{L\alpha^2}{2} \|\theta^\star - \theta_{n-1}\|_2^2 \right], \end{aligned} \tag{16}$$

where the minimization over Θ is replaced by a minimization over the line segment $\alpha\theta^\star + (1-\alpha)\theta_{n-1} : \alpha \in [0,1]$. Then, because the sequence $(f(\theta_n))_{n\geq 0}$ is monotonically decreasing we can use the bounded level set assumption and

$$f(\theta_n) - f^\star \leq \min_{\alpha \in [0,1]} \left[(1-\alpha)(f(\theta_{n-1}) - f^\star) + \frac{LR^2\alpha^2}{2} \right].$$

- if $f(\theta_{n-1}) - f^\star \geq LR^2$, then the optimal value $\alpha^\star$ is 1 and $f(\theta_n) - f^\star \leq \frac{LR^2}{2}$;
- otherwise $\alpha^\star = \frac{f(\theta_{n-1}) - f^\star}{LR^2}$ and $r_n \leq r_{n-1}\left(1 - \frac{r_{n-1}}{2LR^2}\right)$, where $r_n \triangleq f(\theta_n) - f^\star$. Thus, $r_n^{-1} \geq r_{n-1}^{-1}\left(1 - \frac{r_{n-1}}{2LR^2}\right)^{-1} \geq r_{n-1}^{-1} + \frac{1}{2LR^2}$, where the second inequality comes from the convexity inequality $(1-x)^{-1} \geq 1 + x$ for $x \in (0,1)$.

Then, we have seen that if $r_0 \geq LR^2$, then $r_1 \leq \frac{LR^2}{2}$ and thus $r_n^{-1} \geq r_1^{-1} + \frac{n-1}{2LR^2} \geq \frac{n+3}{2LR^2}$. Otherwise, we have $r_n^{-1} \geq r_0^{-1} + \frac{n}{2LR^2} \geq \frac{n+2}{2LR^2}$, which is sufficient to conclude the first part of the proposition.

μ-strongly convex case:

Let us now assume that f is μ-strongly convex, and drop the bounded level sets assumption. The proof again follows [40] for computing the convergence rate of proximal gradient methods. We start from (16). We use the second-order growth property of f (Lemma A.6) which states that $f(\theta_{n-1}) \geq f^\star + \frac{\mu}{2}\|\theta_{n-1} - \theta^\star\|_2^2$ and we obtain

$$f(\theta_n) - f^\star \leq \left(\min_{\alpha \in [0,1]} 1 - \alpha + \frac{L\alpha^2}{\mu} \right) (f(\theta_{n-1}) - f^\star).$$

At this point, it is easy to show that if $\mu \geq 2L$, the previous binomial is minimized for $\alpha^\star = 1$, and if $\mu \leq 2L$, then we have $\alpha^\star = \frac{\mu}{2L}$. This yields the desired result.

□

B.5 Proof of theorem 2.8

Proof.

Non-strongly convex case:

From Lemma 2.3 (with $g = g_n$, $\kappa = \theta_{n-1}$, $\theta' = \theta_n$, $\theta = \theta^\star$), we have for all $n \geq 1$,

$$f(\theta_n) - f(\theta^\star) \leq \frac{L}{2}\|\theta_{n-1} - \theta^\star\|_2^2 - \frac{\rho}{2}\|\theta_n - \theta^\star\|_2^2 \leq \frac{L}{2}\|\theta_{n-1} - \theta^\star\|_2^2 - \frac{L}{2}\|\theta_n - \theta^\star\|_2^2. \tag{17}$$

After summation, we have

$$n(f(\theta_n) - f(\theta^\star)) \leq \sum_{k=1}^{n} (f(\theta_k) - f(\theta^\star)) \leq \frac{L}{2}(\|\theta_0 - \theta^\star\|_2^2 - \|\theta_n - \theta^\star\|_2^2) \leq \frac{L\|\theta_0 - \theta^\star\|_2^2}{2},$$

where the first inequality comes from the fact that $f(\theta_k) \geq f(\theta_n)$ for all $k \leq n$. This is sufficient to prove (2.8). Note that proving convergence rates for first-order methods by finding telescopic sums is a classical technique (see, e.g.,[3]).

μ-strongly convex case:

Let us now prove the second part of the proposition and assume that f is μ-strongly convex. The strong convexity implies the second-order growth property of Lemma A.6: $f(\theta_n) - f^\star \geq \frac{\mu}{2}\|\theta_n - \theta^\star\|_2^2$ for all n. Combined with (17), this yields

$$\frac{\mu + \rho}{2}\|\theta_n - \theta^\star\|_2^2 \leq \frac{L}{2}\|\theta_{n-1} - \theta^\star\|_2^2,$$

and thus

$$f(\theta_n) - f(\theta^\star) \leq \frac{L}{2}\|\theta_{n-1} - \theta^\star\|_2^2 \leq \left(\frac{L}{\rho + \mu}\right)^{n-1} \frac{L\|\theta_0 - \theta^\star\|_2^2}{2}.$$

□

B.6 Proof of theorem 3.1

Proof. We proceed in several steps.

Almost sure convergence of $(f(\theta_n))_{n \geq 0}$:

Let us define $\bar{g}_n \triangleq \frac{1}{T}\sum_{t=1}^{T} g_n^t$. We have the following relation for all $n \geq 1$,

$$\bar{g}_n = \bar{g}_{n-1} + \frac{g_n^{\hat{t}_n} - g_{n-1}^{\hat{t}_n}}{T}, \tag{18}$$

where the surrogates and the index $\hat{t}_n$ are chosen in the algorithm. Then, we obtain the following inequalities, which hold with probability one for all $n \geq 1$,

$$\begin{aligned}\bar{g}_n(\theta_n) \leq \bar{g}_n(\theta_{n-1}) &= \bar{g}_{n-1}(\theta_{n-1}) + \frac{g_n^{\hat{t}_n}(\theta_{n-1}) - g_{n-1}^{\hat{t}_n}(\theta_{n-1})}{T} \\ &= \bar{g}_{n-1}(\theta_{n-1}) + \frac{f^{\hat{t}_n}(\theta_{n-1}) - g_{n-1}^{\hat{t}_n}(\theta_{n-1})}{T} \leq \bar{g}_{n-1}(\theta_{n-1}).\end{aligned}$$

The first inequality is true by definition of θ_n and the second one because $\bar{g}_{n-1}^{\hat{t}_n}$ is a majorizing surrogate of $f^{\hat{t}_n}$. The sequence $(\bar{g}_n(\theta_n))_{n\geq 0}$ is thus monotonically decreasing, bounded below with probability one and thus converges almost surely. By taking the expectation of these previous inequalities, we also obtain that the sequence $(\mathbb{E}[\bar{g}_n(\theta_n)])_{n\geq 0}$ monotonically converges. Thus, the non-positive quantity $\mathbb{E}[f^{\hat{t}_n}(\theta_{n-1}) - g_{n-1}^{\hat{t}_n}(\theta_{n-1})]$ is the summand of a converging sum and we have

$$\begin{aligned}\mathbb{E}\left[\sum_{n=0}^{+\infty} g_n^{\hat{t}_{n+1}}(\theta_n) - f^{\hat{t}_{n+1}}(\theta_n)\right] &= \sum_{n=0}^{+\infty} \mathbb{E}[g_n^{\hat{t}_{n+1}}(\theta_n) - f^{\hat{t}_{n+1}}(\theta_n)] \\ &= \sum_{n=0}^{+\infty} \mathbb{E}[\mathbb{E}[g_n^{\hat{t}_{n+1}}(\theta_n) - f^{\hat{t}_{n+1}}(\theta_n)|\mathcal{F}_n]] \\ &= \sum_{n=0}^{+\infty} \mathbb{E}[\bar{g}_n(\theta_n) - f(\theta_n)] \\ &= \mathbb{E}\left[\sum_{n=0}^{+\infty} \bar{g}_n(\theta_n) - f(\theta_n)\right] < +\infty,\end{aligned}$$

where we use Beppo-Lévy theorem to interchange the expectation and the sum in front of non-negative quantities, and $\mathcal{F}_n$ is the filtration representing all information up to iteration n (including θ_n). As a result, the sequence $(\bar{g}_n(\theta_n) - f(\theta_n))_{n\geq 0}$ converges almost surely to 0, implying the almost sure convergence of $(f(\theta_n))_{n\geq 0}$.

Asymptotic stationary point conditions:

Let us define $\bar{h}_n \triangleq \bar{g}_n - f$, which is L-smooth. Then, for all θ in Θ and $n \geq 1$,

$$\nabla f(\theta_n, \theta - \theta_n) = \nabla \bar{g}_n(\theta_n, \theta - \theta_n) - \nabla \bar{h}_n(\theta_n)^\top (\theta - \theta_n).$$

We have $\nabla \bar{g}_n(\theta_n, \theta - \theta_n) \geq 0$ by definition of θ_n, and $\|\nabla \bar{h}_n(\theta_n)\|_2^2 \leq 2L\bar{h}_n(\theta_n)$, following similar steps as in the proof of Proposition 2.5. Since we have shown that $(\bar{h}_n(\theta_n))_{n\geq 0}$ almost surely converges to zero, we conclude as in the proof of Proposition 2.5, replacing h_n by $\bar{h}_n$ and g_n by $\bar{g}_n$. □

B.7 Proof of theorem 3.2

Proof. We first remark that the first part of the proof of Proposition 3.1 does not exploit the fact that the approximation errors $g_n^t - f^t$ are L-smooth, but only the fact that g_n^t is majorizing f^t for all n and t. Thus, the first part of the proof of Proposition 3.1 is valid, $(f(\theta_n))_{n\geq 0}$ almost surely converges, and the sequence $(\bar{g}_n(\theta_n) - f(\theta_n))_{n\geq 0}$ almost surely converges to zero, where $\bar{g}_n$ is defined in the proof of Proposition 3.1.

It remains to show that the asymptotic stationary point conditions are satisfied. To that effect, we follow the proof of Proposition 2.6. We first have, for all $n \geq 1$,

$$\nabla f(\theta_n, \theta - \theta_n) = \nabla \bar{g}_n(\theta_n, \theta - \theta_n) - \frac{1}{T}\sum_{t=1}^{T} \nabla \bar{h}_n^t(\theta_n, \theta - \theta_n),$$

with $\nabla \bar{g}_n(\theta_n, \theta - \theta_n) \geq 0$ and $\bar{h}_n^t \triangleq \bar{g}_n^t - f^t$. Then, following the proof of Proposition 2.6, it is easy to show that

$$|\nabla \bar{h}_n^t(\theta_n, \theta - \theta_n)| \leq C\|\nabla \bar{h}_n'^t(e^t(\theta_n))\|_2 \|\theta - \theta_n\|_2,$$

where $\bar{h}_n'^t = \bar{g}_n'^t - f'^t$, and we conclude as in Proposition 2.6. □

B.8 Proof of theorem 3.3

Proof. We proceed in several steps.

Preliminaries:

Let us denote by κ_{n-1}^t the point in Θ such that g_n^t is in $\mathcal{S}_{L,\rho}(f^t, \kappa_{n-1}^t)$ for all $n \geq 1$. We remark that such points are drawn recursively according to the following conditional probability distribution:

$$\mathbb{P}(\kappa_{n-1}^t = \theta_{n-1} | \mathcal{F}_{n-1}) = \delta \quad \text{and} \quad \mathbb{P}(\kappa_{n-1}^t = \kappa_{n-2}^t | \mathcal{F}_{n-1}) = 1 - \delta,$$

where $\delta \triangleq 1/T$, $\mathcal{F}_n$ is the filtration representing all information up to iteration n (including θ_n), and $\kappa_0^t \triangleq \theta_0$ for all t. Thus we have for all t and all $n \geq 1$,

$$\mathbb{E}[\|\theta^\star - \kappa_{n-1}^t\|_2^2] = \mathbb{E}[\mathbb{E}[\|\theta^\star - \kappa_{n-1}^t\|_2^2 | \mathcal{F}_{n-1}]] = \delta \mathbb{E}[\|\theta^\star - \theta_{n-1}\|_2^2] + (1-\delta)\mathbb{E}[\|\theta^\star - \kappa_{n-2}^t\|_2^2]. \tag{19}$$

The other relation we need is an extension of Lemma 2.3 to the incremental setting. For all θ in Θ and $n \geq 1$, we have

$$f(\theta_n) \leq f(\theta) + \frac{1}{T} \sum_{t=1}^{T} \left(\frac{L}{2} \|\theta - \kappa_{n-1}^t\|_2^2 - \frac{\rho}{2} \|\theta - \theta_n\|_2^2 \right). \tag{20}$$

The proof of this relation is similar to that of Lemma 2.3, exploiting the ρ-strong convexity of $\bar{g}_n$. We can now study the first part of the proposition.

Non-strongly convex case ($\rho = L$); convergence rate:

Let us define the quantities $A_n \triangleq \mathbb{E}[\frac{1}{2T} \sum_{t=1}^{T} \|\theta^\star - \kappa_n^t\|_2^2]$ and $\xi_n \triangleq \frac{1}{2}\mathbb{E}[\|\theta^\star - \theta_n\|_2^2]$. Then, we have from (20) with $\theta = \theta^\star$, and by taking the expectation

$$\mathbb{E}[f(\theta_n) - f^\star] \leq L A_{n-1} - L \xi_n.$$

It follows from (19) that $A_n = \delta \xi_n + (1-\delta) A_{n-1}$ and thus, for all $n \geq 1$,

$$\mathbb{E}[f(\theta_n) - f^\star] \leq \frac{L}{\delta}(A_{n-1} - A_n).$$

By summing the above inequalities, and using Jensen's inequality, we obtain that

$$\mathbb{E}[f(\bar{\theta}_n) - f^\star] \leq \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[f(\theta_i) - f^\star] \leq \frac{L A_0}{\delta},$$

leading to the convergence rate of Eq. (5), since $A_0 = \frac{1}{2}\|\theta^\star - \theta_0\|_2^2$.

μ-strongly convex case:

Suppose now that the functions f^t are μ-strongly convex. We have from (20) and the second-order growth condition of Lemma A.6 that for all $n \geq 1$,

$$\mu \xi_n \leq \mathbb{E}[f(\theta_n) - f^\star] \leq L A_{n-1} - \rho \xi_n,$$

Combining this last inequality, Eq. (19), we obtain that for all $n \geq 1$,

$$A_n = \delta \xi_n + (1-\delta) A_{n-1} \leq \left(\frac{\delta L}{\mu + \rho} + (1-\delta) \right) A_{n-1}.$$

Thus, $A_n \leq \beta^n A_0$ with $\beta \triangleq \frac{(1-\delta)(\rho+\mu)+\delta L}{\rho+\mu}$. Since $A_0 = \xi_0$ and $\mathbb{E}[f(\theta_n) - f^\star] \leq L A_{n-1}$, we finally have shown the desired convergence rate (6). □

B.9 Proof of theorem 3.4

Proof. As in the proof of Proposition 3.1, we introduce the function $\bar{g}_n \triangleq \frac{1}{T}\sum_{t=1}^{T} g_n^t$, which is minimized by θ_n for $n \geq 1$. Since $\bar{g}_n$ is a lower bound on f, we have the relation $\bar{g}_n(\theta_n) \leq \bar{g}_n(\theta^\star) \leq f^\star$. Inspired by the convergence proof of SDCA [45], which computes an convergence rate of an expected duality gap, we proceed by studying the convergence of the sequence $(f^\star - \mathbb{E}[\bar{g}_n(\theta_n)])_{n\geq 1}$.

On the one hand, we have for all $n \geq 1$,

$$\begin{aligned}\bar{g}_n(\theta_n) &= \bar{g}_n(\theta_{n-1}) - \frac{\mu}{2}\|\theta_n - \theta_{n-1}\|_2^2 \\ &= \bar{g}_{n-1}(\theta_{n-1}) + \delta(g_n^{\hat{t}_n}(\theta_{n-1}) - g_{n-1}^{\hat{t}_n}(\theta_{n-1})) - \frac{\mu}{2}\|\theta_n - \theta_{n-1}\|_2^2,\end{aligned}$$

where the first equality comes from the fact that $\bar{g}_n$ is quadratic and minimized by θ_n, and the second one simply uses Equation (18). By taking the expectation, we have $\mathbb{E}[g_n^{\hat{t}_n}(\theta_{n-1})] = \mathbb{E}[f^{\hat{t}_n}(\theta_{n-1})] = \mathbb{E}[\mathbb{E}[f^{\hat{t}_n}(\theta_{n-1})|\mathcal{F}_{n-1}]] = \mathbb{E}[f(\theta_{n-1})]$, where $\mathcal{F}_n$ is the the filtration representing all information up to iteration n, and $\mathbb{E}[g_{n-1}^{\hat{t}_n}(\theta_{n-1})] = \mathbb{E}[\mathbb{E}[g_{n-1}^{\hat{t}_n}(\theta_{n-1})|\mathcal{F}_{n-1}]] = \mathbb{E}[\bar{g}_{n-1}(\theta_{n-1})]$. Thus,

$$\mathbb{E}[\bar{g}_n(\theta_n)] = (1-\delta)\mathbb{E}[\bar{g}_{n-1}(\theta_{n-1})] + \delta\mathbb{E}[f(\theta_{n-1})] - \frac{\mu}{2}\mathbb{E}\left[\|\theta_n - \theta_{n-1}\|_2^2\right]. \tag{21}$$

On the other hand, for all $n \geq 2$,

$$\begin{aligned}\bar{g}_n(\theta_n) &= \bar{g}_{n-1}(\theta_n) + \delta(g_n^{\hat{t}_n}(\theta_n) - g_{n-1}^{\hat{t}_n}(\theta_n)) \\ &= \bar{g}_{n-1}(\theta_{n-1}) + \frac{\mu - \delta L}{2}\|\theta_n - \theta_{n-1}\|_2^2 + \delta\left(g_n^{\hat{t}_n}(\theta_n) + \frac{L}{2}\|\theta_n - \theta_{n-1}\|_2^2 - g_{n-1}^{\hat{t}_n}(\theta_n)\right) \\ &\geq \bar{g}_{n-1}(\theta_{n-1}) + \frac{\mu - \delta L}{2}\|\theta_n - \theta_{n-1}\|_2^2.\end{aligned}$$

We have used the fact that $\theta \mapsto g_n^{\hat{t}_n}(\theta) + (L/2)\|\theta - \theta_{n-1}\|_2^2$ is a majorizing surrogate of $f^{\hat{t}_n}$, whereas $g_{n-1}^{\hat{t}_n}$ is minorizing $f^{\hat{t}_n}$. By taking the expectation, using the fact that $\mu - \delta L \geq \mu/2$, and combining with (21), we obtain that for all $n \geq 2$,

$$3\mathbb{E}[\bar{g}_n(\theta_n)] \geq (3-\delta)\mathbb{E}[\bar{g}_{n-1}(\theta_{n-1})] + \delta\mathbb{E}[f(\theta_{n-1})]. \tag{22}$$

Since $\mathbb{E}[f(\theta_{n-1})] \geq f^\star$, this immediately gives for $n \geq 2$,

$$f^\star - \mathbb{E}\left[\bar{g}_n(\theta_n)\right] \leq \left(1 - \frac{1}{3T}\right)\left(f^\star - \mathbb{E}\left[\bar{g}_{n-1}(\theta_{n-1})\right]\right).$$

To obtain a convergence rate for $\mathbb{E}[f(\theta_n)] - f^\star$, we use again Eq. (22). For $n \geq 2$,

$$\begin{aligned}\delta(\mathbb{E}[f(\theta_{n-1})] - f^\star) &\leq \delta(\mathbb{E}[f(\theta_{n-1})] - \mathbb{E}[\bar{g}_{n-1}(\theta_{n-1})]) \\ &\leq 3(\mathbb{E}[\bar{g}_n(\theta_n)] - \mathbb{E}[\bar{g}_{n-1}(\theta_{n-1})]) \\ &\leq 3(f^\star - \mathbb{E}[\bar{g}_{n-1}(\theta_{n-1})]) \\ &\leq 3\left(1 - \frac{1}{3T}\right)^{n-2}(f^\star - \bar{g}_1(\theta_1)),\end{aligned} \tag{23}$$

and we obtain the convergence rate (10) by first noticing that

$$\begin{aligned}f^\star - \bar{g}_1(\theta_1) &= f^\star - \bar{g}_1(\theta_0) + \frac{\mu}{2}\|\theta_0 - \theta_1\|_2^2 \\ &= f^\star - f(\theta_0) + \frac{\mu}{2}\left\|\frac{1}{\mu}\nabla f(\theta_0)\right\|_2^2 \\ &\leq \frac{1}{2\mu}\|\nabla f(\theta_0)\|_2^2,\end{aligned}$$

where we use the fact that $\bar{g}_1 = \bar{g}_0$ and $\bar{g}_0(\theta_0) = f(\theta_0)$. Then, we use the fact that $(1 - 1/3T) \geq 5/6$ since $T \geq 2L/\mu \geq 2$, such that $3(1 - 1/3T)^{-1}/(2\mu) \leq 9/(5\mu) \leq 2/\mu$.

To prove the last part of the proposition, we remark that all inequalities we have proved so far for $n \geq 2$, become true for $n = 1$. Thus, the last inequality in (23) is also true when replacing $n-2$ by $n-1$ and $\bar{g}_1(\theta_1)$ by $\bar{g}_0(\theta_0) = 0$. □

References

[1] S. Ahn, J. A. Fessler, D. Blatt, and A. O. Hero, *Convergent incremental optimization transfer algorithms: Application to tomography*, IEEE T. Med. Imaging, 25 (2006), pp. 283–296.

[2] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, *Optimization with sparsity-inducing penalties*, Found. Trends Mach. Learn., 4 (2012), pp. 1–106.

[3] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sci., 2 (2009), pp. 183–202.

[4] D. Bertsekas, *Nonlinear programming*, Athena Scientific Belmont, 1999. 2nd edition.

[5] D. Blatt, A. O. Hero, and H. Gauchman, *A convergent incremental gradient method with a constant step size*, SIAM J. Optimiz., 18 (2007), pp. 29–51.

[6] D. Böhning and B. G. Lindsay, *Monotonicity of quadratic-approximation algorithms*, Ann. I. Stat. Math., 40 (1988), pp. 641–663.

[7] J. M. Borwein and A. S. Lewis, *Convex analysis and nonlinear optimization: Theory and examples*, Springer, 2006.

[8] L. Bottou, *Online algorithms and stochastic approximations*, in Online Learning and Neural Networks, D. Saad, ed., 1998.

[9] S. P. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[10] E. J. Candès, M. Wakin, and S. P. Boyd, *Enhancing sparsity by reweighted ℓ_1 minimization*, J. Fourier Anal. Appl., 14 (2008), pp. 877–905.

[11] A. Choromanska and T. Jebara, *Stochastic bound majorization*, arXiv:1309.5605, (2013).

[12] M. Collins, R. Schapire, and Y. Singer, *Logistic regression, AdaBoost and Bregman distances*, Mach. Learn., 48 (2002), pp. 253–285.

[13] P. L. Combettes and J.-C. Pesquet, *Proximal splitting methods in signal processing*, in Fixed-Point Algorithms for Inverse Problems in Science and Engineering, Springer, 2010.

[14] I. Daubechies, M. Defrise, and C. De Mol, *An iterative thresholding algorithm for linear inverse problems with a sparsity constraint*, Commun. Pur. Appl. Math., 57 (2004), pp. 1413–1457.

[15] S. Della Pietra, V. Della Pietra, and J. Lafferty, *Duality and auxiliary functions for Bregman distances*, tech. rep., CMU-CS-01-109, 2001.

[16] A. P. Dempster, N. M. Laird, and D. B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, J. Roy. Stat. Soc. B, 39 (1977), pp. 1–38.

[17] J. Duchi, E. Hazan, and Y. Singer, *Adaptive subgradient methods for online learning and stochastic optimization*, J. Mach. Learn. Res., 12 (2011), pp. 2121–2159.

[18] J. Duchi and Y. Singer, *Efficient online and batch learning using forward backward splitting*, J. Mach. Learn. Res., 10 (2009), pp. 2899–2934.

[19] H. Erdogan and J. A. Fessler, *Ordered subsets algorithms for transmission tomography*, Phys. Med. Biol., 44 (1999), pp. 2835–2851.

[20] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, *LIBLINEAR: A library for large linear classification*, J. Mach. Learn. Res., 9 (2008), pp. 1871–1874.

[21] M. Fashing and C. Tomasi, *Mean shift is a bound optimization*, IEEE T. Pattern Anal., 27 (2005), pp. 471–474.

[22] G. Gasso, A. Rakotomamonjy, and S. Canu, *Recovering sparse signals with non-convex penalties and DC programming*, IEEE T. Signal Process., 57 (2009), pp. 4686–4698.

[23] S. Ghadimi and G. Lan, *Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework*, SIAM J. Optimiz., 22 (2012), pp. 1469–1492.

[24] E. T. Hale, W. Yin, and Y. Zhang, *Fixed-point continuation for ℓ_1-minimization: Methodology and convergence*, SIAM J. Optimiz., 19 (2008), pp. 1107–1130.

[25] E. Hazan and S. Kale, *Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization*, in Proc. COLT, 2011.

[26] R. Horst and N. V. Thoai, *DC programming: overview*, J. Optim. Theory App., 103 (1999), pp. 1–43.

[27] T. Jebara and A. Choromanska, *Majorization for CRFs and latent likelihoods*, in Adv. Neur. In. (NIPS), 2012.

[28] A. Juditsky and A. Nemirovski, *First order methods for nonsmooth convex large-scale optimization*, in Optimization for Machine Learning, MIT Press, 2011.

[29] E. Khan, B. Marlin, G. Bouchard, and K. Murphy, *Variational bounds for mixed-data factor analysis*, in Adv. Neur. In. (NIPS), 2010.

[30] G. Lan, *An optimal method for stochastic composite optimization*, Math. Program., 133 (2012), pp. 365–397.

[31] K. Lange, D. R. Hunter, and I. Yang, *Optimization transfer using surrogate objective functions*, J. Comput. Graph. Stat., 9 (2000), pp. 1–20.

[32] N. Le Roux, M. Schmidt, and F. Bach, *A stochastic gradient method with an exponential convergence rate for finite training sets*, in Adv. Neur. In. (NIPS), 2012.

[33] D. D. Lee and H. S. Seung, *Algorithms for non-negative matrix factorization*, in Adv. Neur. In. (NIPS), 2001.

[34] J. Mairal, *Optimization with first-order surrogate functions*, in Proc. ICML, 2013.

[35] J. Mairal, *Stochastic majorization-minimization algorithms for large-scale optimization*, in Adv. Neur. In. (NIPS), 2013.

[36] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, *Online learning for matrix factorization and sparse coding*, J. Mach. Learn. Res., 11 (2010), pp. 19–60.

[37] R. M. Neal and G. E. Hinton, *A view of the EM algorithm that justifies incremental, sparse, and other variants*, Learning in graphical models, 89 (1998).

[38] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, *Robust stochastic approximation approach to stochastic programming*, SIAM J. Optimiz., 19 (2009), pp. 1574–1609.

[39] Y. Nesterov, *Introductory lectures on convex optimization*, Kluwer Academic Publishers, 2004.

[40] ——, *Gradient methods for minimizing composite objective functions*, Math. Program., 140 (2012), pp. 125–161.

[41] J. Nocedal and S. J. Wright, *Numerical optimization*, Springer Verlag, 2006. 2nd edition.

[42] M. Razaviyayn, M. Hong, and Z.-Q. Luo, *A unified convergence analysis of block successive minimization methods for nonsmooth optimization*, SIAM J. Optimiz., 23 (2013), pp. 1126–1153.

[43] M. Razaviyayn, M. Sanjabi, and Z.-Q. Luo, *A stochastic successive minimization method for nonsmooth nonconvex optimization*, arXiv:1307.4457v2, (2013).

[44] M. Schmidt, N. L. Roux, and F. Bach, *Minimizing finite sums with the stochastic average gradient*, arXiv:1309.2388, (2013).

[45] S. Shalev-Schwartz and T. Zhang, *Proximal stochastic dual coordinate ascent*, arXiv:1211.2717, (2012).

[46] B. A. Turlach, W. N. Venables, and S. J. Wright, *Simultaneous variable selection*, Technometrics, 47 (2005), pp. 349–363.

[47] M. J. Wainwright and M. I. Jordan, *Graphical models, exponential families, and variational inference*, Found. Trends Mach. Learn., 1 (2008), pp. 1–305.

[48] S. J. Wright, R. D. Nowak, and M. A. Figueiredo, *Sparse reconstruction by separable approximation*, IEEE T. Signal Process., 57 (2009), pp. 2479–2493.

[49] L. Xiao, *Dual averaging methods for regularized stochastic learning and online optimization*, J. Mach. Learn. Res., 11 (2010), pp. 2543–2596.

[50] M. Yuan and Y. Lin, *Model selection and estimation in regression with grouped variables.*, J. Roy. Stat. Soc. B, 68 (2006), pp. 49–67.

[51] X. Zhang, M. Burger, and S. Osher, *A unified primal-dual algorithm framework based on bregman iteration*, Journal of Scientific Computing, 46 (2011), pp. 20–46.

[52] L. W. Zhong and J. T. Kwok, *Fast stochastic alternating direction method of multipliers*, arXiv:1308.3558, (2013).