Journal of Optimization Theory and Applications
https://doi.org/10.1007/s10957-021-01957-1

An Accelerated Coordinate Gradient Descent Algorithm for Non-separable Composite Optimization

Aviad Aberdam[1] · **Amir Beck**[2]

Received: 16 March 2021 / Accepted: 29 September 2021

Abstract

Coordinate descent algorithms are popular in machine learning and large-scale data analysis problems due to their low computational cost iterative schemes and their improved performances. In this work, we define a monotone accelerated coordinate gradient descent-type method for problems consisting of minimizing $f+g$, where f is quadratic and g is *nonsmooth and non-separable* and has a low-complexity proximal mapping. The algorithm is enabled by employing the forward–backward envelope, a composite envelope that possess an exact smooth reformulation of $f+g$. We prove the algorithm achieves a convergence rate of $O(1/k^{1.5})$ in terms of the original objective function, improving current coordinate descent-type algorithms. In addition, we describe an adaptive variant of the algorithm that backtracks the spectral information and coordinate Lipschitz constants of the problem. We numerically examine our algorithms on various settings, including two-dimensional total-variation-based image inpainting problems, showing a clear advantage in performance over current coordinate descent-type methods.

Keywords Coordinate gradient descent · Composite functions · Forward–backward envelope · Convex optimization

Mathematics Subject Classification 90C25 · 65K05

Communicated by Massimo Pappalardo.

✉ Amir Beck
becka@tauex.tau.ac.il

Aviad Aberdam
aaberdam@campus.technion.ac.il

[1] Electrical Engineering, Technion - Israel Institute of Technology, Haifa, Israel

[2] School of Mathematical Sciences, Tel Aviv University, Tel Aviv, Israel

Published online: 23 November 2021

1 Introduction

Coordinate descent-type algorithms are popular in the fields of machine learning and large-scale data analysis due to their low computational cost and their favorable performance [16,17,38]. At each iteration of a coordinate descent-type algorithm, a set of only a few coordinates, called the *working set*, is revised according to some deterministic or random update rule. The update rule can consist of an exact minimization of the working set, as in alternating minimization [9,34], or of a step (or few steps) of a certain minimization algorithm as in block proximal gradient [3,7,15,23]) or block conditional gradient methods [4,26]

A current limitation of coordinate descent-type methods is that they are applicable to nonsmooth problems only under very specific structures. One such form comprises the minimization of the sum of two convex functions—the first is smooth and the second is nonsmooth and *separable* [3,15,23]. Unfortunately, in the non-separable case, these algorithms are not guaranteed to converge to an optimal solution, as coordinate-wise minima points are not necessarily global minimizers. Furthermore, in many non-separable cases, coordinate-descent methods cannot even be directly applied, as a change in one coordinate of a given point in the domain of the function might result in a point that is outside of the domain.

In this paper, we approach the scenario of a composite optimization model, which consists of the sum of a convex quadratic function and a nonsmooth convex function, by replacing it with the *forward–backward envelope function* that was introduced and studied in [36]. This envelope can be considered as a convex and smooth approximation of the objective function, and most importantly, it shares the same set of minimizers with the original problem as long as a certain smoothing parameter used to define the forward–backward envelope is below a certain threshold. Therefore, we can define very simple coordinate descent-type methods that alternate only a single coordinate per iteration for problems in which the proximal operator can be efficiently computed (see Remark 4.2 for examples).

Main contributions Our aim in this work is to extend the applicability of coordinate descent-type methods to non-separable, nonsmooth functions. More specifically,

- By employing the composite envelope as a smooth approximation, we define a monotone accelerated coordinate gradient descent method for general classes of optimization problems. We prove that the algorithm converges at a rate of $O\left(\frac{1}{k^{1.5}}\right)$ in terms of the *original objective function*, rather than in terms of the composite envelope. This convergence rate outperforms current guarantees of coordinate descent-type methods for the considered nonsmooth and non-separable problems.
- We derive a theoretically justified method that backtracks the smoothing parameter and the coordinate Lipschitz constants of the method. More specifically, we prove that using a wrong smoothing parameter must result in the violation of a certain lower bound, and thus frees us from computing spectral information on the problem parameters.
- We examine our algorithms through extensive numerical experiments, studying three types of model constraints: affine sets, ℓ_1-balls, and hyperplanes intersected with boxes, taking portfolio optimization as a specific example. In all these exper-

iments, our method demonstrates significant improvement in performance over existing methods. We further consider the task of natural image inpainting, presenting a single-coordinate-descent algorithm for the two-dimensional total-variation regularized problem.

Related Work In our work, we allow for coordinate-wise algorithms by substituting the non-separable and nonsmooth objective function with the forward–backward envelope [36] which is smooth and shares the same set of minimizers. Here, we discuss two alternative coordinate-descent algorithms for such problems which we also compare to in the experimental section.

The first approach is to reformulate the composite objective as a saddle point problem using the Fenchel conjugate, which enables to utilize primal–dual algorithms [14]. These algorithms share the same $O(n)$ complexity per iteration as our method, and are applicable when the proximal operator of f^*, the Fenchel conjugate of the smooth part f, is easy to compute. It was shown in [14] that the coordinate descent primal–dual approach can guarantee an $O(1/\sqrt{k})$ rate of convergence. In our work, we are able to improve this rate for the sum of a quadratic function and nonsmooth convex function, by proving a convergence rate of $O(1/k^{1.5})$ for our suggested algorithm, while maintaining the same complexity per iteration.

The sketched proximal gradient algorithms of [19,20] can also be an alternative when adopting a coordinate sketch. These methods assume that the available information at each iteration is a random linear transformation (i.e., a sketch) of the gradient. They prove linear rate of convergence for strongly convex objectives [20], and present favorable iteration complexity guarantees when incorporating a momentum term [19]. However, these guarantees do not apply to the non-strongly convex problems we consider in our work.

The forward–backward envelope was used in [27] for the analysis of block-coordinate and incremental-type methods. They proved a linear rate of convergence for certain block-type methods under the celebrated KL condition, when minimizing a model consisting of a separable smooth function and general proper and closed function.

Our work takes the forward–backward Envelope a step forward by utilizing it in the definition of the algorithm. We further prove that despite the fact that the steps of the method aim to minimize the smooth envelope, an $O(1/k^{1.5})$ rate of convergence in terms of the *original* objective function can be guaranteed.

2 Problem Formulation

Our main objective in this paper is to develop coordinate gradient descent-type methods for solving the convex, possibly nonsmooth, optimization model

$$\text{(P)} \quad \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^n} \{F(\mathbf{x}) \equiv f(\mathbf{x}) + g(\mathbf{x})\}, \tag{2.1}$$

where the following underlying assumption is made throughout the paper:

Assumption 1
- $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{M}\mathbf{x} + \mathbf{b}^T\mathbf{x}$ is a quadratic function with a symmetric positive semidefinite matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$ and a vector $\mathbf{b} \in \mathbb{R}^n$.
- $g : \mathbb{R}^n \to (-\infty, \infty]$ is a proper closed and convex function which is possibly *nonsmooth*.
- The optimal set of (P), denoted by X^*, is nonempty.

This model appears frequently in various scientific applications, including portfolio optimization, dual SVM, and regularized least squares problems (see more details in Remark 4.2 and Sect. 5).

2.1 Notations

For a symmetric matrix $\mathbf{A}$, we denote its minimal and maximal eigenvalues by $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$, respectively, its spectral norm by $\|\mathbf{A}\|_2 = \sqrt{\lambda_{\max}(\mathbf{A}^T\mathbf{A})}$, and its ith column by $\mathbf{a}_i$. The vector $\mathbf{e}_i$ has 1 in its ith entry and 0 elsewhere, and $\mathbf{e}$ is the vector of all ones. In addition, for a given proper closed and convex function $h : \mathbb{R}^n \to (-\infty, \infty]$, we denote the *proximal operator* as [31]

$$\mathrm{prox}_h(\mathbf{x}) = \underset{\mathbf{u}}{\mathrm{argmin}} \left\{ h(\mathbf{u}) + \frac{1}{2}\|\mathbf{u} - \mathbf{x}\|_2^2 \right\}. \tag{2.2}$$

The level set $\mathrm{Lev}(f, \alpha)$ of a function $f : \mathbb{R}^n \to \mathbb{R}$ and scalar $\alpha \in \mathbb{R}$ is defined as

$$\mathrm{Lev}(f, \alpha) = \left\{ \mathbf{x} \in \mathbb{R}^n : f(\mathbf{x}) \le \alpha \right\}. \tag{2.3}$$

Given a nonempty closed and convex set $C \subseteq \mathbb{R}^n$, the indicator function δ_C is given by $\delta_C(\mathbf{x}) = 0$ for $\mathbf{x} \in C$ and ∞ otherwise. The orthogonal projection onto such a set C is defined as $P_C(\mathbf{x}) \equiv \mathrm{argmin}_{\mathbf{y}\in C} \|\mathbf{y} - \mathbf{x}\|_2$, and the distance function as $d_C(\mathbf{x}) \equiv \|\mathbf{x} - P_C(\mathbf{x})\|_2 = \min_{\mathbf{y}\in C} \|\mathbf{y} - \mathbf{x}\|_2$. In this paper, we will not use the sup/inf notation but rather use only the min/max notation.

3 The Forward–Backward Envelope

As described above, in order to employ coordinate gradient descent type algorithms, we aim to replace the problem (P) with minimization of a smooth function. One approach would be to find a smooth approximation of the (generally) nonsmooth function g, say g_μ ($\mu > 0$ being a smoothing parameter), and rewrite the problem as

$$\min_{\mathbf{x}\in\mathbb{R}^n} \left\{ f(\mathbf{x}) + g_\mu(\mathbf{x}) \right\}. \tag{3.1}$$

A popular choice (see for example [6,32]) for the smooth approximation function g_μ is the celebrated Moreau envelope, defined as [31]

$$M_g^\mu(\mathbf{x}) \equiv \min_{\mathbf{u}\in\mathbb{R}^n} \left\{ g(\mathbf{u}) + \frac{1}{2\mu}\|\mathbf{u} - \mathbf{x}\|_2^2 \right\}. \tag{3.2}$$

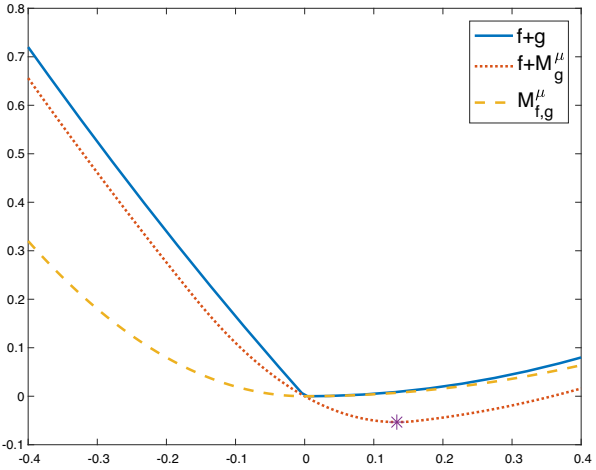


Fig. 1 The composite envelope $M^{\mu}_{f,g}$ compared with the Moreau envelope alternative $f + M^{\mu}_{g}$ for $f(x) = 0.5x^2 - 0.8x$, $g(x) = 0.8|x|$, and $\mu = 0.2$

In general, however, the optimal set of problem (3.1) is different than the optimal set of (P), and thus, the two problems are not equivalent. Therefore, in this paper we employ a different approach and replace the objective function by the so-called *forward–backward envelope function*, which was recently introduced in [36]. The forward–backward envelope function is, in a sense, a combined smoothed version of the composite function $F = f + g$:

Definition 3.1 (*Forward–Backward Envelope Function*, [36, Definition 2.1]) The forward–backward envelope function of (f, g) with a smoothing parameter of $\mu > 0$ is the function:

$$\boxed{M^{\mu}_{f,g}(\mathbf{x}) \equiv f(\mathbf{x}) - \frac{\mu}{2}\|\nabla f(\mathbf{x})\|_2^2 + M^{\mu}_{g}(\mathbf{x} - \mu\nabla f(\mathbf{x})).}$$

where M^{μ}_{g} is the Moreau envelope (Eq. 3.2) of g with a smoothing parameter of $\mu > 0$.

As stated in Theorem 3.1 below and demonstrated in Fig. 1, the fundamental properties of the composite envelope function are that it is (i) convex, (ii) smooth, and (iii) its set of minimizers coincides with the set of minimizers of problem (P). This means that the optimization problem (P) is equivalent to minimizing the smooth unconstrained convex function, $M^{\mu}_{f,g}$, and therefore, can be solved by coordinate descent-type methods. Theorem 3.1 spells out known properties of the forward–backward envelope function that are essential for the analysis that follows.

Theorem 3.1 *Let f and g satisfy Assumption 1, and let $\mu \in \left(0, \frac{1}{\lambda_{\max}(\boldsymbol{M})}\right)$.*[1] *Then,*

(i) [18, Proposition 3.3] *$M^{\mu}_{f,g}$ is convex.*

[1] In the extreme case where $\mathbf{M} = \mathbf{0}$, the condition is $\mu \in (0, \infty)$.

(ii) [36, Proposition 2.3] *The set of minimizers of* $F = f + g$ *coincides with the set of minimizers of* $M^{\mu}_{f,g}$*:*

$$\operatorname*{argmin}_{\boldsymbol{x}\in\mathbb{R}^n} F(\boldsymbol{x}) = \operatorname*{argmin}_{\boldsymbol{x}\in\mathbb{R}^n} M^{\mu}_{f,g}(\boldsymbol{x}),$$

as well as the optimal values:

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} F(\boldsymbol{x}) = \min_{\boldsymbol{x}\in\mathbb{R}^n} M^{\mu}_{f,g}(\boldsymbol{x}).$$

(iii) [18, Proposition 4.4] $M^{\mu}_{f,g}$ *is* $\frac{1}{\mu}$*-smooth, and*

$$\nabla M^{\mu}_{f,g}(\boldsymbol{x}) = (\boldsymbol{I} - \mu\boldsymbol{M})G^{1/\mu}_{f,g}(\boldsymbol{x}), \quad \textit{for all } \boldsymbol{x} \in \mathbb{R}^n,$$

where $G^{1/\mu}_{f,g}$ *is the gradient mapping defined as* [33]*,*

$$G^{1/\mu}_{f,g}(\boldsymbol{x}) \equiv \frac{1}{\mu}\left[\boldsymbol{x} - \operatorname{prox}_{\mu g}\left(\boldsymbol{x} - \mu\nabla f(\boldsymbol{x})\right)\right], \quad \textit{for all } \boldsymbol{x} \in \mathbb{R}^n. \tag{3.3}$$

As an illustration of parts (i) and (ii) of Theorem 3.1, consider problem (P) with $g = \delta_C$, namely, the minimization of $\frac{1}{2}\mathbf{x}^T\mathbf{M}\mathbf{x} + \mathbf{b}^T\mathbf{x}$ over a nonempty closed and convex set C. Utilizing the above theorem and the fact that $M^{\mu}_{\delta_C} = \frac{1}{2\mu}d^2_C$ [31], we can reformulate problem (P) in this setting as the following smooth optimization problem:

$$\operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{2}\mathbf{x}^T(\mathbf{M} - \mu\mathbf{M}^2)\mathbf{x} + \mathbf{b}^T(\mathbf{I} - \mu\mathbf{M})\mathbf{x} + \frac{1}{2\mu}d^2_C\left((\mathbf{I} - \mu\mathbf{M})\mathbf{x} - \mu\mathbf{b}\right), \tag{3.4}$$

where $\mu \in \left(0, \frac{1}{\lambda_{\max}(\mathbf{M})}\right)$. Note that the above formulation bares some resemblance to the following classical penalty-based smooth reformulation ($\eta > 0$ is a penalty parameter):

$$\operatorname*{argmin}_{\mathbf{x}} \left\{\frac{1}{2}\mathbf{x}^T\mathbf{M}\mathbf{x} + \mathbf{b}^T\mathbf{x} + \frac{\eta}{2}d_C(\mathbf{x})^2\right\}.$$

However, the major difference is that the last formulation does not share in general the same minimizers as the original problem. In this sense, problem (3.4) can be seen as an *exact* penalty reformulation. Concrete examples of this reformulation appear in Remark 4.2 and in Sect. 5.

4 Coordinate Descent Methods

In this section, we utilize the properties of the forward–backward envelope (Theorem 3.1) to define a coordinate gradient descent method for the nonsmooth, non-separable

objective (P). We then provide theoretically justified conditions on the algorithm parameters so that these constants can be estimated during run time. We prove that the suggested algorithm converges in the rate of $O\left(\frac{1}{k^{1.5}}\right)$ in terms of the expected values of the original objective function $F = f + g$.

4.1 A Monotone Accelerated Coordinate Gradient Descent Method for Smooth Minimization

Given a continuously differentiable convex function $H : \mathbb{R}^n \to \mathbb{R}$, whose gradient is Lipschitz continuous with a Lipschitz constant $L > 0$, a problem of the form

$$\min_{\mathbf{x}\in\mathbb{R}^n} H(\mathbf{x}), \tag{4.1}$$

can be solved, for example, by accelerated gradient-based methods that enjoy an $O(1/k^2)$ rate of convergence in terms of function values. One example of such an algorithm is the accelerated gradient (AG) method first introduced by in [1], and then generalized in [37]. The method repeats the following steps:
AG

(a) Define $\mathbf{y}^k = (1-\theta^k)\mathbf{x}^k + \theta^k\mathbf{z}^k$.
(b) Set $\mathbf{z}^{k+1} = \mathbf{z}^k - \frac{1}{\theta^k L}\nabla H(\mathbf{y}^k)$.
(c) Set $\mathbf{x}^{k+1} = \mathbf{y}^k - \frac{1}{L}\nabla H(\mathbf{y}^k)$.
(d) Set θ^{k+1} to satisfy $\frac{1-\theta^{k+1}}{(\theta^{k+1})^2} = \frac{1}{(\theta^k)^2}$.

The above scheme is not a monotone method, meaning that the sequence of function values it generates is not necessarily nonincreasing. A monotone sequence can be produced by replacing step (c) with

(c1) Set $\tilde{\mathbf{x}}^{k+1} = \mathbf{y}^k - \frac{1}{L}\nabla H(\mathbf{y}^k)$.
(c2) Choose $\mathbf{x}^{k+1}$ satisfying $H(\mathbf{x}^{k+1}) \le \min\{H(\tilde{\mathbf{x}}^{k+1}), H(\mathbf{x}^k)\}$.

We proceed by presenting a monotone accelerated coordinate gradient descent method for solving problem (4.1). We now assume that ∇H has coordinate Lipschitz constants $L_1, L_2, \ldots, L_n > 0$, which means that

$$\left\| \frac{\partial H(\mathbf{x} + t\mathbf{e}_i)}{\partial x_i} - \frac{\partial H(\mathbf{x})}{\partial x_i} \right\| \le L_i|t|, \quad \text{for any } \mathbf{x} \in \mathbb{R}^n, t \in \mathbb{R}.$$

It is well known (see e.g., [3, Lemma 11.9]) that the above property implies the following "sufficient decrease" property

$$H(\mathbf{x}^k) - H\left(\mathbf{x}^k - \frac{1}{L_i}\nabla_i H(\mathbf{x}^k)\mathbf{e}_i\right) \ge \frac{1}{2L_i}(\nabla_i H(\mathbf{x}^k))^2, \quad \text{for all } i \in \{1, 2, \ldots, n\}. \tag{4.2}$$

The maximal coordinate Lipschitz constant is denoted by

$$L_{\max} = \max\{L_1, L_2, \ldots, L_n\}.$$

Several variants of randomized accelerated coordinate gradient descent methods exist in the literature [15,28]—all share an $O\left(\frac{1}{k^2}\right)$ rate of convergence in terms of expected function values, and are usually designed also to incorporate a nonsmooth *separable* component. The algorithm below is based on a monotone version of these methods; it can be seen as a coordinate descent variant of algorithm AG.

Algorithm 1: Monotone Accelerated Coordinate Gradient Descent (MACGD)

Initialization: $\mathbf{x}^0 \in \mathbb{R}^n, \mathbf{z}^0 = \mathbf{x}^0, \theta^0 = 1$, and coordinate Lipschitz constants of ∇H: $L_1, L_2, \ldots, L_n > 0$.
General Step: For any $k = 0, 1, 2, \ldots$ execute the following steps:
1. Pick i_k at random (uniformly).
2. Define $\mathbf{y}^k = (1-\theta^k)\mathbf{x}^k + \theta^k \mathbf{z}^k$.
3. Set $s = \nabla_{i_k} H(\mathbf{y}^k)$.
4. Set $\tilde{\mathbf{x}}^{k+1} = \mathbf{y}^k - \frac{s}{L_{i_k}}\mathbf{e}_{i_k}$, and $\mathbf{z}^{k+1} = \mathbf{z}^k - \frac{s}{n\theta^k L_{i_k}}\mathbf{e}_{i_k}$.
5. Choose $\mathbf{x}^{k+1} \in \operatorname{argmin}\left\{H(\mathbf{x}) : \mathbf{x} \in \left\{\tilde{\mathbf{x}}^{k+1}, \mathbf{x}^k - \frac{1}{L_{i_k}}\nabla_{i_k} H(\mathbf{x}^k)\mathbf{e}_{i_k}\right\}\right\}$.
6. Set θ^{k+1} to satisfy $\frac{1-\theta^{k+1}}{(\theta^{k+1})^2} = \frac{1}{(\theta^k)^2}$.

By step 5 and the sufficient decrease property (4.2), we conclude that

$$H(\mathbf{x}^k) - H(\mathbf{x}^{k+1}) \geq H(\mathbf{x}^k) - H\left(\mathbf{x}^k - \frac{1}{L_{i_k}}\nabla_{i_k} H(\mathbf{x}^k)\mathbf{e}_{i_k}\right) \geq \frac{1}{2L_{i_k}}(\nabla_{i_k} H(\mathbf{x}^k))^2,$$

which in particular implies the monotonicity of the method. The resulting rate of convergence of this method is given in Theorem 4.1 below. Its proof is almost identical to the derivations in [28], and is provided for the sake of completeness in "Appendix A." We use the notation $\xi_k = \{i_0, i_1, \ldots, i_k\}$.

Theorem 4.1 *Let $\{\mathbf{x}^k\}_{k\geq 0}$ be the sequence generated by the MACGD method, and let $\mathbf{x}^*$ be a minimizer of H over $\mathbb{R}^n$. Then, for any $k \geq 1$,*

$$\mathbb{E}_{\xi_{k-1}}[H(\mathbf{x}^k)] - H(\mathbf{x}^*) \leq \frac{2n^2 \sum_{i=1}^n L_i (x_i^* - x_i^0)^2}{(k+1)^2}. \tag{4.3}$$

4.2 Accelerated Coordinate Gradient Descent for Solving (P)

We return to problem (P) (see (2.1)) under Assumption 1. Obviously, since the objective function in (P) is not smooth, the MACGD method cannot be employed to solve it directly. Instead, we employ it on the *forward–backward envelope function* $M_{f,g}^{\mu}$, and call the devised method MACGD-FB ("FB" for forward–backward). Recalling Theorem 3.1(ii), it follows that the set of minimizers of problem (P) is the same as the set of minimizers of $M_{f,g}^{\mu}$, and thus, the defined procedure, Algorithm 2, also aims at finding optimal solutions of the original problem (P). We note that in the description

of the MACGD-FB method and elsewhere in this paper, $\mathbf{m}_i^T$ is the ith row of $\mathbf{M}$ for any $i \in \{1, 2, \ldots, n\}$.

Algorithm 2: Monotone Accelerated Coordinate Gradient Descent Forward–Backward (MACGD-FB) for problem (P)

Initialization: $\mathbf{x}^0 \in \mathbb{R}^n$, $\mathbf{z}^0 = \mathbf{x}^0$, $\theta^0 = 1$, a smoothing parameter $\mu > 0$, and the coordinate Lipschitz constants of $\nabla M^{\mu}_{f,g}$: $L_1, L_2, \ldots, L_n > 0$.
General Step: For any $k = 0, 1, 2, \ldots$ execute the following steps:
1. Pick $i_k \in \{1, 2, \ldots, n\}$.
2. Define $\mathbf{y}^k = (1 - \theta^k)\mathbf{x}^k + \theta^k \mathbf{z}^k$.
3. Set $s = \frac{\partial M^{\mu}_{f,g}(\mathbf{y}^k)}{\partial y_{i_k}} = \frac{1}{\mu}(\mathbf{e}_{i_k}^T - \mu \mathbf{m}_{i_k}^T)(\mathbf{y}^k - \mathrm{prox}_{\mu g}(\mathbf{y}^k - \mu(\mathbf{M}\mathbf{y}^k + \mathbf{b})))$.
4. Set $\tilde{\mathbf{x}}^{k+1} = \mathbf{y}^k - \frac{s}{L_{i_k}}\mathbf{e}_{i_k}$, and $\mathbf{z}^{k+1} = \mathbf{z}^k - \frac{s}{n\theta^k L_{i_k}}\mathbf{e}_{i_k}$.
5. Set $r = \frac{\partial M^{\mu}_{f,g}(\mathbf{x}^k)}{\partial x_{i_k}} = \frac{1}{\mu}(\mathbf{e}_{i_k}^T - \mu \mathbf{m}_{i_k}^T)(\mathbf{x}^k - \mathrm{prox}_{\mu g}(\mathbf{x}^k - \mu(\mathbf{M}\mathbf{x}^k + \mathbf{b})))$.
6. Set $\mathbf{w}^{k+1} = \mathbf{x}^k - \frac{r}{L_{i_k}}\mathbf{e}_{i_k}$.
7. Update $\theta^{k+1} = \frac{\sqrt{(\theta^k)^4 + 4(\theta^k)^2} - (\theta^k)^2}{2}$.
8. Set $\mathbf{x}^{k+1} = \mathrm{argmin}\left\{M^{\mu}_{f,g}(\mathbf{x}) : \mathbf{x} \in \{\tilde{\mathbf{x}}^{k+1}, \mathbf{w}^{k+1}\}\right\}$.

In the above description, we use the following formula for the ith partial derivative of $M^{\mu}_{f,g}$ that is based on Theorem 3.1(iii):

$$\begin{aligned}\frac{\partial M^{\mu}_{f,g}(\mathbf{x})}{\partial x_i} &= \mathbf{e}_i^T \nabla M^{\mu}_{f,g}(\mathbf{x}) = \mathbf{e}_i^T (\mathbf{I} - \mu\mathbf{M}) G^{1/\mu}_{f,g}(\mathbf{x}) \\ &= (\mathbf{e}_i^T - \mu \mathbf{m}_i^T)\left[\frac{1}{\mu}\left(\mathbf{x} - \mathrm{prox}_{\mu g}\left(\mathbf{x} - \mu\left(\mathbf{M}\mathbf{x} + \mathbf{b}\right)\right)\right)\right].\end{aligned}$$

Remark 4.1 (Complexity of computing $\mathbf{M}\mathbf{x}^k, \mathbf{M}\mathbf{z}^k, \mathbf{M}\mathbf{y}^k$) While a naive implementation of MACGD-FB involves computing $\mathbf{M}\mathbf{x}^k, \mathbf{M}\mathbf{z}^k, \mathbf{M}\mathbf{y}^k$ in $O(n^2)$ operations per iteration, this complexity can be reduced to $O(n)$ by keeping in memory the vectors $\mathbf{M}\mathbf{x}^k$, $\mathbf{M}\mathbf{z}^k$ and $\mathbf{M}\mathbf{y}^k$. Utilizing the fact that $\mathbf{z}^{k+1} = \mathbf{z}^k + (z_{i_k}^{k+1} - z_{i_k}^k)\mathbf{e}_{i_k}$, the vector $\mathbf{M}\mathbf{z}^{k+1}$ can be easily updated in $O(n)$ operations by the formula:

$$\mathbf{M}\mathbf{z}^{k+1} = \mathbf{M}\mathbf{z}^k + (z_{i_k}^{k+1} - z_{i_k}^k)\mathbf{m}_{i_k}.$$

In the same way, the vector $\mathbf{M}\mathbf{x}^{k+1}$ can be computed in $O(n)$ operations as $\mathbf{x}^{k+1}$ differs from $\mathbf{y}^k$ or $\mathbf{x}^k$, depending on step 8, only in its i_kth coordinate. Finally, the vector $\mathbf{M}\mathbf{y}^{k+1}$ is a linear combination of the vectors $\mathbf{M}\mathbf{x}^{k+1}$ and $\mathbf{M}\mathbf{z}^{k+1}$.

Remark 4.2 (Complexity of the prox) The MACGD-FB method makes sense from a computational point of view only when $\mathrm{prox}_{\mu g}$ can be efficiently computed, preferably in $O(n)$ operations. A non-exhaustive list of examples in which this is the case are:

- $g = \delta_{\text{Box}[\boldsymbol{\ell},\mathbf{u}]} = \delta_{\{\mathbf{x}:\boldsymbol{\ell}\le\mathbf{x}\le\mathbf{u}\}}$.
- $g = \delta_{B_2[\mathbf{c},r]} = \delta_{\{\mathbf{x}:\|\mathbf{x}-\mathbf{c}\|_2\le r\}}$ or $g = \delta_{B_1[\mathbf{c},r]} = \delta_{\{\mathbf{x}:\|\mathbf{x}-\mathbf{c}\|_1\le r\}}$ for $\mathbf{c}\in\mathbb{R}^n$, $r>0$, see [13].
- $g = \delta_C$ where $C = \{\mathbf{x}\in\mathbb{R}^n : \mathbf{a}^T\mathbf{x} = b, l \le \mathbf{x} \le u\}$, see [29].
- $g = \delta_C$ where $C = \{\mathbf{A}\mathbf{x} = \mathbf{d}\}$ with a full row rank $\mathbf{A}\in\mathbb{R}^{m\times n}$ and $\mathbf{d}\in\mathbb{R}^m$. Assuming that $m \ll n$, then for any $\mu > 0$, $\text{prox}_{\mu g}(\mathbf{x}) = P_C(\mathbf{x}) = \frac{1}{\mu}\left(\mathbf{x} - \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}(\mathbf{A}\mathbf{x}-\mathbf{d})\right)$ can be computed in $O(n)$ operations (after a pre-process in which $\mathbf{A}\mathbf{A}^T$ is computed).
- $g(\mathbf{x}) = \sum_{i=1}^n g_i(x_i)$, where every g_i is a one-dimensional function whose prox can be computed in $O(1)$ operations (e.g., $g(\mathbf{x}) = \|\mathbf{x}\|_1$).
- $g(\mathbf{x}) = \lambda\text{TV}(\mathbf{x})$ ($\lambda > 0$), the total variation regularization, see [2,24,25].
- $g(\mathbf{x}) = \|\mathbf{x}\|_2$.

A direct application of Theorem 4.1 and the fact that the minimal values of F and $M^{\mu}_{f,g}$ are the same (Theorem 3.1(ii)) is the following $O\left(\frac{1}{k^2}\right)$ rate of convergence of the expected function values of the surrogate function $M^{\mu}_{f,g}(\mathbf{x}^k)$ to the optimal value of (P).

Theorem 4.2 *Let $\{\mathbf{x}^k\}_{k\ge0}$ be the sequence generated by the MACGD-FB method employed on problem (P) with $\mu\in\left(0,\frac{1}{\lambda_{\max}(\mathbf{M})}\right)$ and $L_1, L_2, \ldots, L_n$ being coordinate Lipschitz constants of $\nabla M^{\mu}_{f,g}$. Then for any $k\ge1$ and $\mathbf{x}^*\in X^*$,*

$$\mathbb{E}_{\xi_{k-1}}[M^{\mu}_{f,g}(\mathbf{x}^k)] - F(\mathbf{x}^*) \le \frac{2n^2L_{\max}\|\mathbf{x}^*-\mathbf{x}^0\|_2^2}{(k+1)^2},$$

where $L_{\max} = \max\{L_1, L_2, \ldots, L_n\}$.

An apparent disadvantage of Theorem 4.2 is that the rate of convergence is not given in terms of the original objective function $F = f + g$, but in terms of the forward–backward envelope $M^{\mu}_{f,g}$. We now aim to show that the randomized index selection strategy obtains an $O\left(\frac{1}{k^{1.5}}\right)$ rate *in terms of the objective function of the original problem*. This result requires in addition a bounded level set assumption on the original function F and Lipschitz continuity of g over its domain.

Assumption 2
- The level sets of $F = f + g$ are bounded.
- $\text{dom}(g)$ is closed and g Lipschitz continuous over $\text{dom}(g)$ with constant $\ell_g > 0$.

A direct consequence of Assumption 2 (combined with Assumption 1) is that the level sets of $M^{\mu}_{f,g}$ are also bounded and that F is Lipschitz continuous on its level sets intersected with $\text{dom}(g)$.

Lemma 4.1 *Suppose that Assumptions 1 and 2 hold, then*

(i) *the level sets of $M^{\mu}_{f,g}$ are bounded;*

(ii) *F is Lispchitz continuous on $Lev(M^{\mu}_{f,g},\alpha)\cap\text{dom}(g)$ (see Definition in (2.3)) for any $\alpha\in\mathbb{R}$.*

Proof (i) Denote the optimal value of (P) by F_{opt}, then by Assumption 2 the following level set is bounded

$$\text{Lev}(F, F_{\text{opt}}) = \{\mathbf{x} \in \mathbb{R}^n : F(\mathbf{x}) \leq F_{\text{opt}}\} = X^*.$$

Now, since the sets of minimizers of F and $M^{\mu}_{f,g}$ coincide (Theorem 3.1(ii)), we get that $\text{Lev}(M^{\mu}_{f,g}, F_{\text{opt}}) = X^*$, and thus in particular, bounded. Finally, by [35, Corollary 8.7.1], we conclude that *all* the level sets of $M^{\mu}_{f,g}$ are bounded.

(ii) Let $\alpha \in \mathbb{R}$. By Assumption 2, g is Lipschitz continuous with constant ℓ_g over $\text{dom}(g)$. Following part (i) and the fact that $M^{\mu}_{f,g}$ is closed, we get that $\text{Lev}(M^{\mu}_{f,g}, \alpha)$ is compact. Therefore, by the continuity of ∇f and Weierstrass theorem, there exists $\ell_f > 0$ such that $\|\nabla f(\mathbf{x})\|_2 \leq \ell_f$ for any $\mathbf{x} \in \text{Lev}(M^{\mu}_{f,g}, \alpha) \cap \text{dom}(g)$, which implies that F is $\ell_f + \ell_g$-Lipschitz continuous over $\text{Lev}(M^{\mu}_{f,g}, \alpha) \cap \text{dom}(g)$.

□

Given $\mathbf{x}^0 \in \mathbb{R}^n$, a consequence of Assumption 2 and Lemma 4.1(i) is that there exists R such that

$$\max_{\mathbf{x}, \mathbf{x}^* \in \mathbb{R}^n} \left\{ \|\mathbf{x} - \mathbf{x}^*\|_2 : \mathbf{x} \in \text{Lev}(M^{\mu}_{f,g}, M^{\mu}_{f,g}(\mathbf{x}^0)),\ \mathbf{x}^* \in X^* \right\} \leq R. \tag{4.4}$$

Define for any $\delta > 0$ the set

$$S_\delta \equiv \{\mathbf{x} : \|\mathbf{x} - \mathbf{x}^*\|_2 \leq \delta \text{ for all } \mathbf{x}^* \in X^*\}. \tag{4.5}$$

Obviously, by the convexity and compactness of X^*, it follows that S_δ is also convex and compact. Let then R satisfy (4.4). By the monotonicity of the MACGD-FB method, it follows that $M^{\mu}_{f,g}(\mathbf{x}^k) \leq M^{\mu}_{f,g}(\mathbf{x}^0)$ for any k, and thus,

$$\mathbf{x}^k \in S_R \text{ for any } k. \tag{4.6}$$

Theorem 4.3 below states the $O\left(\frac{1}{k^{1.5}}\right)$ convergence result of the sequence of expected values of the original objective function. It requires two results which are fundamental in the analysis of first-order methods. The first one is the descent lemma.

Lemma 4.2 (Descent Lemma [8, Proposition A.24]) *Suppose that a differentiable function $h : \mathbb{R}^n \to \mathbb{R}$ is L-smooth over $\mathbb{R}^n$, meaning that*

$$\|\nabla h(\boldsymbol{x}) - \nabla h(\boldsymbol{y})\|_2 \leq L \|\boldsymbol{x} - \boldsymbol{y}\|_2 \text{ for any } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n.$$

Then

$$h(\boldsymbol{y}) \leq h(\boldsymbol{x}) + \langle \nabla h(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{L}{2} \|\boldsymbol{y} - \boldsymbol{x}\|_2^2 \quad \text{for any } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n.$$

The second result is the so-called second prox theorem that is an important characterization of the proximal operator given in (2.2).

Lemma 4.3 (Second Prox Theorem [3, Theorem 6.39]) *Suppose that* $h : \mathbb{R}^n \to (-\infty, \infty]$ *is a proper closed and convex function. Then,* $\boldsymbol{u} = \text{prox}_h(\boldsymbol{x})$ *if and only if*

$$\langle \boldsymbol{x} - \boldsymbol{u}, \boldsymbol{y} - \boldsymbol{u} \rangle \leq h(\boldsymbol{y}) - h(\boldsymbol{u}) \quad \text{for any } \boldsymbol{y} \in \mathbb{R}^n.$$

The main convergence result now follows. Note that since the sequence generated by the method $\{\mathbf{x}^k\}_{k\geq 0}$ is not necessarily in $\text{dom}(g)$, the rate of convergence is in terms of the projected sequence[2] $\{P_{\text{dom(g)}}(\mathbf{x}^k)\}_{k\geq 0}$.

Theorem 4.3 *Suppose that Assumptions 1 and 2 hold. Let* $\{\mathbf{x}^k\}_{k\geq 0}$ *be the sequence generated by the MACGD-FB method employed on problem (P) with* $\mu \in \left(0, \frac{1}{\lambda_{\max}(\mathbf{M})}\right)$. *Let R satisfy (4.4) and* ℓ_F *be a Lipschitz constant*[3] *of F over* S_{2R}, *where* S_{2R} *is the set given in (4.5) with* $\delta = 2R$. *Then for any* $K \geq 3$ *and* $\mathbf{x}^* \in X^*$,

$$\min_{k=1,2,\ldots,K} \mathbb{E}_{\xi_{k-1}}[F(P_{\text{dom}(g)}(\mathbf{x}^k))] - F(\boldsymbol{x}^*) \leq \frac{\sqrt{2}C}{(K-2)^{1.5}},$$

where $C = (\mu\ell_F + R)\frac{4n^{1.5}L_{\max}R}{1-\mu\lambda_{\max}(\boldsymbol{M})}$ *and* $L_{\max} = \max\{L_1, L_2, \ldots, L_n\}$ *with* $L_1, L_2, \ldots, L_n$ *being coordinate Lipschitz constants of* $\nabla M^{\mu}_{f,g}$.

Proof Theorem 4.2 states the following convergence rate in terms of the envelope function:

$$\mathbb{E}_{\xi_{k-1}}\left[M^{\mu}_{f,g}(\mathbf{x}^k)\right] - F(\mathbf{x}^*) \leq \frac{2n^2 L_{\max} R^2}{(k+1)^2}. \tag{4.7}$$

In what follows, we first lower bound the left-hand side of (4.7) in terms of the gradient mapping, $G^{1/\mu}_{f,g}(\mathbf{x}^k)$ (see (3.3)). Then, we derive an upper bound on $F(P_{\text{dom}(g)}(\mathbf{x}^k)) - F(\mathbf{x}^*)$ in terms of the gradient mapping which will ultimately result in a rate of convergence in terms of the original objective function F.

We start with lower bounding the left-hand side of (4.7) by the gradient mapping. Since $\mathbf{w}^{k+1} = \mathbf{x}^k - \frac{1}{L_{i_k}}\frac{\partial M^{\mu}_{f,g}(\mathbf{x}^k)}{\partial x_{i_k}}\mathbf{e}_{i_k}$, the following sufficient decrease property holds (see 4.2):

$$M^{\mu}_{f,g}(\mathbf{x}^k) - M^{\mu}_{f,g}(\mathbf{w}^{k+1}) \geq \frac{1}{2L_{i_k}}\left(\frac{\partial M^{\mu}_{f,g}(\mathbf{x}^k)}{\partial x_{i_k}}\right)^2,$$

[2] Recall that for a nonempty closed and convex set C, P_C denotes the orthogonal projection operator.

[3] The existence of such a Lipschitz constant is warranted by Lemma 4.1.

which along with the relation $M^{\mu}_{f,g}(\mathbf{x}^{k+1}) \leq M^{\mu}_{f,g}(\mathbf{w}^{k+1})$ implies that

$$M^{\mu}_{f,g}(\mathbf{x}^k) - M^{\mu}_{f,g}(\mathbf{x}^{k+1}) \geq \frac{1}{2L_{i_k}} \left(\frac{\partial M^{\mu}_{f,g}(\mathbf{x}^k)}{\partial x_{i_k}} \right)^2.$$

Taking expectation over the uniformly distributed random variable i_k results in

$$\begin{aligned} M^{\mu}_{f,g}(\mathbf{x}^k) - \mathbb{E}_{i_k}\left[M^{\mu}_{f,g}(\mathbf{x}^{k+1}) \right] &\geq \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2L_i} \left(\frac{\partial M^{\mu}_{f,g}(\mathbf{x}^k)}{\partial x_i} \right)^2 \\ &\geq \frac{1}{2nL_{\max}} \left\| \nabla M^{\mu}_{f,g}(\mathbf{x}^k) \right\|_2^2. \end{aligned} \tag{4.8}$$

Since $\nabla M^{\mu}_{f,g}(\mathbf{x}) = (\mathbf{I} - \mu\mathbf{M}) G^{1/\mu}_{f,g}(\mathbf{x})$ (Theorem 3.1(iii)), we can deduce that

$$\begin{aligned} \left\| \nabla M^{\mu}_{f,g}(\mathbf{x}^k) \right\|_2^2 &= G^{1/\mu}_{f,g}(\mathbf{x}^k)^T (\mathbf{I} - \mu\mathbf{M})^2 G^{1/\mu}_{f,g}(\mathbf{x}^k) \\ &\geq \lambda_{\min}((\mathbf{I} - \mu\mathbf{M})^2) \left\| G^{1/\mu}_{f,g}(\mathbf{x}^k) \right\|_2^2 \\ &= (1 - \mu\lambda_{\max}(\mathbf{M}))^2 \left\| G^{1/\mu}_{f,g}(\mathbf{x}^k) \right\|_2^2, \end{aligned}$$

where the last equality follows by the fact that $\mu < \frac{1}{\lambda_{\max}(\mathbf{M})}$. Combining the above with (4.8) leads to

$$M^{\mu}_{f,g}(\mathbf{x}^k) - \mathbb{E}_{i_k}\left[M^{\mu}_{f,g}(\mathbf{x}^{k+1}) \right] \geq D \left\| G^{1/\mu}_{f,g}(\mathbf{x}^k) \right\|_2^2,$$

where

$$D \equiv \frac{(1 - \mu\lambda_{\max}(\mathbf{M}))^2}{2nL_{\max}}. \tag{4.9}$$

Now, by taking expectation over ξ_{k-1}, we get

$$\mathbb{E}_{\xi_{k-1}}\left[M^{\mu}_{f,g}(\mathbf{x}^k) \right] - \mathbb{E}_{\xi_k}\left[M^{\mu}_{f,g}(\mathbf{x}^{k+1}) \right] \geq D\mathbb{E}_{\xi_{k-1}}\left[\left\| G^{1/\mu}_{f,g}(\mathbf{x}^k) \right\|_2^2 \right],$$

which, for some $\mathbf{x}^* \in X^*$, can be rewritten equivalently as

$$\begin{aligned} \mathbb{E}_{\xi_{k-1}}\left[M^{\mu}_{f,g}(\mathbf{x}^k) \right] - F(\mathbf{x}^*) \geq \mathbb{E}_{\xi_k}\left[M^{\mu}_{f,g}(\mathbf{x}^{k+1}) \right] \\ - F(\mathbf{x}^*) + D\mathbb{E}_{\xi_{k-1}}\left[\left\| G^{1/\mu}_{f,g}(\mathbf{x}^k) \right\|_2^2 \right]. \end{aligned}$$

Let p be a positive integer. Summing the above inequality over $k = p, p+1, \ldots, 2p-1$ and utilizing the fact that $M^{\mu}_{f,g}(\mathbf{x}^k) \geq M^{\mu}_{f,g}(\mathbf{x}^*) = F(\mathbf{x}^*)$ (Theorem 3.1(ii)) result in

$$\begin{aligned}
\mathbb{E}_{\xi_{p-1}}\left[M^{\mu}_{f,g}(\mathbf{x}^p)\right] - F(\mathbf{x}^*) &\geq \mathbb{E}_{\xi_{2p-1}}\left[M^{\mu}_{f,g}(\mathbf{x}^{2p})\right] \\
&\quad -F(\mathbf{x}^*) + D\sum_{k=p}^{2p-1}\mathbb{E}_{\xi_{k-1}}\left[\left\|G^{1/\mu}_{f,g}(\mathbf{x}^k)\right\|_2^2\right] \\
&\geq \mathbb{E}_{\xi_{2p-1}}\left[M^{\mu}_{f,g}(\mathbf{x}^{2p})\right] \\
&\quad -F(\mathbf{x}^*) + Dp\min_{k=p,p+1,\ldots,2p-1}\mathbb{E}_{\xi_{k-1}}\left[\left\|G^{1/\mu}_{f,g}(\mathbf{x}^k)\right\|_2^2\right] \\
&\geq Dp\min_{k=p,p+1,\ldots,2p-1}\mathbb{E}_{\xi_{k-1}}\left[\left\|G^{1/\mu}_{f,g}(\mathbf{x}^k)\right\|_2^2\right].
\end{aligned}$$

The above inequality provides a lower bound on $\mathbb{E}_{\xi_{p-1}}\left[M^{\mu}_{f,g}(\mathbf{x}^p)\right] - F(\mathbf{x}^*)$ in terms of the gradient mapping. Combining it with (4.7) yields

$$\min_{k=p,p+1,\ldots,2p-1}\mathbb{E}\left[\left\|G^{1/\mu}_{f,g}(\mathbf{x}^k)\right\|_2^2\right] \leq \frac{2n^2L_{\max}R^2}{Dp(p+1)^2}. \tag{4.10}$$

In the next step, we aim to upper bound $F(\mathbf{x}^k) - F(\mathbf{x}^*)$ in terms of the gradient mapping as to combine it later on with (4.10). We first recall the definition of the prox-grad operator:

$$T_{\mu^{-1}}(\mathbf{x}) \equiv \mathrm{prox}_{\mu g}(\mathbf{x} - \mu\nabla f(\mathbf{x})).$$

Then obviously

$$G^{1/\mu}_{f,g}(\mathbf{x}) = \frac{1}{\mu}(\mathbf{x} - T_{\mu^{-1}}(\mathbf{x})). \tag{4.11}$$

We first note that for any k, it holds that $T_{\mu^{-1}}(\mathbf{x}^k) \in S_{2R}$. This is due to the following chain of equalities and inequalities that holds for any $\mathbf{x}^* \in X$:

$$\begin{aligned}
\|T_{\mu^{-1}}(\mathbf{x}^k) - \mathbf{x}^*\|_2 &= \|T_{\mu^{-1}}(\mathbf{x}^k) - T_{\mu^{-1}}(\mathbf{x}^*)\|_2 \\
&= \|\mathrm{prox}_{\mu g}(\mathbf{x}^k - \mu\nabla f(\mathbf{x}^k)) - \mathrm{prox}_{\mu g}(\mathbf{x}^* - \mu\nabla f(\mathbf{x}^*))\|_2 \\
&\overset{(*)}{\leq} \|\mathbf{x}^k - \mu\nabla f(\mathbf{x}^k) - \mathbf{x}^* + \mu\nabla f(\mathbf{x}^*)\|_2 \\
&\leq \|\mathbf{x}^k - \mathbf{x}^*\|_2 + \mu\|\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^*)\|_2 \\
&\overset{(**)}{\leq} 2\|\mathbf{x}^k - \mathbf{x}^*\|_2 \\
&\overset{(***)}{\leq} 2R,
\end{aligned}$$

where (*) is due to the nonexpansiveness property of the prox operator [3, Theorem 6.42], (**) follows by the fact that ∇f is Lipschitz with constant $\lambda_{\max}(\mathbf{M})$ and hence Lipschitz with constant $\frac{1}{\mu} > \lambda_{\max}(\mathbf{M})$. Finally, (***) is a consequence of the relation $\mathbf{x}^k \in S_R$ (see (4.6)).

Now, since ℓ_F is a Lipschitz constant of F over S_{2R}, the following holds:

$$\begin{aligned} F(P_{\mathrm{dom}(g)}(\mathbf{x}^k)) - F(\mathbf{x}^*) &= F(P_{\mathrm{dom}(g)}(\mathbf{x}^k)) - F(T_{\mu^{-1}}(\mathbf{x}^k)) + F(T_{\mu^{-1}}(\mathbf{x}^k)) - F(\mathbf{x}^*) \\ &\leq \ell_F \left\| P_{\mathrm{dom}(g)}(\mathbf{x}^k) - T_{\mu^{-1}}(\mathbf{x}^k) \right\|_2 + F(T_{\mu^{-1}}(\mathbf{x}^k)) - F(\mathbf{x}^*) \\ &\leq \ell_F \left\| \mathbf{x}^k - T_{\mu^{-1}}(\mathbf{x}^k) \right\|_2 + F(T_{\mu^{-1}}(\mathbf{x}^k)) - F(\mathbf{x}^*) \\ &\overset{(4.11)}{=} \mu \ell_F \left\| G^{1/\mu}_{f,g}(\mathbf{x}^k) \right\|_2 + F(T_{\mu^{-1}}(\mathbf{x}^k)) - F(\mathbf{x}^*), \end{aligned} \tag{4.12}$$

where the nonexpansivity of the orthogonal projection operator [3, Theorem 5.4] and the fact that $T_{\mu^{-1}}(\mathbf{x}^k) \in \mathrm{dom}(g)$ were used in the second inequality. To bound $F(T_{\mu^{-1}}(\mathbf{x}^k)) - F(\mathbf{x}^*)$, we first note that since f is $\frac{1}{\mu}$-smooth, it follows by the descent lemma (Lemma 4.2) that

$$\begin{aligned} f(T_{\mu^{-1}}(\mathbf{x}^k)) - f(\mathbf{x}^*) &\leq f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), T_{\mu^{-1}}(\mathbf{x}^k) - \mathbf{x}^k \rangle + \frac{1}{2\mu} \left\| T_{\mu^{-1}}(\mathbf{x}^k) - \mathbf{x}^k \right\|_2^2 - f(\mathbf{x}^*) \\ &\leq \langle \nabla f(\mathbf{x}^k), T_{\mu^{-1}}(\mathbf{x}^k) - \mathbf{x}^* \rangle + \frac{1}{2\mu} \left\| T_{\mu^{-1}}(\mathbf{x}^k) - \mathbf{x}^k \right\|_2^2, \end{aligned} \tag{4.13}$$

where in the last inequality we used the fact that since f is convex, then $f(\mathbf{x}^k) - f(\mathbf{x}^*) \leq \langle \nabla f(\mathbf{x}^k), \mathbf{x}^k - \mathbf{x}^* \rangle$. Noting that $T_{\mu^{-1}}(\mathbf{x}^k) = \mathrm{prox}_{\mu g}(\mathbf{x}^k - \mu \nabla f(\mathbf{x}^k))$, then invoking the second prox theorem (Lemma 4.3) with $h = \mu g$, $\mathbf{x} = \mathbf{x}^k - \mu \nabla f(\mathbf{x}^k)$, $\mathbf{u} = T_{\mu^{-1}}(\mathbf{x}^k)$ and $\mathbf{y} = \mathbf{x}^*$, we have

$$g(T_{\mu^{-1}}(\mathbf{x}^k)) - g(\mathbf{x}^*) \leq \frac{1}{\mu} \left\langle \mathbf{x}^k - \mu \nabla f(\mathbf{x}^k) - T_{\mu^{-1}}(\mathbf{x}^k), T_{\mu^{-1}}(\mathbf{x}^k) - \mathbf{x}^* \right\rangle. \tag{4.14}$$

Combining (4.13) and (4.14), and recalling that $G^{1/\mu}_{f,g}(\mathbf{x}) = \mu^{-1}(\mathbf{x} - T_{\mu^{-1}}(\mathbf{x}))$, we obtain that

$$\begin{aligned} F(T_{\mu^{-1}}(\mathbf{x}^k)) - F(\mathbf{x}^*) &= f(T_{\mu^{-1}}(\mathbf{x}^k)) + g(T_{\mu^{-1}}(\mathbf{x}^k)) - f(\mathbf{x}^*) - g(\mathbf{x}^*) \\ &\leq \frac{1}{\mu} \left\langle \mathbf{x}^k - T_{\mu^{-1}}(\mathbf{x}^k), T_{\mu^{-1}}(\mathbf{x}^k) - \mathbf{x}^* \right\rangle + \frac{1}{2\mu} \left\| T_{\mu^{-1}}(\mathbf{x}^k) - \mathbf{x}^k \right\|_2^2 \\ &= \left\langle G^{1/\mu}_{f,g}(\mathbf{x}^k), T_{\mu^{-1}}(\mathbf{x}^k) - \mathbf{x}^k + \mathbf{x}^k - \mathbf{x}^* \right\rangle + \frac{\mu}{2} \left\| G^{1/\mu}_{f,g}(\mathbf{x}^k) \right\|_2^2 \\ &= \left\langle G^{1/\mu}_{f,g}(\mathbf{x}^k), \mathbf{x}^k - \mathbf{x}^* \right\rangle - \frac{\mu}{2} \left\| G^{1/\mu}_{f,g}(\mathbf{x}^k) \right\|_2^2 \\ &\leq \left\| G^{1/\mu}_{f,g}(\mathbf{x}^k) \right\|_2 \left\| \mathbf{x}^k - \mathbf{x}^* \right\|_2 \\ &\leq R \left\| G^{1/\mu}_{f,g}(\mathbf{x}^k) \right\|_2, \end{aligned}$$

where the last inequality follows by (4.6). Utilizing the above and the inequality in (4.12) implies

$$F(P_{\mathrm{dom}(g)}(\mathbf{x}^k)) - F(\mathbf{x}^*) \leq (\mu\ell_F + R)\left\|G^{1/\mu}_{f,g}(\mathbf{x}^k)\right\|_2.$$

Hence, squaring and taking expectation with respect to ξ_{k-1} results in

$$\mathbb{E}_{\xi_{k-1}}\left[(F(P_{\mathrm{dom}(g)}(\mathbf{x}^k)) - F(\mathbf{x}^*))^2\right] \leq (\mu\ell_F + R)^2\,\mathbb{E}_{\xi_{k-1}}\left[\left\|G^{1/\mu}_{f,g}(\mathbf{x}^k)\right\|_2^2\right], \quad (4.15)$$

which is the desired upper bound on $\mathbb{E}_{\xi_{k-1}}\left[(F(P_{\mathrm{dom}(g)}(\mathbf{x}^k)) - F(\mathbf{x}^*))^2\right]$ in terms of the gradient mapping.

Combining (4.10) with (4.15) implies that the following holds for any $p \geq 1$:

$$\begin{aligned}
&\min_{k=1,2,\ldots,2p-1}\mathbb{E}_{\xi_{k-1}}\left[(F(P_{\mathrm{dom}(g)}(\mathbf{x}^k)) - F(\mathbf{x}^*))^2\right]\\
&\leq \min_{k=p,\ldots,2p-1}\mathbb{E}_{\xi_{k-1}}\left[(F(P_{\mathrm{dom}(g)}(\mathbf{x}^k)) - F(\mathbf{x}^*))^2\right]\\
&\leq (\mu\ell_F + R)^2\min_{k=p,\ldots,2p-1}\mathbb{E}_{\xi_{k-1}}\left[\left\|G^{1/\mu}_{f,g}(\mathbf{x}^k)\right\|_2^2\right]\\
&\overset{(4.10)}{\leq} (\mu\ell_F + R)^2\,\frac{2n^2L_{\max}R^2}{Dp(p+1)^2}.
\end{aligned}$$

Using the inequality $\mathbb{E}[Y]^2 \leq \mathbb{E}[Y^2]$ for any random vector Y, taking the square root of both sides, recalling the definition of D (see 4.9), and using the inequality $p(p+1)^2 \geq p^3$ for all $p \geq 1$ yield

$$\begin{aligned}
\min_{k=1,2,\ldots,2p-1}\mathbb{E}_{\xi_{k-1}}\left[F(P_{\mathrm{dom}(g)}(\mathbf{x}^k))\right] - F(\mathbf{x}^*) &\leq (\mu\ell_F + R)\,\frac{\sqrt{2L_{\max}}nR}{\sqrt{D}\sqrt{p^3}}\\
&= \underbrace{(\mu\ell_F + R)\,\frac{4n^{1.5}L_{\max}R}{1-\mu\lambda_{\max}(\mathbf{M})}}_{C}\cdot\frac{1}{2p^{1.5}}\\
&= \frac{C}{2p^{1.5}}.
\end{aligned}$$

Finally, let $K \geq 3$ be a positive integer, then by the trivial inequatlity $K \geq 2\lfloor K/2\rfloor - 1$, it follows that

$$\begin{aligned}
&\min_{k=1,2,\ldots,K}\mathbb{E}_{\xi_{k-1}}\left[F(P_{\mathrm{dom}(g)}(\mathbf{x}^k))\right] - F(\mathbf{x}^*)\\
&\leq \min_{k=1,2,\ldots,2\lfloor K/2\rfloor-1}\mathbb{E}_{\xi_{k-1}}\left[F(P_{\mathrm{dom}(g)}(\mathbf{x}^k))\right] - F(\mathbf{x}^*)\\
&\leq \frac{C}{2\lfloor K/2\rfloor^{1.5}} \overset{\lfloor K/2\rfloor\geq K/2-1}{\leq} \frac{\sqrt{2}C}{(K-2)^{1.5}}.
\end{aligned}$$

□

4.3 Smoothing Parameter and Coordinate Lipschitz Constants

In several large-scale settings, choosing the smoothing parameter μ by computing $\lambda_{\max}(\mathbf{M})$ exactly is an inapplicable task. In this part, we offer an alternative scheme that computes μ during the execution of the MACGD-FB algorithm. For this goal, we provide a lower bound on $M^{\mu}_{f,g}$ that must be violated for large enough μ.

Theorem 4.4 *Let f and g satisfy Assumption 1 and suppose that in addition g is nonnegative.*

(a) *If $\mu \in \left(0, \frac{1}{\lambda_{\max}(\boldsymbol{M})}\right)$, then, for all $\boldsymbol{x} \in \mathbb{R}^n$,*

$$M^{\mu}_{f,g}(\boldsymbol{x}) \geq \phi_{\mu}(\boldsymbol{x}), \tag{4.16}$$

where

$$\phi_{\mu}(\boldsymbol{x}) \equiv \boldsymbol{b}^T(\boldsymbol{I} - \mu \boldsymbol{M})\boldsymbol{x} - \frac{\mu}{2}\|\boldsymbol{b}\|_2^2.$$

(b) *If $\mu > \frac{1}{\lambda_{\max}(\boldsymbol{M})}$ and $\boldsymbol{M} \neq \boldsymbol{0}$, then,*

$$\min_{\boldsymbol{x}\in\mathbb{R}^n}\{M^{\mu}_{f,g}(\boldsymbol{x}) - \phi_{\mu}(\boldsymbol{x})\} = -\infty. \tag{4.17}$$

Proof (a) Since $\mu < \frac{1}{\lambda_{\max}(\mathbf{M})}$ which results in that $\mathbf{M} - \mu\mathbf{M}^2 \succeq \mathbf{0}$, and since the nonnegativity of g implies the nonnegativity of M^{μ}_{g}, we can lower bound the composite envelope function as follows:

$$\begin{aligned} M^{\mu}_{f,g}(\mathbf{x}) &= \frac{1}{2}\mathbf{x}^T(\mathbf{M} - \mu\mathbf{M}^2)\mathbf{x} + \mathbf{b}^T(\mathbf{I} - \mu\mathbf{M})\mathbf{x} + M^{\mu}_{g}((\mathbf{I} - \mu\mathbf{M})\mathbf{x} - \mu\mathbf{b}) - \frac{\mu}{2}\|\mathbf{b}\|_2^2 \\ &\geq \mathbf{b}^T(\mathbf{I} - \mu\mathbf{M})\mathbf{x} - \frac{\mu}{2}\|\mathbf{b}\|_2^2 = \phi_{\mu}(\mathbf{x}), \end{aligned}$$

Therefore, if $\mu \in \left(0, \frac{1}{\lambda_{\max}(\mathbf{M})}\right)$ then $M^{\mu}_{f,g}(\mathbf{x}) \geq \phi_{\mu}(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^n$.

(b) Suppose that $\mu > \frac{1}{\lambda_{\max}(\mathbf{M})}$ and let $\lambda = \lambda_{\max}(\mathbf{M})$. Then, there exists a normalized eigenvector $\mathbf{v}$ such that $\mathbf{M}\mathbf{v} = \lambda\mathbf{v}$. Then

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^n}\{M^{\mu}_{f,g}(\mathbf{x}) - \phi_{\mu}(\mathbf{x})\} &\leq \min_{\mathbf{x}=\alpha\mathbf{v},\alpha\in\mathbb{R}}\{M^{\mu}_{f,g}(\mathbf{x}) - \phi_{\mu}(\mathbf{x})\} \\ &= \min_{\alpha\in\mathbb{R}}\left\{(\lambda - \mu\lambda^2)\frac{\alpha^2}{2} + M^{\mu}_{g}\left(\alpha(1-\mu\lambda)\mathbf{v} - \mu\mathbf{b}\right)\right\} \end{aligned} \tag{4.18}$$

Take $\mathbf{x}_0 \in \mathrm{dom}(g)$ (whose existence is guaranteed by the properness of g). Then for any $\mathbf{x} \in \mathbb{R}^n$,

$$M^{\mu}_{g}(\mathbf{x}) = \min_{\mathbf{u}\in\mathbb{R}^n}\left\{g(\mathbf{u}) + \frac{1}{2\mu}\|\mathbf{u} - \mathbf{x}\|_2^2\right\} \leq g(\mathbf{x}_0) + \frac{1}{2\mu}\|\mathbf{x} - \mathbf{x}_0\|_2^2.$$

Utilizing the above inequality with $\mathbf{x} = \alpha(1-\mu\lambda)\mathbf{v} - \mu\mathbf{b}$ in (4.18), we obtain that

$$\begin{aligned}&\min_{\mathbf{x}\in\mathbb{R}^n}\left\{M^{\mu}_{f,g}(\mathbf{x}) - \phi_{\mu}(\mathbf{x})\right\}\\ &\leq \min_{\alpha\in\mathbb{R}}\left\{h(\alpha) \equiv (\lambda - \mu\lambda^2)\frac{\alpha^2}{2} + \frac{1}{2\mu}\left\|\alpha(1-\mu\lambda)\mathbf{v} - \mu\mathbf{b} - \mathbf{x}_0\right\|_2^2 + g(\mathbf{x}_0)\right\}.\end{aligned}$$

The function h is a one-dimensional quadratic function in α where the coefficient of α^2 is $\frac{1}{2}\left((\lambda - \mu\lambda^2) + \frac{(1-\mu\lambda)^2}{\mu}\right)$. The minimal value of h is guaranteed to be $-\infty$ if the coefficient is negative, meaning if

$$(\lambda - \mu\lambda^2) + \frac{(1-\mu\lambda)^2}{\mu} < 0,$$

which after some simple arrangement can be seen to be the same as

$$\frac{1-\mu\lambda}{\mu} < 0,$$

a valid inequality as we assume here that $\mu > \frac{1}{\lambda}$.

□

Theorem 4.4 indicates that $M^{\mu}_{f,g}(\mathbf{x})$ is lower bounded by $\phi_{\mu}(\mathbf{x})$; nevertheless, this is guaranteed only if μ is smaller than $1/\lambda_{\max}(\mathbf{M})$. As revealed by (4.17), if $\mu > 1/\lambda_{\max}(\mathbf{M})$, then this lower bound must be violated.

Remark 4.3 (Lower Boundedness of g) The nonnegativity condition on g is quite common in applications, and in fact, holds for all the examples in Remark 4.2. That said, Theorem 4.4 and our algorithm are applicable also for cases in which g is lower bounded by some constant C_g, and in these cases, $\phi_{\mu}(\mathbf{x})$ is modified to $\phi_{\mu}(\mathbf{x}) \equiv \mathbf{b}^T(\mathbf{I} - \mu\mathbf{M})\mathbf{x} - \frac{\mu}{2}\|\mathbf{b}\|_2^2 - C_g$.

To avoid calculating the exact coordinate Lipschitz constants $\{L_i\}_{i=1}^n$, we backtrack them as to provide a sufficient descent in the objective function. Specifically, for $\mathbf{x}^{k+1} = \mathbf{x}^k - \frac{1}{L_i}\frac{\partial M^{\mu}_{f,g}(\mathbf{x}^k)}{\partial x_i}$, the following holds:

$$M^{\mu}_{f,g}(\mathbf{x}^k) - M^{\mu}_{f,g}(\mathbf{x}^{k+1}) \geq \frac{1}{2L_i}\left(\frac{\partial M^{\mu}_{f,g}(\mathbf{x}^k)}{\partial x_i}\right)^2, \quad \text{for all } i \in \{1, 2, \ldots, n\}. \tag{4.19}$$

Based on Theorem 4.4 and inequality (4.19), we derive Algorithm 3 below which does not require pre-calculations, and is a variant of the MACGD-FB method with backtracking. We initialize the smoothing parameter $\mu > 0$ and set the coordinate Lipschitz constants $L_i = \alpha/\mu$ with some $\alpha > 0$. To compute the coordinate Lipschitz

constants, we adopt a backtracking procedure (step 8) that guarantees the validity of (4.19). As to the smoothing parameter μ, we decrease it either if the lower bound condition in (4.16) fails for the current iterates $\mathbf{x} = \{\tilde{\mathbf{x}}^{k+1}, \mathbf{y}^k, \mathbf{w}^{k+1}\}$ (step 9), or if the decrease condition (inequality (4.19)) does not hold while $L_i \geq \frac{1}{\mu}$. We note that our numerical results appearing in Sect. 5 show that practically, Algorithm 3 finds a smoothing parameter μ that is smaller than $1/\lambda_{\max}(\mathbf{M})$ after at most two mega iterations (mega iteration = n iterations of a coordinate descent-type algorithm).

5 Numerical Experiments

To demonstrate the effectiveness of our approach, we performed extensive numerical experiments, where our goal was twofold. First, we examined the MACGD-FB method (with backtracking—Algorithm 3) on a variety of synthetic data scenarios, considering common hard constraints that are nonsmooth and non-separable. We compared the performance of the algorithm with several state-of-the-art alternatives. Second, we performed a natural image inpainting experiment, showcasing our method on the two-dimensional total-variation regularization.

5.1 Synthetic Experiments

In our synthetic experiments, we consider problem (P) with the following instances:[4]

1. *Affine set* We examine a least-squares problem under an affine set constraint:

$$\min_{\mathbf{x}} \left\{ \frac{1}{2} \|\mathbf{f} - \mathbf{A}\mathbf{x}\|_2^2, \text{ s.t. } \mathbf{D}\mathbf{x} = \mathbf{c} \right\},$$

   where $\mathbf{A} \in \mathbb{R}^{120\times 100}$, $\mathbf{f} \in \mathbb{R}^{120}$, $\mathbf{D} \in \mathbb{R}^{70\times 100}$, $\mathbf{c} \in \mathbb{R}^{70}$ are all generated i.i.d. from a Gaussian distribution with $A_{i,j}$, $f_i \sim \mathcal{N}(0, \frac{1}{120})$, $D_{i,j} \sim \mathcal{N}(0, \frac{1}{100})$, and $c_i \sim \mathcal{N}(0, \frac{1}{70})$.
2. ℓ_1*-ball* We study the hard-constrained Lasso ([22]) for the same parameters above and $R_0 = \frac{1}{2}$:

$$\min_{\mathbf{x}} \left\{ \frac{1}{2} \|\mathbf{f} - \mathbf{A}\mathbf{x}\|_2^2, \text{ s.t. } \|\mathbf{x}\|_1 \leq R_0 \right\},$$

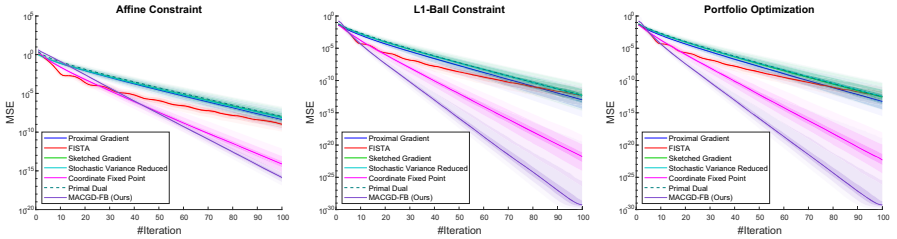
3. *Intersection of a hyperplane and a box* We explore the Markowitz portfolio optimization problem [30] of the form of

$$\min_{\mathbf{x}} \left\{ \frac{1}{2} \mathbf{x}^T \Sigma \mathbf{x} - {}^T\mathbf{x}, \text{ s.t. } \mathbf{1}^T\mathbf{x} = 1, \mathbf{x} \geq \mathbf{0} \right\},$$

[4] We describe below how to reproduce these synthetic datasets, and they are available from the authors on reasonable request.

Algorithm 3: MACGD-FB with Backtracking

Initialization: Pick $\mathbf{x}^0 \in \mathbb{R}^n$, a smoothing parameter $\mu > 0$, some $0 < \alpha < 1$, and backtracking parameters $\gamma_\mu < 1, \gamma_L > 1$.
Set $L_1 = L_2 = \cdots = L_n = \frac{\alpha}{\mu}$, $\mathbf{z}^0 = \mathbf{x}^0$, and $\theta^0 = 1$.
General Step: For any $k = 0, 1, 2, \ldots$ execute the following steps:
1. Pick $i_k \in \{1, 2, \ldots, n\}$.
2. Define $\mathbf{y}^k = (1 - \theta^k)\mathbf{x}^k + \theta^k \mathbf{z}^k$.
3. Set $s = \frac{\partial M^\mu_{f,g}(\mathbf{y}^k)}{\partial y_{i_k}} = \frac{1}{\mu}(\mathbf{e}^T_{i_k} - \mu \mathbf{m}^T_{i_k})(\mathbf{y}^k - \text{prox}_{\mu g}(\mathbf{y}^k - \mu(\mathbf{M}\mathbf{y}^k + \mathbf{b})))$.
4. Set $\tilde{\mathbf{x}}^{k+1} = \mathbf{y}^k - \frac{s}{L_{i_k}}\mathbf{e}_{i_k}$, and $\mathbf{z}^{k+1} = \mathbf{z}^k - \frac{s}{n\theta^k L_{i_k}}\mathbf{e}_{i_k}$.
5. Set $r = \frac{\partial M^\mu_{f,g}(\mathbf{x}^k)}{\partial x_{i_k}} = \frac{1}{\mu}(\mathbf{e}^T_{i_k} - \mu \mathbf{m}^T_{i_k})(\mathbf{x}^k - \text{prox}_{\mu g}(\mathbf{x}^k - \mu(\mathbf{M}\mathbf{x}^k + \mathbf{b})))$.
6. Set $\mathbf{w}^{k+1} = \mathbf{x}^k - \frac{r}{L_{i_k}}\mathbf{e}_{i_k}$.
7. Update $\theta^{k+1} = \frac{\sqrt{(\theta^k)^4 + 4(\theta^k)^2} - (\theta^k)^2}{2}$.
8. While $M^\mu_{f,g}(\tilde{\mathbf{x}}^{k+1}) > M^\mu_{f,g}(\mathbf{y}^k) - \frac{1}{2L_{i_k}}s^2$ or $M^\mu_{f,g}(\mathbf{w}^{k+1}) > M^\mu_{f,g}(\mathbf{x}^k) - \frac{1}{2L_{i_k}}r^2$:
   i. If $L_{i_k} \geq \frac{1}{\mu}$:
      - Set $\mu = \gamma_\mu \mu$, $L_1 = L_2 = \cdots = L_n = \frac{\alpha}{\mu}$.
      - Set $\mathbf{x}^{k+1} = \mathbf{x}^0$ and $\mathbf{z}^{k+1} = \mathbf{x}^{k+1}$.
      - Move to Step 1.
   ii Else:
      - Set $L_{i_k} = \gamma_L L_{i_k}$.
      -Set $\tilde{\mathbf{x}}^{k+1} = \mathbf{y}^k - \frac{s}{L_{i_k}}\mathbf{e}_{i_k}$, $\mathbf{z}^{k+1} = \mathbf{z}^k - \frac{s}{n\theta^k L_{i_k}}\mathbf{e}_{i_k}$, and $\mathbf{w}^{k+1} = \mathbf{x}^k - \frac{r}{L_{i_k}}\mathbf{e}_{i_k}$.
9. If $M^\mu_{f,g}(\tilde{\mathbf{x}}^{k+1}) < \phi_\mu(\tilde{\mathbf{x}}^{k+1})$ or $M^\mu_{f,g}(\mathbf{y}^k) < \phi_\mu(\mathbf{y}^k)$ or $M^\mu_{f,g}(\mathbf{w}^{k+1}) < \phi_\mu(\mathbf{w}^{k+1})$:
   - Set $\mu = \gamma_\mu \mu$, $L_1 = L_2 = \cdots = L_n = \frac{\alpha}{\mu}$.
   - Set $\mathbf{x}^{k+1} = \mathbf{x}^0$ and $\mathbf{z}^{k+1} = \mathbf{x}^{k+1}$.
   - Move to Step 1.
10. Set $\mathbf{x}^{k+1} = \text{argmin}\left\{M^\mu_{f,g}(\mathbf{x}) : \mathbf{x} \in \{\tilde{\mathbf{x}}^{k+1}, \mathbf{w}^{k+1}\}\right\}$



(a) An affine set. (b) ℓ_1-ball. (c) Portfolio optimization.

Fig. 2 A comparison between current coordinate descent-type algorithms and the proposed monotone accelerated coordinate gradient descent algorithm (Algorithm 3). The central line is the median over the 1000 runs and the ribbons show 90%, 75%, 60%, 40%, 25%, and 10% quantiles

where $\in \mathbb{R}^{100}$, $\Sigma = \mathbf{H}^T\mathbf{H}$, $\mathbf{H} \in \mathbb{R}^{100\times 100}$, with $\alpha_i \sim \mathcal{N}(0, \frac{1}{100})$ and $H_{i,j} \sim \mathcal{N}(0, \frac{1}{100})$. Apart from finance applications, this type of constraint is also used in the formulation of the dual support vector machine problem (SVM) [10].

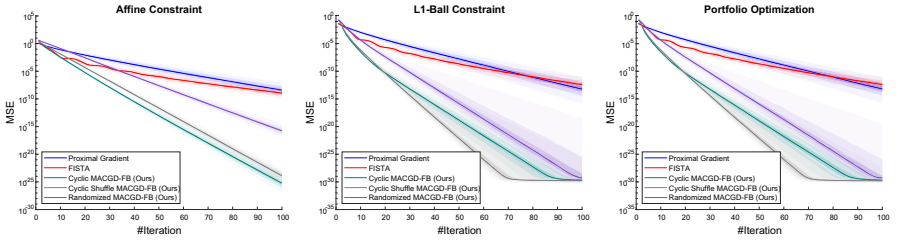
We compare our method to non-coordinate descent-type algorithms employed directly on problem (P), as well as several coordinate descent-type algorithms. Specifically, we consider the following methods:

- Proximal gradient (also known as ISTA) [12].
- Fast iterative shrinkage thresholding (FISTA) [5].
- Sketched gradient algorithm (SEGA) [20].
- Stochastic variance reduced coordinate descent (SVRCD) [21].
- Coordinate fixed point algorithm [11,27], which is similar to the coordinate descent primal–dual algorithm [14] with $h = 0$, using their notations.
- Coordinate descent primal–dual algorithm [14] with $h = \|\cdot\|_2^2$.
- The proposed monotone accelerated coordinate gradient descent forward–backward (MACGD-FB) with backtracking, as detailed in Algorithm 3.

For proximal gradient, FISTA and coordinate fixed point algorithms, we use a step size of $1/\lambda_{\max}(\mathbf{M})$. In the sketched gradient and stochastic variance reduced coordinate descent, we use $\alpha = 1/(n\lambda_{\max}(\mathbf{M}))$, $p_i = 1/n$, and $\theta = n$. In the coordinate descent primal–dual algorithm with $h = \|\cdot\|_2^2$, we use $\sigma_i = \mathbf{1}$, $\tau_i = 1/\lambda_{\max}(\mathbf{M})$, and $\pi_i = 1/n$ for all coordinates. Note that all these algorithms require the spectral information of $\mathbf{M}$.

In our proposed MAFGD-FB method, we backtrack the algorithm parameters and it is thus free of computing $\lambda_{\max}(\mathbf{M})$ and the coordinate Lipschitz constants. We initialize $\mathbf{x}^0 = \mathbf{0}$, $\mu = 0.9$, and $L_i = 0.1/\mu$, and use the parameters of $\gamma_\mu = 0.5$ and $\gamma_L = 1.5$ for the smoothing parameter and the Lipschitz constants, respectively.

Figure 2 presents the mean squared error between the methods' output and the optimal solution of problem (P) obtained by running FISTA directly on problem (P) for 1000 iterations. For a fair comparison, we counted every n iterations of the coordinate descent-type algorithms as one iteration, thus making the computational effort per iteration of each method similar. As shown in Fig. 2, the MACGD-FB algorithm starts with larger errors, a natural artifact of backtracking the algorithm parameters; however,



(a) An affine set. (b) ℓ_1-ball. (c) Portfolio optimization.

Fig. 3 MSE performance of three types of index selection strategies for the MACGD-FB algorithm (Algorithm 3)

after few iterations, it outperforms current algorithms. An empirical check reveals that out of 1000 runs of the MACGD-FB for the portfolio optimization task, 796 finished updating μ in the first mega iteration (n iterations of CGD), and the rest finished in the second iteration.

Till now, we have only discussed a randomized index selection; nevertheless, the literature contains several options on how to choose the index i_k. In Fig. 3, we examine empirically three of them:

- *Cyclic* $i_k = (k \mod n) + 1$.
- *Randomized* At each iteration i_k is randomly generated from a uniform distribution on $U\{1, 2, \ldots, n\}$.
- *Cyclic shuffle* At the beginning of each batch of n iterations, the order of the chosen indices in the next n iterations is picked by a random permutation.

As shown in Fig. 3, the cyclic and the cyclic-shuffle variants have a clear advantage over the randomized version in our simulations. This observation, however, is purely empirical, and yet there is no theoretical explanation for this phenomenon.

5.2 Total-Variation

We demonstrate our method on a natural image inpainting task using the two-dimensional ℓ_1 total-variation regularization [2]. We evaluate the performance of proximal gradient and MACGD-FB for 20 iterations on images from the popular `Set11`,[5] corrupted with the noise of 10dB SNR and with 50% missing pixels. In this case, $\mathbf{M}$ is a diagonal matrix with binary values indicating if a pixel is missing, and therefore, $\lambda_{\max}(\mathbf{M}) = 1$ and does not need to be estimated. For the Lipschitz constants estimation, we use a learning rate of $\gamma_L = 1.2$ and initializations of $L = 0.6/\mu$ and $\mathbf{x}^0 = \mathbf{0}$. The index selection strategy in this experiment is the cyclic shuffle. As shown in Figs. 4 and 5, the MACGD-FB algorithm that alternates a single pixel per iteration achieves improved results and is especially efficient in highly corrupted areas where many pixels are missing. Moreover, in Table 1 we present the PSNR results of our method while changing a patch of 8×8 at every iteration, showing a clear advantage over the proximal gradient alternative.

[5] This standard dataset is available in https://github.com/aaberdam/AdaLISTA.

Table 1 Image inpainting task with SNR of 10dB and 50% missing pixels for 20 iterations, comparing PSNR of proximal gradient and MACGD-FB (Algorithm 3) with working set of 8×8 patches

	Barbara	Boat	House	Lena	Peppers	C.man	Couple	Finger	Hill	Man	Montage	Average
Proximal gradient	24.378	26.857	27.524	28.511	25.908	24.656	27.181	22.54	27.361	27.687	23.187	25.981
MACGD-FB	24.838	27.6	30.272	29.905	26.827	25.327	27.694	24.101	28.81	28.221	24.424	27.093

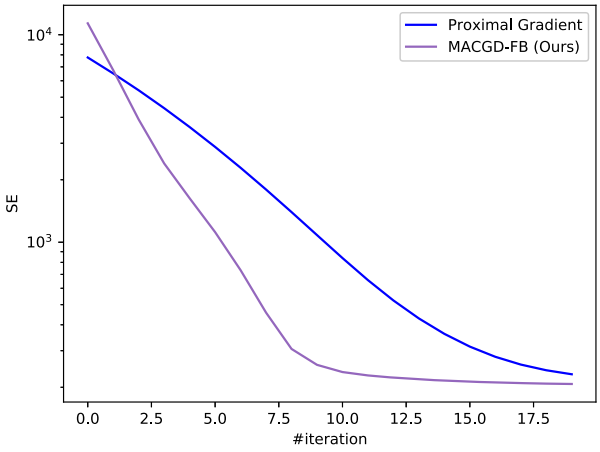


Fig. 4 TV convergence rate

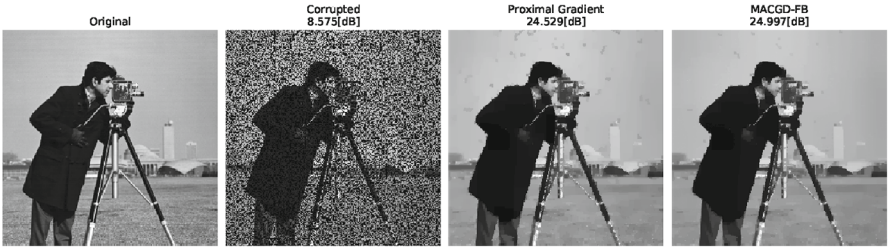


Fig. 5 Image inpainting with 10 dB SNR and 50% missing pixels. From left to right: original image, corrupted image (8.587 dB), proximal gradient (24.529 dB), and MACGD-FB (24.997 dB)

6 Conclusions

In this work, we utilized the forward–backward envelope to develop a monotone accelerated coordinate gradient decent algorithm for problem (P) which is nonsmooth and non-separable. Our scheme achieves a convergence rate of $O(1/k^{1.5})$ which improves current coordinate descent-types methods. We further suggested a backtracking variant of our algorithm which is free of computing the spectral information and coordinate Lipschitz constants of the problem. As demonstrated through an extensive numerical study, our method outperforms current coordinate descent-type methods in various settings.

Acknowledgements The research of A. Beck is supported by the ISF Grant 926-21. A. Aberdam thanks the Azrieli foundation for providing additional research support. The authors would like to thank two anonymous reviewers for their valuable suggestions that improved the final manuscript.

Appendix A: Proof of Theorem 4.1

Throughout the proof, we use the notation: $\mathbf{L} = \mathrm{diag}(\{L_i\}_{i=1}^n)$ and denote the $\mathbf{L}$-norm and the $\mathbf{L}$-inner product by

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{L}} \equiv \sum_{i=1}^n L_i x_i y_i; \ \ \|\mathbf{x}\|_{\mathbf{L}} \equiv \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_{\mathbf{L}}} = \sqrt{\sum_{i=1}^n L_i x_i^2}.$$

By the definition of step 4 and the block descent lemma [3, Lemma 11.8], it follows that

$$H(\tilde{\mathbf{x}}^{k+1}) \leq H(\mathbf{y}^k) + \nabla_{i_k} H(\mathbf{y}^k)(\tilde{x}_{i_k}^{k+1} - y_{i_k}^k) + \frac{L_{i_k}}{2}(\tilde{x}_{i_k}^{k+1} - y_{i_k}^k)^2$$

Taking the expectation with respect to i_k, and recalling that $\tilde{\mathbf{x}}^{k+1} = \mathbf{y}^k - \frac{1}{L_{i_k}} \nabla_{i_k} H(\mathbf{y}^k)\mathbf{e}_{i_k}$, we obtain

$$\mathbb{E}_{i_k} H(\tilde{\mathbf{x}}^{k+1}) \leq H(\mathbf{y}^k) + \nabla H(\mathbf{y}^k)^T(\mathbf{s}^{k+1} - \mathbf{y}^k) + \frac{n}{2}\|\mathbf{s}^{k+1} - \mathbf{y}^k\|_{\mathbf{L}}^2, \tag{7.1}$$

where $\mathbf{s}^{k+1} = \mathbf{y}^k - \frac{1}{n}\mathbf{L}^{-1}\nabla H(\mathbf{y}^k)$. Define

$$\begin{aligned} \mathbf{t}^{k+1} &\equiv \mathbf{z}^k - \frac{1}{n\theta^k}\mathbf{L}^{-1}\nabla H(\mathbf{y}^k) \\ &= \operatorname*{argmin}_{\mathbf{y}} \left\{ \nabla H(\mathbf{y}^k)^T(\mathbf{y} - \mathbf{z}^k) + \frac{n\theta^k}{2}\|\mathbf{y} - \mathbf{z}^k\|_{\mathbf{L}}^2 \right\}. \end{aligned} \tag{7.2}$$

Obviously, $\mathbf{s}^{k+1} - \mathbf{y}^k = \theta^k(\mathbf{t}^{k+1} - \mathbf{z}^k)$. Thus, by (7.1) and the fact that $H(\mathbf{x}^{k+1}) \leq H(\tilde{\mathbf{x}}^{k+1})$ (step 5), it follows that

$$\begin{aligned} \mathbb{E}_{i_k} H(\mathbf{x}^{k+1}) &\leq \mathbb{E}_{i_k} H(\tilde{\mathbf{x}}^{k+1}) \leq H(\mathbf{y}^k) + \nabla H(\mathbf{y}^k)^T(\mathbf{s}^{k+1} - \mathbf{y}^k) + \frac{n}{2}\|\mathbf{s}^{k+1} - \mathbf{y}^k\|_{\mathbf{L}}^2 \\ &= H(\mathbf{y}^k) + \theta^k \left[\nabla H(\mathbf{y}^k)^T(\mathbf{t}^{k+1} - \mathbf{z}^k) + \frac{n\theta^k}{2}\|\mathbf{t}^{k+1} - \mathbf{z}^k\|_{\mathbf{L}}^2 \right]. \end{aligned} \tag{7.3}$$

By Tseng's three-points property [37, Property 1] and the relation (7.2), we have

$$\begin{aligned} &\nabla H(\mathbf{y}^k)^T(\mathbf{x}^* - \mathbf{z}^k) + \frac{n\theta^k}{2}\|\mathbf{x}^* - \mathbf{z}^k\|_{\mathbf{L}}^2 - \nabla H(\mathbf{y}^k)^T(\mathbf{t}^{k+1} - \mathbf{z}^k) \\ &\qquad - \frac{n\theta^k}{2}\|\mathbf{t}^{k+1} - \mathbf{z}^k\|_{\mathbf{L}}^2 \geq \frac{n\theta^k}{2}\|\mathbf{x}^* - \mathbf{t}^{k+1}\|_{\mathbf{L}}^2. \end{aligned} \tag{7.4}$$

Combining the above with (7.3) yields

$$\begin{aligned}\mathbb{E}_{i_k} H(\mathbf{x}^{k+1}) &\leq H(\mathbf{y}^k) + \theta^k \left[\nabla H(\mathbf{y}^k)^T (\mathbf{x}^* - \mathbf{z}^k) + \frac{n\theta^k}{2} \|\mathbf{x}^* - \mathbf{z}^k\|_{\mathbf{L}}^2 - \frac{n\theta^k}{2} \|\mathbf{x}^* - \mathbf{t}^{k+1}\|_{\mathbf{L}}^2 \right] \\ &= H(\mathbf{y}^k) + \theta^k \left[\nabla H(\mathbf{y}^k)^T (\mathbf{x}^* - \mathbf{z}^k) + \frac{n^2\theta^k}{2} \|\mathbf{x}^* - \mathbf{z}^k\|_{\mathbf{L}}^2 - \frac{n^2\theta^k}{2} \mathbb{E}_{i_k} \|\mathbf{x}^* - \mathbf{z}^{k+1}\|_{\mathbf{L}}^2 \right], \end{aligned} \tag{7.5}$$

where the equality follows by the following argument:

$$\begin{aligned}\|\mathbf{x}^* - \mathbf{z}^k\|_{\mathbf{L}}^2 - \|\mathbf{x}^* - \mathbf{t}^{k+1}\|_{\mathbf{L}}^2 &= 2\langle \mathbf{t}^{k+1} - \mathbf{z}^k, \mathbf{x}^* - \mathbf{z}^k \rangle_{\mathbf{L}} - \|\mathbf{t}^{k+1} - \mathbf{z}^k\|_{\mathbf{L}}^2 \\ &= 2n\mathbb{E}_{i_k} \langle \mathbf{z}^{k+1} - \mathbf{z}^k, \mathbf{x}^* - \mathbf{z}^k \rangle_{\mathbf{L}} - n\mathbb{E}_{i_k} \|\mathbf{z}^{k+1} - \mathbf{z}^k\|_{\mathbf{L}}^2 \\ &= n\mathbb{E}_{i_k} (\|\mathbf{x}^* - \mathbf{z}^k\|_{\mathbf{L}}^2 - \|\mathbf{x}^* - \mathbf{z}^{k+1}\|_{\mathbf{L}}^2) \end{aligned}$$

Now, using the update formula in step 2, we have

$$\begin{aligned}\nabla H(\mathbf{y}^k)^T (\theta^k \mathbf{x}^* - \theta^k \mathbf{z}^k) &= \nabla H(\mathbf{y}^k)^T (\theta^k \mathbf{x}^* - \mathbf{y}^k + (1 - \theta^k)\mathbf{x}^k) \\ &= \theta^k \nabla H(\mathbf{y}^k)^T (\mathbf{x}^* - \mathbf{y}^k) + (1 - \theta^k) \nabla H(\mathbf{y}^k)^T (\mathbf{x}^k - \mathbf{y}^k). \end{aligned} \tag{7.6}$$

Thus, combining (7.5) and (7.6) along with the gradient inequality, the following is implied:

$$\begin{aligned}\mathbb{E}_{i_k} H(\mathbf{x}^{k+1}) \leq (1 - \theta^k) H(\mathbf{x}^k) + \theta^k H(\mathbf{x}^*) + \frac{n^2(\theta^k)^2}{2} \|\mathbf{x}^* \\ - \mathbf{z}^k\|_{\mathbf{L}}^2 - \frac{n^2(\theta^k)^2}{2} \mathbb{E}_{i_k} \|\mathbf{x}^* - \mathbf{z}^{k+1}\|_{\mathbf{L}}^2, \end{aligned} \tag{7.7}$$

which is the same as

$$\begin{aligned}\mathbb{E}_{i_k} H(\mathbf{x}^{k+1}) - H(\mathbf{x}^*) \leq (1 - \theta^k)(H(\mathbf{x}^k) - H(\mathbf{x}^*)) + \frac{n^2(\theta^k)^2}{2} \|\mathbf{x}^* \\ - \mathbf{z}^k\|_{\mathbf{L}}^2 - \frac{n^2(\theta^k)^2}{2} \mathbb{E}_{i_k} \|\mathbf{x}^* - \mathbf{z}^{k+1}\|_{\mathbf{L}}^2. \end{aligned} \tag{7.8}$$

Taking expectation over ξ_{k-1} leads to

$$\begin{aligned}\mathbb{E}_{\xi_k} H(\mathbf{x}^{k+1}) - H(\mathbf{x}^*) \leq (1 - \theta^k)(\mathbb{E}_{\xi_{k-1}} H(\mathbf{x}^k) - H(\mathbf{x}^*)) \\ + \frac{n^2(\theta^k)^2}{2} \mathbb{E}_{\xi_{k-1}} \|\mathbf{x}^* - \mathbf{z}^k\|_{\mathbf{L}}^2 - \frac{n^2(\theta^k)^2}{2} \mathbb{E}_{\xi_k} \|\mathbf{x}^* - \mathbf{z}^{k+1}\|_{\mathbf{L}}^2. \end{aligned} \tag{7.9}$$

Denoting $e_k \equiv \mathbb{E}_{\xi_{k-1}} H(\mathbf{x}^k) - H(\mathbf{x}^*)$ and $\Delta_k \equiv \frac{n^2}{2} \mathbb{E}_{\xi_{k-1}} \|\mathbf{x}^* - \mathbf{z}^k\|_{\mathbf{L}}^2$, we can rewrite (7.9) as

$$e_{k+1} \leq (1 - \theta^k) e_k + (\theta^k)^2 \Delta_k - (\theta^k)^2 \Delta_{k+1}.$$

Dividing the inequality by $(\theta^k)^2$ yields

$$\frac{1}{(\theta^k)^2}e_{k+1} \leq \frac{1-\theta^k}{(\theta^k)^2}e_k + \Delta_k - \Delta_{k+1}.$$

By the definition of the sequence θ^k (Step 6), the above is the same as

$$\frac{1}{(\theta^k)^2}e_{k+1} \leq \frac{1}{(\theta^{k-1})^2}e_k + \Delta_k - \Delta_{k+1},$$

and hence,

$$\frac{1}{(\theta^k)^2}e_{k+1} + \Delta_{k+1} \leq \frac{1}{(\theta^{k-1})^2}e_k + \Delta_k.$$

Since $\theta^0 = 1$ the above inequality results in that $\frac{1}{(\theta^{k-1})^2}e_k \leq \Delta_0$, which by the facts that $\Delta_0 = \frac{n^2}{2}\|\mathbf{x}^* - \mathbf{x}^0\|_{\mathbf{L}}^2$ and $\theta^k \leq \frac{2}{k+2}$ (see [37]) leads to the desired result (4.3).

References

1. Auslender, A., Teboulle, M.: Interior gradient and proximal methods for convex and conic optimization. SIAM J. Optim. **16**(3), 697–725 (2006)
2. Barbero, A., Sra, S.: Modular proximal optimization for multidimensional total-variation regularization. arXiv preprint arXiv:1411.0589 (2014)
3. Beck, A.: First-Order Methods in Optimization, vol. 25. SIAM (2017)
4. Beck, A., Pauwels, E., Sabach, S.: The cyclic block conditional gradient method for convex optimization problems. SIAM J. Optim. **25**(4), 2024–2049 (2015)
5. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**(1), 183–202 (2009)
6. Beck, A., Teboulle, M.: Smoothing and first order methods: a unified framework. SIAM J. Optim. **22**(2), 557–580 (2012)
7. Beck, A., Tetruashvili, L.: On the convergence of block coordinate descent type methods. SIAM J. Optim. **23**(4), 2037–2060 (2013)
8. Bertsekas, D.P.: Nonlinear Program. Athena Scientific Optimization and Computation Series, 2nd edn. Athena Scientific, Belmont (1999)
9. Bertsekas, D.P., Tsitsiklis, J.N.: Parallel and Distributed Computation: Numerical Methods, vol. 23. Prentice Hall, Englewood Cliffs (1989)
10. Burges, C.J.: A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Discov. **2**(2), 121–167 (1998)
11. Combettes, P.L., Pesquet, J.C.: Stochastic quasi-Fejér block-coordinate fixed point iterations with random sweeping. SIAM J. Optim. **25**(2), 1221–1248 (2015)
12. Daubechies, I., Defrise, M., De Mol, C.: An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. Commun. Pure Appl. Math. J. Issued Courant Inst. Math. Sci. **57**(11), 1413–1457 (2004)
13. Duchi, J., Shalev-Shwartz, S., Singer, Y., Chandra, T.: Efficient projections onto the l 1-ball for learning in high dimensions. In: Proceedings of the 25th International Conference on Machine Learning, pp. 272–279 (2008)
14. Fercoq, O., Bianchi, P.: A coordinate-descent primal-dual algorithm with large step size and possibly nonseparable functions. SIAM J. Optim. **29**(1), 100–134 (2019)
15. Fercoq, O., Richtárik, P.: Accelerated, parallel, and proximal coordinate descent. SIAM J. Optim. **25**(4), 1997–2023 (2015)
16. Friedman, J., Hastie, T., Höfling, H., Tibshirani, R.: Pathwise coordinate optimization. Ann. Appl. Stat. **1**(2), 302–332 (2007)

17. Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. **33**(1), 1 (2010)
18. Giselsson, P., Fält, M.: Envelope functions: unifications and further properties. J. Optim. Theory Appl. **178**(3), 673–698 (2018). https://doi.org/10.1007/s10957-018-1328-z
19. Hanzely, F., Kovalev, D., Richtárik, P.: Variance reduced coordinate descent with acceleration: new method with a surprising application to finite-sum problems. arXiv preprint arXiv:2002.04670 (2020)
20. Hanzely, F., Mishchenko, K., Richtárik, P.: SEGA: Variance reduction via gradient sketching. In: Advances in Neural Information Processing Systems, vol. 31, pp. 2082–2093 (2018)
21. Hanzely, F., Richtárik, P.: One method to rule them all: variance reduction for data, parameters and many new methods. arXiv preprint arXiv:1905.11266 (2019)
22. Hastie, T., Tibshirani, R., Wainwright, M.: Statistical Learning with Sparsity: The Lasso and Generalizations. CRC Press (2015)
23. Hong, M., Wang, X., Razaviyayn, M., Luo, Z.Q.: Iteration complexity analysis of block coordinate descent methods. Math. Program. **163**(1–2), 85–114 (2017)
24. Johnson, N.A.: A dynamic programming algorithm for the fused Lasso and L_0-segmentation. J. Comput. Graph. Stat. **22**(2), 246–260 (2013)
25. Kolmogorov, V., Pock, T., Rolinek, M.: Total variation on a tree. SIAM J. Imaging Sci. **9**(2), 605–636 (2016)
26. Lacoste-Julien, S., Jaggi, M., Schmidt, M., Pletscher, P.: Block-coordinate Frank-Wolfe optimization for structural SVMs. In: Dasgupta, S., McAllester, D. (eds.) Proceedings of the 30th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 28, pp. 53–61. PMLR (2013)
27. Latafat, P., Themelis, A., Patrinos, P.: Block-coordinate and incremental aggregated proximal gradient methods for nonsmooth nonconvex problems. Math. Program. 1–30 (2021)
28. Lu, H., Freund, R., Mirrokni, V.: Accelerating greedy coordinate descent methods. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 3257–3266. PMLR (2018)
29. Maculan, N., Santiago, C.P., Macambira, E., Jardim, M.: An O(n) algorithm for projecting a vector on the intersection of a hyperplane and a box in R n. J. Optim. Theory Appl. **117**(3), 553–574 (2003)
30. Markowitz, H.: Portfolio selection. J. Finance **7**(1), 77–91 (1952)
31. Moreau, J.J.: Proximité et dualité dans un espace hilbertien. Bull. de la Société mathématique de France **93**, 273–299 (1965)
32. Nesterov, Y.: Efficiency of coordinate descent methods on huge-scale optimization problems. SIAM J. Optim. **22**(2), 341–362 (2012)
33. Nesterov, Y.: Gradient methods for minimizing composite functions. Math. Program. **140**(1), 125–161 (2013)
34. Ortega, J.M., Rheinboldt, W.C.: Iterative Solution of Nonlinear Equations in Several Variables, vol. 30. SIAM (1970)
35. Rockafellar, R.T.: Convex Analysis. Princeton Mathematical Series, No. 28, Princeton University Press, Princeton (1970)
36. Stella, L., Themelis, A., Patrinos, P.: Forward–backward quasi-newton methods for nonsmooth optimization problems. Comput. Optim. Appl. **67**(3), 443–487 (2017)
37. Tseng, P.: On accelerated proximal gradient methods for convex–concave optimization. Unpublished manuscript (2008)
38. Wright, S.J.: Coordinate descent algorithms. Math. Program. **151**(1), 3–34 (2015)