Stochastic Majorization-Minimization Optimization with First-Order Surrogate Functions

Julien Mairal

Stochastic Majorization-Minimization Optimization with First-Order Surrogate Functions

Julien Mairal
LEAR Project-Team - INRIA
julien.mairal@inria.fr

June 19, 2013

Abstract

Majorization-minimization algorithms consist of iteratively minimizing a majorizing surrogate of an objective function. Because of its simplicity and its wide applicability, this principle has been very popular in statistics and in signal processing. In this paper, we intend to make this principle scalable. We introduce and study a stochastic majorization-minimization scheme, which is able to deal with large-scale or possibly infinite data sets. When applied to convex optimization problems under suitable assumptions, we show that it achieves an expected convergence rate of $O(1/\sqrt{n})$ after n iterations, and $O(1/n)$ for strongly convex functions. Equally important, our scheme almost surely converges to stationary points for a large class of non-convex problems. We derive from our framework several efficient algorithms. First, we propose a new stochastic proximal gradient method, which experimentally matches state-of-the-art solvers for large-scale ℓ_1-logistic regression. Second, we develop an online DC programming algorithm for non-convex sparse estimation. Finally, we demonstrate the effectiveness of our technique for solving large-scale structured matrix factorization problems.

1 Introduction

Majorization-minimization [19] is a simple optimization principle for minimizing an objective function. It consists of iteratively minimizing a surrogate that upper-bounds the objective, thus monotonically driving the objective function value downhill. This methodology is quite general, and encompasses many existing procedures. For instance, the expectation-maximization (EM) algorithm (see [8, 25]) builds a surrogate for a likelihood model by using Jensen's inequality. Other approaches can also be interpreted under the majorization-minimization point of view, such as DC programming, often used in the sparse optimization literature [7, 13], variational Bayes techniques [35], gradient-based or proximal algorithms [1, 28, 36], and some matrix factorization methods [22].

In this paper, we propose a stochastic majorization-minimization algorithm, which is suitable for solving large-scale problems arising in machine learning and signal processing. More precisely, we address the minimization of an expected cost—that is, an objective function that can be represented as an expectation over a data distribution. In such a situation, online techniques based on stochastic approximations have proven to be particularly efficient, and have drawn a lot of attraction in machine learning, statistics, and optimization [4, 5, 9, 10, 14–16, 18, 20, 21, 26, 31–33, 37].

Our scheme follows this line of research. It consists of approximating a surrogate of the expected cost when only a single data point is observed at each iteration; this data point is used to update the surrogate, which is in turn minimized to obtain a new estimate. Some previous works are closely related to this scheme: the online EM algorithm for latent data models [8, 25] and the online matrix factorization technique of [23] involve for instance surrogate functions updated in a similar fashion. Compared to these two approaches, our method is targeted to more general optimization problems.

Another related work is the incremental majorization-minimization algorithm of [22] for finite training sets; it was indeed shown to be efficient for solving machine learning problems where storing dense information about the past iterates can be afforded. Concretely, this incremental scheme requires to store at least $O(pn)$ values, where p is the variable size, and n is the size of the training set.[1] This issue was the main motivation for us for proposing a stochastic scheme, with a memory load independent of n, thus allowing us to possibly deal with infinite data sets, or problems where p is huge. Such an extension is far from simple, and requires a more involved analysis than [22].

We study the convergence properties of our algorithm by focusing on strongly convex *first-order surrogate functions* introduced in [22], which consist of approximating a possibly non-smooth objective up to a smooth error. By making such assumptions on the surrogates, we present two analyses. For convex optimization problems, we combine our approach with an averaging scheme [14, 26], and obtain classical convergence rates which are optimal up to a multiplicative constant for the class of functions we consider [26]. More precisely, the convergence rate is of order $O(1/\sqrt{n})$ for convex functions in a finite horizon setting, and $O(1/n)$ for strongly convex functions in an infinite horizon setting. Our second analysis shows that for non-convex problems, our method asymptotically provides a stationary point with probability one under suitable assumptions. We believe that this result is equally valuable as convergence rates for convex optimization. To the best of our knowledge, the literature on stochastic non-convex optimization is rather scarce, and we are not aware of any method that exhibits such a guarantee in a setting as general as ours—see for instance [4] for the stochastic gradient descent algorithm, [8] for the online EM, or [23] for online matrix factorization.

From our framework, we derive several practical algorithms. The first one is a new stochastic proximal gradient method for composite or constrained optimization, which is related to a long series of work in the convex optimization literature [9, 10, 14, 16, 18, 20, 26, 32, 37]. We demonstrate that this algorithm matches state-of-the-art solvers for large-scale ℓ_1-logistic regression [11]. The second one is an online DC programming technique ("DC" stands for difference of convex functions), which we prove to be better than batch alternatives for large-scale non-convex sparse estimation [13]. Finally, we apply our scheme to online matrix factorization in a more general setting than [23], with various loss and regularization functions, opening up possibilities that were not allowed by [23].

This paper is organized as follows. Section 2 introduces first-order surrogate functions for batch optimization. Section 3 is devoted to our stochastic approach and its convergence analysis. Section 4 presents several applications and numerical experiments, and Section 5 concludes the paper.

2 Optimization with First-Order Surrogate Functions

Throughout the paper, we are interested in the minimization of a continuous function $f : \mathbb{R}^p \to \mathbb{R}$:

$$\min_{\theta \in \Theta} f(\theta), \tag{1}$$

where $\Theta \subseteq \mathbb{R}^p$ is a convex set. The majorization-minimization principle consists of computing a surrogate g_n of f at iteration n and updating the current estimate by $\theta_n \in \arg\min_{\theta\in\Theta} g_n(\theta)$. The success of such a scheme depends on how well the surrogates approximate f. In this paper, we consider a particular class of surrogate functions introduced by [22] and defined as follows:

Definition 2.1 (Strongly Convex First-Order Surrogate Functions).
Let κ be in Θ. We denote by $\mathcal{S}_{L,\rho}(f, \kappa)$ the set of ρ-strongly convex functions g such that $g \geq f$, $g(\kappa) = f(\kappa)$, the approximation error $h \triangleq g - f$ is differentiable, and the gradient ∇h is L-Lipschitz continuous. We call the functions g in $\mathcal{S}_{L,\rho}(f, \kappa)$ "first-order surrogate functions".

Among the first-order surrogate functions presented in [22], we should mention the following ones:

- **Lipschitz Gradient Surrogates.**

[1] To alleviate this issue, the author of [22] proposes to cut the dataset into η mini-batches, reducing the memory load to $O(p\eta)$, which remains cumbersome when p is very large.

When f is differentiable and ∇f is L-Lipschitz, f admits the following surrogate g in $\mathcal{S}_{2L,L}(f,\kappa)$:

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

When f is convex, g is in $\mathcal{S}_{L,L}(f,\kappa)$, and when f is μ-strongly convex, g is in $\mathcal{S}_{L-\mu,L}(f,\kappa)$. Minimizing g amounts to performing a classical classical gradient descent step $\theta' \leftarrow \kappa - \frac{1}{L}\nabla f(\kappa)$.

- **Proximal Gradient Surrogates.**

Assume that f splits into $f = f_1 + f_2$, where f_1 is differentiable, ∇f_1 is L-Lipschitz, and f_2 is convex. Then, the function g below is in $\mathcal{S}_{2L,L}(f,\kappa)$:

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

As before, when f_1 is convex, g is in $\mathcal{S}_{L,L}(f,\kappa)$. If f_1 is μ-strongly convex, g is in $\mathcal{S}_{L-\mu,L}(f,\kappa)$. Minimizing g amounts to performing a proximal gradient step [1,28,36].

- **DC Programming Surrogates.**

Assume that $f = f_1 + f_2$, where f_2 is concave and differentiable, ∇f_2 is L_2-Lipschitz, and g_1 is in $\mathcal{S}_{L_1,\rho_1}(f_1,\kappa)$, Then, the following function g is a surrogate in $\mathcal{S}_{L_1+L_2,\rho_1}(f,\kappa)$:

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

When f_1 is convex, $f_1 + f_2$ is a difference of convex functions, leading to a DC program [13].

With the definition of first-order surrogates and a basic "batch" algorithm in hand, we now introduce our main contribution—that is, a fully stochastic scheme for solving large-scale problems.

3 Stochastic Optimization

As pointed out in [5], one is usually not interested in the minimization of an *empirical cost* on a finite training set, but instead in minimizing an *expected cost*. Thus, we assume from now on that f has the form of an expectation:

$$\min_{\theta \in \Theta} \left[f(\theta) \triangleq \mathbb{E}_{\mathbf{x}}[\ell(\mathbf{x},\theta)] \right], \tag{2}$$

where $\mathbf{x}$ from some set $\mathcal{X}$ represents a data point, which is drawn according to some unknown distribution, and ℓ is a continuous loss function. As often done in the literature [26], we assume that the expectations are well defined and finite valued; we also assume that f is bounded below.

We present our approach for tackling (2) in Algorithm 1. At each iteration, we draw a training point $\mathbf{x}_n$, assuming that these points are i.i.d. samples from the data distribution. Note that in practice, since it is often difficult to obtain true i.i.d. samples, the points $\mathbf{x}_n$ are computed by cycling on a randomly permuted training set [5]. Then, we compute a first-order surrogate g_n for the function $\theta \mapsto \ell(\mathbf{x}_n,\theta)$, and we use it to update an approximate surrogate $\bar{g}_n$; the update involves a sequence of weights $(w_n)_{n\geq 1}$ that will be discussed later. We minimize $\bar{g}_n$, and obtain a new estimate θ_n. We also use an averaging scheme, which is a classical technique for improving theoretical convergence rates in convex optimization [14,26], for reasons that are clear in the convergence proofs.

We remark that Algorithm 1 is only practical when the functions $\bar{g}_n$ can be parameterized with a small number of parameters, and when they can be easily minimized over Θ. Practical examples are discussed in Section 4. Before that, we proceed with the convergence analysis.

3.1 Convergence Analysis - Convex Case

We start by studying the case of convex functions f_n, and we make the following assumptions:

(A) for all θ in Θ, the functions f_n are R-Lipschitz. Note that for convex functions, this is equivalent to saying that subgradients are uniformly bounded by R;

Algorithm 1 Stochastic Majorization-Minimization Scheme

Inputs: $\theta_0 \in \Theta$ (initial estimate); N (number of iterations); $(w_n)_{n\geq 1}$, weights in $(0,1)$;
initialize the approximate surrogate: $\bar{g}_0 : \theta \mapsto \frac{\rho}{2}\|\theta - \theta_0\|_2^2$; $\bar{\theta}_0 = \theta_0$; $\hat{\theta}_0 = \theta_0$;
for $n = 1, \ldots, N$ **do**
    draw a training point $\mathbf{x}_n$; define $f_n : \theta \mapsto \ell(\mathbf{x}_n, \theta)$;
    choose a surrogate function g_n in $\mathcal{S}_{L,\rho}(f_n, \theta_{n-1})$;
    update the approximate surrogate: $\bar{g}_n = (1 - w_n)\bar{g}_{n-1} + w_n g_n$;
    update the current estimate:

$$\theta_n \in \operatorname*{arg\,min}_{\theta \in \Theta} \bar{g}_n(\theta).$$

    for option 2, update the averaged iterate: $\hat{\theta}_n \triangleq (1 - w_{n+1})\hat{\theta}_{n-1} + w_{n+1}\theta_n$;
    for option 3, update the averaged iterate: $\bar{\theta}_n \triangleq \frac{(1-w_{n+1})\bar{\theta}_{n-1} + w_{n+1}\theta_n}{\sum_{k=1}^{n+1} w_k}$;
end for
Outputs: (option 1): θ_N (current estimate);
Outputs: (option 2): $\bar{\theta}_N$ (first averaging scheme);
Outputs: (option 3): $\hat{\theta}_N$ (second averaging scheme);

(B) the functions g_n are surrogates in $\mathcal{S}_{L,\rho}(f_n, \theta_{n-1})$ for all $n \geq 1$.

Assumption **(A)** is classical in the stochastic optimization literature, even though slightly weaker assumptions are sometimes made [26]. Assumption **(B)** was explicitly made in the algorithm. Our first result shows that with an appropriate averaging scheme [14, 26], we obtain an expected convergence rate making explicit the role of the weight sequence $(w_n)_{n\geq 1}$.

Proposition 3.1 (**Convergence Rate).**
*When the functions f_n are convex, under assumptions **(A)** and **(B)**, and when $\rho = L$, we have*

$$\mathbb{E}[f(\bar{\theta}_{n-1}) - f^\star] \leq \frac{L\|\theta^\star - \theta_0\|_2^2 + \frac{R^2}{L}\sum_{k=1}^n w_k^2}{2\sum_{k=1}^n w_k} \qquad \text{for all } n \geq 1, \tag{3}$$

where $\bar{\theta}_{n-1}$ is defined in Algorithm 1, $\theta^\star$ is a minimizer of f on Θ, and $f^\star \triangleq f(\theta^\star)$.

For space limitation details, all proofs and details of the analysis are provided in the supplemental material. Such a rate is similar to the one of stochastic gradient descent with averaging, see [26] for example. Note that the constraint $\rho = L$ here is compatible with the proximal gradient surrogate of the previous section. From Proposition 3.1, we immediately obtain the following $O(1/\sqrt{n})$ bound for finite horizon (meaning the total number of iterations is known in advance) and constant weights.

Corollary 3.1 (**Convergence Rate - Finite Horizon - Constant Weights).**
Let us make the same assumptions as in Proposition 3.1 and use the same notation. For a fixed $n \geq 1$, we define $w_k \triangleq \min\left(\frac{L\|\theta^\star - \theta_0\|_2}{R\sqrt{n}}, 1\right)$ for all $k \leq n$, and we have,

$$\mathbb{E}[f(\bar{\theta}_{n-1}) - f^\star] \leq \max\left(\frac{R\|\theta^\star - \theta_0\|_2}{\sqrt{n}}, \frac{L\|\theta^\star - \theta_0\|_2^2}{n}\right).$$

The convergence rate above exhibits an interesting, but well known fact: the upper bound $O(1/\sqrt{n})$ cannot be improved in general without making further assumptions on the objective function [26]. Even though one could choose surrogates with a very small approximation error—that is, with a small Lipschitz constant L, the upper bound remains of the order $O(1/\sqrt{n})$. The next corollary shows that in an infinite horizon setting and with decreasing weights, we lose a logarithmic factor.

Corollary 3.2 (Convergence Rate - Infinite Horizon - Decreasing Weights).
Let us make the same assumptions as in Proposition 3.1 and use the same notation. For all $n \geq 2$,

$$\mathbb{E}[f(\bar{\theta}_{n-1}) - f^\star] \leq \max\left(R\|\theta^\star - \theta_0\|_2, L\|\theta^\star - \theta_0\|_2^2\right) \frac{1 + \log\sqrt{n}}{\sqrt{n}},$$

for $w_n \triangleq \min\left(\frac{L\|\theta^\star - \theta_0\|_2}{R\sqrt{n}}, \frac{1}{\sqrt{n}}\right)$. *When instead* $w_n \triangleq \frac{\gamma}{\sqrt{n}}$ *with* $\gamma \leq 1$, *the rate is bounded by* $O(\frac{\log n}{\sqrt{n}})$.

3.2 Convergence Analysis - Strongly Convex Case

In this section, we introduce an additional assumption:

(C) the functions f_n are μ-strongly convex.

Note that this assumption, when combined with **(B)**, implies that Θ is necessarily compact, according to lemma A.5 in the appendix. Our main result is that our method achieves a rate $O(1/n)$, which is optimal up to a multiplicative constant for strongly convex functions (see [26]).

Proposition 3.2 (Convergence Rate).
Under assumptions ***(A)***, ***(B)***, *and* ***(C)***, *with* $\rho = L + \mu$. *Define* $\beta \triangleq \frac{\mu}{\rho}$ *and* $w_n \triangleq \frac{1+\beta}{1+\beta n}$. *Then,*

$$\mathbb{E}[f(\hat{\theta}_{n-1}) - f^\star] + \frac{\rho}{2}\mathbb{E}[\|\theta^\star - \theta_n\|_2^2] \leq \max\left(\frac{2R^2}{\mu}, \rho\|\theta^\star - \theta_0\|_2^2\right)\frac{1}{\beta n + 1} \quad \textit{for all } n \geq 1,$$

where $\hat{\theta}_n$ *is recursively defined in Algorithm 1.*

The averaging scheme is slightly different than in the previous section and the weights decrease at a different speed. Again, this rate applies to the proximal gradient surrogates, which satisfy the constraint $\rho = L + \mu$. In the next section, we analyze our scheme in a non-convex setting.

3.3 Convergence Analysis - Non-Convex Case

Convergence results for non-convex problems are by nature weak, and difficult to obtain for stochastic optimization [5]. In such a context, proving convergence to a global (or local) minimum is out of reach, and classical analyses study instead asymptotic stationary point conditions, which involve directional derivatives (see [3, 22]). Concretely, we introduce the following assumptions:

(D) Θ and the support $\mathcal{X}$ of the data are compact;

(E) The functions f_n are uniformly bounded by some constant M;

(F) The weights w_n are such that $w_1 = 1$, $\sum_{n\geq 1} w_n = +\infty$, and $\sum_{n \geq 1} w_n^2 \sqrt{n} < +\infty$;

(G) The directional derivatives $\nabla f_n(\theta, \theta' - \theta)$, and $\nabla f(\theta, \theta' - \theta)$ exist for all θ and θ' in Θ.

The assumptions **(D)** and **(E)** combined with **(A)** are useful because they allow us to use some uniform convergence results from the theory of empirical processes [34]. In a nutshell, these assumptions ensure that the function class $\{\mathbf{x} \mapsto \ell(\mathbf{x}, \theta) : \theta \in \Theta\}$, is "simple enough", such that a uniform law of large numbers applies. The assumption **(F)** is more technical: it resembles classical conditions used for proving the convergence of stochastic gradient descent algorithms, usually stating that the weights w_n should be the summand of a diverging sum while the sum of w_n^2 should be finite; the constraint $\sum_{n \geq 1} w_n^2 \sqrt{n} < +\infty$ is slightly stronger. Finally, **(G)** is a mild assumption, which is useful to characterize the stationary points of the problem. A classical necessary first-order condition [3] for θ to be a local minimum of f is indeed to have $\nabla f(\theta, \theta' - \theta)$ non-negative for all θ' in Θ. We call such points θ the stationary points of the function f. The next proposition is a generalization of a convergence result obtained in [23] in the context of sparse matrix factorization.

Proposition 3.3 (**Non-Convex Analysis - Almost Sure Convergence).**
*Under assumptions **(A)**, **(B)**, **(D)**, **(E)**, **(F)**, $(f(\theta_n))_{n\geq 0}$ converges with probability one. Moreover, under assumption **(G)**, we also have*

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

where the function $\bar{f}_n$ uniformly converges to f.

The function $\bar{f}_n$ is in fact a weighted empirical cost at time n. Even though $\bar{f}_n$ converges uniformly to the expected cost f, we need another assumption to characterize the limit points of the sequence $(\theta_n)_{n\geq 1}$ as stationary points of f. We consider the case of surrogates that are parameterized, an assumption always satisfied when Algorithm 1 is used in practice.

Proposition 3.4 (**Non-Convex Analysis - Parameterized Surrogates).**
Let us make the same assumptions as in Proposition 3.3, and let us assume that the approximate surrogates $\bar{g}_n$ are parameterized by some variables κ_n living in some compact set $\mathcal{K}$ of $\mathbb{R}^d$. In other words, $\bar{g}_n$ can be written as g_{κ_n}, with κ_n in $\mathcal{K}$. Suppose there exists a constant $K>0$ such that $|g_\kappa(\theta) - g_{\kappa'}(\theta)| \leq K\|\kappa-\kappa'\|_2$ for all θ in Θ and κ, κ' in $\mathcal{K}$.

Then, every limit point θ_∞ of the sequence $(\theta_n)_{n\geq 1}$ is a stationary point of f—that is, for all θ in Θ,

$$\nabla f(\theta_\infty, \theta-\theta_\infty) \geq 0.$$

Finally, we conclude our non-convex convergence analysis with a useful extension, which we apply in Section 4 to a non-convex sparse estimation formulation.

Proposition 3.5 (**Non-Convex Analysis - Extension).**
Assume that the functions f_n split into $f_n(\theta) = f_{0,n}(\theta) + \sum_{k=1}^K f_{k,n}(\gamma_k(\theta))$, where the functions $\gamma_k : \mathbb{R}^p \to \mathbb{R}$ are convex and R-Lipschitz, and the $f_{k,n}$ are non-decreasing for $k \geq 1$. Consider $g_{n,0}$ in $\mathcal{S}_{L_0,\rho_1}(f_{0,n}, \theta_{n-1})$, and some non-decreasing functions $g_{k,n}$ in $\mathcal{S}_{L_k,0}(f_{k,n}, \gamma_k(\theta_{n-1}))$. Instead of choosing g_n in $\mathcal{S}_{L,\rho}(f_n, \theta_{n-1})$, define $g_n \triangleq \theta \mapsto g_{0,n}(\theta) + g_{k,n}(\gamma_k(\theta))$.

*Then, under the same assumptions as in Proposition 3.3, and Proposition 3.4 except **(B)**, the respective conclusions of these two propositions hold.*

4 Applications and Experimental Validation

In this section, we introduce different applications, and provide some numerical experiments. An open-source C++ software package containing our implementations will be publicly released. All experiments were performed on a single core of a 2GHz Intel CPU with 64GB of RAM.

4.1 Stochastic Proximal Gradient Descent

One of our main application is a stochastic proximal gradient descent method, which we call SMM (Stochastic Majorization-Minimization), for solving problems of the form:

$$\min_{\theta\in\Theta} \mathbb{E}_{\mathbf{x}}[\ell(\mathbf{x}, \theta)] + \psi(\theta), \tag{4}$$

where ψ is a convex deterministic regularization function, and the functions $\theta \mapsto \ell(\mathbf{x}, \theta)$ are differentiable and their gradients are L-Lipschitz continuous. We can thus use the proximal gradient surrogate presented in Section 2. Assume that $w_1 = 1$, and define the weights w_n^i recursively as $w_n^i \triangleq (1-w_n)w_n^{i-1}$ for $i<n$ and $w_n^n \triangleq w_n$. Then, our scheme yields the update rule:

$$\theta_n \leftarrow \operatorname*{arg\,min}_{\theta\in\Theta} \sum_{i=1}^n w_n^i \left[\nabla f_i(\theta_{i-1})^\top \theta + \tfrac{L}{2}\|\theta-\theta_{i-1}\|_2^2 + \psi(\theta)\right]. \tag{SMM}$$

Our algorithm is related to FOBOS [9], to SMIDAS [32] or the truncated gradient method [20] (when ψ is the ℓ_1-norm). These three algorithms use indeed the following update rule:

$$\theta_n \leftarrow \arg\min_{\theta\in\Theta} \nabla f_n(\theta_{n-1})^\top \theta + \tfrac{1}{2\eta_n}\|\theta-\theta_{n-1}\|_2^2 + \psi(\theta), \qquad \text{(FOBOS)}$$

Another related scheme is the regularized dual averaging (RDA) of [37], which can be written as

$$\theta_n \leftarrow \arg\min_{\theta\in\Theta} \tfrac{1}{n}\textstyle\sum_{i=1}^n \nabla f_i(\theta_{i-1})^\top\theta + \tfrac{1}{2\eta_n}\|\theta\|_2^2 + \psi(\theta). \qquad \text{(RDA)}$$

Like RDA, our scheme includes an averaging of the gradients seen in the past iterations, but also involves the average of the past iterates $\sum_{i=1}^n w_n^i \theta_{i-1}$. Some links can also be drawn with approaches such as the "approximate follow the leader" algorithm of [14] and other works [16, 18].

We now evaluate the performance of our method for ℓ_1-logistic regression. In summary, the datasets consist of pairs $(y_i, \mathbf{x}_i)_{i=1}^N$, where the y_i's are in $\{-1, +1\}$, and the $\mathbf{x}_i$'s are in $\mathbb{R}^p$ with unit ℓ_2-norm. The function ψ in (4) is the ℓ_1-norm: $\psi(\theta) \triangleq \lambda\|\theta\|_1$, and λ is a regularization parameter; the functions f_i are logistic losses: $f_i(\theta) \triangleq \log(1+e^{-y_i\mathbf{x}_i^\top\theta})$. One part of each dataset is devoted to training, and another part to testing. We used weights of the form $w_n \triangleq \sqrt{(n_0+1)/(n+n_0)}$, where n_0 is automatically adjusted at the beginning of each experiment by performing one pass on 5% of the training data. We implemented SMM in C++ and exploited the sparseness of the datasets, such that each update has a computational complexity of the order $O(s)$, where s is the number of non zeros in $\nabla f_n(\theta_{n-1})$; such an implementation is non trivial but proved to be very efficient.

We consider three datasets described in the table below. rcv1 and webspam are obtained from the 2008 Pascal large scale learning challenge.[2] kdd2010 is available from the LIBSVM website.[3]

name	N_{tr} (train)	N_{te} (test)	p	density (%)	size (GB)
rcv1	781 265	23 149	47 152	0.161	0.95
webspam	250 000	100 000	16 091 143	0.023	14.95
kdd2010	10 000 000	9 264 097	28 875 157	10^{-4}	4.8

We compare our implementation with state-of-the-art publicly available solvers: the batch algorithm FISTA of [1] implemented in the C++ SPAMS toolbox [23], and LIBLINEAR v1.93 [11]. LIBLINEAR is based on a working-set algorithm, and, to the best of our knowledge, is one of the most efficient available solver for ℓ_1-logistic regression with sparse datasets. Because p is large, the incremental majorization-minimization method of [22] could not run for memory reasons. We run every method on $1, 2, 3, 4, 5, 10$ and 25 epochs, for three regularization regimes, respectively yielding a solution with approximately $100, 1\,000$ and $10\,000$ non-zero coefficients. We report the results for the medium regularization in Figure 1 and provide the rest as supplemental material. For visibility purposes, FISTA is not represented in this figure since it required more than 25 epochs to achieve reasonable values. Our conclusion is clear from these plots: *SMM often provides a reasonable solution after one epoch, and outperforms LIBLINEAR in the low-precision regime. For high-precision regimes, LIBLINEAR should be preferred.* Such a conclusion is classical when comparing batch and stochastic algorithms [5], but matching the performance of LIBLINEAR is very challenging.

4.2 Online DC Programming for Non-Convex Sparse Estimation

We now consider the same experimental setting as in the previous section, but with a non-convex regularizer $\psi : \theta \mapsto \lambda\sum_{j=1}^p \log(|\theta[j]|+\varepsilon)$, where $\theta[j]$ is the j-th entry in θ. A classical way for minimizing the regularized empirical cost $\frac{1}{N}\sum_{i=1}^N f_i(\theta) + \psi(\theta)$ is to resort to DC programming [13]. It consists of solving a sequence of reweighted-ℓ_1 problems [7]. A current estimate θ_{n-1} is updated as a solution of $\min_{\theta\in\Theta} \frac{1}{N}\sum_{i=1}^N f_i(\theta) + \lambda\sum_{j=1}^p w_j|\theta[j]|$, where $w_j \triangleq 1/(|\theta_{n-1}[j]|+\varepsilon)$.

[2] `http://largescale.ml.tu-berlin.de`.
[3] `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`.

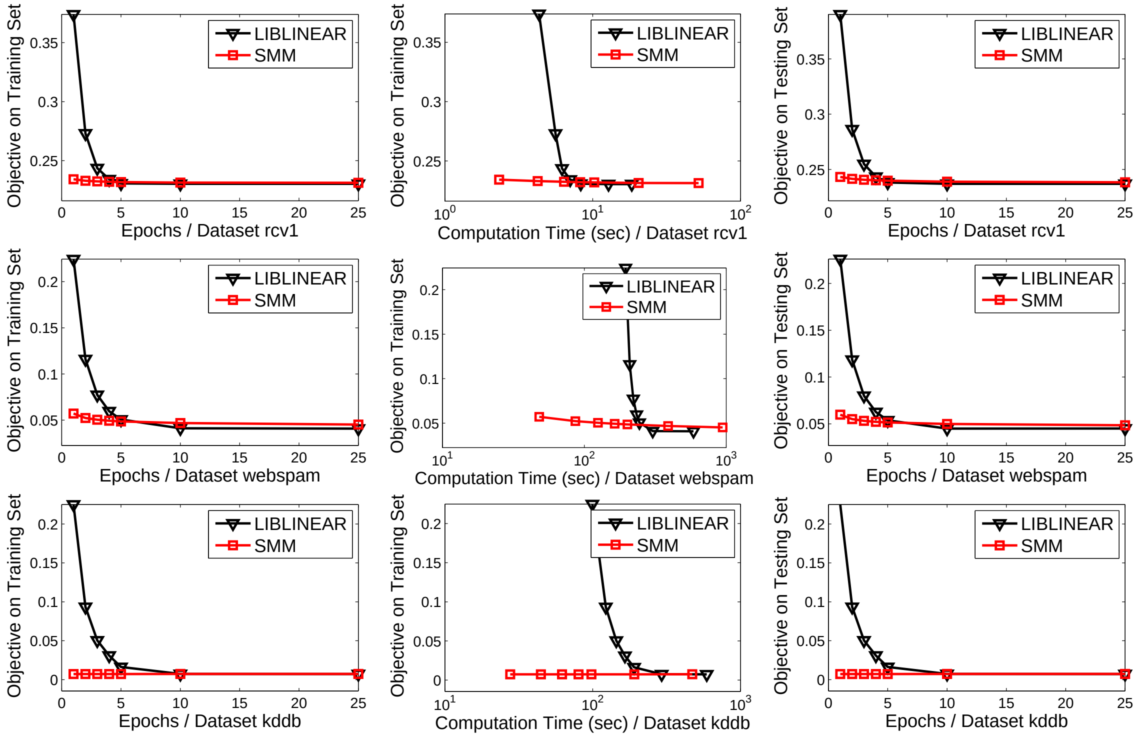

Figure 1: Comparison between LIBLINEAR and SMM for the medium regularization regime.

In contrast to this "batch" methodology, we can use our framework to address the problem online. At iteration n of Algorithm 1, we define the surrogate:

$$g_n : \theta \mapsto f_n(\theta_{n-1}) + \nabla f_n(\theta_{n-1})^\top(\theta - \theta_{n-1}) + \frac{L}{2}\|\theta - \theta_{n-1}\|_2^2 + \lambda \sum_{j=1}^p \frac{|\theta[j]|}{|\theta_{n-1}[j]|+\varepsilon},$$

and it is possible to show that the conditions of Proposition 3.5 are satisfied. We compare our online DC programming algorithm against the batch one, and report the results in Figure 2. We set the parameter ε to 0.01. The most interesting conclusion of our study is that *the batch reweighted-ℓ_1 algorithm always converges after 2 or 3 weight updates, but suffers from local minima issues. The stochastic algorithm exhibits a slower convergence, but provides significantly better solutions.* Whether or not there are good theoretical reasons for this fact remains to be investigated.

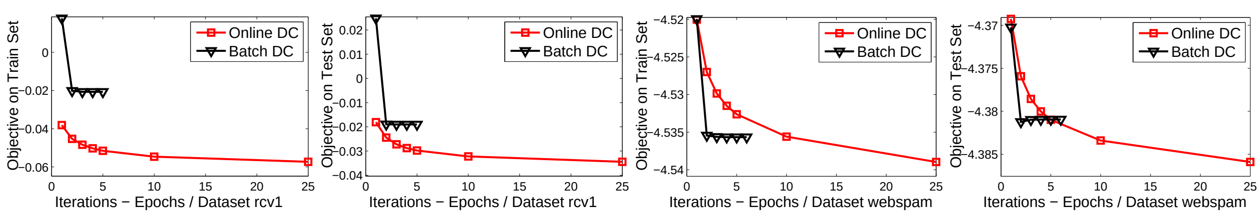

Figure 2: Comparison between batch and online DC programming, with medium regularization for the datasets rcv1 and webspam. Additional plots are provided in the supplemental material. Note that each iteration in the batch setting can perform several epochs (passes over training data).

4.3 Online Structured Sparse Coding

In this section, we show that we can bring new functionalities to existing matrix factorization techniques [17, 22]. We are given a large collection of signals $(\mathbf{x}_i)_{i=1}^N$ in $\mathbb{R}^m$, and we want to find a dictionary $\mathbf{D}$ in $\mathbb{R}^{m\times K}$ that can represent these signals in a sparse way. The quality of $\mathbf{D}$ is measured through the loss $\ell(\mathbf{x},\mathbf{D}) \triangleq \min_{\boldsymbol{\alpha}\in\mathbb{R}^K} \frac{1}{2}\|\mathbf{x}-\mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda_1\|\boldsymbol{\alpha}\|_1 + \frac{\lambda_2}{2}\|\boldsymbol{\alpha}\|_2^2$, where the ℓ_1-norm can be replaced by any convex regularizer, and the square loss by any convex smooth loss.

Then, we are interested in minimizing the following expected cost:

$$\min_{\mathbf{D}\in\mathbb{R}^{m\times K}} \mathbb{E}_{\mathbf{x}}\left[\ell(\mathbf{x},\mathbf{D})\right] + \gamma\varphi(\mathbf{D}),$$

where φ is a regularizer for $\mathbf{D}$. In the online learning approach of [22], the only way to regularize $\mathbf{D}$ is to use a constraint set, on which we need to be able to project efficiently; this is unfortunately not always possible. In the matrix factorization framework of [17], it is argued that some applications can benefit from a structured penalty φ, but the approach of [17] is not easily amenable to stochastic optimization. Our approach makes it possible by using the proximal gradient surrogate

$$g_n : \mathbf{D} \mapsto \ell(\mathbf{x}_n, \mathbf{D}_{n-1}) + \mathrm{Tr}(\mathbf{D}^\top \nabla_{\mathbf{D}}\ell(\mathbf{x}_n,\mathbf{D}_{n-1})) + \frac{L}{2}\|\mathbf{D}-\mathbf{D}_{n-1}\|_F^2 + \varphi(\mathbf{D}). \tag{5}$$

It is indeed possible to show that $\mathbf{D} \mapsto \ell(\mathbf{x}_n,\mathbf{D})$ is differentiable, and its gradient is Lipschitz continuous with a constant L that can be explicitly computed [22, 23].

We now design a proof-of-concept experiment. We consider a set of $N = 400\,000$ whitened natural image patches $\mathbf{x}_n$ of size $m = 20 \times 20$ pixels. We visualize some elements from a dictionary $\mathbf{D}$ trained by SPAMS [22] on the left of Figure 3; the dictionary elements are almost sparse, but have some residual noise among the small coefficients. Following [17], we propose to use a regularization function φ encouraging neighbor pixels to be set to zero together, thus leading to a sparse structured dictionary. We consider the collection $\mathcal{G}$ of all groups of variables corresponding to squares of 4 neighbor pixels in $\{1,\ldots,m\}$. Then, we define $\varphi(\mathbf{D}) \triangleq \gamma_1 \sum_{j=1}^K \sum_{g\in\mathcal{G}} \max_{k\in g} |\mathbf{d}_j[k]| + \gamma_2\|\mathbf{D}\|_F^2$, where $\mathbf{d}_j$ is the j-th column of $\mathbf{D}$. Since a fast solver for the proximal operator of φ is available in the toolbox SPAMS, it is then easy to use the surrogates (5). We set $\lambda_1 = 0.15$ and $\lambda_2 = 0.01$; After trying a few values for γ_1 and γ_2 at a reasonable computational cost, we obtain dictionaries with the desired regularization effect, as shown in Figure 3. Learning one dictionary of size $K = 256$ took a few minutes when performing one pass on the training data with mini-batches of size 100. This experiment demonstrates that our approach is more flexible and general than [22] and [17], opening new possibilities for these two works.

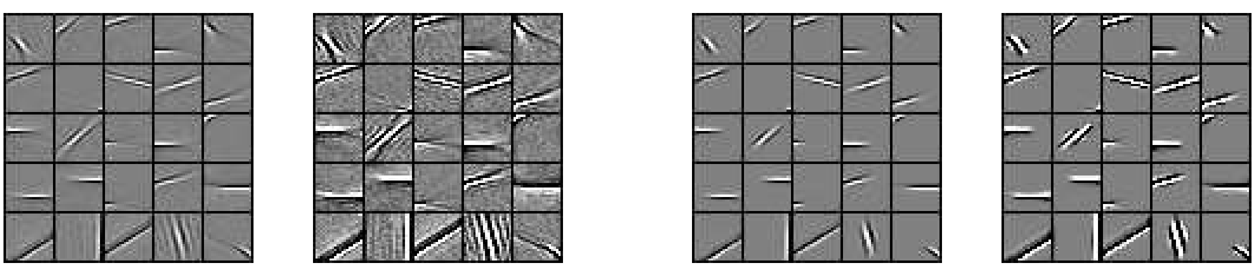

Figure 3: Left: Two visualizations of 25 elements from a larger dictionary obtained by the toolbox SPAMS [22]; the second view amplifies the small coefficients. Right: the corresponding views of the dictionary elements obtained by our approach after initialization with the dictionary on the left.

5 Conclusion

In this paper, we have introduced a stochastic majorization-minimization algorithm that gracefully scales to millions of training samples. We have shown that it has strong theoretical properties and some practical value in the context of machine learning. We have derived from our framework several new algorithms, which have shown to match or outperform the state of the art for solving large-scale convex problems, and to open up new possibilities for non-convex ones. In the future, we would like to study surrogate functions that can exploit the curvature of the objective function, which we believe is a crucial issue to deal with badly conditioned datasets.

Acknowledgments

JM would like to thank Zaid Harchaoui for fruitful discussions. This work was supported by the Gargantua project (program Mastodons - CNRS).

A Mathematical Background and Useful Results

In this paper, we use subdifferential calculus for convex functions. The definition of subgradients and directional derivatives can be found in classical textbooks, see, e.g., [3], [30]. We denote by $\partial f(\theta)$ the subdifferential of a convex function f at a point θ. Other definitions, for example the concept of strong convexity, can be found in the appendix of [22], which uses similar notation as us.

In this section, we present several classical optimization and probabilistic tools, which we use in our paper. The first lemma is a classical quadratic upper-bound for differentiable functions with a Lipschitz gradient. It can be found for instance in Lemma 1.2.3 of [27], or in the appendix of [22].

Lemma A.1 (**Convex Surrogate for Functions with Lipschitz Gradient).**
Let $f : \mathbb{R}^p \to \mathbb{R}$ be differentiable and ∇f be L-Lipschitz continuous. Then, for all θ, θ' in $\mathbb{R}^p$,

$$|f(\theta') - f(\theta) - \nabla f(\theta)^\top(\theta' - \theta)| \leq \frac{L}{2}\|\theta - \theta'\|_2^2. \tag{6}$$

The next lemma is a simple relation, which will allow us to identify the subdifferential of a convex function with the one of its surrogate at a particular point.

Lemma A.2 (**Surrogate Functions and Subdifferential).**
Assume that $f, g : \mathbb{R}^p \to \mathbb{R}$ are convex, and that $h \triangleq g - f$ is differentiable at θ in $\mathbb{R}^p$ with $\nabla h(\theta) = 0$. Then, $\partial f(\theta) = \partial g(\theta)$.

Proof. It is easy to show that g and f have the same directional derivatives at θ since h is differentiable and $\nabla h(\theta) = 0$. This is sufficient to conclude that $\partial g(\theta) = \partial f(\theta)$ by using Proposition 3.1.6 of [3], a simple lemma relating directional derivatives and subgradients. □

The following lemma is a lower bound for strongly convex functions. It can be found for instance in [29].

Lemma A.3 (**Lower Bound for Strongly Convex Functions).**
Let $f : \mathbb{R}^p \to \mathbb{R}$ be a μ-strongly convex function. Let z be in $\partial f(\kappa)$ for some κ in $\mathbb{R}^p$. Then, the following inequality holds for all θ in $\mathbb{R}^p$:

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

Proof. $l : \theta' \mapsto f(\theta) - \frac{\mu}{2}\|\theta - \kappa\|_2^2$ is convex by definition of strong convexity, and $l - f$ is differentiable with $\nabla(l - f)(\theta) = 0$. We apply Lemma A.2, which tells us that z is in $\partial l(\theta)$. This is sufficient to conclude, by noticing that a convex function is always above its tangents. □

The next lemma is also classical (see the appendix of [22]).

Lemma A.4 (Second-Order Growth Property).
Let $f : \mathbb{R}^p \to \mathbb{R}$ be a μ-strongly convex function 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:

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

Finally, we briefly recall that a strongly convex function can only have uniformly bounded subgradients on a compact set.

Lemma A.5 (Link Between Strong Convexity and Bounded Subgradients).
Let $f : \mathbb{R}^p \to \mathbb{R}$ be a μ-strongly convex function and $\Theta \subseteq \mathbb{R}^p$ be a convex set. Assume that the subgradients of f are uniformly bounded by R. Let us call $\theta^\star$ a minimizer of f on Θ. Then, for all θ in Θ, we have

$$\|\theta^\star - \theta\|_2 \leq \frac{R}{\mu}.$$

Proof. Let us call $\theta^\star$ a minimizer of f on Θ and let us define $f^\star \triangleq f(\theta^\star)$. According to Lemma A.3, if z is in $\partial f(\theta)$,

$$\begin{aligned} f^\star &\geq f(\theta) + z^\top(\theta^\star - \theta) + \frac{\mu}{2}\|\theta^\star - \theta\|_2^2 \\ &\geq f(\theta) - R\|\theta^\star - \theta\|_2 + \frac{\mu}{2}\|\theta^\star - \theta\|_2^2 \\ &\geq f^\star - R\|\theta^\star - \theta\|_2 + \mu\|\theta^\star - \theta\|_2^2, \end{aligned}$$

where the second inequality is obtained by using Lemma A.4. This is sufficient to conclude. □

We now introduce a sequence of probabilistic tools, which we use in our convergence analysis for non-convex functions. The first one is a classical theorem on quasi-martingales, which was used in [4] for proving the convergence of the stochastic gradient descent algorithm.

Theorem A.1 (Convergence of Quasi-Martingales.).
This presentation follows [4] and Proposition 9.5 and Theorem 9.4 of [24]. The original theorem is due to [12]. Let $(\mathcal{F}_n)_{n\geq 0}$ be an increasing family of σ-fields. Let $(X_n)_{n\geq 0}$ be a real stochastic process such that every random variable X_n is bounded below by a constant independent of n, and $\mathcal{F}_n$-measurable.

Let

$$\delta_n \triangleq \begin{cases} 1 & \text{if } \mathbb{E}[X_{n+1} - X_n|\mathcal{F}_n] > 0, \\ 0 & \text{otherwise.} \end{cases}$$

If for all n, the series $\sum_{n=0}^{\infty} \mathbb{E}[\delta_n(X_{n+1} - X_n)]$ converges, then $(X_n)_{n\geq 0}$ is a quasi-martingale and converges almost surely to an integrable random variable X_∞. Moreover,

$$\sum_{n=0}^{\infty} \mathbb{E}\big[|\mathbb{E}[X_{n+1} - X_n|\mathcal{F}_n]|\big] < \infty.$$

The next lemma is simple, but useful to prove the convergence of deterministic algorithms.

Lemma A.6. Deterministic Lemma on Non-negative Converging Series.
Let $(a_n)_{n\geq 0}$, $(b_n)_{n\geq 0}$ be two non-negative real sequences such that the series $\sum_{n=1}^{\infty} a_n$ diverges, the series $\sum_{n=1}^{\infty} a_n b_n$ converges, and there exists $K > 0$ such that $|b_{n+1} - b_n| \leq K a_n$. Then, the sequence $(b_n)_{n\geq 0}$ converges to 0.

Proof. The proof is inspired by the one of Proposition 1.2.4 of [2]. Since the series $\sum_{n\geq 0} a_n$ diverges, we necessarily have $\liminf_{n\to+\infty} b_n = 0$. Otherwise, it would be easy to contradict the assumption $\sum_{n\geq 0} a_n b_n < +\infty$.

Let us now proceed by contradiction and assume that $\limsup_{n\to+\infty} b_n = \lambda > 0$. We can then build two sequence of indices m_j and n_j such that

- $m_j < n_j < m_{j+1}$,
- $\frac{\lambda}{3} < b_k$, for $m_j \leq k < n_j$,
- $b_k \leq \frac{\lambda}{3}$, for $n_j < k \leq m_{j+1}$.

Let $\varepsilon = \frac{\lambda^2}{9K}$ and $\tilde{\jmath}$ be large enough so that

$$\sum_{n=m_{\tilde{\jmath}}}^{\infty} a_n b_n < \varepsilon.$$

Then, we have for all $j \geq \tilde{\jmath}$ and all m with $m_j \leq m \leq n_j - 1$,

$$\begin{aligned}|b_{n_j} - b_m| &\leq \sum_{k=m}^{n_j-1} |b_{k+1} - b_k| \leq \frac{3K}{\lambda} \sum_{k=m}^{n_j-1} a_k \frac{\lambda}{3} \leq \frac{3K}{\lambda} \sum_{k=m}^{n_j-1} a_k b_k \leq \frac{3K}{\lambda} \sum_{k=m}^{+\infty} a_k b_k \\ &\leq \frac{3K\varepsilon}{\lambda} \leq \frac{\lambda}{3}.\end{aligned}$$

Therefore, using the triangle inequality,

$$b_m \leq b_{n_j} + \frac{\lambda}{3} \leq \frac{2\lambda}{3}.$$

and finally, for all $m \geq \tilde{\jmath}$,

$$b_m \leq \frac{2\lambda}{3},$$

which contradicts $\limsup_{n\to+\infty} b_n = \lambda > 0$. Therefore, $b_n \underset{n\to+\infty}{\longrightarrow} 0$.

□

We now provide a stochastic version of Lemma A.7.

Lemma A.7. Stochastic Lemma on Non-negative Converging Series.
Let $(X_n)_{n\geq 1}$ be a sequence of non-negative measurable random variables on a probability space. Let also a_n, b_n be two non-negative sequences such that $\sum_{n\geq 1} a_n = +\infty$ and $\sum_{n\geq 1} a_n b_n < +\infty$. Assume that there exists a constant C such that for all $n \geq 1$, $\mathbb{E}[X_n] \leq b_n$ and $|X_{n+1} - X_n| \leq C a_n$ almost surely. Then X_n almost surely converges to zero.

Proof. The following series is convergent

$$\mathbb{E}\left[\sum_{n\geq 1} a_n X_n\right] = \sum_{n\geq 1} \mathbb{E}\left[a_n X_n\right] \leq \sum_{n\geq 1} a_n b_n < +\infty,$$

where we use the fact that the random variables are non-negative to interchange the sum and the expectation. We thus have that $\sum_{n\geq 1} a_n X_n$ converges with probability one. Then, let us call $a'_n = a_n$ and $b'_n = X_n$; the conditions of Lemma A.6 are satisfied for a'_n and b'_n with probability one, and X_n almost surely converges to zero. □

B Auxiliary Lemmas

In this section, we present auxiliary lemmas for our convex and non-convex analyses. We start by presenting a lemma which is useful for both of them, and which is in fact a core component of all proofs presented in [22].

Lemma B.1 (**Basic Properties of First-Order Surrogate Functions).**
Let g be in $\mathcal{S}_{L,\rho}(f,\kappa)$ for some κ in Θ. Define the approximation error function $h \triangleq g - f$ and let θ' be the minimizer of g over Θ. Then, for all θ in Θ,

- $\nabla h(\kappa) = 0$;
- $|h(\theta)| \leq \frac{L}{2}\|\theta - \kappa\|_2^2$;
- $f(\theta') \leq g(\theta') \leq f(\theta) + \frac{L}{2}\|\theta - \kappa\|_2^2 - \frac{\rho}{2}\|\theta - \theta'\|_2^2$.

Proof. h is differentiable and minimized by 0. Thus, $\nabla h(\kappa) = 0$. see [22] for the rest of the proof. □

B.1 Convex Analysis

We introduce, for all $n \geq 0$, the quantity $\xi_n \triangleq \frac{1}{2}\mathbb{E}[\|\theta^\star - \theta_n\|_2^2]$, where $\theta^\star$ is a minimizer of f on Θ. Our analysis also involves several quantities that are defined recursively for all $n \geq 1$:

$$\left\{\begin{array}{lcl} A_n & \triangleq & (1-w_n)A_{n-1} + w_n\xi_{n-1} \\ B_n & \triangleq & (1-w_n)B_{n-1} + w_n\mathbb{E}[f(\theta_{n-1})] \\ C_n & \triangleq & (1-w_n)C_{n-1} + \frac{(Rw_n)^2}{2\rho} \\ \bar{g}_n & \triangleq & (1-w_n)\bar{g}_{n-1} + w_n g_n \\ \bar{f}_n & \triangleq & (1-w_n)\bar{f}_{n-1} + w_n f_n \end{array}\right. , \tag{7}$$

where $A_0 \triangleq \frac{1}{L}(\rho\xi_0 - f^\star)$, $B_0 \triangleq 0$, $C_0 \triangleq 0$, $\bar{g}_0 = \bar{f}_0 \triangleq \theta \mapsto \frac{\rho}{2}\|\theta - \theta_0\|_2^2$. Note that $\bar{g}_0$ is ρ-strongly convex, and is minimized by θ_0. The choice for A_0, B_0, C_0 is driven by technical reasons, which appear in the proof of Lemma B.4, a stochastic version of Lemma B.1.

Note that we also assume here that all the expectation above are well defined and finite valued. In other words, we do not deal with measurability or integrability issues for simplicity, as often done in the literature [26].

Lemma B.2 (**Auxiliary Lemma for Convex Analysis).**
*When the functions f_n are convex, under assumptions **(A)** and **(B)**, we have for all $n \geq 1$,*

$$\bar{g}_n(\theta_{n-1}) \leq \bar{g}_n(\theta_n) + \frac{(Rw_n)^2}{2\rho}.$$

Proof. First, we remark that the subdifferentials of g_n and f_n at θ_{n-1} coincide by applying Lemma A.2. Then, we choose z_n in $\partial g_n(\theta_{n-1}) = \partial f_n(\theta_{n-1})$, which is bounded by R according to assumption **(A)**.

$$\begin{aligned} \bar{g}_n(\theta_n) &= (1-w_n)\bar{g}_{n-1}(\theta_n) + w_n g_n(\theta_n) \\ &\geq (1-w_n)\left(\bar{g}_{n-1}(\theta_{n-1}) + \frac{\rho}{2}\|\theta_n - \theta_{n-1}\|_2^2\right) + w_n\left(g_n(\theta_{n-1}) + z_n^\top(\theta_n - \theta_{n-1}) + \frac{\rho}{2}\|\theta_n - \theta_{n-1}\|_2^2\right) \\ &= \bar{g}_n(\theta_{n-1}) + w_n z_n^\top(\theta_n - \theta_{n-1}) + \frac{\rho}{2}\|\theta_n - \theta_{n-1}\|_2^2 \\ &\geq \bar{g}_n(\theta_{n-1}) - Rw_n\|\theta_n - \theta_{n-1}\|_2 + \frac{\rho}{2}\|\theta_n - \theta_{n-1}\|_2^2 \\ &\geq \bar{g}_n(\theta_{n-1}) - \frac{(Rw_n)^2}{2\rho}. \end{aligned}$$

The first inequality uses Lemma A.4 and Lemma A.3 since g_n is ρ-strongly convex according to assumption **(B)** (and by induction $\bar{g}_n$ is ρ-strongly convex as well); the second inequality uses Cauchy-Schwarz's inequality and the fact that the subgradients of the functions f_n are bounded by R. □

Lemma B.3 (**Another Auxiliary Lemma for Convex Analysis).**
*When the functions f_n are convex, under assumptions **(A)** and **(B)**, for all $n \geq 0$,*

$$B_n \leq \mathbb{E}[\bar{g}_n(\theta_n)] + C_n, \tag{8}$$

Proof. We proceed by induction, and start by showing that Eq. (8) is true for $n = 0$.

$$B_0 = 0 = \mathbb{E}[\bar{g}_0(\theta_0)] = \mathbb{E}[\bar{g}_0(\theta_0)] + C_0.$$

Let us now assume that it is true for $n-1$, and show that it is true for n.

$$\begin{aligned} B_n &= (1-w_n)B_{n-1} + w_n\mathbb{E}[f(\theta_{n-1})] \\ &\leq (1-w_n)(\mathbb{E}[\bar{g}_{n-1}(\theta_{n-1})] + C_{n-1}) + w_n\mathbb{E}[f(\theta_{n-1})] \\ &= (1-w_n)(\mathbb{E}[\bar{g}_{n-1}(\theta_{n-1})] + C_{n-1}) + w_n\mathbb{E}[f_n(\theta_{n-1})] \\ &\leq (1-w_n)(\mathbb{E}[\bar{g}_{n-1}(\theta_{n-1})] + C_{n-1}) + w_n\mathbb{E}[g_n(\theta_{n-1})] \\ &= \mathbb{E}[\bar{g}_n(\theta_{n-1})] + (1-w_n)C_{n-1} \\ &\leq \mathbb{E}[\bar{g}_n(\theta_n)] + C_n. \end{aligned}$$

The first inequality uses the induction hypothesis; the last inequality uses Lemma B.2 and the definition of C_n. We also used the fact that $\mathbb{E}[f_n(\theta_{n-1})] = \mathbb{E}[\mathbb{E}[f_n(\theta_{n-1})]|\mathcal{F}_{n-1}]] = \mathbb{E}[\mathbb{E}[f(\theta_{n-1})]|\mathcal{F}_{n-1}]] = \mathbb{E}[f(\theta_{n-1})]$, where $\mathcal{F}_{n-1}$ corresponds to the filtration induced by the past information before time n, such that θ_{n-1} is deterministic given $\mathcal{F}_{n-1}$. □

The next lemma is important; it is the stochastic version of Lemma B.1 for first-order surrogates.

Lemma B.4 (**Basic Properties of Stochastic First-Order Surrogates).**
*When the functions f_n are convex, under assumptions **(A)** and **(B)**, for all $n \geq 0$,*

$$B_n \leq f^\star + LA_n - \rho\xi_n + C_n,$$

Proof. According to Lemma B.3, it remains to show that $\mathbb{E}[\bar{g}_n(\theta_n)] \leq f^\star + LA_n - \rho\xi_n$ for all $n \geq 0$. Since $\bar{g}_n$ is ρ-strongly convex, we have $\mathbb{E}[\bar{g}_n(\theta_n)] \leq \mathbb{E}[\bar{g}_n(\theta^\star)] - \rho\xi_n$, by using Lemma A.4. Thus, it is in fact sufficient to show that $\mathbb{E}[\bar{g}_n(\theta^\star)] \leq f^\star + LA_n$. For $n = 0$, this inequality holds since $\mathbb{E}[\bar{g}_0(\theta^\star)] = \rho\xi_0 = f^\star + LA_0$. We can then proceed again by induction: assume that $\mathbb{E}[\bar{g}_{n-1}(\theta^\star)] \leq f^\star + LA_{n-1}$. Then,

$$\begin{aligned} \mathbb{E}[\bar{g}_n(\theta^\star)] &= (1-w_n)\mathbb{E}[\bar{g}_{n-1}(\theta^\star)] + w_n\mathbb{E}[g_n(\theta^\star)] \\ &\leq (1-w_n)(f^\star + LA_{n-1}) + w_n\mathbb{E}[f_n(\theta^\star) + L\xi_{n-1}] \\ &= (1-w_n)(f^\star + LA_{n-1}) + w_n(f^\star + L\xi_{n-1}) \\ &= f^\star + LA_n, \end{aligned}$$

where we have used Lemma B.1 to upper-bound the difference $g_n(\theta^\star) - f_n(\theta^\star)$. □

For strongly-convex function, another simple but useful relation between A_n and B_n can be obtained.

Lemma B.5 (**Relation between A_n and B_n).**
*Under assumption **(C)**, if $w_1 = 1$, we have for all $n \geq 1$,*

$$f^\star + \mu A_n \leq B_n.$$

Proof. This relation is true For $n = 1$ since we have $f^\star + \mu A_0 = f^\star + \mu\xi_0 \leq f(\theta_0) = B_1$ by applying Lemma A.4, since f is μ-strongly convex according to assumption **(C)**. The rest follows by induction. □

B.2 Non-convex Analysis

When the functions f_n are not convex, the convergence analysis becomes more involved. One key tool we use is a uniform convergence result when the function class $\{f_n\}_{n \geq 0}$ is "simple enough" (in terms of entropy). Under the assumptions made in our paper, it is indeed possible to use some results from empirical processes [34], which provides us the following lemma.

Lemma B.6 (**Uniform Convergence**).
*Under assumptions **(A)**, **(D)**, and **(E)**, we have the following uniform law of large numbers:*

$$\mathbb{E}\left[\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n} f_i(\theta) - f(\theta)\right|\right] \leq \frac{C}{\sqrt{n}}, \tag{9}$$

where C is a constant, and $\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n} f_i(\theta) - f(\theta)\right|$ converges almost surely to zero.

Proof. We simply refer to Lemma 19.36 and Example 19.7 of [34], where assumptions **(D)** and **(E)** ensure uniform boundness and squared integrability conditions. Note that we assume that the quantities $\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n} f_i(\theta) - f(\theta)\right|$ are measurable. This assumption does not incur a loss of generality, since measurability issues for empirical processes can be dealt with rigorously [34]. □

The next lemma shows that uniform convergence applies to $\bar{f}_n$, defined in Eq. (7), but with a different rate.

Lemma B.7 (**Uniform Convergence for $\bar{f}_n$**).
*Under assumptions **(A)**, **(D)**, **(E)**, and **(F)**, we have for all $n \geq 1$,*

$$\mathbb{E}\left[\sup_{\theta\in\Theta}\left|\bar{f}_n(\theta) - f(\theta)\right|\right] \leq C w_n \sqrt{n},$$

where C is the same as in Lemma B.6, and $\sup_{\theta\in\Theta}\left|\bar{f}_n(\theta) - f(\theta)\right|$ converges almost surely to zero.

Proof. We prove the two parts of the lemma separately. As in Lemma B.6, we assume all the quantities of interest to be measurable.

First part of the lemma:

Let us fix $n > 0$. It is easy to show that $\bar{f}_n$ can be written as $\bar{f}_n = \sum_{i=1}^{n} w_n^i f_i$ for some non-negative weights w_n^i with $w_n^n = w_n$. Let us also define the empirical cost $F_i \triangleq \frac{1}{n-i+1}\sum_{j=i}^{n} f_j$. According to (9), we have $\mathbb{E}\left[\sup_{\theta\in\Theta}|F_i(\theta) - f(\theta)|\right] \leq \frac{C}{\sqrt{n-i+1}}$. We now remark that

$$\bar{f}_n - f = \sum_{i=1}^{n}(w_n^i - w_n^{i-1})(n-i+1)(F_i - f),$$

where we have defined $w_n^0 \triangleq 0$. This relation can be proved by simple calculation. We obtain the first part by using the triangle inequality, and the fact that $x_n^i \geq w_n^{i-1}$ for all i:

$$\begin{aligned}
\mathbb{E}\left[\sup_{\theta\in\Theta}|\bar{f}_n(\theta) - f(\theta)|\right] &\leq \mathbb{E}\left[\sum_{i=1}^{n}(w_n^i - w_n^{i-1})(n-i+1)\sup_{\theta\in\Theta}|F_i(\theta) - f(\theta)|\right] \\
&= \sum_{i=1}^{n}(w_n^i - w_n^{i-1})(n-i+1)\mathbb{E}\left[\sup_{\theta\in\Theta}|F_i(\theta) - f(\theta)|\right] \\
&\leq \sum_{i=1}^{n}(w_n^i - w_n^{i-1})C\sqrt{n-i+1} \\
&\leq \sqrt{n}C\sum_{i=1}^{n}(w_n^i - w_n^{i-1}) \\
&= C\sqrt{n}w_n.
\end{aligned}$$

This is unfortunately not sufficient to show that $\mathbb{E}\left[\sup_{\theta\in\Theta}|\bar{f}_n(\theta) - f(\theta)|\right]$ converges to zero almost surely. We will show this fact by using Lemma A.7.

Second part of the lemma:
We call $X_n = \sup_{\theta \in \Theta} |\bar{f}_n(\theta) - f(\theta)|$. We have

$$X_n - X_{n-1} = \sup_{\theta\in\Theta} |(1-w_n)(\bar{f}_{n-1}(\theta) - f(\theta)) + w_n(f_n(\theta) - f(\theta))| - X_{n-1}$$
$$\leq \sup_{\theta\in\Theta} w_n|f_n(\theta) - f(\theta)| - w_n X_{n-1} \leq 2M w_n$$

Let us denote by $\theta_n^\star$ a point in Θ such that $X_n = |\bar{f}_n(\theta_n^\star) - f(\theta_n^\star)|$. We also have

$$\begin{aligned} X_n - X_{n-1} &= \sup_{\theta\in\Theta} |(1-w_n)(\bar{f}_{n-1}(\theta) - f(\theta)) + w_n(f_n(\theta) - f(\theta))| - X_{n-1} \\ &\geq (1-w_n)X_{n-1} + w_n(f_n(\theta_{n-1}^\star) - f(\theta_{n-1}^\star)) - X_{n-1} \\ &\geq -w_n X_{n-1} + w_n(f_n(\theta_{n-1}^\star) - f(\theta_{n-1}^\star)) \\ &\geq -w_n 4M, \end{aligned}$$

where we use again the fact that all functions f_n, $\bar{f}_n$ and f are bounded by M. Thus, we have shown that $|X_n - X_{n-1}| \leq 4Mw_n$. Call $a_n = w_n$ and $b_n = w_n\sqrt{n}$, then the conditions of Lemma A.7 are satisfied, and X_n converges almost surely to zero. □

Finally, the next lemma illustrates why the strong convexity of the surrogates is important.

Lemma B.8 (Stability of the Estimates).
*Under assumption **(B)**,*

$$\|\theta_n - \theta_{n-1}\|_2 \leq \frac{2Rw_n}{\rho}.$$

Proof. Because the surrogates are ρ-strongly convex, we have from Lemma A.4

$$\begin{aligned} \frac{\rho}{2}\|\theta_n - \theta_{n-1}\|_2^2 &\leq \bar{g}_n(\theta_{n-1}) - \bar{g}_n(\theta_n) \\ &= w_n\left(g_n(\theta_{n-1}) - g_n(\theta_n)\right) + (1-w_n)\left(\bar{g}_{n-1}(\theta_{n-1}) - \bar{g}_{n-1}(\theta_n)\right) \\ &\leq w_n\left(g_n(\theta_{n-1}) - g_n(\theta_n)\right) \\ &\leq w_n\left(f_n(\theta_{n-1}) - f_n(\theta_n)\right) \\ &\leq Rw_n\|\theta_n - \theta_{n-1}\|_2. \end{aligned}$$

The second inequality comes from the fact that θ_{n-1} is a minimizer of $\bar{g}_{n-1}$; the third inequality is because $g_n(\theta_{n-1}) = f_n(\theta_{n-1})$ and $g_n \geq f_n$. This is sufficient to conclude. □

C Proofs of the Main Lemmas and Propositions

C.1 Proof of Proposition 3.1

Proof. According to Lemma B.4, we have for all $n \geq 1$,

$$w_n B_{n-1} \leq w_n f^\star + Lw_n A_{n-1} - Lw_n \xi_{n-1} + w_n C_{n-1}.$$

By using the relations (7), this is equivalent to

$$B_{n-1} - B_n + w_n \mathbb{E}[f(\theta_{n-1})] \leq w_n f^\star + L(A_{n-1} - A_n) + C_{n-1} - C_n + \frac{(Rw_n)^2}{2L}.$$

By summing this inequalities between 1 and n, we obtain

$$B_0 - B_n + \sum_{k=1}^n w_k f(\theta_{k-1}) \leq \left(\sum_{k=1}^n w_k\right) f^\star + LA_0 - LA_n - C_n + \sum_{k=1}^n \frac{(Rw_k)^2}{2L}.$$

Note that we also have

$$B_n \leq f^\star + LA_n + C_n = LA_n + C_n + B_0 - LA_0 + L\xi_0.$$

Therefore, by combining the two previous inequalities,

$$\sum_{k=1}^{n} w_k \mathbb{E}[f(\theta_{k-1})] \leq \left(\sum_{k=1}^{n} w_k\right) f^\star + L\xi_0 + \sum_{k=1}^{n} \frac{(Rw_k)^2}{2L}.$$

and by using Jensen's inequality,

$$\mathbb{E}[f(\bar{\theta}_{n-1}) - f^\star] \leq \frac{L\xi_0 + \frac{R^2}{2L}\sum_{k=1}^{n} w_k^2}{\sum_{k=1}^{n} w_k}.$$

□

C.2 Proof of Corollary 3.1

Proof. When assuming weights w_k to be constant, we can minimize the right side of Eq. (3) under the constraint $w_k \in [0,1]$. It yields $w_k \triangleq \min\left(\frac{L\|\theta^\star - \theta_0\|_2}{R\sqrt{n}}, 1\right)$. Plugging in this quantity in Eq. (3) gives the desired rates after a few calculations. □

C.3 Proof of Corollary 3.2

Proof.
We choose weights of the form $w_n \triangleq \frac{\gamma}{\sqrt{n}}$. Then, we have

$$\sum_{k=1}^{n} w_k^2 \leq \gamma^2(1 + \log n),$$

by using the fact that $\sum_{k=1}^{n} \frac{1}{k} \leq 1 + \log(n)$. We also have for $n \geq 2$,

$$\sum_{k=1}^{n} w_k \geq 2\gamma(\sqrt{n+1} - 1) \geq \gamma\sqrt{n},$$

where we use the fact that $\sum_{k=1}^{n} \frac{1}{\sqrt{k}} \geq 2(\sqrt{n+1} - 1)$, and the fact that $2(\sqrt{n+1} - 1) \geq \sqrt{n}$ for all $n \geq 2$.
if $R \geq L\|\theta^\star - \theta_0\|_2$:
By using the above inequalities in Eq. (3) with $\gamma = \frac{L\|\theta^\star - \theta_0\|_2}{R}$, we get for $n \geq 2$,

$$\mathbb{E}[f(\bar{\theta}_{n-1}) - f^\star] \leq R\|\theta^\star - \theta_0\|_2 \frac{1 + \log\sqrt{n}}{\sqrt{n}}.$$

if $R < L\|\theta^\star - \theta_0\|_2$:
We choose the weights $w_n \triangleq 1$, and we get for $n \geq 2$,

$$\mathbb{E}[f(\bar{\theta}_{n-1}) - f^\star] \leq L\|\theta^\star - \theta_0\|_2^2 \frac{1 + \log\sqrt{n}}{\sqrt{n}},$$

by observing that $\frac{R}{L} \leq \|\theta^\star - \theta_0\|_2$.
Asymptotic rate for any γ:
This is straightforward by proceeding as before, using the same inequalities. □

C.4 Proof of Proposition 3.2

Proof. We proceed in several steps, proving the convergence rates of several quantities of interest.
Convergence rate of C_n:
Let us show by induction that we have $C_n \leq \frac{R^2}{\rho} w_n$ for all $n \geq 1$. This is obviously true for $n = 1$ by definitions of $w_1 = 1$ and $C_1 = \frac{R^2}{2\rho}$. Let us now assume that it is true for $n-1$. We have

$$\begin{aligned}
C_n &= (1-w_n)C_{n-1} + \frac{R^2}{2\rho}w_n^2 \\
&\leq \frac{R^2}{\rho}w_n\left((1-w_n)\frac{w_{n-1}}{w_n} + \frac{w_n}{2}\right) \\
&\leq \frac{R^2}{\rho}w_n\left(\frac{\beta(n-1)}{\beta n+1}\frac{\beta n+1}{\beta(n-1)+1} + \frac{1}{\beta n+1}\right) \\
&\leq \frac{R^2}{\rho}w_n\left(\frac{\beta(n-1)}{\beta(n-1)+1} + \frac{1}{\beta(n-1)+1}\right) \\
&= \frac{R^2}{\rho}w_n.
\end{aligned} \tag{10}$$

We conclude by induction that this is true for all $n \geq 1$.
Convergence rate of A_n:
From Lemma B.5 and B.4, we have for all $n \geq 2$,

$$\mu A_{n-1} \leq L A_{n-1} - \rho\xi_{n-1} + C_{n-1}.$$

Multiplying this inequality by w_n,

$$2\mu w_n A_{n-1} \leq \rho w_n (A_{n-1} - \xi_{n-1}) + w_n C_{n-1},$$

where the factor 2 comes from the fact that $\rho = L + \mu$. By using the definition of A_n in Eq. (7), we obtain the relation

$$A_n \leq \left(1 - \frac{2\mu w_n}{\rho}\right) A_{n-1} + \frac{w_n}{\rho} C_{n-1}.$$

Let us now show by induction that we have, for all $n \geq 1$, the convergence rate $A_n \leq \delta w_n$, where $\delta \triangleq \max\left(\frac{R^2}{\rho\mu}, \xi_0\right)$. For $n = 1$, $A_1 = \xi_0 \leq \delta$ and $w_1 = 1$. Assume now that we have $A_{n-1} \leq \delta w_{n-1}$ for some $n \geq 1$. Then, by using the convergence rate (10) and the induction hypothesis,

$$\begin{aligned}
A_n &\leq \delta w_n \left(\left(1 - \frac{2\mu w_n}{\rho}\right)\frac{w_{n-1}}{w_n} + \frac{R^2 w_{n-1}}{\rho\delta}\right) \\
&\leq \delta w_n \left(\left(1 - \frac{2\mu w_n}{\rho}\right)\frac{w_{n-1}}{w_n} + \mu\frac{w_{n-1}}{\rho}\right) \\
&\leq \delta w_n \left(\frac{\beta n + 1 - \frac{2\mu(1+\beta)}{\rho}}{\beta n+1}\frac{\beta n+1}{\beta(n-1)+1} + \frac{\frac{\mu(1+\beta)}{\rho}}{\beta(n-1)+1}\right) \\
&= \delta w_n \left(\frac{\beta n + 1 - \frac{\mu(1+\beta)}{\rho}}{\beta(n-1)+1}\right) \\
&\leq \delta w_n.
\end{aligned}$$

The last inequality uses the fact that $\frac{\mu(1+\beta)}{\rho} \geq \beta$ because $\beta \leq \frac{\mu}{L}$. we conclude by induction that $A_n \leq \delta w_n$ for all $n \geq 1$.

Convergence rate of $\mathbb{E}[f(\hat{\theta}_n) - f^\star] + \rho\xi_n$:
We use again Lemma B.4:

$$B_n - f^\star + \rho\xi_n \leq LA_n + C_n,$$

and we consider two possible cases

- If $\frac{R^2}{\rho\mu} \geq \xi_0$, then

$$\begin{aligned} B_n - f^\star + \rho\xi_n &\leq \frac{R^2}{\rho}\left(1 + \frac{L}{\mu}\right) w_n \\ &= \frac{R^2}{\mu} w_n \\ &\leq \frac{2R^2}{\mu(\beta n + 1)}, \end{aligned}$$

  where we simply use the convergence rates of A_n and C_n computed before.

- If instead $\frac{R^2}{\rho\mu} < \xi_0$, then

$$\begin{aligned} B_n - f^\star + \rho\xi_n &\leq \left(\frac{R^2}{\rho} + L\xi_0\right) w_n \\ &\leq \rho\xi_0 w_n \\ &\leq \frac{2\rho\xi_0}{\beta n + 1}. \end{aligned}$$

  It is then easy to prove that $\mathbb{E}[f(\hat{\theta}_n) - f^\star] \leq B_n$ by using Jensen's inequality, which allows us to conclude.

□

C.5 Proof of Proposition 3.3

Proof. We generalize the proof of convergence for online matrix factorization of [23]. The proof exploits Theorem A.1 about the convergence of quasi-martingales [12], similarly as [4] for proving the convergence of the stochastic gradient descent algorithm for non-convex functions.
Almost sure convergence of $(\bar{g}_n(\theta_n))_{n\geq 1}$:
The first step consists of applying a convergence theorem for the sequence $(\bar{g}_n(\theta_n))_{n\geq 1}$ by bounding its positive expected variations. Define $Y_n \triangleq \bar{g}_n(\theta_n)$. For $n \geq 2$, we have

$$\begin{aligned} Y_n - Y_{n-1} &= \bar{g}_n(\theta_n) - \bar{g}_n(\theta_{n-1}) + \bar{g}_n(\theta_{n-1}) - \bar{g}_{n-1}(\theta_{n-1}) \\ &= (\bar{g}_n(\theta_n) - \bar{g}_n(\theta_{n-1})) + w_n(g_n(\theta_{n-1}) - \bar{g}_{n-1}(\theta_{n-1})) \\ &= (\bar{g}_n(\theta_n) - \bar{g}_n(\theta_{n-1})) + w_n(\bar{f}_{n-1}(\theta_{n-1}) - \bar{g}_{n-1}(\theta_{n-1})) + w_n(g_n(\theta_{n-1}) - \bar{f}_{n-1}(\theta_{n-1})) \\ &= (\bar{g}_n(\theta_n) - \bar{g}_n(\theta_{n-1})) + w_n(\bar{f}_{n-1}(\theta_{n-1}) - \bar{g}_{n-1}(\theta_{n-1})) + w_n(f_n(\theta_{n-1}) - \bar{f}_{n-1}(\theta_{n-1})) \\ &\leq w_n(f_n(\theta_{n-1}) - \bar{f}_{n-1}(\theta_{n-1})). \end{aligned} \tag{11}$$

The final inequality comes from the inequality $\bar{g}_n \geq \bar{f}_n$, which is easy to show by induction starting from $n = 1$ since $w_1 = 1$. It follows,

$$\begin{aligned} \mathbb{E}[\bar{g}_n(\theta_n) - \bar{g}_{n-1}(\theta_{n-1})|\mathcal{F}_{n-1}] &\leq w_n \mathbb{E}[f_n(\theta_{n-1}) - \bar{f}_{n-1}(\theta_{n-1})|\mathcal{F}_{n-1}] \\ &= w_n(f(\theta_{n-1}) - \bar{f}_{n-1}(\theta_{n-1})) \\ &\leq w_n \sup_{\theta\in\Theta} |f(\theta) - \bar{f}_{n-1}(\theta)|, \end{aligned}$$

where $\mathcal{F}_{n-1}$ is the filtration representing the past information before time n. Call now

$$\delta_n \triangleq \begin{cases} 1 & \text{if } \mathbb{E}[\bar{g}_n(\theta_n) - \bar{g}_{n-1}(\theta_{n-1})|\mathcal{F}_{n-1}] > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Then, the series below with non-negative summands converges:

$$\begin{aligned}\sum_{n=1}^{\infty} \mathbb{E}[\delta_n(\bar{g}_n(\theta_n) - \bar{g}_{n-1}(\theta_{n-1}))] &= \sum_{n=1}^{\infty} \mathbb{E}[\delta_n \mathbb{E}[(\bar{g}_n(\theta_n) - \bar{g}_{n-1}(\theta_{n-1}))|\mathcal{F}_{n-1}]] \\ &\leq \sum_{n=1}^{\infty} \mathbb{E}\left[w_n \sup_{\theta\in\Theta} |f(\theta) - \bar{f}_{n-1}(\theta)|\right] \\ &\leq \sum_{n=1}^{\infty} C w_n^2 \sqrt{n} < +\infty,\end{aligned}$$

The second inequality comes from Lemma B.7. Since in addition $\bar{g}_n$ is bounded below by some constant independent of n, we can apply Theorem A.1. This theorem tells us that $(\bar{g}_n(\theta_n))_{n\geq 1}$ converges almost surely to an integrable random variable $g^\star$ and that $\sum_{n=1}^{\infty} \mathbb{E}[|\mathbb{E}[\bar{g}_n(\theta_n) - \bar{g}_{n-1}(\theta_{n-1})|\mathcal{F}_{n-1}]|]$ converges almost surely.
Almost sure convergence of $(\bar{f}_n(\theta_n))_{n\geq 1}$:
We will show by using Lemma A.6 that the non-positive term $\bar{f}_n(\theta_n) - \bar{g}_n(\theta_n)$ almost surely converges to zero, and thus $(\bar{f}_n(\theta_n))_{n\geq 1}$ is also converging almost surely to $g^\star$.

We observe that

$$\sum_{n=1}^{\infty} \mathbb{E}[|\mathbb{E}[\bar{g}_n(\theta_n) - \bar{g}_{n-1}(\theta_{n-1})|\mathcal{F}_{n-1}]|] = \mathbb{E}\left[\sum_{n=1}^{\infty} |\mathbb{E}[\bar{g}_n(\theta_n) - \bar{g}_{n-1}(\theta_{n-1})|\mathcal{F}_{n-1}]|\right] < +\infty.$$

Thus, the series $\sum_{n=1}^{\infty} |\mathbb{E}[\bar{g}_n(\theta_n) - \bar{g}_{n-1}(\theta_{n-1})|\mathcal{F}_{n-1}]|$ is absolutely convergent with probability one, and the series $\sum_{n=1}^{\infty} \mathbb{E}[\bar{g}_n(\theta_n) - \bar{g}_{n-1}(\theta_{n-1})|\mathcal{F}_{n-1}]$ is also almost surely convergent.

We also remark that, using Lemma B.7,

$$\mathbb{E}\left[\sum_{n=1}^{+\infty} w_n |f(\theta_{n-1}) - \bar{f}_{n-1}(\theta_{n-1})|\right] \leq C \sum_{n=1}^{+\infty} w_n^2 \sqrt{n} < +\infty,$$

and thus $w_n(f(\theta_{n-1}) - \bar{f}_{n-1}(\theta_{n-1}))$ is the summand of an absolutely convergent series with probability one.

Taking the expectation of Eq. (11) conditioned on $\mathcal{F}_{n-1}$, it remains that the non-positive term $w_n(\bar{f}_{n-1}(\theta_{n-1}) - \bar{g}_{n-1}(\theta_{n-1}))$ is also necessarily the summand of an almost surely convergent series, since all other terms in the equation are summands of almost surely converging sums. This is not sufficient to immediately conclude that $\bar{f}_n(\theta_n) - \bar{g}_n(\theta_n)$ converges to zero almost surely, and thus we will use Lemma A.6. We have that $\sum_{n=1}^{+\infty} w_n$ diverges, that $\sum_{n=1}^{+\infty} w_n(\bar{g}_{n-1}(\theta_{n-1}) - \bar{f}_{n-1}(\theta_{n-1}))$ converges almost surely. Define $X_n \triangleq (\bar{g}_{n-1}(\theta_{n-1}) - \bar{f}_{n-1}(\theta_{n-1}))$. By definition of the surrogate functions, the differences $h_n \triangleq g_n - f_n$ are differentiable and their gradients are L-Lipschitz continuous. Since in addition Θ is compact and $\nabla h_n(\theta_{n-1}) = 0$, ∇h_n is bounded by some constant R' independent of n, and the function h_n is R'-Lipschitz. This is therefore also the case for $\bar{h}_n = \bar{g}_n - \bar{f}_n$.

$$\begin{aligned}|X_{n+1} - X_n| &= |\bar{h}_n(\theta_n) - \bar{h}_{n-1}(\theta_{n-1})| \\ &\leq |\bar{h}_n(\theta_n) - \bar{h}_n(\theta_{n-1})| + |\bar{h}_n(\theta_{n-1}) - \bar{h}_{n-1}(\theta_{n-1})| \\ &\leq R'\|\theta_n - \theta_{n-1}\|_2 + |\bar{h}_n(\theta_{n-1}) - \bar{h}_{n-1}(\theta_{n-1})| \\ &\leq \frac{2RR'}{\rho} w_n + w_n |h_n(\theta_{n-1}) - \bar{h}_{n-1}(\theta_{n-1})| \\ &= \frac{2RR'}{\rho} w_n + w_n |\bar{h}_{n-1}(\theta_{n-1})| \\ &\leq O(w_n).\end{aligned}$$

The second inequality uses the fact that $\bar{h}_n$ is R'-Lipschitz; The second inequality uses Lemma B.8; the last equality uses the fact that the functions h_n are also bounded by some constant independent of n (using the fact that ∇h_n is uniformly bounded). We can now apply Lemma A.6, and X_n converges to zero with probability one. Thus, $(\bar{f}_n(\theta_n))_{n\geq 1}$ converges almost surely to $g^\star$.

Almost sure convergence of $(f(\theta_n))_{n\geq 1}$:

Since $(\bar{f}_n(\theta_n))_{n\geq 1}$ converges almost surely, we simply use Lemma A.7, which tells us that $\bar{f}_n$ converges uniformly to f. Then, $(f(\theta_n))_{n\geq 1}$ converges almost surely to $g^\star$.

Asymptotic Stationary Point Condition:

Let us call $\bar{h}_n \triangleq \bar{g}_n - \bar{f}_n$, which can be shown to be differentiable with a L-Lipschitz gradient, according to assumption **(B)**. For all θ in Θ,

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

Since θ_n is the minimizer of $\bar{g}_n$, we have $\nabla \bar{g}_n(\theta_n, \theta - \theta_n) \geq 0$.

Since $\bar{h}_n$ is differentiable and its gradient is L-Lipschitz continuous, we can apply Lemma A.1 to $\theta = \theta_n$ and $\theta' = \theta_n - \frac{1}{L}\nabla \bar{h}_n(\theta_n)$, which gives $\bar{h}_n(\theta') \leq \bar{h}_n(\theta_n) - \frac{1}{2L}\|\nabla \bar{h}_n(\theta_n)\|_2^2$. Since we have shown that $\bar{h}_n(\theta_n) = \bar{g}_n(\theta_n) - \bar{f}(\theta_n)$ converges to zero and $\bar{h}_n(\theta') \geq 0$, we have that $\|\nabla \bar{h}_n(\theta_n)\|_2$ converges to zero. Thus,

$$\inf_{\theta \in \Theta} \frac{\nabla \bar{f}_n(\theta_n, \theta - \theta_n)}{\|\theta - \theta_n\|_2} \geq -\|\nabla \bar{h}_n(\theta_n)\|_2 \underset{n \to +\infty}{\longrightarrow} 0 \text{ a.s.}$$

□

C.6 Proof of Proposition 3.4

Proof. Since Θ is compact according to assumption **(D)**, the sequence $(\theta_n)_{n\geq 1}$ admits limit points. Let us consider a converging subsequence $(n_k)_{k\geq 1}$ to a limit point θ_∞ in Θ. In this converging subsequence, we can also find a subsequence $(n_{k'})_{k'\geq 1}$ such that $\kappa_{n_{k'}}$ converges to a point κ_∞ in $\mathcal{K}$ (which is compact). For the sake of simplicity, and without loss of generality, we remove the indices k and k' from the notation and assume that θ_n converges to θ_∞, while κ_n converges to κ_∞. It is then easy to see that the functions $\bar{g}_n$ converge uniformly to $\bar{g}_\infty \triangleq g_{\kappa_\infty}$, given the assumptions made in the Proposition.

Defining $\bar{h}_\infty \triangleq \bar{g}_\infty - f$, we have for all θ in Θ:

$$\nabla f(\theta_\infty, \theta - \theta_\infty) = \nabla \bar{g}_\infty(\theta_\infty, \theta - \theta_\infty) - \nabla \bar{h}_\infty(\theta_\infty, \theta - \theta_\infty).$$

To prove the proposition, we will first show that $\nabla \bar{g}_\infty(\theta_\infty, \theta - \theta_\infty) \geq 0$ and then that $\nabla \bar{h}_\infty(\theta_\infty, \theta - \theta_\infty) = 0$.

Proof of $\nabla \bar{g}_\infty(\theta_\infty, \theta - \theta_\infty) \geq 0$:

It is sufficient to show that θ_∞ is a minimizer of $\bar{g}_\infty$. This is straightforward, by taking the limit when n goes to infinity of

$$\bar{g}_n(\theta) \geq \bar{g}_n(\theta_n),$$

where we use the uniform convergence of $\bar{g}_n$.

Proof of $\nabla \bar{h}_\infty(\theta_\infty, \theta - \theta_\infty) = 0$:

Since both $\bar{f}_n$ and $\bar{g}_n$ converges uniformly (according to Lemma B.7 for $\bar{f}_n$), we have that $\bar{h}_n$ converges uniformly to $\bar{h}_\infty$. Since $\bar{h}_n$ is differentiable with a L-Lipschitz gradient, we have for all vector $\mathbf{z}$ in $\mathbb{R}^p$,

$$\bar{h}_n(\theta_n + \mathbf{z}) = \bar{h}_n(\theta_n) + \nabla \bar{h}_n(\theta_n)^\top \mathbf{z} + O(\|\mathbf{z}\|_2^2),$$

where the constant in O is independent of n. By taking the limit when n goes to infinity, it remains

$$\bar{h}_\infty(\theta_\infty + \mathbf{z}) = \bar{h}_\infty(\theta_\infty) + O(\|\mathbf{z}\|_2^2),$$

since we have shown in the proof of Proposition 3.3 that $\|\nabla \bar{h}_n(\theta_n)\|_2$ converges to zero. Since $\bar{h}_\infty$ admits a first order extension around θ_∞ it is differentiable at this point and furthermore, $\nabla \bar{h}_\infty(\theta_\infty) = 0$. This is sufficient to conclude. □

C.7 Proof of Proposition 3.5

Proof. First we notice that

- $g_n \geq f_n$;
- $g_n(\theta_{n-1}) = f_n(\theta_{n-1})$;
- g_n is ρ_1-strongly convex since $\theta \mapsto g_{k,n}(\gamma_k(\theta))$ can be shown to be convex, following elementary composition rules for convex functions (see [6], Section 3.2.4).

Thus, the only property missing is the smoothness of the approximation error $h_n \triangleq g_n - f_n$. Rather than writing again a full proof, we now simply review the different places where this property is used, and which modifications should be made to the proofs of Propositions 3.3 and 3.4.

In the second step of this proof, we require the functions h_n to be uniformly Lipschitz and uniformly bounded. It easy to check that it is still the case with the assumptions we made in Proposition 3.5.

The last step about the asymptotic point condition is however more problematic, where we cannot show anymore that the quantity $\nabla \bar{h}_n(\theta_n)$ converges to zero (since $\bar{h}_n$ is not differentiable anymore). Instead, we need to show that the directional derivative $\frac{\nabla \bar{h}_n(\theta_n, \theta - \theta_n)}{\|\theta - \theta_n\|}$ uniformly converges to zero on Θ.

We will show the result for $K = 1$; it will be easy to extend it to any arbitrary $K > 2$. We remark that

$$\nabla \bar{h}_n(\theta_n, \theta - \theta_n) = \nabla \bar{h}_{0,n}(\theta_n)^\top(\theta - \theta_n) + \lim_{t \to 0^+} \frac{\bar{h}_{1,n}(\gamma_1(\theta_n + t(\theta - \theta_n))) - \bar{h}_{1,n}(\gamma_1(\theta_n))}{t},$$

where $\bar{h}_{0,n}$ and $\bar{h}_{1,n}$ are defined similarly as $\bar{h}_n$ for the functions $h_{0,n} \triangleq g_{0,n} - f_{0,n}$ and $h_{1,n} \triangleq g_{1,n} - f_{1,n}$ respectively. Since $\bar{h}_n(\theta_n)$ is shown to converge to zero, we have that the non-negative quantities $\bar{h}_{0,n}(\theta_n)$ and $\bar{h}_{1,n}(\gamma_1(\theta_n))$ converge to zero as well. Since $\bar{h}_{0,n}$ and $\bar{h}_{1,n}$ are differentiable and their gradients are Lipschitz, we use similar arguments as in the proof of Proposition 3.3, and we have that $\nabla \bar{h}_{0,n}(\theta_n)$ and $\bar{h}'_{1,n}(\gamma_1(\theta_n))$ converge to zero (where $\bar{h}'_{1,n}$ is the derivative of $\bar{h}_{1,n}$. Concerning the second term, we can make the following Taylor expansion for $\bar{h}_{1,n}$:

$$\bar{h}_{1,n}(\gamma_1(\theta_n + \mathbf{z})) = \bar{h}_{1,n}(\gamma_1(\theta_n)) + \bar{h}'_{1,n}(\gamma_1(\theta_n))(\gamma_1(\theta_n + \mathbf{z}) - \gamma_1(\theta_n)) + O\left((\gamma_1(\theta_n + \mathbf{z}) - \gamma_1(\theta_n))^2\right),$$

where the constant in the O notation is independent of θ_n and $\mathbf{z}$ (since the derivative is L_1-Lipschitz). Plugging $\mathbf{z} \triangleq t(\theta - \theta_n)$ in this last equation, and using the Lipschitz property of γ_1, we have

$$\lim_{t \to 0^+} \left| \frac{\bar{h}_{1,n}(\gamma_1(\theta_n + t(\theta - \theta_n))) - \bar{h}_{1,n}(\gamma_1(\theta_n))}{t} \right| \leq |\bar{h}'_{1,n}(\gamma_1(\theta_n))| \|\theta - \theta_n\|.$$

Since $\bar{h}'_{1,n}(\gamma_1(\theta_n))$ converges to zero, we can conclude the proof of the modified Proposition 3.3.

The proof of Proposition 3.4 can be modified with very similar arguments. □

D Additional Experimental Results

We present in Figures 4 and 5 some additional experimental comparisons, which complement the ones of Section 4.1. Figures 6 and 7 present additional plots from the experiment of Section 4.2. Finally, we present three dictionaries corresponding to the experiment of Section 4.3 in Figures 8, 9 and 10.

References

[1] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

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

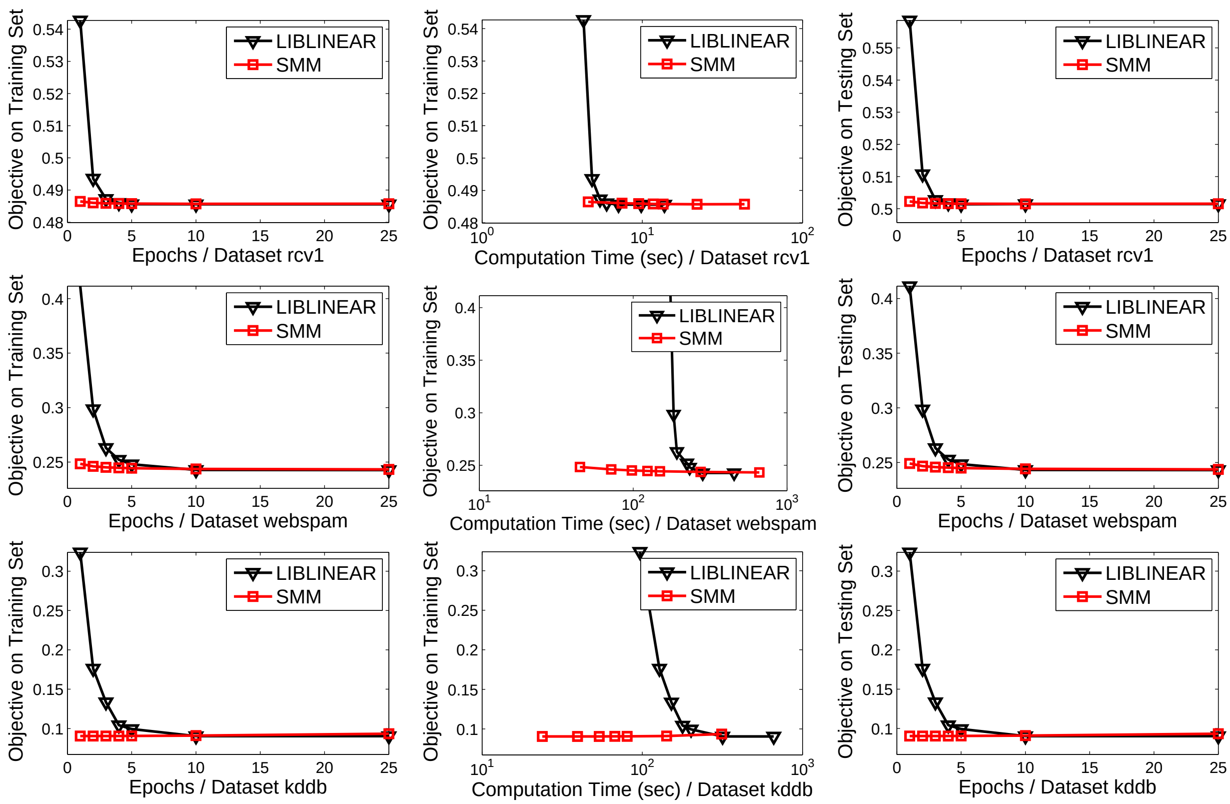

Figure 4: Comparison between LIBLINEAR and SMM in the high regularization regime for ℓ_1-logistic regression.

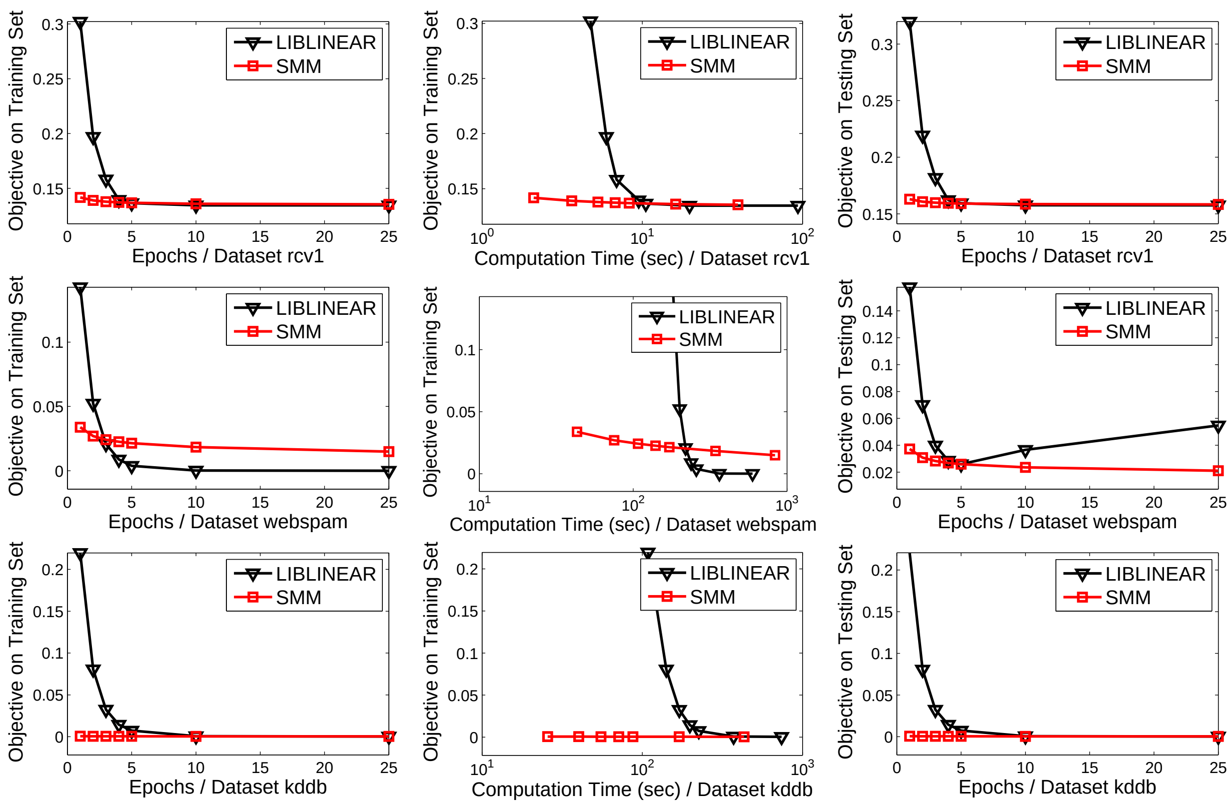

Figure 5: Comparison between LIBLINEAR and SMM in the low regularization regime for ℓ_1-logistic regression.

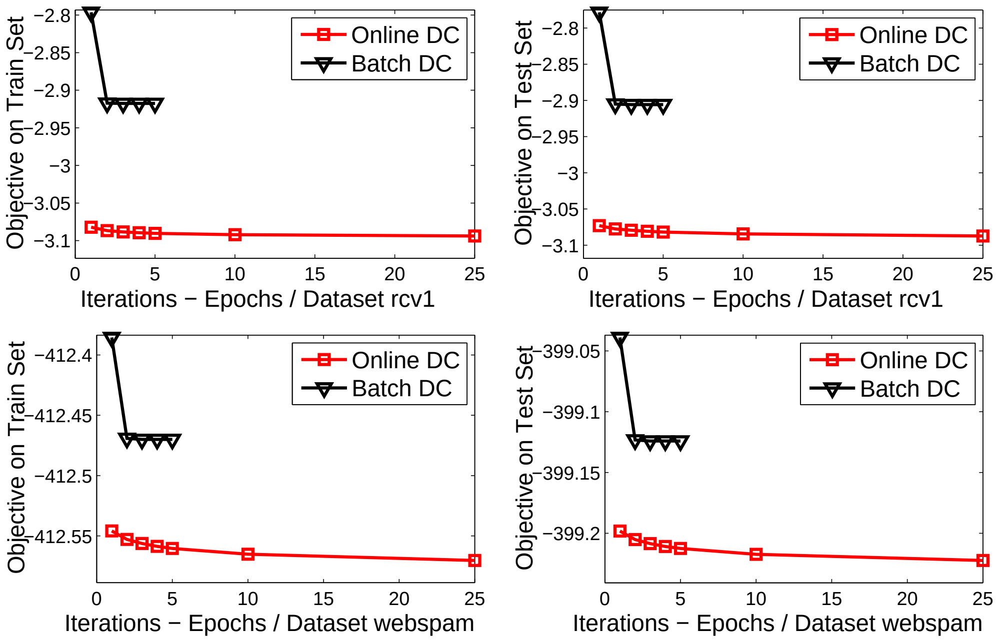

Figure 6: Comparison between batch and online DC programming, with high regularization for the datasets rcv1 and webspam. Note that each iteration in the batch setting can perform several epochs.

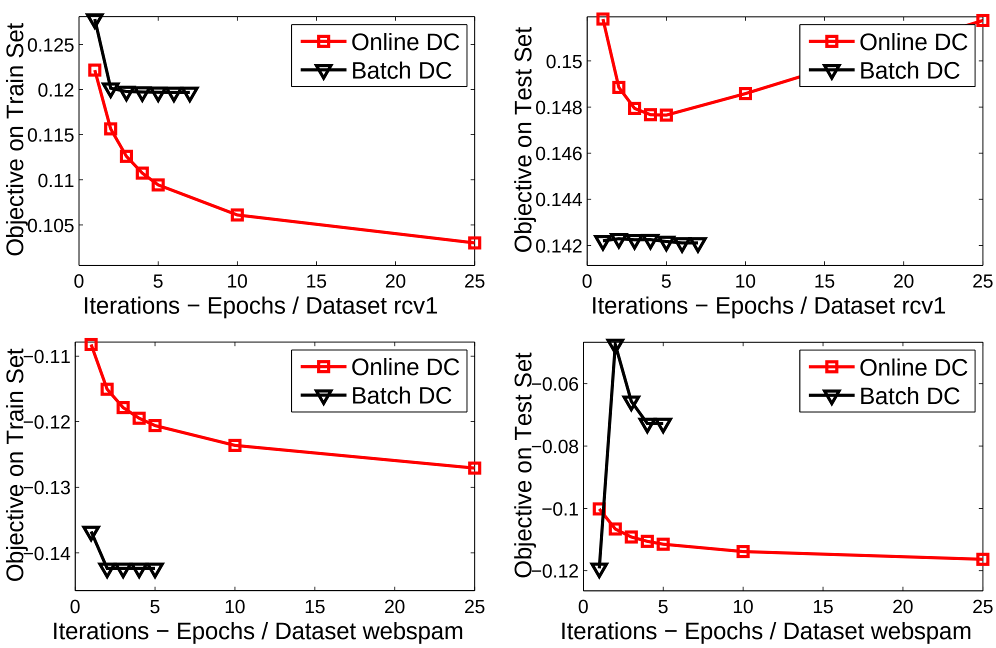

Figure 7: Comparison between batch and online DC programming, with low regularization for the datasets rcv1 and webspam. Note that each iteration in the batch setting can perform several epochs.

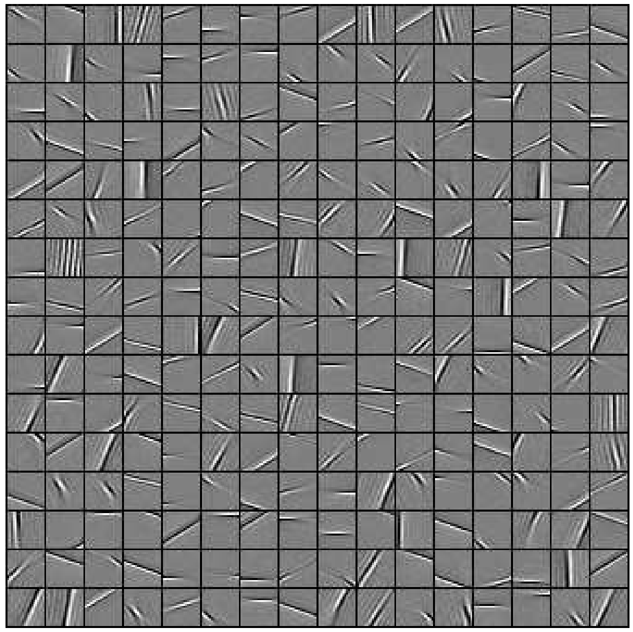

Figure 8: Dictionary obtained using the toolbox SPAMS [22].

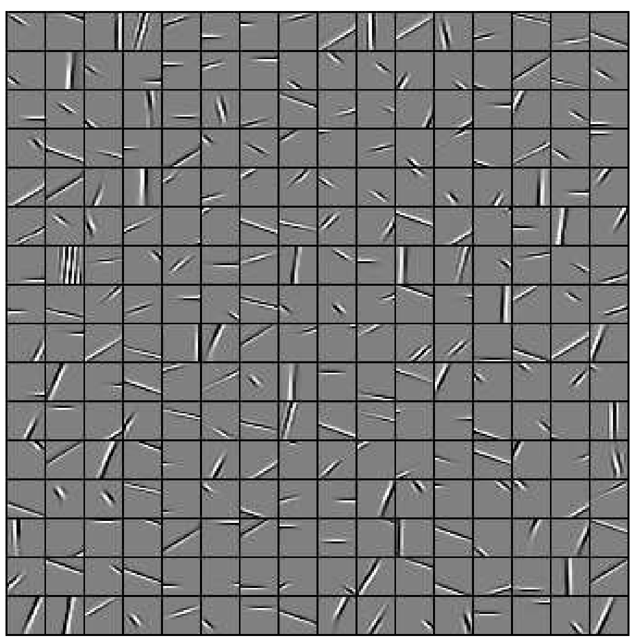

Figure 9: Sparse dictionary obtained by our approach, using the dictionary of Figure 8 as an initialization.

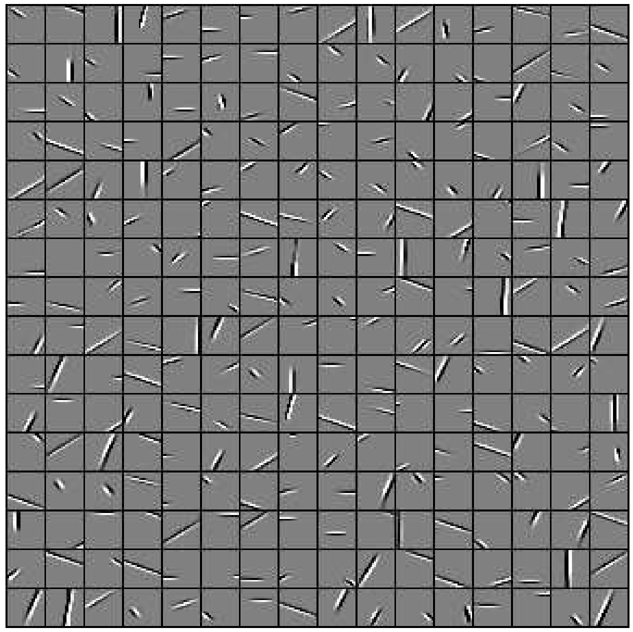

Figure 10: Sparse dictionary obtained by our approach, using the dictionary of Figure 8 as an initialization, and with a higher regularization parameter than in Figure 9.

[3] J.M. Borwein and A.S. Lewis. *Convex analysis and nonlinear optimization*. Springer, 2006.

[4] L. Bottou. Online algorithms and stochastic approximations. In David Saad, editor, *Online Learning and Neural Networks*. 1998.

[5] L. Bottou and O. Bousquet. The trade-offs of large scale learning. In *Adv. NIPS*, 2008.

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

[7] E.J. Candès, M. Wakin, and S.P. Boyd. Enhancing sparsity by reweighted ℓ_1 minimization. *J. Fourier Anal. Appl.*, 14(5):877–905, 2008.

[8] O. Cappé and E. Moulines. On-line expectation–maximization algorithm for latent data models. *J. Roy. Stat. Soc. B*, 71(3):593–613, 2009.

[9] J. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari. Composite objective mirror descent. In *Proc. COLT*, 2010.

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

[11] 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.*, 9:1871–1874, 2008.

[12] D. L. Fisk. Quasi-martingales. *T. Am. Math. Soc.*, 120(3):359–388, 1965.

[13] G. Gasso, A. Rakotomamonjy, and S. Canu. Recovering sparse signals with non-convex penalties and DC programming. *IEEE T. Signal Process.*, 57(12):4686–4698, 2009.

[14] E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Mach. Learn.*, 69(2-3):169–192, 2007.

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

[16] C. Hu, J. Kwok, and W. Pan. Accelerated gradient methods for stochastic optimization and online learning. In *Adv. NIPS*, 2009.

[17] R. Jenatton, G. Obozinski, and F. Bach. Structured sparse principal component analysis. In *Proc. AISTATS*, 2010.

[18] G. Lan. An optimal method for stochastic composite optimization. *Math. Program.*, 133:365–397, 2012.

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

[20] J. Langford, L. Li, and T. Zhang. Sparse online learning via truncated gradient. *J. Mach. Learn. Res.*, 10:777–801, 2009.

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

[22] J. Mairal. Optimization with first-order surrogate functions. In *Proc. ICML*, 2013. arXiv:1305.3120.

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

[24] M. Métivier. *Semi-martingales*. Walter de Gruyter, 1983.

[25] 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.

[26] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optimiz.*, 19(4):1574–1609, 2009.

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

[28] Y. Nesterov. Gradient methods for minimizing composite objective functions. Technical report, CORE Discussion Paper, 2007.

[29] Y. Nesterov and J.-P. Vial. Confidence level solutions for stochastic programming. *Automatica*, 44(6):1559–1568, 2008.

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

[31] S. Shalev-Schwartz and T. Zhang. Proximal stochastic dual coordinate ascent. *arXiv 1211.2717v1*, 2012.

[32] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Stochastic convex optimization. In *Proc. COLT*, 2009.

[33] S. Shalev-Shwartz and A. Tewari. Stochastic methods for ℓ_1 regularized loss minimization. In *Proc. ICML*, 2009.

[34] A. W. Van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

[35] M.J. Wainwright and M.I. Jordan. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1(1-2):1–305, 2008.

[36] S. Wright, R. Nowak, and M. Figueiredo. Sparse reconstruction by separable approximation. *IEEE T. Signal Process.*, 57(7):2479–2493, 2009.

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