Optimization with First-Order Surrogate Functions

Julien Mairal

Optimization with First-Order Surrogate Functions

Julien Mairal JULIEN.MAIRAL@INRIA.FR
INRIA LEAR Project-Team, Grenoble, France

Abstract

In this paper, we study optimization methods consisting of iteratively minimizing surrogates of an objective function. By proposing several algorithmic variants and simple convergence analyses, we make two main contributions. First, we provide a unified viewpoint for several first-order optimization techniques such as accelerated proximal gradient, block coordinate descent, or Frank-Wolfe algorithms. Second, we introduce a new incremental scheme that experimentally matches or outperforms state-of-the-art solvers for large-scale optimization problems typically arising in machine learning.

1. Introduction

The principle of iteratively minimizing a majorizing surrogate of an objective function is often called *majorization-minimization* (Lange et al., 2000). Each iteration drives the objective function downhill, thus giving the hope of finding a local optimum. A large number of existing procedures can be interpreted from this point of view. This is for instance the case of gradient-based or proximal methods (see Nesterov, 2007; Beck & Teboulle, 2009; Wright et al., 2009), EM algorithms (see Neal & Hinton, 1998), DC programming (Horst & Thoai, 1999), boosting (Collins et al., 2002; Della Pietra et al., 2001), and some variational Bayes techniques (Wainwright & Jordan, 2008; Seeger & Wipf, 2010). The concept of "surrogate" has also been used successfully in the signal processing literature about sparse optimization (Daubechies et al., 2004; Gasso et al., 2009) and matrix factorization (Lee & Seung, 2001; Mairal et al., 2010).

In this paper, we are interested in generalizing the majorization-minimization principle. Our goal is both to discover new algorithms, and to draw connections with existing methods. We focus our study on "first-order surrogate functions", which consist of approximating a possibly non-smooth objective function up to a smooth error. We present several schemes exploiting such surrogates, and analyze their convergence properties: asymptotic stationary point conditions for non-convex problems, and convergence rates for convex ones. More precisely, we successively study:

- a generic majorization-minimization approach;
- a randomized block coordinate descent algorithm (see Tseng & Yun, 2009; Shalev-Shwartz & Tewari, 2009; Nesterov, 2012; Richtárik & Takáč, 2012);
- an accelerated variant for convex problems inspired by Nesterov (2004); Beck & Teboulle (2009);
- a generalization of the "Frank-Wolfe" conditional gradient method (see Zhang, 2003; Harchaoui et al., 2013; Hazan & Kale, 2012; Zhang et al., 2012);
- a new incremental scheme, which we call MISO.[1]

We present in this work a unified view for analyzing a large family of algorithms with simple convergence proofs and strong guarantees. In particular, all the above optimization methods except Frank-Wolfe have linear convergence rates for minimizing strongly convex objective functions. This is remarkable for MISO, the new incremental scheme derived from our framework; to the best of our knowledge, only two recent incremental algorithms share such a property: the *stochastic average gradient* method (SAG) of Le Roux et al. (2012), and the *stochastic dual coordinate ascent* method (SDCA) of Shalev-Schwartz & Zhang (2012). Our scheme MISO is inspired in part by these two works, but yields different update rules than SAG or SDCA.

After we present and analyze the different optimization schemes, we conclude the paper with numerical experiments focusing on the scheme MISO. We show that in most cases MISO matches or outperforms cutting-edge solvers for large-scale ℓ_2- and ℓ_1-regularized logistic regression (Bradley et al., 2011; Beck & Teboulle, 2009; Le Roux et al., 2012; Fan et al., 2008; Bottou, 2010).

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

[1] *Minimization by Incremental Surrogate Optimization.*

2. Basic Optimization Scheme

Given a convex subset Θ of $\mathbb{R}^p$ and a continuous function $f : \mathbb{R}^p \to \mathbb{R}$, we are interested in solving

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

where we assume, to simplify, that f is bounded below. Our goal is to study the majorization-minimization scheme presented in Algorithm 1 and its variants. This procedure relies on the concept of surrogate functions, which are minimized instead of f at every iteration.[2]

Algorithm 1 Basic Scheme

input $\theta_0 \in \Theta$; 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: Update solution: $\theta_n \in \arg\min_{\theta \in \Theta} g_n(\theta)$.
4: **end for**
output θ_N (final estimate);

For this approach to be successful, we intuitively need surrogates that approximate well the objective f and that are easy to minimize. In this paper, we focus on "first-order surrogate functions" defined below, which will be shown to have "good" theoretical properties.

Definition 2.1 (**First-Order Surrogate**). *A function $g : \mathbb{R}^p \to \mathbb{R}$ is a first-order surrogate of f near κ in Θ when the following conditions are satisfied:*

- ***Majorization**: we have $g(\theta') \geq f(\theta')$ for all θ' in $\arg\min_{\theta \in \Theta} g(\theta)$. When the more general condition $g \geq f$ holds, we say that g is a **majorant** function;*
- ***Smoothness**: the approximation error $h \triangleq g - f$ is differentiable, and its gradient is L-Lipschitz continuous. Moreover, we have $h(\kappa) = 0$ and $\nabla h(\kappa) = 0$.*

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

First-order surrogates have a few simple properties, which form the building block of our analyses:

Lemma 2.1 (**Basic Properties - Key Lemma**). *Let g be in $\mathcal{S}_L(f, \kappa)$ for some κ in Θ. Define $h \triangleq g - f$ and let θ' be in $\arg\min_{\theta \in \Theta} g(\theta)$. 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 in $\mathcal{S}_{L,\rho}(f, \kappa)$, then, for all θ in Θ,

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

[2] Note that this concept differs from the machine learning terminology, where a "surrogate" often denotes a fixed convex upper bound of the nonconvex $(0-1)$-loss.

The proof of this lemma is relatively simple but for space limitation reasons, all proofs in this paper are provided as supplemental material. With Lemma 2.1 in hand, we now study the properties of Algorithm 1.

2.1. Convergence Analysis

For general non-convex problems, proving convergence to a global (or local) minimum is out of reach, and classical analyses study instead asymptotic stationary point conditions (see, e.g., Bertsekas, 1999). To do so, we make the mild assumption that for all θ, θ' in Θ, the directional derivative $\nabla f(\theta, \theta' - \theta)$ of f at θ in the direction $\theta' - \theta$ exists. A classical necessary first-order condition (see Borwein & Lewis, 2006) for θ to be a local minimum of f is to have $\nabla f(\theta, \theta' - \theta)$ non-negative for all θ' in Θ. This naturally leads us to consider the following asymptotic condition to assess the quality of a sequence $(\theta_n)_{n \geq 0}$ for non-convex problems:

Definition 2.2 (**Asymptotic Stationary Point**). *A sequence $(\theta_n)_{n \geq 0}$ satisfies an asymptotic stationary point condition if*

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

In particular, if f is differentiable on $\mathbb{R}^p$ and $\Theta = \mathbb{R}^p$, this condition implies $\lim_{n \to +\infty} \|\nabla f(\theta_n)\|_2 = 0$.

Building upon this definition, we now give a first convergence result about Algorithm 1.

Proposition 2.1 (**Non-Convex Analysis**). *Assume that the surrogates g_n from Algorithm 1 are in $\mathcal{S}_L(f, \theta_{n-1})$ and are majorant or strongly convex. Then, $(f(\theta_n))_{n \geq 0}$ monotonically decreases and $(\theta_n)_{n \geq 0}$ satisfies an asymptotic stationary point condition.*

Convergence results for non-convex problems are by nature weak. This is not the case when f is convex. In the next proposition, we obtain convergence rates by following a proof technique from Nesterov (2007) originally designed for proximal gradient methods.

Proposition 2.2 (**Convex Analysis for $\mathcal{S}_L(f, \kappa)$**). *Assume that f is convex and that for some $R > 0$,*

$$\|\theta - \theta^\star\|_2 \leq R \quad \text{for all } \theta \in \Theta \ \text{ s.t. } \ f(\theta) \leq f(\theta_0), \quad (1)$$

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

$$f(\theta_n) - f^\star \leq \frac{2LR^2}{n+2} \quad \text{for all } n \geq 1,$$

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

$$f(\theta_n) - f^\star \leq \beta^n (f(\theta_0) - f^\star) \quad \text{for all } n \geq 1,$$

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

The result of Proposition 2.2 is interesting in the sense that it provides sharp theoretical results without making strong assumption on the surrogate functions. The next proposition shows that slightly better rates can be obtained when the surrogates are strongly convex.

Proposition 2.3 (Convex Analysis for $\mathcal{S}_{L,\rho}(f,\kappa)$). *Assume that f is convex 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 $n \geq 1$,*

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

When f is μ-strongly convex, we have for all $n \geq 1$,

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

Note that the condition $\rho \geq L$ is relatively strong; it can indeed be shown that f is necessarily $(\rho - L)$-strongly convex if $\rho > L$, and convex if $\rho = L$. The fact that making stronger assumptions yields better convergence rates suggests that going beyond first-order surrogates could provide even sharper results. This is confirmed in the next proposition:

Proposition 2.4 (Second-Order Surrogates). *Make similar assumptions as in Proposition 2.2, and also assume that the error functions $h_n \triangleq g_n - f$ are twice differentiable, that their Hessians $\nabla^2 h_n$ are M-Lipschitz, and that $\nabla^2 h_n(\theta_{n-1}) = 0$ for all n. Then,*

$$f(\theta_n) - f^\star \leq \frac{9MR^3}{2(n+3)^2} \quad \text{for all } n \geq 1.$$

If f is μ-strongly convex, the convergence rate is superlinear with order 3/2.

Consistently with this proposition, similar rates were obtained by Nesterov & Polyak (2006) for the Newton method with cubic regularization, which involve second-order surrogates. In the next section, we focus again on first-order surrogates, and present simple mechanisms to build them. The proofs of the different claims are provided in the supplemental material.

2.2. Examples of Surrogate Functions

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

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

In addition, when f is convex, g is in $\mathcal{S}_{L,L}(f,\kappa)$, and when f is μ-strongly convex, g is in $\mathcal{S}_{L-\mu,L}(f,\kappa)$. Note also that 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 with a L-Lipschitz gradient. Then, f admits the following majorant surrogate in $\mathcal{S}_{2L}(f,\kappa)$:

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

The approximation error $g - f$ is indeed the same as in the previous paragraph 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 a proximal gradient step (see Nesterov, 2007; Beck & Teboulle, 2009).

DC Programming Surrogates.
Assume that $f = f_1 + f_2$, where f_2 is concave and differentiable with a L_2-Lipschitz gradient. Then, the following function g is a majorant surrogate in $\mathcal{S}_{L_2}(f,\kappa)$:

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

Such a surrogate forms the root of DC- (difference of convex functions)-programming (see Horst & Thoai, 1999). It is also indirectly used in reweighted-ℓ_1 algorithms (Candès et al., 2008) for minimizing on $\mathbb{R}_+^p$ a cost function of the form $\theta \mapsto f_1(\theta) + \lambda \sum_{i=1}^p \log(\theta_i + \varepsilon)$.

Variational Surrogates.
Let f be a real-valued function 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. Define $\tilde{f}$ as $\tilde{f}(\theta_1) \triangleq \min_{\theta_2 \in \Theta_2} f(\theta_1, \theta_2)$ and 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 for all θ_1 in $\mathbb{R}^{p_1}$;[3]
- $\theta_1 \mapsto \nabla_1 f(\theta_1, \theta_2)$ is L'-Lipschitz for all θ_2 in Θ_2;
- $\theta_2 \mapsto f(\theta_1, \theta_2)$ is μ-strongly convex for all θ_1 in $\mathbb{R}^{p_1}$.

Let us fix κ_1 in Θ_1. Then, the following function is a majorant surrogate in $\mathcal{S}_{2L''}(\tilde{f},\kappa)$ for some $L'' > 0$:

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

When f is jointly convex in θ_1 and θ_2, $\tilde{f}$ is itself convex and we can choose $L'' = L'$. Algorithm 1 becomes a block-coordinate descent procedure with two blocks.

Saddle Point Surrogates.
Let us make the same assumptions as in the previous paragraph but with the following differences:

[3] The notation ∇_1 denotes the gradient w.r.t. θ_1.

- $\theta_2 \mapsto f(\theta_1, \theta_2)$ is μ-strongly concave for all θ_1 in $\mathbb{R}^{p_1}$;
- $\theta_1 \mapsto f(\theta_1, \theta_2)$ is convex for all θ_2 in Θ_2;
- $\tilde{f}(\theta_1) \triangleq \max_{\theta_2 \in \Theta_2} f(\theta_1, \theta_2)$.

Then, $\tilde{f}$ is convex and the function below is a majorant surrogate in $\mathcal{S}_{2L''}(\tilde{f}, \kappa_1)$:

$$g : \theta_1 \mapsto f(\theta_1, \kappa_2^\star) + \frac{L''}{2}\|\theta_1 - \kappa_1\|_2^2,$$

where $L'' \triangleq \max(2L^2/\mu, L')$. When $\theta_1 \mapsto f(\theta_1, \theta_2)$ is affine, we can instead choose $L'' \triangleq L^2/\mu$.

Jensen Surrogates.
Jensen's inequality provides a natural mechanism to obtain surrogates for convex functions. Following the presentation of Lange et al. (2000), 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 differentiable with an L-Lipschitz gradient, and $\mathbf{w}_i \triangleq |\mathbf{x}_i|^\nu / \|\mathbf{x}\|_\nu^\nu$, then 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 et al. (2001) interpret boosting procedures under this point of view through the concept of *auxiliary functions*.

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 majorant 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 (Böhning & Lindsay, 1988; Khan et al., 2010).

We have shown that there are many rules to build first-order surrogates. Choosing one instead of another mainly depends on how easy it is to build the surrogate (do we need to estimate an a priori unknown Lipschitz constant?), and on how cheaply it can be minimized.

3. Block Coordinate Scheme

In this section, we introduce a block coordinate descent extension of Algorithm 1 under the assumptions that

- Θ is separable—that is, it can be written as a Cartesian product $\Theta = \Theta^1 \times \Theta^2 \times \ldots \times \Theta^k$;
- the surrogates g_n are separable into k components:

$$g_n(\theta) = \sum_{i=1}^k g_n^i(\theta^i) \quad \text{for} \quad \theta = (\theta^1, \ldots, \theta^k) \in \Theta.$$

We present a randomized procedure in Algorithm 2 following Tseng & Yun (2009); Shalev-Shwartz & Tewari (2009); Nesterov (2012); Richtárik & Takáč (2012).

Algorithm 2 Block Coordinate Descent Scheme

input $\theta_0 = (\theta_0^1, \ldots, \theta_0^k) \in \Theta = (\Theta^1 \times \ldots \times \Theta^k)$; N.
1: **for** $n = 1, \ldots, N$ **do**
2: Choose a separable surrogate g_n of f near θ_{n-1};
3: Randomly pick up one block $\hat{\imath}_n$ and update $\theta_n^{\hat{\imath}_n}$:

$$\theta_n^{\hat{\imath}_n} \in \operatorname*{arg\,min}_{\theta^{\hat{\imath}_n} \in \Theta^{\hat{\imath}_n}} g_n^{\hat{\imath}_n}(\theta^{\hat{\imath}_n}).$$

4: **end for**
output $\theta_N = (\theta_N^1, \ldots, \theta_N^k)$ (final estimate);

As before, we first study the convergence for non-convex problems. The next proposition shows that similar guarantees as for Algorithm 1 can be obtained.

Proposition 3.1 (Non-Convex Analysis). *Assume that the functions g_n are majorant surrogates in $\mathcal{S}_L(f, \theta_{n-1})$. Assume also that θ_0 is the minimizer of a majorant surrogate function in $\mathcal{S}_L(f, \theta_{-1})$ for some θ_{-1} in Θ. Then, the conclusions of Proposition 2.1 hold with probability one.*

Under convexity assumptions on f, the next two propositions give us expected convergence rates.

Proposition 3.2 (Convex Analysis for $\mathcal{S}_L(f, \kappa)$). *Make the same assumptions as in Proposition 2.2 and define $\delta \triangleq \frac{1}{k}$. When the surrogate functions g_n in Algorithm 2 are majorant and in $\mathcal{S}_L(f, \theta_{n-1})$, the sequence $(f(\theta_n))_{n \geq 0}$ almost surely converges to $f^\star$ and*

$$\mathbb{E}[f(\theta_n) - f^\star] \leq \frac{2LR^2}{2 + \delta(n - n_0)} \quad \text{for all } n \geq n_0,$$

where $n_0 \triangleq \left\lceil \log\left(\frac{2(f(\theta_0) - f^\star)}{LR^2} - 1\right) / \log\left(\frac{1}{1-\delta}\right) \right\rceil$ *if* $f(\theta_0) - f^\star > LR^2$ *and* $n_0 \triangleq 0$ *otherwise. Assume now that f is μ-strongly convex. Then, we have instead an expected linear convergence rate*

$$\mathbb{E}[f(\theta_n) - f^\star] \leq ((1 - \delta) + \delta\beta)^n (f(\theta_0) - f^\star),$$

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

Proposition 3.3 (Convex Analysis for $\mathcal{S}_{L,\rho}(f,\kappa)$). *Assume that f is convex. Define $\delta \triangleq \frac{1}{k}$. Choose majorant surrogates g_n in $\mathcal{S}_{L,\rho}(f,\theta_{n-1})$ with $\rho \geq L$, then $(f(\theta_n))_{n\geq 0}$ almost surely converges to $f^\star$ and we have*

$$\mathbb{E}[f(\theta_n) - f^\star] \leq \frac{C_0}{(1-\delta)+\delta n} \quad \text{for all} \quad n \geq 1,$$

with $C_0 \triangleq (1-\delta)(f(\theta_0)-f^\star)+\frac{(1-\delta)\rho+\delta L}{2}\|\theta_0-\theta^\star\|_2^2$. Assume now that f is μ-strongly convex, then we have an expected linear convergence rate

$$\left\{ \begin{array}{rcl} \frac{L}{2}\mathbb{E}[\|\theta^\star - \theta_n\|_2^2] & \leq & C_0 \left((1-\delta)+\delta\frac{L}{\rho+\mu}\right)^n \\ \mathbb{E}[f(\theta_n) - f^\star] & \leq & \frac{C_0}{\delta}\left((1-\delta)+\delta\frac{L}{\rho+\mu}\right)^{n-1} \end{array} \right. .$$

The quantity $\delta = 1/k$ represents the probability for a block to be updated during an iteration. Note that updating all blocks ($\delta = 1$) gives the same results as in Section 2. Linear convergence for strongly convex objectives with block coordinate descent is classical since the works of Tseng & Yun (2009); Nesterov (2012). Results of the same nature have also been obtained by Richtárik & Takáč (2012) for composite functions.

4. Frank-Wolfe Scheme

In this section, we show how to use surrogates to generalize the Frank-Wolfe method, an old convex optimization technique that has regained some popularity in machine learning (Zhang, 2003; Harchaoui et al., 2013; Hazan & Kale, 2012; Zhang et al., 2012). We present this approach in Algorithm 3.

Algorithm 3 Frank-Wolfe Scheme

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

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

2: Let g_n be a majorant surrogate in $\mathcal{S}_{L,L}(f,\theta_{n-1})$.

3: Compute a search direction:

$$\nu_n \in \arg\min_{\theta\in\Theta}\left[g_n(\theta) - \frac{L}{2}\|\theta-\theta_{n-1}\|_2^2\right].$$

4: Line search: $\alpha^\star \triangleq \arg\min_{\alpha\in[0,1]} g_n(\alpha\nu_n+(1-\alpha)\theta_{n-1})$.

5: Update solution: $\theta_n \triangleq \alpha^\star\nu_n + (1-\alpha^\star)\theta_{n-1}$.

6: **end for**

output θ_N (final estimate);

When f is smooth and the "gradient Lipschitz based surrogates" from Section 2.2 are used, Algorithm 3 becomes the classical Frank-Wolfe method.[4] Our point of view is however more general since it allows for example to use "proximal gradient surrogates". The next proposition gives a convergence rate.

[4]Note that the classical Frank-Wolfe algorithm performs in fact the line search over the function f and not g_n.

Proposition 4.1 (Convex Analysis). *Assume that f is convex and that Θ is bounded. Call $R \triangleq \max_{\theta_1,\theta_2\in\Theta}\|\theta_1-\theta_2\|_2$ the diameter of Θ. Then, the sequence $(f(\theta_n))_{n\geq 0}$ provided by Algorithm 3 converges to the minimum $f^\star$ of f over Θ and*

$$f(\theta_n) - f^\star \leq \frac{2LR^2}{n+2} \quad \text{for all } n \geq 1.$$

Other extensions of Algorithm 3 can also easily be designed by using our framework. We present for instance in the supplemental material a randomized block Frank-Wolfe algorithm, revisiting the recent work of Lacoste-Julien et al. (2013).

5. Accelerated Scheme

A popular scheme for convex optimization is the accelerated proximal gradient method (Nesterov, 2007; Beck & Teboulle, 2009). By using surrogate functions, we exploit similar ideas in Algorithm 4. When using the "Lipschitz gradient surrogates" of Section 2.2, Algorithm 4 is exactly the scheme 2.2.19 of Nesterov (2004). When using the "proximal gradient surrogate" and when $\mu = 0$, it is equivalent to the FISTA method of Beck & Teboulle (2009). Algorithm 4 consists of iteratively minimizing a surrogate computed at a point κ_{n-1} extrapolated from θ_{n-1} and θ_{n-2}. It results in better convergence rates, as shown in the next proposition by adapting a proof technique of Nesterov (2004).

Algorithm 4 Accelerated Scheme

input $\theta_0 \in \Theta$; N; μ (strong convexity parameter);

1: Initialization: $\kappa_0 \triangleq \theta_0$; $a_0 = 1$;

2: **for** $n = 1, \dots, N$ **do**

3: Choose a surrogate g_n in $\mathcal{S}_{L,L+\mu}(f,\kappa_{n-1})$;

4: Update solution: $\theta_n \triangleq \arg\min_{\theta\in\Theta} g_n(\theta)$;

5: Compute $a_n \geq 0$ such that:

$$a_n^2 = (1-a_n)a_{n-1}^2 + \tfrac{\mu}{L+\mu}a_n;$$

6: Set $\beta_n \triangleq \frac{a_{n-1}(1-a_{n-1})}{a_{n-1}^2+a_n}$ and update κ:

$$\kappa_n \triangleq \theta_n + \beta_n(\theta_n - \theta_{n-1});$$

7: **end for**

output θ_N (final estimate);

Proposition 5.1 (Convex Analysis). *Assume that f is convex. When $\mu = 0$, the sequence $(\theta_n)_{n\geq 0}$ provided by Algorithm 4 satisfies for all $n \geq 1$,*

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

When f is μ-strongly convex, we have instead a linear

convergence rate: for $n \geq 1$,

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

6. Incremental Scheme

This section is devoted to objective functions f that split into many components:

$$f(\theta) = \frac{1}{T}\sum_{t=1}^{T} f^t(\theta). \tag{2}$$

The most classical method exploiting such a structure when f is smooth is probably the stochastic gradient descent (SGD) and its variants (see Bottou, 2010). It consists of drawing at iteration n an index $\hat{t}_n$ and updating the solution as $\theta_n \leftarrow \theta_{n-1} - \eta_n \nabla f^{\hat{t}_n}(\theta_{n-1})$ with a scalar η_n. Another popular algorithm is the *stochastic mirror descent* (see Juditsky & Nemirovski, 2011) for general non-smooth convex problems, a setting we do not consider in this paper since non-smooth functions do not always admit first-order surrogates.

Recently, it was shown by Shalev-Schwartz & Zhang (2012) and Le Roux et al. (2012) that linear convergence rates could be obtained for strongly convex functions f^t. The SAG algorithm of Le Roux et al. (2012) for smooth unconstrained optimization is an approximate gradient descent strategy, where an estimate of ∇f is incrementally updated at each iteration. The work of Shalev-Schwartz & Zhang (2012) for composite optimization is a dual coordinate ascent method called SDCA which performs incremental updates in the primal (2). Unlike SGD, both SAG and SDCA require storing information about past iterates.

In a different context, incremental EM algorithms have been proposed by Neal & Hinton (1998), where surrogates of a log-likelihood are incrementally updated. By using similar ideas, we present in Algorithm 5 a scheme for solving (2), which we call MISO. In the next propositions, we study its convergence properties.

Algorithm 5 Incremental Scheme MISO

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

1: Choose 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 $t \neq \hat{t}_n$;
4: Update 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);

Proposition 6.1 (Non-Convex Analysis). *Assume that the surrogates* $g_n^{\hat{t}_n}$ *from Algorithm 5 are majorant and are in* $\mathcal{S}_L(f^{\hat{t}_n}, \theta_{n-1})$. *Then, the conclusions of Proposition 2.1 hold with probability one.*

Proposition 6.2 (Convex Analysis). *Assume that* f *is convex. Define* $f^\star \triangleq \min_{\theta\in\Theta} f(\theta)$ *and* $\delta \triangleq \frac{1}{T}$. *When the surrogates* g_n^t *in Algorithm 5 are majorant and in* $\mathcal{S}_{L,\rho}(f^t, \theta_{n-1})$ *with* $\rho \geq L$, *we have*

$$\mathbb{E}[f(\theta_n) - f^\star] \leq \frac{L\|\theta^\star - \theta_0\|_2^2}{2\delta n} \quad \textit{for all } n \geq 1.$$

Assume now that f *is* μ*-strongly convex. For all* $n \geq 1$,

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

Interestingly, the proof and the convergence rates of Proposition 6.2 are similar to those of the block coordinate scheme. For both schemes, the current iterate θ_n can be shown to be the minimizer of an approximate surrogate function which splits into different parts. Each iteration randomly picks up one part, and updates it. Like SAG or SDCA, we obtain linear convergence for strongly convex functions f, even though the upper bounds obtained for SAG and SDCA are better than ours.

It is also worth noticing that for smooth unconstrained problems, MISO and SAG yield different, but related, update rules. Assume for instance that "Lipschitz gradient surrogates" are used. At iteration n of MISO, each function g_n^t is a surrogate of f^t near some κ_{n-1}^t. The update rule of MISO can be shown to be $\theta_n \leftarrow \frac{1}{T}\sum_{t=1}^{T}\kappa_{n-1}^t - \frac{1}{TL}\sum_{t=1}^{T}\nabla f^t(\kappa_{n-1}^t)$; in comparison, the update rule of SAG is $\theta_n \leftarrow \theta_{n-1} - \frac{1}{TL}\sum_{t=1}^{T}\nabla f^t(\kappa_{n-1}^t)$.

The next section complements the theoretical analysis of the scheme MISO by numerical experiments and practical implementation heuristics.

7. Experiments

In this section, we show that MISO is efficient for solving large-scale machine learning problems.

7.1. Experimental Setting

We consider ℓ_2- and ℓ_1- logistic regression without intercept, and denote by m the number of samples and by p the number of features. The corresponding optimization problem can be written

$$\min_{\theta\in\mathbb{R}^p} \frac{1}{m}\sum_{t=1}^{m} \log(1 + e^{-y_t \mathbf{x}^{t\top}\theta}) + \lambda\psi(\theta), \tag{3}$$

where the regularizer ψ is either the ℓ_1- or squared ℓ_2-norm. The y_t's are in $\{-1,+1\}$ and the $\mathbf{x}^t$'s are vectors in $\mathbb{R}^p$ with unit ℓ_2-norm. We use four classical datasets described in the following table:

name	m	p	storage	size (GB)
alpha	250 000	500	dense	1
rcv1	781 265	47 152	sparse	0.95
covtype	581 012	54	dense	0.11
ocr	2 500 000	1 155	dense	23.1

Three datasets, alpha, rcv1 and ocr were obtained from the 2008 Pascal large scale learning challenge.[5] The dataset covtype is available from the LIBSVM website.[6] We have chosen to test several software packages including LIBLINEAR 1.93 (Fan et al., 2008), the ASGD and SGD implementations of L. Bottou (version 2)[7], an implementation of SAG kindly provided to us by the authors of Le Roux et al. (2012), the FISTA method of Beck & Teboulle (2009) implemented in the SPAMS toolbox[8], and SHOTGUN (Bradley et al., 2011). All these softwares are coded in C++ and were compiled using gcc. Experiments were run on a single core of a 2.00GHz Intel Xeon CPU E5-2650 using 64GB of RAM, and all computations were done in double precision. All the timings reported do not include data loading into memory. Note that we could not run the softwares SPAMS, LIBLINEAR and SHOTGUN on the dataset ocr because of index overflow issues.

7.2. On Implementing MISO

The objective function (3) splits into m components $f^t : \theta \mapsto \log(1+e^{-y_t \mathbf{x}^{t\top}\theta}) + \lambda\psi(\theta)$. It is thus natural to consider the incremental scheme of Section 6 together with the proximal gradient surrogates of Section 2.2. Concretely, we build at iteration n of MISO a surrogate $g_n^{\hat{t}_n}$ of $f^{\hat{t}_n}$ as follows: $g_n^{\hat{t}_n} : \theta \mapsto l^{\hat{t}_n}(\theta_{n-1}) + \nabla l^{\hat{t}_n}(\theta_{n-1})^\top(\theta-\theta_{n-1}) + \frac{L}{2}\|\theta-\theta_{n-1}\|_2^2 + \lambda\psi(\theta)$, where l^t is the logistic function $\theta \mapsto \log(1+e^{-y_t \mathbf{x}^{t\top}\theta})$.

After removing the dependency over n to simplify the notation, all the surrogates can be rewritten as $g^t : \theta \mapsto a_t + \mathbf{z}^{t\top}\theta + \frac{L}{2}\|\theta\|_2^2 + \lambda\psi(\theta)$, where a_t is a constant and $\mathbf{z}^t$ is a vector in $\mathbb{R}^p$. Therefore, all surrogates can be "summarized" by the pair $(a_t, \mathbf{z}^t)$, quantities which we keep into memory during the optimization. Then, finding the estimate θ_n amounts to minimizing a function of the form $\theta \mapsto \bar{\mathbf{z}}_n^\top\theta + \frac{L}{2}\|\theta\|_2^2 + \lambda\psi(\theta)$, where $\bar{\mathbf{z}}_n$ is the average value of the quantities $\mathbf{z}^t$ at iteration n. It is then easy to see that obtaining $\bar{\mathbf{z}}_{n+1}$ from $\bar{\mathbf{z}}_n$ can be done in $O(p)$ operations with the following update: $\bar{\mathbf{z}}_{n+1} \leftarrow \bar{\mathbf{z}}_n + (\mathbf{z}_{\text{new}}^{\hat{t}_n} - \mathbf{z}_{\text{old}}^{\hat{t}_n})/m$.

One issue is that building the surrogates g^t requires choosing some constant L. An upper bound on the Lipschitz constants of the gradients ∇l^t could be used here. However, we have observed that significantly faster convergence could be achieved by using a smaller value, probably because a local Lipschitz constant may be better adapted than a global one. By studying the proof of Proposition 6.2, we notice indeed that our convergence rates can be obtained without majorant surrogates, when we simply have: $\mathbb{E}[f^t(\theta_n)] \leq \mathbb{E}[g_n^t(\theta_n)]$ for all t and n. This motivates the following heuristics:

- MISO1: start by performing one pass over $\eta = 5\%$ of the data to select a constant L' yielding the smallest decrease of the objective, and set $L = L'\eta$;
- MISO2: in addition to MISO1, check the inequalities $f^{\hat{t}_n}(\theta_{n-1}) \leq g_{n-1}^{\hat{t}_n}(\theta_{n-1})$ during the optimization. After each pass over the data, if the rate of satisfied inequalities drops below 50%, double the value of L.

Following these strategies, we have implemented the scheme MISO in C++. The resulting software package will be publicly released with an open source license.

7.3. ℓ_2-Regularized Logistic Regression

We compare LIBLINEAR, FISTA, SAG, ASGD, SGD, MISO1, MISO2 and MISO2 with $T = 1000$ blocks (grouping some observations into minibatches). LIBLINEAR was run using the option -s 0 -e 0.000001. The implementation of SAG includes a heuristic line search in the same spirit as MISO2, introduced by Le Roux et al. (2012). Every method was stopped after 50 passes over the data. We considered three regularization regimes, high ($\lambda = 10^{-3}$), medium ($\lambda = 10^{-5}$) and low ($\lambda = 10^{-7}$). We present in Figure 1 the values of the objective function during the optimization for the regime medium, both in terms of passes over the data and training time. The regimes low and high are provided as supplemental material only. Note that to reduce the memory load, we used a minibatch strategy for the dataset rcv1 with $T = 10\,000$ blocks.

Overall, there is no clear winner from this experiment, and the preference for an algorithm depends on the dataset, the required precision, or the regularization level. The best methods seem to be consistently MISO, ASGD and SAG and the slowest one FISTA. Note that this apparently mixed result is a significant achievement. We have indeed focused on state-of-the-art solvers, which already significantly outperform a large number of other baselines (see Bottou, 2010; Fan et al., 2008; Le Roux et al., 2012).

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

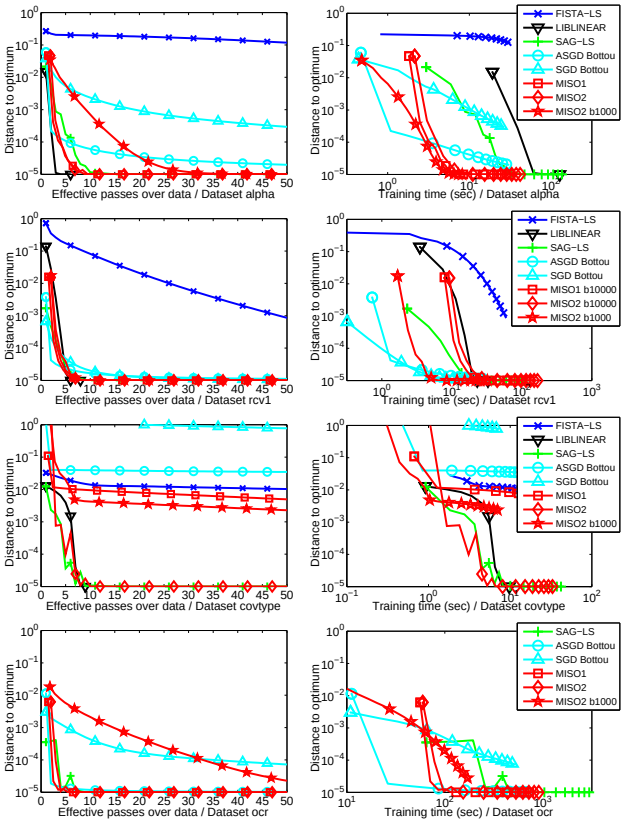

Figure 1. Results for ℓ_2-logistic regression with $\lambda = 10^{-5}$.

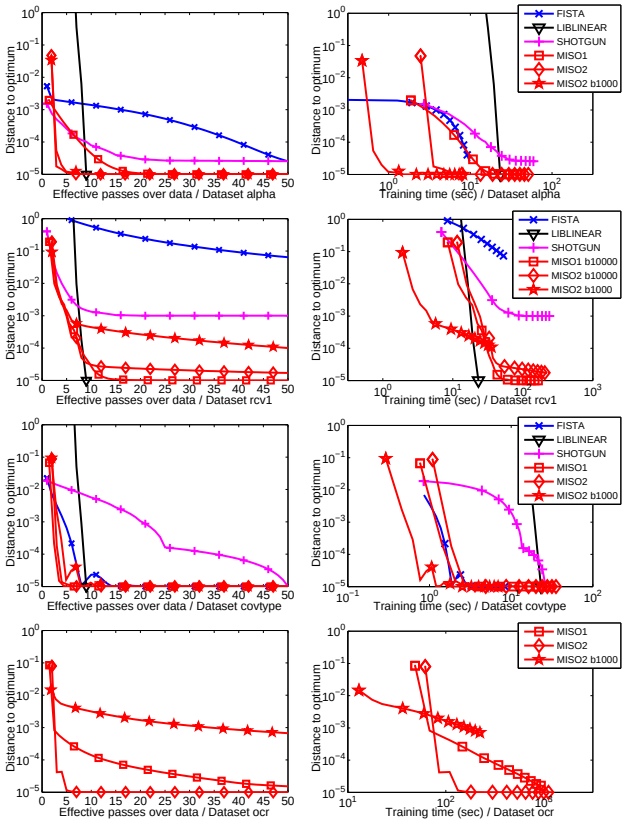

Figure 2. Benchmarks for ℓ_1-logistic regression. λ was chosen to obtain a solution with 10% nonzero coefficients.

7.4. ℓ_1-Regularized Logistic Regression

Since SAG, SGD and ASGD cannot deal with ℓ_1-regularization, we compare here LIBLINEAR, FISTA, SHOTGUN and MISO. We use for LIBLINEAR the option `-s 6 -e 0.000001`. We proceed as in Section 7.3, considering three regularization regimes yielding different sparsity levels. We report the results for one of them in Figure 2 and provide the rest as supplemental material. In this experiment, our method outperforms other competitors, except LIBLINEAR on the dataset `rcv1` when a high precision is required (and the regularization is low). We also remark that a low precision solution is often achieved quickly using the minibatch scheme (MISO2 b1000), but this strategy is outperformed by MISO1 and MISO2 for high precisions.

8. Conclusion

In this paper, we have introduced a flexible optimization framework based on the computation of "surrogate functions". We have revisited numerous schemes and discovered new ones. For each of them, we have studied convergence guarantees for non-convex problems and convergence rates for convex ones. Our methodology led us in particular to the design of an incremental algorithm, which has theoretical properties and empirical performance matching state-of-the-art solvers for large-scale machine learning problems.

In the future, we are planning to study fully stochastic or memoryless variants of our framework. As in the incremental setting, it consists of drawing a single training point at each iteration, but the algorithm does not keep track of all past information. This is essentially a strategy followed by Neal & Hinton (1998) and Mairal et al. (2010) in the respective contexts of EM and sparse coding algorithms. This would be particularly important for processing sparse datasets with a large number of features, where storing (dense) information about the past surrogates is cumbersome.

Acknowledgments

JM would like to thank Zaid Harchaoui, Francis Bach, Simon Lacoste-Julien, Mark Schmidt, Martin Jaggi, and Bin Yu for fruitful discussions. This work was supported by Quaero, (funded by OSEO, the French state agency for innovation), by the Gargantua project (program Mastodons - CNRS), and by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-0939370.

References

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

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

Böhning, D. and Lindsay, B. G. Monotonicity of quadratic-approximation algorithms. *Ann. I. Stat. Math.*, 40(4): 641–663, 1988.

Borwein, J.M. and Lewis, A.S. *Convex analysis and nonlinear optimization: theory and examples*. Springer, 2006.

Bottou, L. Large-scale machine learning with stochastic gradient descent. In *Proc. COMPSTAT*, 2010.

Bradley, J.K., Kyrola, A., Bickson, D., and Guestrin, C. Parallel coordinate descent for l1-regularized loss minimization. In *Proc. ICML*, 2011.

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

Collins, M., Schapire, R.E., and Singer, Y. Logistic regression, AdaBoost and Bregman distances. *Mach. Learn.*, 48(1):253–285, 2002.

Daubechies, I., Defrise, M., and De Mol, C. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Commun. Pur. Appl. Math.*, 57 (11):1413–1457, 2004.

Della Pietra, S., Della Pietra, V., and Lafferty, J. Duality and auxiliary functions for Bregman distances. Technical report, CMU-CS-01-109, 2001.

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

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

Harchaoui, Z., Juditsky, A., and Nemirovski, A. Conditional gradient algorithms for norm-regularized smooth convex optimization. *preprint arXiv:1302.2325v4*, 2013.

Hazan, E. and Kale, S. Projection-free online learning. In *Proc. ICML*, 2012.

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

Juditsky, A. and Nemirovski, A. First order methods for nonsmooth convex large-scale optimization, I: General purpose methods. In *Optimization for Machine Learning*. MIT Press, 2011.

Khan, E., Marlin, B., Bouchard, G., and Murphy, K. Variational bounds for mixed-data factor analysis. In *Adv. NIPS*, 2010.

Lacoste-Julien, S., Jaggi, M., Schmidt, M., and Pletscher, P. Block-coordinate Frank-Wolfe optimization for structural SVMs. In *Proc. ICML*, 2013.

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

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

Lee, D.D. and Seung, H.S. Algorithms for non-negative matrix factorization. In *Adv. NIPS*, 2001.

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

Neal, R.M. and Hinton, G.E. A view of the EM algorithm that justifies incremental, sparse, and other variants. *Learning in graphical models*, 89:355–368, 1998.

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

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

Nesterov, Y. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM J. Optimiz.*, 22 (2):341–362, 2012.

Nesterov, Y. and Polyak, B.T. Cubic regularization of Newton method and its global performance. *Math. Program.*, 108(1):177–205, 2006.

Richtárik, P. and Takáč, M. Iteration complexity of randomized block coordinate descent methods for minimizing a composite function. *Math. Program.*, 2012.

Seeger, M.W. and Wipf, D.P. Variational Bayesian inference techniques. *IEEE Signal Proc. Mag.*, 27(6):81–91, 2010.

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

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

Tseng, P. and Yun, S. A coordinate gradient descent method for nonsmooth separable minimization. *Math. Program.*, 117:387–423, 2009.

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

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

Zhang, T. Sequential greedy approximation for certain convex optimization problems. *IEEE T. Inform. Theory*, 49 (3):682–691, 2003.

Zhang, X., Yu, Y., and Schuurmans, D. Accelerated training for matrix-norm regularization: a boosting approach. In *Adv. NIPS*, 2012.

Supplementary Material
Optimization with First-Order Surrogate Functions

Outline. In Appendix A, we present simple mathematical definitions. Appendix B contains useful mathematical results, which are used in the paper. In Appendix C, we present various mechanisms to build first-order surrogate functions; it is in fact a more rigorous version of Section 2.2, where all claims are proved. In Appendix D, we present the block Frank-Wolfe optimization scheme. Finally, all proofs of propositions are given in Appendix E, and Appendix F contains additional experimental results.

A. Mathematical Background

For self-containedness purposes, we introduce in this section some mathematical definitions. Most of them can be found in classical textbooks on optimization (e.g., Bertsekas, 1999; Boyd & Vandenberghe, 2004; Borwein & Lewis, 2006; Nocedal & Wright, 2006; Nesterov, 2004).

Definition A.1 (**Directional Derivative**).
Let us consider a function $f : \Theta \subseteq \mathbb{R}^p \to \mathbb{R}$, where Θ is a convex set, and θ, θ' be in Θ. When it exists, the following limit is called the directional derivative of f at θ in the direction $\theta' - \theta$:

$$\nabla f(\theta, \theta' - \theta) \triangleq \lim_{t \to 0^+} \frac{f(\theta + t(\theta' - \theta)) - f(\theta)}{t}.$$

When f is differentiable at θ, directional derivatives always exist and we have $\nabla f(\theta, \theta' - \theta) = \nabla f(\theta)^\top (\theta' - \theta)$.

Definition A.2 (**Feasible Direction**).
Let $\Theta \subseteq \mathbb{R}^p$ be a convex set and θ be a point in Θ. A vector $\mathbf{z}$ in $\mathbb{R}^p$ is a feasible direction if $\theta + \mathbf{z}$ is in Θ. In other words, $\mathbf{z}$ can be written as $\theta' - \theta$, where θ' is in Θ.

Definition A.3 (**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 directional derivatives everywhere in Θ for every feasible direction. Let θ be a point in Θ. We say that θ is a stationary point if for all $\theta' \neq \theta$ in Θ,

$$\nabla f(\theta, \theta' - \theta) \geq 0. \tag{4}$$

When f is differentiable and θ is in the interior of Θ, this condition reduces to $\nabla f(\theta) = 0$. When f is convex and θ is also in the interior of Θ, this condition reduces to $0 \in \partial f(\theta)$, where ∂f is the subdifferential of f.

Proof. Let us assume that θ is a stationary point and f is differentiable at θ. Then, for all θ' in Θ, $\nabla f(\theta, \theta' - \theta) = \nabla f(\theta)^\top (\theta' - \theta) \geq 0$. In particular, since θ is in the interior of Θ, we can find θ' such that $\theta' - \theta = -\delta \nabla f(\theta)$ for some $\delta > 0$ small enough. Thus, we necessarily have $\nabla f(\theta) = 0$. The converse is trivial.

The equivalence between (4) and $0 \in \partial f(\theta)$ when f is convex but non-differentiable can be found in Borwein & Lewis (2006, Proposition 3.1.6). □

Definition A.4 (**Lipschitz Continuity**).
A function $f : \Theta \subseteq \mathbb{R}^p \to \mathbb{R}$ is called Lipschitz if there exists a constant $L > 0$ such that for all θ, θ' in Θ, we have

$$|f(\theta') - f(\theta)| \leq L \|\theta - \theta'\|_2.$$

In that case, we say that the function is L-Lipschitz.

Definition A.5 (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. This definition is equivalent to having for all α in $[0, 1]$ and θ, θ' in Θ,

$$f(\alpha\theta + (1-\alpha)\theta') \leq \alpha f(\theta) + (1-\alpha)f(\theta') - \frac{\mu}{2}\alpha(1-\alpha)\|\theta - \theta'\|_2^2. \tag{5}$$

Note that the value $\mu = 0$ leads to the classical definition of convex functions.

Proof. Let us consider θ' in Θ and define the function $g : \theta \mapsto f(\theta) - \frac{\mu}{2}\|\theta - \theta'\|_2^2$. This function is convex if and only if for all α in $[0, 1]$, we have

$$g(\alpha\theta + (1-\alpha)\theta') \leq \alpha g(\theta) + (1-\alpha)g(\theta').$$

In other words, if and only if

$$f(\alpha\theta + (1-\alpha)\theta') - \frac{\mu}{2}\alpha^2\|\theta - \theta'\|_2^2 \leq \alpha\left(f(\theta) - \frac{\mu}{2}\|\theta - \theta'\|_2^2\right) + (1-\alpha)f(\theta'),$$

which is equivalent to (5). □

B. Useful Mathematical Results

We provide in this section a few propositions and lemmas which are used in this paper.

Lemma B.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 $\mathbb{R}^p$,

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

Proof. This lemma is classical (see Nesterov, 2004, Lemma 1.2.3 and its proof). □

Note that Eq. (6) does not imply the gradient of a differentiable function f to be L-Lipschitz continuous. The equivalence is only true in some cases, as shown in the following lemma.

Lemma B.2 (Relation between Quadratic Surrogates and Lipschitz Constants).
Let $f : \mathbb{R}^p \to \mathbb{R}$ be a differentiable function. Assume that for all θ, θ' in $\mathbb{R}^p$, inequality (6) holds. Then, ∇f is L-Lipschitz continuous when one of the following conditions is true:

1. *f is convex;*
2. *f is twice differentiable;*
3. *∇f is Lipschitz continuous (the lemma then provides the Lipschitz constant L).*

Proof.
First point:
a proof of the first point can be found in Nesterov (2004, Theorem 2.1.5).

Second point:
To prove the second point, we upper-bound the extremal eigenvalues of the Hessian matrix. Let us fix θ in $\mathbb{R}^p$. Since f is twice differentiable at θ, we have for all θ' in $\mathbb{R}^p$

$$\nabla f(\theta') - \nabla f(\theta) = \nabla^2 f(\theta)(\theta' - \theta) + o(\|\theta' - \theta\|_2),$$

and thus

$$(\theta' - \theta)^\top(\nabla f(\theta') - \nabla f(\theta)) = (\theta' - \theta)^\top \nabla^2 f(\theta)(\theta' - \theta) + o(\|\theta' - \theta\|_2^2). \tag{7}$$

Summing twice Eq. (6) without the absolute values when exchanging the roles of θ and θ' gives

$$(\theta' - \theta)^\top (\nabla f(\theta') - \nabla f(\theta)) \leq L\|\theta - \theta'\|_2^2.$$

Plugging Eq. (7) into this inequality yields $\|\nabla^2 f(\theta)\|_2 \leq L$. To conclude, we use a mean value theorem, as done for example in Nesterov (2004, Lemma 1.2.2)

$$\begin{aligned}\|\nabla f(\theta') - \nabla f(\theta)\|_2 &= \left\| \int_{t=0}^1 \nabla^2 f(\theta + t(\theta' - \theta))(\theta' - \theta) dt \right\|_2 \\ &\leq \int_{t=0}^1 \|\nabla^2 f(\theta + t(\theta' - \theta))\|_2 dt \|\theta' - \theta\|_2 \\ &\leq L\|\theta - \theta'\|_2.\end{aligned}$$

Third point:
Proving the third point is more difficult due to the lack of smoothness assumptions on f. However, when making the explicit assumption that ∇f is Lipschitz continuous, we can show that Eq. (6) provides us a Lipschitz constant. The proof exploits some results from nonsmooth analysis developed by Clarke (1983). We essentially use a mean value theorem for multi-dimensional Lipschitz functions (Clarke, 1983, Proposition 2.6.5), exploiting the fact that a Lipschitz function is differentiable almost everywhere (Rademacher theorem). This allows us to follow a similar proof as for the twice differentiable case.

More precisely, we have that ∇f is Lipschitz continuous and thus differentiable almost everywhere on Θ. Let us call the Hessian matrix $\nabla^2 f(\theta)$ at a point θ in $\mathbb{R}^p$, when it exists. Then, we have at such a point $\|\nabla^2 f(\theta)\|_2 \leq L$, following the beginning of the second point's proof. Then, it turns out that for all θ in $\mathbb{R}^p$, the following mean value theorem holds for almost all θ' (see Clarke, 1983, proof of Proposition 2.6.5):

$$\nabla f(\theta') - \nabla f(\theta) = \int_{t=0}^1 \nabla^2 f(\theta + t(\theta' - \theta))(\theta' - \theta) dt.$$

This comes from the fact that for almost all θ', the intersection of the line segment $[\theta, \theta']$ and the set where $\nabla^2 f$ is not defined has 0 one-dimensional measure (see again Clarke, 1983, Proposition 2.6.5). We therefore have for almost all θ' (and a fixed θ), $\|\nabla f(\theta) - \nabla f(\theta')\|_2 \leq L\|\theta - \theta'\|_2$ and the general result comes from a continuity argument. □

Lemma B.3 (Surrogate for Functions with Lipschitz Hessian).
Let $f : \mathbb{R}^p \to \mathbb{R}$ be a twice differentiable function with M-Lipschitz continuous Hessian. Then for all θ, θ' in $\mathbb{R}^p$,

$$\left| f(\theta') - f(\theta) - \nabla f(\theta)^\top (\theta' - \theta) - \frac{1}{2}(\theta' - \theta)^\top \nabla^2 f(\theta)(\theta' - \theta) \right| \leq \frac{M}{6}\|\theta - \theta'\|_2^3.$$

Proof. This is again a classical lemma (see Nesterov, 2004, Lemma 1.2.4). □

Lemma B.4 (Lower Surrogate for Strongly Convex Functions).
Let $f : \mathbb{R}^p \to \mathbb{R}$ be a μ-strongly convex function. Suppose that f is differentiable, then the following inequality holds for all θ, θ' in $\mathbb{R}^p$:

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

Proof. $\theta' \mapsto f(\theta') - \frac{\mu}{2}\|\theta - \theta'\|_2^2$ is convex and differentiable and is therefore above its tangent at θ, immediately leading to the desired inequality. □

Lemma B.5 (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:

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

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 looking at first-order optimality conditions based on directional derivatives. For all θ in Θ, we have

$$\begin{aligned}\nabla g(\theta^\star, \theta - \theta^\star) &= \lim_{t \to 0^+} \frac{f(\theta^\star + t(\theta - \theta^\star)) - f(\theta^\star) - \frac{\mu t^2}{2}\|\theta - \theta^\star\|_2^2}{t} \\ &= \lim_{t \to 0^+} \frac{f(\theta^\star + t(\theta - \theta^\star)) - f(\theta^\star)}{t} = \nabla f(\theta^\star, \theta - \theta^\star) \geq 0,\end{aligned}$$

where $\nabla f(\theta^\star, \theta - \theta^\star)$ is non-negative because $\theta^\star$ is a stationary point of f on Θ. Thus, $\theta^\star$ is also a stationary point of the function g on Θ, and is a minimizer of g on Θ since g is convex (Borwein & Lewis, 2006, Proposition 2.1.2). This is sufficient to conclude. □

Lemma B.6 (**Lipschitz Continuity of Minimizers for Parameterized 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 $\mathbb{R}^{p_1}$.*

Then, the function $\theta_1 \mapsto \arg\min_{\theta_2 \in \Theta_2} f(\theta_1, \theta_2)$ is well defined and $\frac{L}{\mu}$-Lipschitz.

Proof. Let us consider θ_1, θ_1' in $\mathbb{R}^{p_1}$ and the corresponding (unique by strong convexity) solutions $\theta_2^\star \triangleq \arg\min_{\theta_2 \in \Theta_2} f(\theta_1, \theta_2)$ and $\theta_2'^\star \triangleq \arg\min_{\theta_2 \in \Theta_2} f(\theta_1', \theta_2)$. From the second-order growth condition of Lemma B.5, we have

$$\frac{\mu}{2}\|\theta_2^\star - \theta_2'^\star\|_2^2 \leq f(\theta_1, \theta_2'^\star) - f(\theta_1, \theta_2^\star),$$

and

$$\frac{\mu}{2}\|\theta_2^\star - \theta_2'^\star\|_2^2 \leq f(\theta_1', \theta_2^\star) - f(\theta_1', \theta_2'^\star),$$

Define the function $g : \kappa \mapsto f(\kappa, \theta_2'^\star) - f(\kappa, \theta_2^\star)$ and sum the above inequalities. We obtain

$$\mu\|\theta_2^\star - \theta_2'^\star\|_2^2 \leq g(\theta_1) - g(\theta_1').$$

We notice that the gradient of g is bounded: for all κ in $\mathbb{R}^{p_1}$, $\|\nabla g(\kappa)\|_2 = \|\nabla_1 f(\kappa, \theta_2'^\star) - \nabla_1 f(\kappa, \theta_2^\star)\| \leq L\|\theta_2'^\star - \theta_2^\star\|_2$. We use here the fact that $\nabla_1 f$ is L-Lipschitz with respect to its second argument. Thus, g is Lipschitz with constant $L\|\theta_2'^\star - \theta_2^\star\|_2$ and

$$\mu\|\theta_2^\star - \theta_2'^\star\|_2^2 \leq L\|\theta_2^\star - \theta_2'^\star\|_2\|\theta_1 - \theta_1'\|_2.$$

This is sufficient to conclude. □

Lemma B.7 (**Differentiability of Optimal Value Functions**). *Let us consider a function f defined as in Lemma B.6 and with the same properties. Define the optimal value function $\tilde{f}(\theta_1) \triangleq \min_{\theta_2 \in \Theta_2} f(\theta_1, \theta_2)$. 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,*

1. when f is convex and $\theta_1 \mapsto \nabla_1 f(\theta_1, \theta_2)$ is L'-Lipschitz continuous for all θ_2 in Θ_2, the function $\tilde{f}$ is convex and $\nabla \tilde{f}$ is Lipschitz continuous with constant L';

2. when $\theta_1 \mapsto f(\theta_1, \theta_2)$ is concave for all θ_2 in Θ_2, the function $\tilde{f}$ is concave and $\nabla \tilde{f}$ is Lipschitz continuous with constant $\frac{2L^2}{\mu}$;

3. when $\theta_1 \mapsto f(\theta_1, \theta_2)$ is affine for all θ_2 in Θ_2, the function $\tilde{f}$ is concave and $\nabla \tilde{f}$ is Lipschitz continuous with constant $\frac{L^2}{\mu}$.

Proof. Note that this lemma is a variant of a theorem introduced by Danskin (1967). We first prove the differentiability of f before detailing how to obtain the Lipschitz constants.

Differentiability of f:
Let us consider θ_1 and θ_1' in $\mathbb{R}^{p_1}$, and let us use the same notation and definitions as in the proof of Lemma B.6. Then, we have

$$\begin{aligned}\tilde{f}(\theta_1') - \tilde{f}(\theta_1) &= f(\theta_1', \theta_2'^\star) - f(\theta_1, \theta_2^\star) \\ &= f(\theta_1', \theta_2'^\star) - f(\theta_1', \theta_2^\star) + f(\theta_1', \theta_2^\star) - f(\theta_1, \theta_2^\star) \\ &= g(\theta_1') + f(\theta_1', \theta_2^\star) - f(\theta_1, \theta_2^\star) \\ &= g(\theta_1') + \nabla_1 f(\theta_1, \theta_2^\star)^\top (\theta_1' - \theta_1) + o(\|\theta_1' - \theta_1\|_2),\end{aligned} \tag{8}$$

where g is defined in the proof of Lemma B.6. Recall that the function g is Lipschitz with constant $L\|\theta_2'^\star - \theta_2^\star\|_2$ (see the proof of Lemma B.6). Thus,

$$|g(\theta_1')| \leq |g(\theta_1) - g(\theta_1')| \leq L\|\theta_2'^\star - \theta_2^\star\|_2 \|\theta_1' - \theta_1\|_2 \leq \frac{L^2}{\mu}\|\theta_1' - \theta_1\|_2^2, \tag{9}$$

where the first inequality uses the fact that $g(\theta_1') \leq 0$ and $g(\theta_1) \geq 0$. The last inequality uses Lemma B.6. We can now show that

$$\tilde{f}(\theta_1') = \tilde{f}(\theta_1) + \nabla_1 f(\theta_1, \theta_2^\star)^\top (\theta_1' - \theta_1) + o(\|\theta_1' - \theta_1\|_2).$$

The function $\tilde{f}$ thus admits a first-order Taylor expansion and is differentiable. Moreover, we have $\nabla \tilde{f}(\theta_1) = \nabla_1 f(\theta_1, \theta_2^\star)$.

Proof of the first point:
When f is jointly convex in θ_1 and θ_2, it is easy to show that $\tilde{f}$ is also convex (Boyd & Vandenberghe, 2004, Section 3.2.5).

By explicitly upper-bounding the quantity $o(\|\theta_1' - \theta_1\|_2)$ in Eq. (8) using the L'-Lipschitz continuity of $\nabla_1 f$ in its first argument and the inequality $g(\theta_1') \leq 0$, we have

$$0 \leq \tilde{f}(\theta_1') - \tilde{f}(\theta_1) - \nabla \tilde{f}(\theta_1)^\top (\theta_1' - \theta_1) \leq \frac{L'}{2}\|\theta_1' - \theta_1\|_2^2,$$

we can apply Lemma B.2 to ensure that $\nabla \tilde{f}$ is L'-Lipschitz continuous.

Proof of the second point:
$-\tilde{f}$ is a pointwise supremum of convex functions and is therefore convex (see Boyd & Vandenberghe, 2004, Section 3.2.3). Then, we have from Eq. (8) and using the concavity of $\theta_1 \mapsto f(\theta_1, \theta_2^\star)$:

$$\tilde{f}(\theta_1') - \tilde{f}(\theta_1) \geq g(\theta_1') + \nabla \tilde{f}(\theta_1)^\top (\theta_1' - \theta_1).$$

Thus,

$$0 \leq -\tilde{f}(\theta_1') + \tilde{f}(\theta_1) + \nabla \tilde{f}(\theta_1)^\top (\theta_1' - \theta_1) \leq |g(\theta_1')| \leq \frac{L^2}{\mu}\|\theta_1 - \theta_1'\|_2^2,$$

where the last inequality was shown in Eq. (9). We can then apply Lemma B.2 to the convex function $-\tilde{f}$ and we obtain the desired Lipschitz constant $\frac{2L^2}{\mu}$.

Proof of the third point:
When $\theta_1 \mapsto f(\theta_1, \theta_2)$ is affine, $\nabla_1 f(\theta_1, \theta_2)$ is independent of θ_1.

$$\begin{aligned}\|\nabla \tilde{f}(\theta_1') - \nabla \tilde{f}(\theta_1)\|_2 &= \|\nabla_1 f(\theta_1', \theta_2'^\star) - \nabla_1 \tilde{f}(\theta_1, \theta_2^\star)\|_2 \\ &= \|\nabla g(\theta_1)\|_2 \leq L\|\theta_2 - \theta_2^\star\|_2 \leq \frac{L^2}{\mu}\|\theta_1 - \theta_1'\|_2,\end{aligned}$$

where the upper-bound on the gradient of g was shown in the proof of Lemma B.6. □

Lemma B.8 (Pythagoras Relation).
Let θ, κ, ν in $\mathbb{R}^p$. Then

$$\|\kappa - \theta\|_2^2 + 2(\kappa - \theta)^\top(\theta - \nu) = \|\kappa - \nu\|_2^2 - \|\theta - \nu\|_2^2.$$

Lemma B.9 (Regularity of Residual Functions).
Let $f, g : \mathbb{R}^p \to \mathbb{R}$ be two functions. Define the difference function $h \triangleq g - f$. Then,

1. if g is ρ-strongly convex and f differentiable with L-Lipschitz continuous gradient, with $\rho \geq L$, the function h is $(\rho - L)$-strongly convex;

2. if g and f are convex and differentiable with L-Lipschitz continuous gradient, ∇h is L-Lipschitz continuous.

3. if g and f are μ-strongly convex and differentiable with L-Lipschitz continuous gradient, ∇h is $(L - \mu)$-Lipschitz continuous.

Proof.
Proof of the first point:
Let θ' be in $\mathbb{R}^p$, and define $l : \theta \mapsto g(\theta) - f(\theta) - \frac{(\rho - L)}{2}\|\theta - \theta'\|_2^2$. Then,

$$l(\theta) = \left(g(\theta) - \frac{\rho}{2}\|\theta - \theta'\|_2^2\right) + \left(\frac{L}{2}\|\theta - \theta'\|_2^2 - f(\theta)\right),$$

The left term inside parentheses is convex by definition of strong convexity. Let us call the right term $l' : \theta \mapsto \frac{L}{2}\|\theta - \theta'\|_2^2 - f(\theta)$. The function l' is differentiable and we can show that it is above its tangent, therefore convex. Let us indeed fix κ in $\mathbb{R}^p$:

$$\begin{aligned} l'(\theta) = \frac{L}{2}\|\theta - \theta'\|_2^2 - f(\theta) &\geq -f(\kappa) - \nabla f(\kappa)^\top(\theta - \kappa) - \frac{L}{2}\|\theta - \kappa\|_2^2 + \frac{L}{2}\|\theta - \theta'\|_2^2 \\ &= \left(\frac{L}{2}\|\kappa - \theta'\|_2^2 - f(\kappa)\right) + L(\kappa - \theta')^\top(\theta - \kappa) - \nabla f(\kappa)^\top(\theta - \kappa) \\ &= l'(\kappa) + \nabla l'(\kappa)^\top(\theta - \kappa). \end{aligned}$$

The first inequality comes from Lemma B.1 applied to the function f at κ. The second equality is simply due to the trivial relation described in Lemma B.8.

Proof of the second and third points:
We simply prove the third point, and then obtain the second point by choosing $\mu = 0$. We have for all θ and θ' in $\mathbb{R}^p$, according to Lemma B.1 and B.4

$$\frac{\mu}{2}\|\theta - \theta'\|_2^2 \leq f(\theta') - f(\theta) - \nabla f(\theta)^\top(\theta' - \theta) \leq \frac{L}{2}\|\theta - \theta'\|_2^2,$$

and

$$-\frac{L}{2}\|\theta - \theta'\|_2^2 \leq -g(\theta') + g(\theta) + \nabla g(\theta)^\top(\theta' - \theta) \leq -\frac{\mu}{2}\|\theta - \theta'\|_2^2.$$

Summing the two inequalities we have that

$$|h(\theta') - h(\theta) - \nabla h(\theta)^\top(\theta' - \theta)| \leq \frac{L - \mu}{2}\|\theta - \theta'\|_2^2.$$

where $h \triangleq g - f$. Since h is differentiable with a Lipschitz gradient, the result follows from Lemma B.2 (whether h is convex or not). □

C. Mechanisms to Construct First-Order Surrogate Functions

We provide here some details and justifications to Section 2.2. We start with a basic lemma, which gives us elementary techniques to combine surrogate functions.

Lemma C.1 (Combination Rules for Majorant First-Order Surrogates).
Let us consider two functions $f : \mathbb{R}^p \to \mathbb{R}$ and $f' : \mathbb{R}^p \to \mathbb{R}$, and majorant surrogate functions g in $\mathcal{S}_L(f, \kappa)$ and g' in $\mathcal{S}_L(f', \kappa)$ for some κ in Θ. Then, the following combination rules hold:

- ***Linear combination**: for all $\alpha, \beta > 0$, $\alpha g + \beta g'$ is a majorant surrogate function in $\mathcal{S}_{\alpha L + \beta L'}(\alpha f + \beta f', \kappa)$;*
- ***Transitivity**: consider g'' a majorant surrogate in $\mathcal{S}_{L''}(g, \kappa)$. Then, g'' is a majorant surrogate in $\mathcal{S}_{L+L''}(f, \kappa)$;*
- ***Negation**: the function $g'' : \theta \mapsto -g(\theta) + \frac{L}{2}\|\theta - \kappa\|_2^2$ is a majorant surrogate in $\mathcal{S}_{2L}(-f, \kappa)$.*

Proof. The first two points are easy to check. For the last one, we have for all θ in Θ, $g(\theta) - f(\theta) \leq \frac{L}{2}\|\theta - \kappa\|_2^2$ according to Lemma 2.1. The proposed surrogate is therefore majorant for $-f$. We can now define the approximation error function $h'' : \theta \mapsto f(\theta) - g(\theta) + \frac{L}{2}\|\theta - \kappa\|_2^2$, which is differentiable with $2L$-Lipschitz continuous gradient and g'' is in $\mathcal{S}_{2L}(-f, \kappa)$ (we have used the fact that $\theta \mapsto \frac{L}{2}\|\theta - \kappa\|_2^2$ and $h \triangleq g - f$ are both differentiable and their gradients are L-Lipschitz). □

In the next paragraphs, we justify the different surrogates we have introduced in Section 2.2.

Lipschitz Gradient Surrogates.
When f is differentiable and ∇f is L-Lipschitz, we consider the following surrogate:

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

By applying Lemma B.1 and studying the approximation error $h \triangleq g - f$, we immediately obtain that g is a majorant surrogate in $\mathcal{S}_{2L,L}(f, \kappa)$. When f is convex, we can use Lemma B.9 to prove that g is in $\mathcal{S}_{L,L}(f, \kappa)$ and $\mathcal{S}_{L-\mu,L}(f, \kappa)$ when f is μ-strongly convex.

Proximal Gradient Surrogates.
Assume that f splits into $f = f_1 + f_2$, where f_1 is differentiable with a L-Lipschitz gradient. Then, we have presented the following surrogate

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

Following the same arguments as in the previous paragraph, we have that g is in $\mathcal{S}_{2L}(f, \kappa)$. Moreover, 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 $\mathcal{S}_{L-\mu,L}(f, \kappa)$.

DC Programming Surrogates.
Assume that $f = f_1 + f_2$, where f_2 is concave and differentiable with a L_2-Lipschitz gradient. Then, we have presented the following surrogate

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

It is easy to see that g is a majorant surrogate since f_2 is concave and below its tangents. It is also easy to see that the approximation error $g - f$ has a L_2-Lipschitz continuous gradient.

Variational Surrogates.
Let f be a function 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. Define $\tilde{f}$ as $\tilde{f}(\theta_1) \triangleq \min_{\theta_2 \in \Theta_2} f(\theta_1, \theta_2)$ and 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 for all θ_1 in $\mathbb{R}^{p_1}$;
- $\theta_1 \mapsto \nabla_1 f(\theta_1, \theta_2)$ is L'-Lipschitz for all θ_2 in Θ_2;

- $\theta_2 \mapsto f(\theta_1, \theta_2)$ is μ-strongly convex for all θ_1 in $\mathbb{R}^{p_1}$.

Let us fix κ_1 in Θ_1. Then, we can show that the following function is a majorant surrogate in $\mathcal{S}_{L''}(\tilde{f}, \kappa)$ for some $L'' > 0$:

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

We can indeed apply Lemma B.7 which ensures that $\tilde{f}$ is differentiable with $\nabla \tilde{f}(\theta_1) = \nabla_1 f(\theta_1, \theta_2^\star)$ and $\theta_2^\star \triangleq \arg\min f(\theta_1, \theta_2)$. Considering the approximation error function $h \triangleq g - \tilde{f}$, we indeed have that $h(\kappa_1) = 0$, $\nabla h(\kappa_1) = 0$ and since $\theta_2^\star$ as a function of θ_1 is Lipschitz according to Lemma B.6, we also have that ∇h is Lipschitz continuous.

When f is jointly convex in θ_1 and θ_2, $\tilde{f}$ is itself convex and $\nabla \tilde{f}$ is L'-Lipschitz continuous according to Lemma B.7. We can then apply Lemma B.9 to obtain that ∇h is L'-Lipschitz continuous such that we can choose $L'' = L'$.

Saddle Point Surrogates.
Let us make the same assumptions as in the previous paragraph with the following exceptions

- $\theta_2 \mapsto f(\theta_1, \theta_2)$ is μ-strongly concave for all θ_1 in $\mathbb{R}^{p_1}$;
- $\theta_1 \mapsto f(\theta_1, \theta_2)$ is convex for all θ_2 in Θ_2;
- $\tilde{f}(\theta_1) \triangleq \max_{\theta_2 \in \Theta_2} f(\theta_1, \theta_2)$.

Then, $\tilde{f}$ is convex as the pointwise supremum of convex functions (see Boyd & Vandenberghe, 2004) and we can show that the function below is a majorant surrogate in $\mathcal{S}_{2L''}(\tilde{f}, \kappa_1)$:

$$g : \theta_1 \mapsto f(\theta_1, \kappa_2^\star) + \frac{L''}{2} \|\theta_1 - \kappa_1\|_2^2,$$

where $L'' \triangleq \max(2L^2/\mu, L')$. When $\theta_1 \mapsto f(\theta_1, \theta_2)$ is affine, we can instead choose $L'' \triangleq L^2/\mu$.

We indeed apply the same methodology as in the previous paragraph. Lemma B.7 tells us that the function $-\tilde{f}$ is differentiable with $2L^2/\mu$-Lipschitz continuous gradient (only L^2/μ in the affine case). Then, we have that the function $\theta_1 \mapsto -f(\theta_1, \kappa_2^\star)$ is in $\mathcal{S}_{L''}(-\tilde{f}, \kappa_1)$ by using Lemma B.9. We then apply the negation rule of Lemma C.1 to conclude.

Jensen Surrogates.
Let us recall the definition of Jensen surrogates. Following Lange et al. (2000), 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 $\mathbf{w}$ in $\mathbb{R}_+^p$ be a weight vector such that $\mathbf{w} \geq 0$, $\|\mathbf{w}\|_1 = 1$ and $\mathbf{w}_i \neq 0$ whenever $\mathbf{x}_i \neq 0$. Then, we consider the following function g for any κ in $\mathbb{R}^p$

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

Assume that f is differentiable with a L-Lipschitz gradient and $\mathbf{w}_i \triangleq |\mathbf{x}_i|^\nu / \|\mathbf{x}\|_\nu^\nu$ for some $\nu \geq 0$.[9] ∇f is obviously Lipschitz with constant $L\|\mathbf{x}\|_2^2$. g is also convex, differentiable with Lipschitz continuous gradient with constant L' obtained below with simple calculations:

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

The fact that g is majorant is a simple application of Jensen inequality. It is also obvious that $g(\kappa) = f(\kappa)$ and that $\nabla g(\kappa) = \nabla f(\kappa)$. We now apply Lemma B.9, noticing that we always have L' greater than $L\|\mathbf{x}\|_2^2$, and we have that g is in $\mathcal{S}_{L'}(\tilde{f}, \kappa)$.

[9] With an abuse of notation, $\|\mathbf{x}\|_0^0$ denotes the ℓ_0-pseudo norm, also denoted by $\|\mathbf{x}\|_0$.

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

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

The fact that it is majorant is simply an application of the mean-value theorem.

D. Additional Optimization Scheme: Block Frank-Wolfe

We provide in this section an additional optimization scheme, combining the ideas of Sections 4 and 3 with separability assumptions on the surrogates g_n and Θ. It results in a block coordinate version of the Frank-Wolfe optimization scheme presented in Algorithm 6 generalizing a procedure recently introduced by Lacoste-Julien et al. (2013). More precisely, the algorithm of Lacoste-Julien et al. (2013) corresponds to using a quadratic surrogate as provided by Lemma B.1 when f is smooth with L-Lipschitz gradient, and performing a line search on the function f instead of g_n. Our approach on the other hand can afford to have a non-smooth component in f and in that sense is more general. Note that Lacoste-Julien et al. (2013) also presents duality gap guarantees and various extensions and applications, which we do not consider in our paper.

Algorithm 6 Block Frank-Wolfe Scheme

input $\theta_0 = (\theta_0^1, \dots, \theta_0^k) \in \Theta = \Theta^1 \times \dots \times \Theta^k$ (initial point); N (number of iterations).

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

2: Compute a separable majorant surrogate function $g_n = \sum_{i=1}^k g_n^i$ in $\mathcal{S}_{L,L}(f, \theta_{n-1})$;

3: Randomly pick one block $\hat{\imath}_n$ in $\{1, \dots, k\}$ and compute a search direction:

$$\nu_n^{\hat{\imath}_n} \in \operatorname*{arg\,min}_{\theta^{\hat{\imath}_n} \in \Theta^{\hat{\imath}_n}} \left[g_n^{\hat{\imath}_n}(\theta^{\hat{\imath}_n}) - \frac{L}{2} \|\theta^{\hat{\imath}_n} - \theta_{n-1}^{\hat{\imath}_n}\|_2^2 \right].$$

4: Line search:

$$\alpha^\star \triangleq \operatorname*{arg\,min}_{\alpha \in [0,1]} g_n^{\hat{\imath}_n}((1-\alpha)\theta_{n-1}^{\hat{\imath}_n} + \alpha \nu_n^{\hat{\imath}_n}).$$

5: Update $\theta_n^{\hat{\imath}_n}$:

$$\theta_n^{\hat{\imath}_n} \triangleq (1 - \alpha^\star)\theta_{n-1}^{\hat{\imath}_n} + \alpha^\star \nu_n^{\hat{\imath}_n}.$$

6: **end for**

output $\theta_N = (\theta_N^1, \dots, \theta_N^k)$ (final estimate);

Proposition D.1 (Convergence Rate for Algorithm 6).
Let f be convex, bounded below and $f^\star$ be the minimum of f on $\Theta = \Theta^1 \times \dots \times \Theta^k$. Assume that Θ is bounded and call $R \triangleq \max_{\theta_1, \theta_2 \in \Theta} \|\theta_1 - \theta_2\|_2$ its diameter. The sequence $(f(\theta_n))_{n \geq 0}$ provided by Algorithm 6 converges almost surely to $f^\star$ and we have for all $n \geq 1$,

$$\mathbb{E}[f(\theta_n) - f^\star] \leq \frac{2LR^2}{2 + \delta(n - n_0)},$$

where $\delta \triangleq 1/k$ and $n_0 \triangleq \left\lceil \log\left(\frac{2(f(\theta_0) - f^\star)}{LR^2} - 1\right) / \log\left(\frac{1}{1-\delta}\right) \right\rceil$ if $f(\theta_0) - f^\star > LR^2$ and $n_0 \triangleq 0$ otherwise.

The proof is given in Appendix E.

E. Proofs of Lemmas and Propositions

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

E.1. Proof of Lemma 2.1

Proof. The first inequality is a direct applications of Lemma B.1 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 can in addition exploit the second-order growth property of g presented in Lemma B.5, 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 second one. □

E.2. Proof of Proposition 2.1

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

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

where the first inequality and the last equality come from Definition 2.1. The second inequality comes from the definition of θ_n. Denote by $f^\star$ the limit of the sequence $(f(\theta_n))_{n\geq 0}$ and by $h_n \triangleq g_n - f$ the approximation error functions. The latter are differentiable and their gradient are L-Lipschitz continuous according to the definitions of the surrogate functions. Then,

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

and thus, by summing over n,

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

and the non-negative sequence $(h_n(\theta_n))_{n\geq 0}$ necessarily converges to zero.

We have then two possibilities (according to the assumptions made in the proposition):

- if the g_n's are majorant surrogates, plugging $\theta' = \theta_n - \frac{1}{L}\nabla h_n(\theta_n)$ in Lemma B.1 yields

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

and therefore,

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

where we use the fact that $h_n(\theta') \geq 0$ because g_n is majorant.

- otherwise, the functions g_n are ρ-strongly convex and we can exploit some inequalities of Lemma 2.1. Notably,

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

Summing this inequality over n yields that $\|\theta_n - \theta_{n-1}\|_2^2$ necessarily 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 2.1.

We can now compute directional derivatives of f at a point θ_n and a direction $\theta - \theta_n$, where 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,

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

using Cauchy-Schwarz inequality. Then,

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

□

E.3. Proof of Proposition 2.2

Proof. We separately prove the two parts of the proposition.

Non-strongly convex case:

Let us define $h_n \triangleq g_n - f$ the approximation error function at iteration n. From Lemma 2.1 (with $g = g_n, \kappa = \theta_{n-1}, \theta' = \theta_n$), 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 (2007, Theorem 4), we have

$$\begin{aligned} f(\theta_n) &\leq \min_{\alpha \in [0,1]} f(\alpha \theta^\star + (1-\alpha)\theta_{n-1}) + \frac{L\alpha^2}{2} \|\theta^\star - \theta_{n-1}\|_2^2, \\ &\leq \min_{\alpha \in [0,1]} \alpha f(\theta^\star) + (1-\alpha) f(\theta_{n-1}) + \frac{L\alpha^2}{2} \|\theta^\star - \theta_{n-1}\|_2^2, \end{aligned} \tag{10}$$

where the minimization over Θ in the previous equation is replaced by a minimization on 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, which yields

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

- if $f(\theta_{n-1}) - f^\star \geq LR^2$, then we have the optimal value $\alpha^\star = 1$ and $f(\theta_n) - f^\star \leq \frac{LR^2}{2}$;
- otherwise $\alpha^\star = \frac{f(\theta_{n-1}) - f^\star}{LR^2}$. Denoting by $r_n \triangleq f(\theta_n) - f^\star$, we have

$$r_n \leq r_{n-1} \left(1 - \frac{r_{n-1}}{2LR^2}\right).$$

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.

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 Nesterov (2007) for computing the convergence rate of proximal gradient methods. We start from (10). We can then use the second-order growth property of f (Lemma B.5) which states that $f(\theta_{n-1}) \geq f^\star + \frac{\mu}{2}\|\theta_{n-1} - \theta^\star\|_2^2$ and 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 $\mu \geq 2L$, then the previous binomial is minimized for $\alpha^\star = 1$ and

$$f(\theta_n) - f^\star \leq \frac{L}{\mu}(f(\theta_{n-1}) - f^\star);$$

- if $\mu \leq 2L$, then we have $\alpha^\star = \frac{\mu}{2L}$ and

$$f(\theta_n) - f^\star \leq \left(1 - \frac{\mu}{4L}\right)(f(\theta_{n-1}) - f^\star),$$

which is sufficient to conclude.

□

E.4. Proof of Proposition 2.3

Proof. We separately prove the two parts of the proposition.

Non-strongly convex case:
From Lemma 2.1 (with $g{=}g_n$, $\kappa{=}\theta_{n-1}$, $\theta'{=}\theta_n$, $\theta{=}\theta^\star$), we have

$$f(\theta_n) - f(\theta^\star) \leq \frac{L}{2}\|\theta_{n-1} - \theta^\star\|_2^2 - \frac{\rho}{2}\|\theta_n - \theta^\star\|_2^2 \leq \frac{L}{2}\|\theta_{n-1} - \theta^\star\|_2^2 - \frac{L}{2}\|\theta_n - \theta^\star\|_2^2. \tag{11}$$

By summing this inequality, 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 (2.3). Note that finding telescopic sums to prove convergence rates is a classical technique (see Beck & Teboulle, 2009).

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 B.5: $f(\theta_n) - f^\star \geq \frac{\mu}{2}\|\theta_n - \theta^\star\|_2^2$ for all n. Combined with (11), 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

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

□

E.5. Proof of Proposition 2.4

Proof. We separately prove the two parts of the proposition.

Non-strongly convex case:
Following a similar scheme as in Proposition 2.2 and using Lemma B.3 on the approximation error functions h_n instead of Lemma 2.1, we have

$$f(\theta_n) \leq \min_{\alpha \in [0,1]} f(\alpha\theta^\star + (1-\alpha)\theta_{n-1}) + \frac{M\alpha^3}{6}\|\theta^\star - \theta_{n-1}\|_2^3, \tag{12}$$

Then, again following the proof of Proposition 2.2,

$$f(\theta_n) - f^\star \leq \min_{\alpha \in [0,1]} (1-\alpha)(f(\theta_{n-1}) - f^\star) + \frac{\alpha^3 M R^3}{6}.$$

Denoting by $r_n \triangleq f(\theta_n) - f^\star$ and by $\alpha^\star$ the solution of this optimization problem, we have

- if $r_{n-1} \geq MR^3/2$, then $\alpha^\star = 1$ and $r_n \leq MR^3/6$;
- otherwise, $\alpha^\star = \sqrt{2r_{n-1}/(MR^3)}$, and

$$r_n \leq r_{n-1}\left(1 - \sqrt{\frac{8r_{n-1}}{9MR^3}}\right).$$

It follows that $r_n^{-1/2} \geq r_{n-1}^{-1/2}\left(1 - \sqrt{\frac{8r_{n-1}}{9MR^3}}\right)^{-1/2} \geq r_{n-1}^{-1/2} + \sqrt{\frac{2}{9MR^3}}$, where the last inequality comes from the convexity inequality $(1-x)^{-1/2} \geq 1 + x/2$.

Then, we have seen that if $r_0 \geq MR^3/2$, then $r_1 \leq MR^3/6$ and thus $r_n^{-1/2} \geq r_1^{-1/2} + (n-1)\sqrt{\frac{2}{9MR^3}} \geq \sqrt{\frac{2}{9MR^3}}(n - 1 + 3\sqrt{3}) \geq \sqrt{\frac{2}{9MR^3}}(n+4)$. Otherwise, we have $r_n^{-1/2} \geq r_0^{-1/2} + n\sqrt{\frac{2}{9MR^3}} \geq \sqrt{\frac{2}{9MR^3}}(n+3)$. This is sufficient to obtain 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. Starting again from (12),

$$f(\theta_n) \leq \min_{\alpha \in [0,1]} f(\alpha\theta^\star + (1-\alpha)\theta_{n-1}) + \frac{M\alpha^3}{6}\|\theta^\star - \theta_{n-1}\|_2^3,$$

and using the second-order growth property of f, we have

$$r_n \leq \min_{\alpha \in [0,1]} (1-\alpha)r_{n-1} + \frac{\gamma\alpha^3}{3} r_{n-1}^{3/2}.$$

- when $\gamma\sqrt{r_{n-1}} \geq 1$, we have $\alpha^\star = \frac{1}{\sqrt{\gamma\sqrt{r_{n-1}}}}$ and the desired inequality follows;
- otherwise, $\alpha^\star = 1$ and $r_n \leq \frac{3}{2} r_{n-1}^{3/2} \leq \frac{r_{n-1}}{3}$.

□

E.6. Proof of Proposition 3.1

Proof. We proceed in several steps and adapt the convergence proof of Proposition 2.1 to our new setting.

Definition of an approximate surrogate $\bar{g}_n$:
We define recursively the sequence of functions $(\bar{g}_n)_{n\geq 0}$ as follows:

$$\bar{g}_n \triangleq \bar{g}_{n-1} + g_n^{\hat{\imath}_n} - \bar{g}_{n-1}^{\hat{\imath}_n},$$

where the surrogate g_n and the index $\hat{\imath}_n$ are chosen in the algorithm. We also define $\bar{g}_{-1}$ as a majorant separable surrogate function such that $\theta_0 \in \arg\min_{\theta\in\Theta} \bar{g}_{-1}(\theta)$ (we have assumed in the proposition that such a surrogate function exists). Then, it is easy to see that $\bar{g}_n$ is constructed in such a way that θ_n is a minimizer of $\bar{g}_n$ over Θ for all $n \geq 0$ and that $\bar{g}_n \geq f$.

Almost sure convergence of $(f(\theta_n))_{n\geq 0}$ and consequences:
We have $f(\theta_n) \leq g_n(\theta_n) = \sum_{i=1}^k g_n^k(\theta_n^k) \leq \sum_{i=1}^k g_n^k(\theta_{n-1}^k) = g_n(\theta_{n-1}) = f(\theta_{n-1})$ since we have $g_n^{\hat{\imath}_n}(\theta_n^{\hat{\imath}_n}) \leq g_n^{\hat{\imath}_n}(\theta_{n-1}^{\hat{\imath}_n})$ and $g_n^i(\theta_n^i) = g_n^i(\theta_{n-1}^i)$ for $i \neq \hat{\imath}_n$. Thus, $(f(\theta_n))_{n\geq 0}$ is monotonically decreasing and converges almost

surely. We also have

$$\begin{aligned}\mathbb{E}[\bar{g}_n(\theta_n) - \bar{g}_{n-1}(\theta_{n-1})] &= \mathbb{E}[\bar{g}_n(\theta_n) - \bar{g}_n(\theta_{n-1})] + \mathbb{E}[\bar{g}_n(\theta_{n-1}) - \bar{g}_{n-1}(\theta_{n-1})] \\ &= \mathbb{E}[\bar{g}_n(\theta_n) - \bar{g}_n(\theta_{n-1})] + \mathbb{E}[g_n^{\hat{\imath}_n}(\theta_{n-1}^{\hat{\imath}_n}) - \bar{g}_{n-1}^{\hat{\imath}_n}(\theta_{n-1}^{\hat{\imath}_n})] \\ &= \mathbb{E}[\bar{g}_n(\theta_n) - \bar{g}_n(\theta_{n-1})] + \mathbb{E}[\mathbb{E}[g_n^{\hat{\imath}_n}(\theta_{n-1}^{\hat{\imath}_n}) - \bar{g}_{n-1}^{\hat{\imath}_n}(\theta_{n-1}^{\hat{\imath}_n})|\theta_{n-1}]] \\ &= \mathbb{E}[\bar{g}_n(\theta_n) - \bar{g}_n(\theta_{n-1})] + \frac{1}{k}\mathbb{E}[g_n(\theta_{n-1}) - \bar{g}_{n-1}(\theta_{n-1})] \\ &= \mathbb{E}[\bar{g}_n(\theta_n) - \bar{g}_n(\theta_{n-1})] + \frac{1}{k}\mathbb{E}[f(\theta_{n-1}) - \bar{g}_{n-1}(\theta_{n-1})].\end{aligned}$$

Note that both terms $\bar{g}_n(\theta_n) - \bar{g}_n(\theta_{n-1})$ and $f(\theta_{n-1}) - \bar{g}_{n-1}(\theta_{n-1})$ are non-positive with probability one and thus the sequence $(\mathbb{E}[\bar{g}_n(\theta_n)])_{n\geq 0}$ is non-increasing, bounded below and convergent. The term $\mathbb{E}[\bar{g}_n(\theta_n) - \bar{g}_{n-1}(\theta_{n-1})]$ is therefore the summand of a converging sum, and so are $\mathbb{E}[\bar{g}_n(\theta_n) - \bar{g}_n(\theta_{n-1})]$ and $\mathbb{E}[f(\theta_{n-1}) - \bar{g}_{n-1}(\theta_{n-1})]$. Then, we have by using Beppo-Levi theorem

$$\sum_{n=0}^{+\infty} \mathbb{E}[\bar{g}_n(\theta_n) - f(\theta_n)] = \mathbb{E}\left[\sum_{n=0}^{+\infty} \bar{g}_n(\theta_n) - f(\theta_n)\right] < +\infty.$$

Thus, the term $\bar{g}_n(\theta_n) - f(\theta_n)$ converges almost surely to 0.

Asymptotic stationary point conditions:
Let us denote by $\bar{h}_n \triangleq \bar{g}_n - f$ which is differentiable with L-Lipschitz continuous gradient. Then, for all θ in Θ,

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

We have $\nabla \bar{g}_n(\theta_n, \theta - \theta_n) \geq 0$ since θ_n is a minimizer of $\bar{g}_n$, and $\|\nabla \bar{h}_n(\theta_n)\|_2^2 \leq 2L\bar{h}_n(\theta_n)$, following a similar argument as in the proof of Proposition 2.1. Since we have shown that $\bar{h}_n(\theta_n)$ almost surely converges to zero, we conclude using the Cauchy-Schwarz inequality as in the proof of Proposition 2.1.

□

E.7. Proof of Proposition 3.2

Proof. The fact that $(f(\theta_n))_{n\geq 0}$ almost surely converges follows the beginning of Proposition 3.1. To show the convergence rates of $(\mathbb{E}[f(\theta_n)])_{n\geq 0}$, we adapt the proof of Proposition 2.2 to out stochastic block setting. Let us denote by $\theta_n^\star$ a minimizer of the surrogate function g_n over Θ. Since the indices $\hat{\imath}_n$ are picked up uniformly at random, we have the following conditional probabilities

$$\mathbb{P}(\theta_n^i = \theta_n^{\star i} | \theta_{n-1}) = \delta \quad \text{and} \quad \mathbb{P}(\theta_n^i = \theta_{n-1}^i | \theta_{n-1}) = 1 - \delta.$$

We can then obtain the following inequalities for all θ in Θ

$$\begin{aligned}\mathbb{E}[f(\theta_n)|\theta_{n-1}] \leq \mathbb{E}[g_n(\theta_n)|\theta_{n-1}] &= \sum_{i=1}^k \mathbb{E}[g_n^i(\theta_n^i)|\theta_{n-1}] = \sum_{i=1}^k (1-\delta) g_n^i(\theta_{n-1}^i) + \delta g_n^i(\theta_n^{\star i}) \\ &= (1-\delta) g_n(\theta_{n-1}) + \delta g_n(\theta_n^\star) \\ &\leq (1-\delta) f(\theta_{n-1}) + \delta g_n(\theta) \\ &\leq (1-\delta) f(\theta_{n-1}) + \delta \left(f(\theta) + \frac{L}{2}\|\theta - \theta_{n-1}\|_2^2 \right),\end{aligned} \tag{13}$$

where we have used the conditional probabilities computed above and the fact that $|g_n(\theta) - f(\theta)| \leq \frac{L}{2}\|\theta - \theta_{n-1}\|_2^2$ according to Lemma 2.1. Let us now follow the proof of Proposition 2.2:

$$\mathbb{E}[f(\theta_n)|\theta_{n-1}] \leq (1-\delta) f(\theta_{n-1}) + \delta \left(\min_{\alpha \in [0,1]} f(\alpha\theta^\star + (1-\alpha)\theta_{n-1}) + \frac{L\alpha^2}{2}\|\theta^\star - \theta_{n-1}\|_2^2 \right).$$

We can now proceed by considering two different cases.

Case 1: without strong convexity:
To simplify the notation, we now introduce the quantities $r_n \triangleq f(\theta_n) - f^\star$ and following again the proof of Proposition 2.2, we have

$$\mathbb{E}[r_n|\theta_{n-1}] \leq (1-\delta)r_{n-1} + \delta\left(\min_{\alpha\in[0,1]}(1-\alpha)r_{n-1} + \frac{LR^2\alpha^2}{2}\right).$$

The term in parenthesis on the right is a concave function of r_{n-1} as a pointwise infimum of concave functions (in fact, pointwise infimum of linear functions). By taking the expectation and using Jensen inequality, we thus have

$$\mathbb{E}[r_n] \leq (1-\delta)\mathbb{E}[r_{n-1}] + \delta\left(\min_{\alpha\in[0,1]}(1-\alpha)\mathbb{E}[r_{n-1}] + \frac{LR^2\alpha^2}{2}\right).$$

By following again the proof of Proposition 2.2, we have

$$\mathbb{E}[r_n] \leq (1-\delta)\mathbb{E}[r_{n-1}] + \delta\begin{cases} \frac{LR^2}{2} & \text{if } \mathbb{E}[r_{n-1}] > LR^2 \\ \mathbb{E}[r_{n-1}]\left(1 - \frac{\mathbb{E}[r_{n-1}]}{2LR^2}\right) & \text{otherwise.}\end{cases}$$

We also notice that the inequality $\mathbb{E}[r_n] \leq (1-\delta)\mathbb{E}[r_{n-1}] + \delta\frac{LR^2}{2}$ is always true. This yields after simple calculations $\mathbb{E}[r_n] \leq (1-\delta)^n r_0 + (1-(1-\delta)^n)\frac{LR^2}{2}$ for all $n \geq 1$. We also remark that the definition of n_0 in the proposition implies that $(1-\delta)^{n_0}r_0 + (1-(1-\delta)^{n_0})\frac{LR^2}{2} \leq LR^2$ after some short calculations. Thus, we have for all $n > n_0$ $\mathbb{E}[r_n] \leq \mathbb{E}[r_{n-1}]\left(1 - \frac{\delta\mathbb{E}[r_{n-1}]}{2LR^2}\right)$ and thus $\mathbb{E}[r_n]^{-1} \geq \mathbb{E}[r_{n_0}]^{-1} + \frac{(n-n_0)\delta}{2LR^2} \geq \frac{2+(n-n_0)\delta}{2LR^2}$, following similar derivations as in Proposition 2.2. This is sufficient to conclude.

Case 2: under strong convexity assumptions:
We proceed similarly as in case 1, but upper-bound instead $\|\theta^\star - \theta_{n-1}\|_2^2$ by $2r_{n-1}/\mu$. This leads us to a similar relation as in the proof of Proposition 2.2:

$$\mathbb{E}[r_n] \leq (1-\delta)\mathbb{E}[r_{n-1}] + \delta\left(\min_{\alpha\in[0,1]} 1 - \alpha + \frac{L\alpha^2}{\mu}\right)\mathbb{E}[r_{n-1}],$$

and following again the proof of Proposition 2.2, we have

$$\mathbb{E}[r_n] \leq ((1-\delta) + \delta\beta)\mathbb{E}[r_{n-1}],$$

yielding the desired convergence rate. □

E.8. Proof of Proposition 3.3

Proof. The fact that $(f(\theta_n))_{n\geq 0}$ almost surely converges follows the beginning of Proposition 3.1. We then separately prove the two remaining parts of the proposition.

Without strong convexity assumptions:
Using the same notation as in the proof of Proposition 3.2, we can replace the inequality $g_n(\theta_n^\star) \leq g_n(\theta)$ in Eq. (13) by $g_n(\theta_n^\star) \leq g_n(\theta) - \frac{\rho}{2}\|\theta_n^\star - \theta\|_2^2$ (using Lemma B.5), and we obtain

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

We also remark that

$$\begin{aligned}\mathbb{E}\left[\|\theta^\star - \theta_n\|_2^2|\theta_{n-1}\right] &= \sum_{i=1}^k \mathbb{E}\left[\|\theta^{\star i} - \theta_n^i\|_2^2|\theta_{n-1}\right] = \sum_{i=1}^k (1-\delta)\|\theta^{\star i} - \theta_{n-1}^i\|_2^2 + \delta\|\theta^{\star i} - \theta_n^{\star i}\|_2^2 \\ &= (1-\delta)\|\theta^\star - \theta_{n-1}\|_2^2 + \delta\|\theta^\star - \theta_n^\star\|_2^2.\end{aligned}$$

Combining the two previous inequalities yields

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

Let us now define $r_n \triangleq \mathbb{E}[f(\theta_n) - f^\star]$. Taking the expectation in the previous inequality gives

$$\begin{aligned} r_n - (1-\delta) r_{n-1} &\leq \frac{\delta L + (1-\delta)\rho}{2} \mathbb{E}[\|\theta^\star - \theta_{n-1}\|_2^2] - \frac{\rho}{2}\mathbb{E}[\|\theta^\star - \theta_n^\star\|_2^2] \\ &\leq \frac{\delta L + (1-\delta)\rho}{2} \left(\mathbb{E}[\|\theta^\star - \theta_{n-1}\|_2^2] - \mathbb{E}[\|\theta^\star - \theta_n\|_2^2] \right). \end{aligned} \tag{14}$$

Summing these inequalities and using the fact that $r_n \leq r_{n-1}$ yields

$$n\delta r_n + (1-\delta)(r_n - r_0) \leq \sum_{k=1}^{n} r_k - (1-\delta) r_{k-1} \leq \frac{\delta L + (1-\delta)\rho}{2} \|\theta^\star - \theta_0\|_2^2,$$

which gives the desired convergence rate.

With strong convexity assumptions:
Assume now that f is μ-strongly convex. To simplify the notation, we introduce the quantity $\xi_n \triangleq \frac{1}{2}\mathbb{E}[\|\theta^\star - \theta_n\|_2^2]$. We can now rewrite the first inequality in (14) as

$$r_n + \rho \xi_n \leq (1-\delta) r_{n-1} + ((1-\delta)\rho + \delta L)\xi_{n-1}.$$

We are going to exploit two inequalities. Since we have the second-order growth property $r_n \geq \mu \xi_n$, for all β in $[0,1]$,

$$\beta r_n + (\rho + (1-\beta)\mu)\xi_n \leq (1-\delta) r_{n-1} + ((1-\delta)\rho + \delta L)\xi_{n-1}.$$

By choosing $\beta \triangleq \frac{(1-\delta)(\rho+\mu)}{(1-\delta)(\rho+\mu)+\delta L}$, it is easy to show that

$$(1-\delta) r_n + ((1-\delta)\rho + \delta L)\xi_n \leq \frac{(1-\delta)(\rho+\mu) + \delta L}{\rho + \mu} \left((1-\delta) r_{n-1} + ((1-\delta)\rho + \delta L)\xi_{n-1} \right).$$

Thus, we have by induction

$$(1-\delta) r_n + ((1-\delta)\rho + \delta L)\xi_n \leq \left(\frac{(1-\delta)(\rho+\mu) + \delta L}{\rho + \mu} \right)^n \left((1-\delta) r_0 + ((1-\delta)\rho + \delta L)\xi_0 \right),$$

and again, since $\mu \xi_n \leq r_n$, we obtain the convergence rate of $(\xi_n)_{n \geq 0}$

$$\xi_n \leq C \left(\frac{(1-\delta)(\rho+\mu) + \delta L}{\rho + \mu} \right)^n \leq \alpha^n \left(\frac{(1-\delta) r_0}{L} + \xi_0 \right),$$

where we have defined the quantities $\alpha \triangleq \frac{(1-\delta)(\rho+\mu)+\delta L}{\rho+\mu}$ and $C \triangleq \frac{(1-\delta) r_0 + ((1-\delta)\rho + \delta L)\xi_0}{(1-\delta)(\rho+\mu)+\delta L}$. We now compute the convergence rate of $(r_n)_{n \geq 0}$ by induction. Suppose that $r_{n-1} \leq C' \alpha^{n-2}$ for some constant C' and some $n \geq 2$. We have shown in (14) that $r_n \leq (1-\delta) r_{n-1} + L \xi_{n-1}$. By using the induction hypothesis, we have $r_n \leq ((1-\delta)C'/\alpha + LC)\alpha^{n-1}$. We therefore study under which conditions we have both $((1-\delta)C'/\alpha + LC) \leq C'$ and $r_1 \leq C'$, which are sufficient conditions to have by induction $r_n \leq C' \alpha^{n-1}$ for all n. It is easy to show that the quantity $C' \triangleq \frac{(1-\delta) r_0 + ((1-\delta)\rho + \delta L)\xi_0}{\delta}$ satisfies such conditions. □

E.9. Proof of Proposition 4.1

Proof. We have from the strong convexity of g_n:

$$f(\theta_n) \leq g_n(\theta_n) \leq \min_{\alpha \in [0,1]} (1-\alpha) g_n(\theta_{n-1}) + \alpha g_n(\nu_n) - \frac{L}{2}\alpha(1-\alpha)\|\theta_{n-1} - \nu_n\|_2^2,$$

where ν_n is defined in Algorithm 3. Let us now consider $\theta^\star$ such that $f(\theta^\star) = f^\star$. Then, we have

$$\begin{aligned} f(\theta_n) &\leq \min_{\alpha \in [0,1]} (1-\alpha) f(\theta_{n-1}) + \alpha \left(g_n(\nu_n) - \frac{L}{2}\|\nu_n - \theta_{n-1}\|_2^2 \right) + \frac{L}{2}(\alpha - \alpha(1-\alpha))\|\nu_n - \theta_{n-1}\|_2^2. \\ &\leq \min_{\alpha \in [0,1]} (1-\alpha) f(\theta_{n-1}) + \alpha \left(g_n(\theta^\star) - \frac{L}{2}\|\theta^\star - \theta_{n-1}\|_2^2 \right) + \frac{\alpha^2 L R^2}{2} \\ &\leq \min_{\alpha \in [0,1]} (1-\alpha) f(\theta_{n-1}) + \alpha f^\star + \frac{\alpha^2 L R^2}{2}, \end{aligned} \tag{15}$$

where we have first used the equality $g_n(\theta_{n-1}) = f(\theta_{n-1})$, then the second inequality exploits $g_n(\nu_n) - \frac{L}{2}\|\nu_n - \theta_{n-1}\|_2^2 \leq g_n(\theta^\star) - \frac{L}{2}\|\theta^\star - \theta_{n-1}\|_2^2$ from the definition of ν_n. Finally, we use the fact that $g_n(\theta^\star) = f^\star + h_n(\theta^\star)$ where h_n is the approximation error function $g_n - f$ with $|h_n(\theta^\star)| \leq \frac{L}{2}\|\theta^\star - \theta_{n-1}\|_2^2$ is ensured by Lemma 2.1. Minimizing (15) with respect to α and denoting by $r_n \triangleq f(\theta_n) - f^\star$ yields

$$r_n \leq \begin{cases} \frac{LR^2}{2} & \text{if } \; r_{n-1} \leq LR^2 \\ r_{n-1}\left(1 - \frac{r_{n-1}}{2LR^2}\right) & \text{otherwise} \end{cases} .$$

These are the same relations used in the proof of Proposition 2.2, leading therefore to the same convergence rate. □

E.10. Proof of Proposition 5.1

Proof. We follow the proof techniques introduced by Nesterov (2004) using the so called "estimate sequences", and more precisely we adapt the proof of Nesterov (2004, Theorem 2.2.8) to deal with our surrogate functions.

Preliminaries:
We rely heavily on Lemma 2.1, which we recall and expand here. Let us define $\rho \triangleq L + \mu$. Then, for all θ in Θ,

$$\begin{aligned} f(\theta_n) &\leq f(\theta) + \frac{L}{2}\|\theta - \kappa_{n-1}\|_2^2 - \frac{\rho}{2}\|\theta - \theta_n\|_2^2 \\ &= f(\theta) + \frac{L}{2}\|\theta - \kappa_{n-1}\|_2^2 - \frac{\rho}{2}\|\theta - \kappa_{n-1} + \kappa_{n-1} - \theta_n\|_2^2 \\ &= f(\theta) - \frac{\mu}{2}\|\theta - \kappa_{n-1}\|_2^2 - \frac{\rho}{2}\|\theta_n - \kappa_{n-1}\|_2^2 + \rho(\theta - \kappa_{n-1})^\top(\theta_n - \kappa_{n-1}). \end{aligned} \tag{16}$$

To simplify the notation in the sequel, we introduce the quantity $\xi_n \triangleq \theta_n - \kappa_{n-1}$, which Nesterov (2004) calls "gradient mapping", up to a multiplicative constant. Then, (16) can be rewritten

$$f(\theta_n) \leq f(\theta) - \frac{\mu}{2}\|\theta - \kappa_{n-1}\|_2^2 - \frac{\rho}{2}\|\xi_n\|_2^2 + \rho(\theta - \kappa_{n-1})^\top \xi_n. \tag{17}$$

Definition of the estimate sequence by induction:
Keeping in mind this key quantity, let us now proceed by induction to prove the main result. The recursion hypothesis $\mathcal{H}_n$ for $n \geq 1$ is the existence of a function $\bar{g}_n : \mathbb{R}^p \to \mathbb{R}$ such that

$$\left\{ \begin{array}{lll} \bar{g}_n(\theta) & = \bar{g}_n^\star + \frac{\gamma_n}{2}\|\theta - v_n\|_2^2 & \\ \bar{g}_n(\theta) & \leq f(\theta) + \frac{A_n}{2}\|\theta - \theta_0\|_2^2 & \forall \theta \in \Theta \\ f(\theta_n) & \leq \bar{g}_n^\star & \\ (\rho a_n + \gamma_n)\kappa_n & = \rho a_n \theta_n + \gamma_n v_n & \end{array} \right. , \tag{$\mathcal{H}_n$}$$

for some v_n and some values A_n, γ_n recursively defined as follows: $A_k = A_{k-1}(1 - a_{k-1})$ and $\gamma_k = (1 - a_{k-1})\gamma_{k-1} + \mu a_{k-1}$ for all $k \geq 2$, and $A_1 = L$, $\gamma_1 = \rho$. We recall that the scalars a_k are also defined in the algorithm. The functions $\bar{g}_n$ which we are going to recursively define are related to the "estimate sequences" introduced by Nesterov (2004). Along with the quantity A_n, they indeed reflect the convergence rate of the algorithm, since $\mathcal{H}_n$ implies that $f(\theta_n) - f^\star \leq \frac{A_n}{2}\|\theta^\star - \theta_0\|_2^2$.

Initialization of the induction for $n = 1$:
Let us first initialize the induction, by showing that $\mathcal{H}_1$ is true. We remark that $A_1 = L$ and $\gamma_1 = \rho$ are chosen such that We can thus define

$$\bar{g}_1(\theta) = f(\theta_1) + \frac{\gamma_1}{2}\|\theta - \theta_1\|_2^2.$$

In other words, we define $v_1 \triangleq \theta_1$ and $\bar{g}_1^\star \triangleq f(\theta_1)$, and we obviously have the first and third conditions of $\mathcal{H}_1$. The second one is simply an application of Lemma 2.1, when noticing that $\kappa_0 = \theta_0$. The last condition is also satisfied because $\kappa_1 = \theta_1 = v_1$ (since $\beta_1 = 0$ in the algorithm).

Induction argument:
Since we have shown that $\mathcal{H}_1$ is true, we now assume $\mathcal{H}_{n-1}$ for $n \geq 2$ and show $\mathcal{H}_n$. We define $\bar{g}_n : \mathbb{R}^p \to \mathbb{R}$ such that for all θ in $\mathbb{R}^p$

$$\bar{g}_n(\theta) = (1 - a_{n-1})\bar{g}_{n-1}(\theta) + a_{n-1}\left(f(\theta_n) + \frac{\mu}{2}\|\theta - \kappa_{n-1}\|_2^2 + \frac{\rho}{2}\|\xi_n\|_2^2 - \rho(\theta - \kappa_{n-1})^\top \xi_n\right). \tag{18}$$

Because of (17), the term between parenthesis on the right is smaller than $f(\theta)$ and thus, we have $\bar{g}_n(\theta) \leq f(\theta) + (1-a_{n-1})\frac{A_{n-1}}{2}\|\theta-\theta_0\|_2^2 = f(\theta) + \frac{A_n}{2}\|\theta-\theta_0\|_2^2$, by definition of A_n. Thus, the second condition of $\mathcal{H}_n$ is true. The function $\bar{g}_n$ is moreover quadratic and the first condition is easy to check (using appropriate values for $\bar{g}_n^\star$ and v_n). Let us now check the third condition, namely that $f(\theta_n) \leq \min_{\theta\in\Theta}\bar{g}_n(\theta)$. We first remark that

$$\begin{aligned}\bar{g}_{n-1}(\theta) &= \bar{g}_{n-1}^\star + \frac{\gamma_{n-1}}{2}\|\theta - v_{n-1}\|_2^2 \\ &\geq f(\theta_{n-1}) + \frac{\gamma_{n-1}}{2}\|\theta - v_{n-1}\|_2^2 \\ &\geq f(\theta_n) + \frac{\rho}{2}\|\xi_n\|_2^2 - \rho(\theta_{n-1}-\kappa_{n-1})^\top\xi_n + \frac{\gamma_{n-1}}{2}\|\theta - v_{n-1}\|_2^2,\end{aligned}$$

The first inequality comes from the induction hypothesis $\mathcal{H}_{n-1}$ and the second inequality comes from (17). Then, we can combine this inequality with (18).

$$\bar{g}_n(\theta) \geq f(\theta_n) + \frac{\rho}{2}\|\xi_n\|_2^2 - (1-a_n)\rho(\theta_{n-1}-\kappa_{n-1})^\top\xi_n + B(\theta), \tag{19}$$

where

$$B(\theta) \triangleq \frac{(1-a_{n-1})\gamma_{n-1}}{2}\|\theta - v_{n-1}\|_2^2 + \frac{a_{n-1}\mu}{2}\|\theta-\kappa_{n-1}\|_2^2 - \rho a_{n-1}(\theta-\kappa_{n-1})^\top\xi_n.$$

Note that $B(\theta)$ is also the part of $\bar{g}_n$ dependent on θ, and such that $v_n = \arg\min_{\theta\in\mathbb{R}^p} B(\theta)$. Minimizing $B(\theta)$ yields

$$v_n = \frac{1}{\gamma_n}\left((1-a_{n-1})\gamma_{n-1}v_{n-1} + a_{n-1}\mu\kappa_{n-1}\right) + \frac{\rho a_{n-1}}{\gamma_n}\xi_n, \tag{20}$$

where we recall that $\gamma_n = (1-a_{n-1})\gamma_{n-1} + \mu a_{n-1}$. Moreover, we have the convexity inequality

$$\begin{aligned}B(\theta) &= \frac{\gamma_n}{2}\left(\frac{(1-a_{n-1})\gamma_{n-1}}{\gamma_n}\|\theta - v_{n-1}\|_2^2 + \frac{a_{n-1}\mu}{\gamma_n}\|\theta-\kappa_{n-1}\|_2^2\right) - \rho a_{n-1}(\theta-\kappa_{n-1})^\top\xi_n \\ &\geq \frac{\gamma_n}{2}\left\|\theta - \left(\frac{(1-a_{n-1})\gamma_{n-1}}{\gamma_n}v_{n-1} + \frac{a_{n-1}\mu}{\gamma_n}\kappa_{n-1}\right)\right\|_2^2 - \rho a_{n-1}(\theta-\kappa_{n-1})^\top\xi_n.\end{aligned}$$

and thus, using the closed form of v_n computed in (20), we have

$$\begin{aligned}B(v_n) &\geq \frac{\gamma_n}{2}\left\|\frac{\rho a_{n-1}}{\gamma_n}\xi_n\right\|_2^2 - \frac{\rho a_{n-1}(1-a_{n-1})\gamma_{n-1}}{\gamma_n}\left(v_{n-1}-\kappa_{n-1}\right)^\top\xi_n - \frac{\rho^2 a_{n-1}^2}{\gamma_n}\|\xi_n\|_2^2 \\ &= -\frac{\rho^2 a_{n-1}^2}{2\gamma_n}\|\xi_n\|_2^2 - \frac{\rho a_{n-1}(1-a_{n-1})\gamma_{n-1}}{\gamma_n}\left(v_{n-1}-\kappa_{n-1}\right)^\top\xi_n.\end{aligned}$$

We can now obtain the following lower-bound on $\bar{g}_n^\star \triangleq \min_{\theta\in\mathbb{R}^p}\bar{g}_n(\theta)$, plugging the value of $B(v_n)$ into (19),

$$\bar{g}_n^\star \geq f(\theta_n) + \left(\frac{\rho}{2} - \frac{\rho^2 a_{n-1}^2}{2\gamma_n}\right)\|\xi_n\|_2^2 - (1-a_{n-1})\rho\left(\theta_{n-1} - \kappa_{n-1} + \frac{a_{n-1}\gamma_{n-1}}{\gamma_n}(v_{n-1}-\kappa_{n-1})\right)^\top\xi_n.$$

Given the definitions of γ_n and a_n, and the fact that $\rho a_0^2 = \gamma_1$, we also obviously have the relation $\rho a_{n-1}^2 = \gamma_n$ for all $n \geq 0$. This cancels the factor in front of $\|\xi_n\|_2^2$. It is also easy to show that the fourth condition of $\mathcal{H}_{n-1}$ implies $\theta_{n-1} - \kappa_{n-1} + \frac{a_{n-1}\gamma_{n-1}}{\gamma_n}(v_{n-1}-\kappa_{n-1}) = 0$.

Since we have shown the three first conditions of $\mathcal{H}_n$, it remains to show the last one, namely that $(\rho a_n + \gamma_n)\kappa_n = \rho a_n\theta_n + \gamma_n v_n$. We first remark that (20) can be rewritten

$$\gamma_n v_n = (1-a_{n-1})\gamma_{n-1}v_{n-1} + a_{n-1}(\mu-\rho)\kappa_{n-1} + \rho a_{n-1}\theta_n.$$

Combining with the fourth condition of $\mathcal{H}_{n-1}$, we have

$$\begin{aligned}\gamma_n v_n &= (1-a_{n-1})\left((\rho a_{n-1} + \gamma_{n-1})\kappa_{n-1} - \rho a_{n-1}\theta_{n-1}\right) + a_{n-1}(\mu-\rho)\kappa_{n-1} + \rho a_{n-1}\theta_n \\ &= -(1-a_{n-1})a_{n-1}\rho\theta_{n-1} + \rho a_{n-1}\theta_n \\ &= \gamma_n\left(\theta_{n-1} + \frac{1}{a_{n-1}}(\theta_n - \theta_{n-1})\right),\end{aligned}$$

where we use the relation $\rho a_{n-1}^2 = \gamma_n$ and the recursive relation between γ_n and γ_{n-1} to remove the terms depending on κ_{n-1} in the first equation. Now that we have a simple form describing v_n, we can finally show,

$$\frac{\rho a_n \theta_n + \gamma_n v_n}{\rho a_n + \gamma_n} = \theta_n + \frac{\gamma_n(1/a_{n-1} - 1)}{\rho a_n + \gamma_n}(\theta_n - \theta_{n-1}).$$

And some simple computation shows that the right part of this equation is equal to κ_n. In other words, the factor in front of $(\theta_n - \theta_{n-1})$ is equal to β_n, and the last condition of $\mathcal{H}_n$ is satisfied.

Obtaining the convergence rate:
Since $\mathcal{H}_n$ is true for all $n \geq 1$, we have $f(\theta_n) - f^\star \leq \frac{A_n}{2}\|\theta^\star - \theta_0\|_2^2$ and thus it remains to compute the convergence rate of the sequence A_n to prove the main result. We follow here the proof of Nesterov (2004, Lemma 2.2.4). Let us first look at the case $\mu = 0$. It is easy to show by induction that for all $n \geq 1$, we have $A_n = La_{n-1}^2$. Moreover

$$\frac{1}{a_n} - \frac{1}{a_{n-1}} = \frac{a_{n-1} - a_n}{a_{n-1}a_n} = \frac{a_{n-1}^2 - a_n^2}{a_{n-1}a_n(a_{n-1} + a_n)} = \frac{a_{n-1}^2 a_n}{a_{n-1}a_n(a_{n-1} + a_n)} = \frac{a_{n-1}}{a_{n-1} + a_n} \geq \frac{1}{2}$$

where we use the relation $a_n^2 = (1 - a_n)a_{n-1}^2$ and and the fact that $a_n \leq a_{n-1}$ for all $n \geq 1$. Thus, we have

$$\frac{1}{a_n} - \frac{1}{a_0} = \frac{1}{a_n} - 1 \geq \frac{n}{2},$$

and $a_n \leq 2/(n+2)$. Since $A_n = La_{n-1}^2$, this gives us the desired convergence rate.

When $\mu > 0$, we have the relation $a_n^2 = (1 - a_n)a_{n-1}^2 - \frac{\mu}{\rho}a_n$. It is then easy to show by induction that for all $n \geq 0$, we have $a_n \geq \sqrt{\frac{\mu}{\rho}}$. Thus, $A_n \leq \left(1 - \sqrt{\frac{\mu}{\rho}}\right)^{n-1} A_1$. Since $A_1 = L$, we have obtain the second convergence rate. □

E.11. Proof of Proposition 6.1

Proof. The proof is very similar to the one of Proposition 3.1. We proceed in several steps.

Almost sure convergence of $f(\theta_n)$:
Let us denote by $\bar{g}_n \triangleq \frac{1}{T}\sum_{t=1}^T g_n^t$. We have the following recursion relation

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

where the surrogates and the index $\hat{t}_n$ are chosen in the algorithm. This allows us to obtain the following inequalities, which hold with probability one

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

The first inequality is true by definition of θ_n and the second one because $\bar{g}_{n-1}^{\hat{t}_n}$ is a majorant 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. Note now that the previous inequalities imply

$$\mathbb{E}[\bar{g}_n(\theta_n)] - \mathbb{E}[\bar{g}_{n-1}(\theta_{n-1})] \leq \mathbb{E}[f^{\hat{t}_n}(\theta_{n-1}) - g_{n-1}^{\hat{t}_n}(\theta_{n-1})]. \tag{21}$$

The non-positive term $\mathbb{E}[\bar{g}_n(\theta_n)] - \mathbb{E}[\bar{g}_{n-1}(\theta_{n-1})]$ is the summand of a converging sum. Thus, the non-positive

terms $\mathbb{E}[f^{\hat{t}_n}(\theta_{n-1}) - g_{n-1}^{\hat{t}_n}(\theta_{n-1})]$ is also the summand of a converging sum and we have

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

where we use two times Beppo-Lévy theorem to exchange the expectation and the sum signs in front of non-negative quantities. As a result, the term $\bar{g}_n(\theta_n) - f(\theta_n)$ converges almost surely to 0, implying the almost sure convergence of $f(\theta_n)$.

Asymptotic stationary point conditions:
Let us denote by $\bar{h}_n \triangleq \bar{g}_n - f$ which is differentiable with L-Lipschitz continuous gradient. Then, for all θ in Θ,

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

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 a similar argument as in the proof of Proposition 2.1. Since we have shown that $\bar{h}_n(\theta_n)$ almost surely converges to zero, we conclude as in the proof of Proposition 3.1. □

E.12. Proof of Proposition 6.2

Proof. The almost sure convergence of $f(\theta_n)$ was shown in Proposition 6.1. We now prove the proposition in several steps and start with some preliminaries.

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

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

where $\delta \triangleq 1/T$. Thus we have for all t in $\{1, \dots, T\}$ and all $n \geq 1$,

$$\mathbb{E}[\|\theta^\star - \kappa_{n-1}^t\|_2^2] = \mathbb{E}[\mathbb{E}[\|\theta^\star - \kappa_{n-1}^t\|_2^2|\theta_{n-1}]] = \delta\mathbb{E}[\|\theta^\star - \theta_{n-1}\|_2^2] + (1-\delta)\mathbb{E}[\|\theta^\star - \kappa_{n-2}^t\|_2^2]. \tag{22}$$

The other relation we need is an extension of Lemma 2.1 to the incremental setting. For all θ in Θ, we have

$$f(\theta_n) \leq f(\theta) + \frac{1}{T}\sum_{t=1}^{T}\left(\frac{L}{2}\|\theta - \kappa_{n-1}^t\|_2^2 - \frac{\rho}{2}\|\theta - \theta_n\|_2^2\right). \tag{23}$$

The proof of this relation is similar to that of Lemma 2.1, exploiting to ρ-strong convexity of $\bar{g}_n$. We can now study the first part of the proposition.

Monotonic decrease of $\mathbb{E}[f(\theta_n)]$:
Note that $\mathbb{E}[g_{n-1}^{\hat{t}_n}(\theta_{n-1})] = \mathbb{E}[\mathbb{E}[g_{n-1}^{\hat{t}_n}(\theta_{n-1})|\theta_{n-1}]] = \mathbb{E}[\bar{g}_{n-1}(\theta_{n-1})]$. Applying this relation to Eq. (21), we have

$$\mathbb{E}[f(\theta_n)] \leq \mathbb{E}[\bar{g}_n(\theta_n)] \leq \mathbb{E}[f^{\hat{t}_n}(\theta_{n-1})] = \mathbb{E}[\mathbb{E}[f^{\hat{t}_n}(\theta_{n-1})|\theta_{n-1}]] = \mathbb{E}[f(\theta_{n-1})],$$

where the first inequality comes from the fact that $f \leq \bar{g}_n$ (see proof of Proposition 6.1).

Non-strongly convex case ($\rho = L$); convergence rate:
Denote by $A_n \triangleq \mathbb{E}[\frac{1}{2T}\sum_{t=1}^{T}\|\theta^\star - \kappa_n^t\|_2^2]$ and by $\xi_n \triangleq \frac{1}{2}\mathbb{E}[\|\theta^\star - \theta_n\|_2^2]$. Then, we have from (23) and by taking the expectation

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

It follows from (22) that $A_n = \delta\xi_n + (1-\delta)A_{n-1}$ and thus

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

By summing the above inequalities, and using the fact that $\mathbb{E}[f(\theta_n) - f^\star]$ is monotonically decreasing, we obtain that

$$n\mathbb{E}[f(\theta_n) - f^\star] \leq \frac{LA_0}{\delta},$$

leading to the convergence rate of Eq. (6.2), since $A_0 = \frac{1}{2}\|\theta^\star - \theta_0\|_2^2$.

μ-strongly convex case:
Suppose now that f is μ-strongly convex. We will prove the proposition by induction. Assume that for some $n \geq 1$, we have $A_{n-1} \leq \beta^{n-1}\xi_0$ with $\beta \triangleq \frac{(1-\delta)(\rho+\mu)+\delta L}{\rho+\mu}$. We have from (23) and the second-order growth condition of Lemma B.5

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

which is true for all $n \geq 1$. Combining the previous inequality, Eq. (22), and the induction hypothesis, we have

$$A_n = \delta\xi_n + (1-\delta)A_{n-1} \leq \left(\frac{\delta L}{\mu+\rho} + (1-\delta)\right)\beta^{n-1}\xi_0 = \beta^n\xi_0.$$

Since we have $A_0 = \xi_0$, the induction hypothesis is true for all $n \geq 0$. Since we have from (23) $\xi_n \leq \frac{L}{\rho+\mu}A_{n-1}$, and $\mathbb{E}[f(\theta_n) - f^\star] \leq LA_{n-1}$, we finally have shown the desired convergence rate (6.2). □

E.13. Proof of Proposition D.1

Proof. Let us denote by $\nu_n^\star \triangleq \arg\min_{\theta\in\Theta}\left[g_n(\theta) - \frac{L}{2}\|\theta - \theta_{n-1}\|_2^2\right]$. Because of the separability of the surrogate function g_n, we have after a few calculations

$$\mathbb{E}[f(\theta_n)|\theta_{n-1}] \leq \mathbb{E}[g_n(\theta_n)|\theta_{n-1}] \leq (1-\delta)f(\theta_{n-1}) + \delta\min_{\alpha\in[0,1]} g_n((1-\alpha)\theta_{n-1} + \alpha\nu_n^\star).$$

Following the proof of Proposition 4.1, we have

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

Taking the expectation and defining $r_n \triangleq \mathbb{E}[f(\theta_n) - f^\star]$, we have

$$r_n \leq (1-\delta)r_{n-1} + \delta\min_{\alpha\in[0,1]}(1-\alpha)r_{n-1} + \frac{\alpha^2 LR^2}{2},$$

where we have used Jensen inequality similarly as in the proof of Proposition 3.2.

Minimizing with respect to α yields

$$r_n \leq (1-\delta)r_{n-1} + \delta\begin{cases}\frac{LR^2}{2} & \text{if } r_{n-1} > LR^2\\ r_{n-1}\left(1 - \frac{r_{n-1}}{2LR^2}\right) & \text{otherwise.}\end{cases}$$

This is the same recursive relations as in the proof of Proposition 3.2, and we therefore obtain the same convergence rate. □

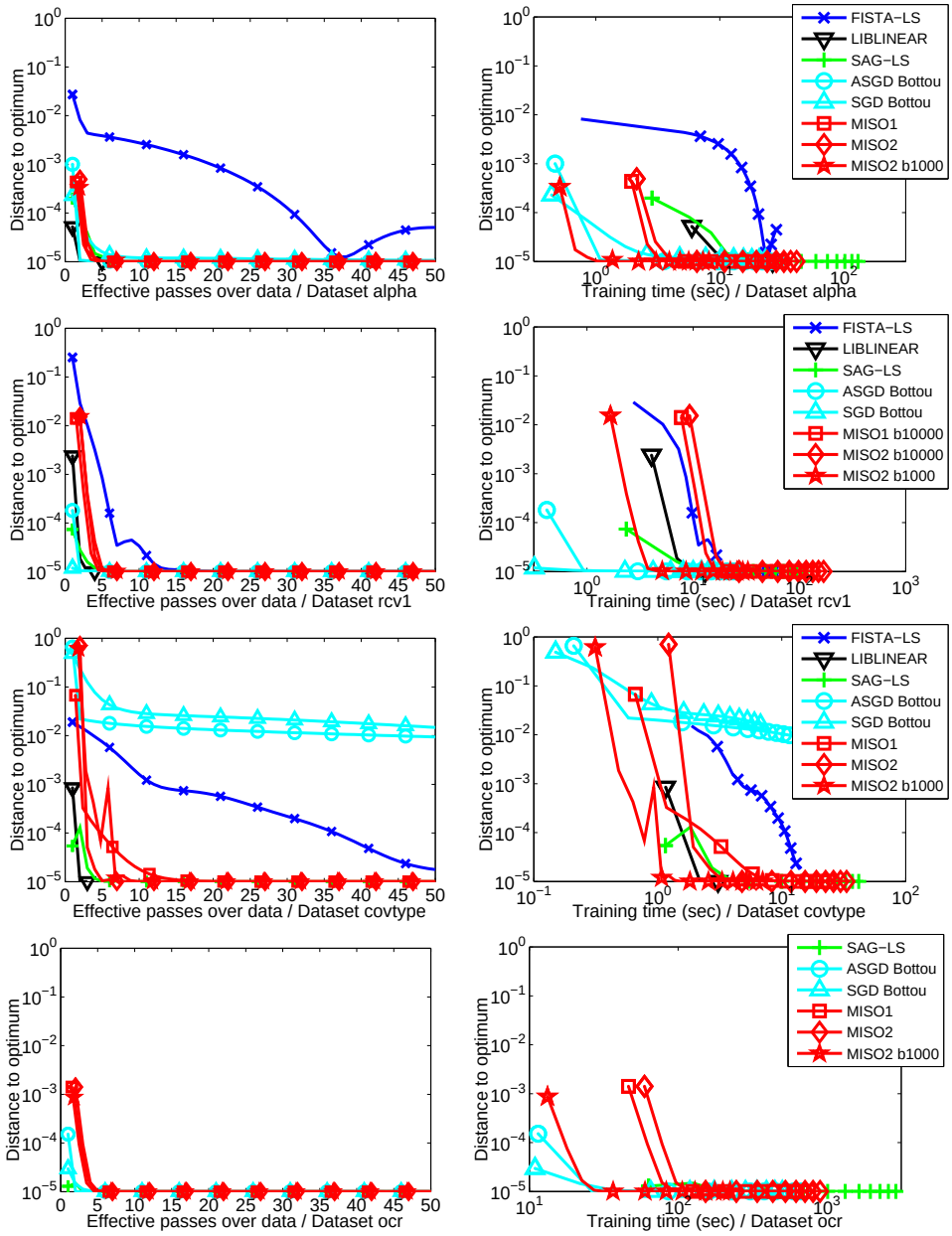

Figure 3. Benchmarks for ℓ_2-logistic regression with $\lambda = 10^{-3}$.

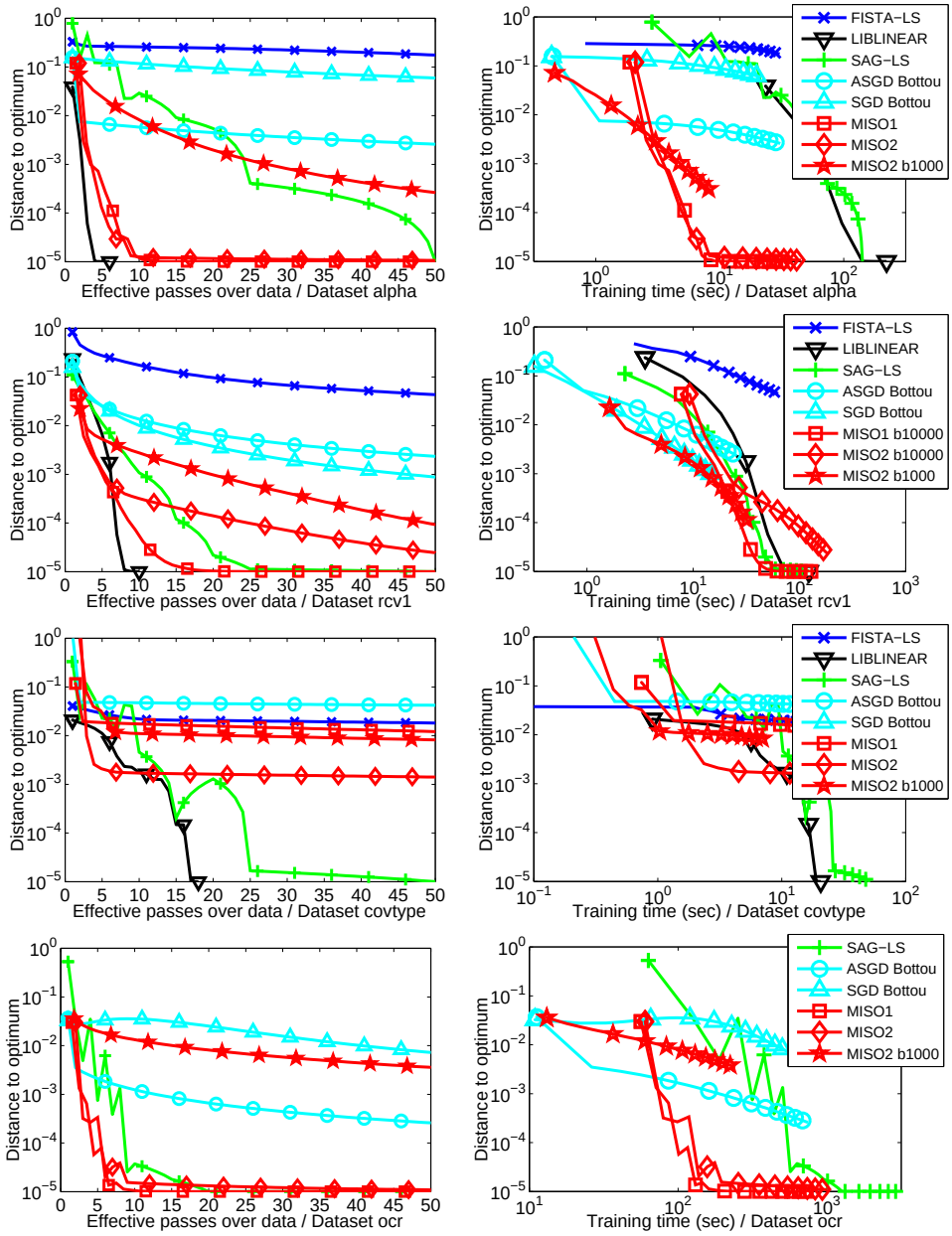

Figure 4. Benchmarks for ℓ_2-logistic regression with $\lambda = 10^{-7}$.

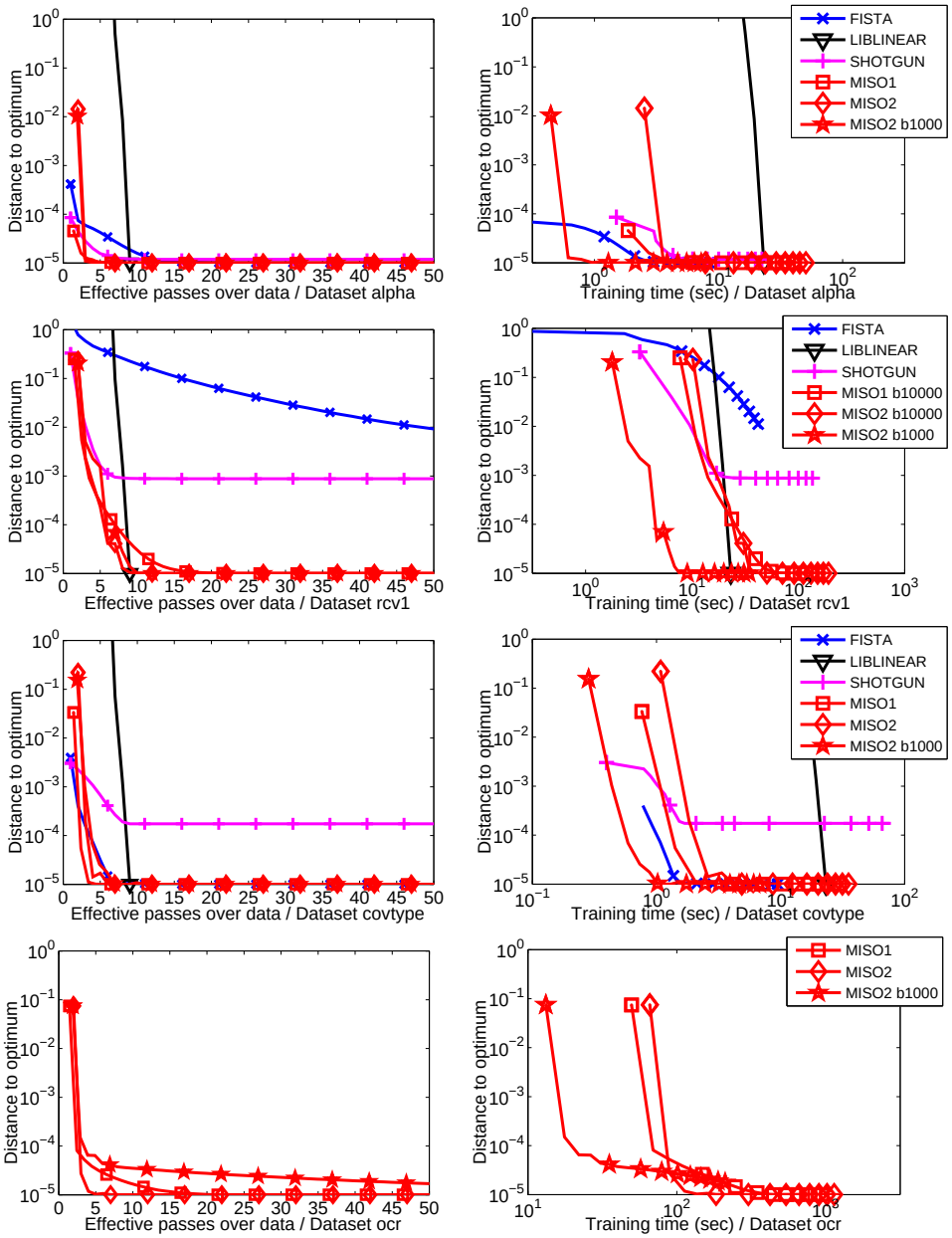

Figure 5. Benchmarks for ℓ_1-logistic regression. The regularization parameter λ was chosen to obtain a solution with about 3% nonzero coefficients.

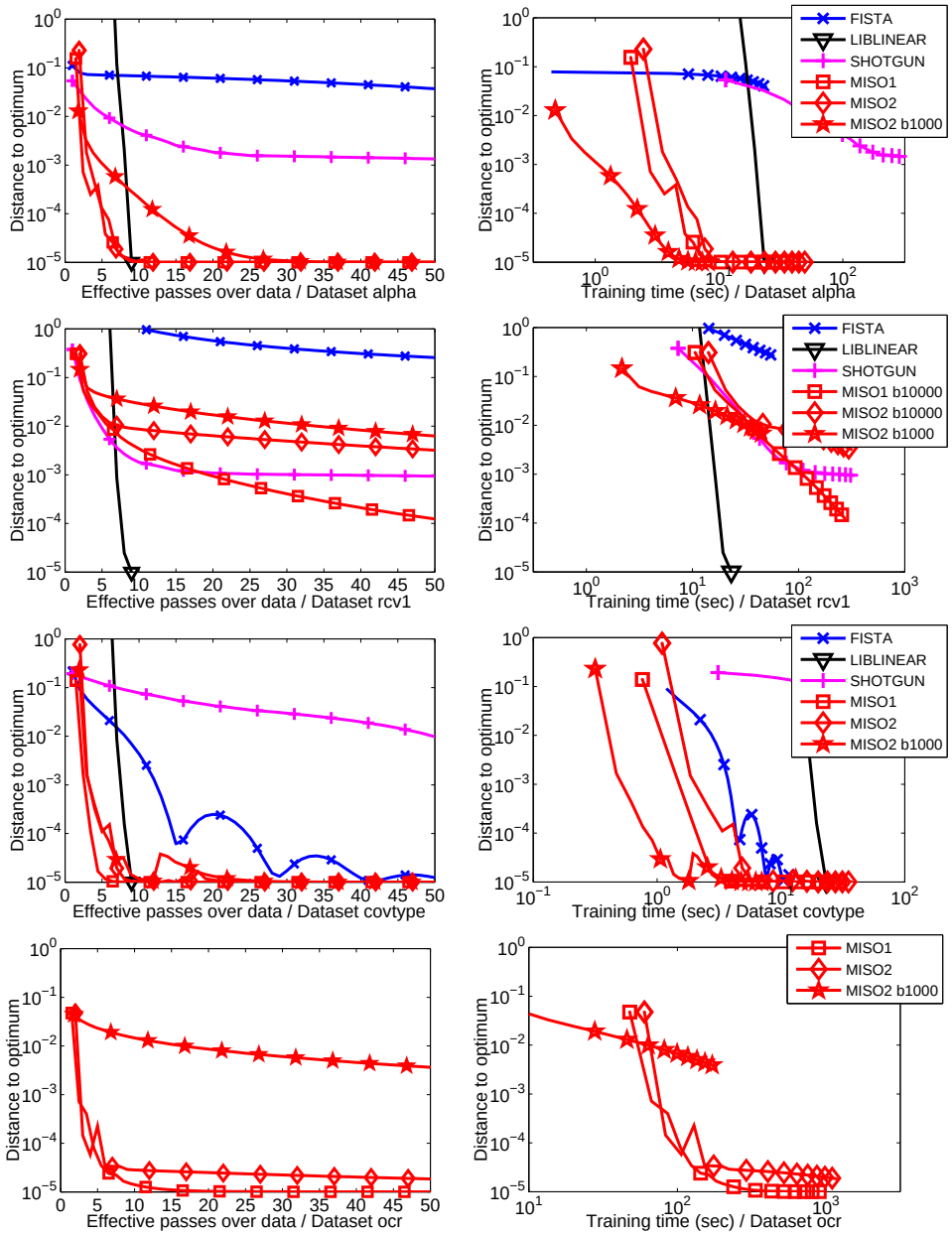

Figure 6. Benchmarks for ℓ_1-logistic regression. The regularization parameter λ was chosen to obtain a solution with about 50% nonzero coefficients.

F. Additional Experimental Results

Figures 3 and 4 presents benchmarks for ℓ_2-logistic regressions with a different regularization parameter than Figure 1. Similarly, we present ℓ_1-logistic regressions benchmarks in Figures 5 and 6 with a different sparsity level than Figure 2.

Supplementary References

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

Clarke, F.H. *Optimization and Nonsmooth Analysis*. John Wiley, 1983.

Danskin, J. M. The theory of max-min, and its application to weapons allocation problems. *Ökonometrie und Unternehmensforschung*, 1967.

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