Block-coordinate and incremental aggregated proximal gradient methods for nonsmooth nonconvex problems

Latafat, Puya
Department of Electrical Engineering (ESAT-STADIUS), KU Leuven

Themelis, Andreas
Faculty of Information Science and Electrical Engineering (ISEE), Kyushu University

Patrinos, Panagiotis
Department of Electrical Engineering (ESAT-STADIUS), KU Leuven

出版情報：Mathematical Programming, 2021-01-13. Springer
バージョン：
権利関係：

Mathematical Programming manuscript No.
(will be inserted by the editor)

Block-coordinate and incremental aggregated proximal gradient methods for nonsmooth nonconvex problems

Puya Latafat · Andreas Themelis · Panagiotis Patrinos

Received: date / Accepted: date

Abstract This paper analyzes block-coordinate proximal gradient methods for minimizing the sum of a separable smooth function and a (nonseparable) nonsmooth function, both of which are allowed to be nonconvex. The main tool in our analysis is the forward-backward envelope (FBE), which serves as a particularly suitable continuous and real-valued Lyapunov function. Global and linear convergence results are established when the cost function satisfies the Kurdyka-Łojasiewicz property without imposing convexity requirements on the smooth function. Two prominent special cases of the investigated setting are regularized finite sum minimization and the sharing problem; in particular, an immediate byproduct of our analysis leads to novel convergence results and rates for the popular Finito/MISO algorithm in the nonsmooth and nonconvex setting with very general sampling strategies.

Keywords Nonsmooth nonconvex optimization, block-coordinate updates, forward-backward envelope, KL inequality

Mathematics Subject Classification (2010) 90C06, 90C25, 90C26, 49J52, 49J53.

This work was supported by the Research Foundation Flanders (FWO) PhD grant 1196820N and research projects G0A0920N, G086518N and G086318N; Research Council KU Leuven C1 project No. C14/18/068; Fonds de la Recherche Scientifique – FNRS and the Fonds Wetenschappelijk Onderzoek – Vlaanderen under EOS project no 30468160 (SeLMA).

P. Latafat
Tel.: +32 (0)16 374408
E-mail: puya.latafat@kuleuven.be

A. Themelis
Tel.: +32 (0)16 374573
E-mail: andreas.themelis@kuleuven.be

P. Patrinos
Tel.: +32 (0)16 374445
E-mail: panos.patrinos@esat.kuleuven.be

Department of Electrical Engineering (ESAT-STADIUS) – KU Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium.

1 Introduction

This paper addresses block-coordinate (BC) proximal gradient methods for problems of the form

$$\underset{\boldsymbol{x}=(x_1,\dots,x_N)\in\mathbb{R}^{\sum_i n_i}}{\text{minimize}} \quad \Phi(\boldsymbol{x}) := F(\boldsymbol{x}) + G(\boldsymbol{x}), \quad \text{where} \quad F(\boldsymbol{x}) := \tfrac{1}{N}\textstyle\sum_{i=1}^{N} f_i(x_i), \tag{1.1}$$

in the following setting.

Assumption I (problem setting)**.** *In problem* (1.1) *the following hold:*

A1 *function* $f_i : \mathbb{R}^{n_i} \to \mathbb{R}$ *is* L_{f_i}*-smooth (Lipschitz differentiable with modulus* L_{f_i}*),* $i \in [N] := \{1, \dots, N\}$*;*

A2 *function* $G : \mathbb{R}^{\sum_i n_i} \to \overline{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$ *is proper and lower semicontinuous (lsc);*

A3 *a solution exists:* $\arg\min \Phi \neq \emptyset$*.*

Unlike typical cases analyzed in the literature where G is separable [57,60,40, 6,13,49,33,15,27,63], we here consider the complementary case where it is only the smooth term F that is assumed to be separable. The main challenge in analyzing convergence of BC schemes for (1.1) especially in the nonconvex setting is the fact that even in expectation the cost does not necessarily decrease along the trajectories. Instead, we demonstrate that the forward-backward envelope (FBE) [43,56] is a suitable Lyapunov function for such problems.

Several BC-type algorithms that allow for a nonseparable nonsmooth term have been considered in the literature, all however in convex settings. In [59,61] a class of convex composite problems is studied that involves a linear constraint as the nonsmooth nonseparable term. A BC algorithm with a Gauss-Southwell-type rule is proposed and the convergence is established using the cost as Lyapunov function by exploiting linearity of the constraint to ensure feasibility. A refined analysis in [38,39] extends this to a random coordinate selection strategy. Another approach in the convex case is to consider randomized BC updates applied to general averaged operators. Although this approach can allow for a fully nonseparable structure, usually separable nonsmooth functions are considered in the literature. The convergence analysis of such methods relies on establishing quasi-Fejér monotonicity [29,18,45,10,44,31]. In a primal-dual setting in [23] a combination of Bregman and Euclidean distance is employed as Lyapunov function. In [26] a BC algorithm is proposed for strongly convex functions that involves coordinate updates for the gradient followed by a full proximal step, and the distance from the (unique) solution is used as Lyapunov function. The analysis and the Lyapunov functions in all of the above mentioned works rely heavily on convexity and are not suitable for nonconvex settings.

Thanks to the nonconvexity and nonseparability of G, many machine learning problems can be formulated as in (1.1), a primary example being constrained and/or regularized finite sum problems [7,53,21,20,36,48,47,52]

$$\text{minimize}_{x\in\mathbb{R}^n} \; \varphi(x) := \tfrac{1}{N}\textstyle\sum_{i=1}^{N} f_i(x) + g(x), \tag{1.2}$$

where $f_i : \mathbb{R}^n \to \mathbb{R}$ are smooth functions and $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ is possibly nonsmooth, and everything here can be nonconvex. One way to cast (1.2) into the form of problem

(1.1) is by setting

$$G(\boldsymbol{x}) \coloneqq \tfrac{1}{N}\textstyle\sum_{i=1}^{N} g(x_i) + \delta_C(\boldsymbol{x}), \tag{1.3}$$

where $C \coloneqq \left\{\boldsymbol{x} \in \mathbb{R}^{nN} \mid x_1 = x_2 = \cdots = x_N\right\}$ is the consensus set, and δ_C is the indicator function of set C, namely $\delta_C(\boldsymbol{x}) = 0$ for $\boldsymbol{x} \in C$ and ∞ otherwise. Since the nonsmooth term g is allowed to be nonconvex, formulation (1.2) can account for nonconvex constraints such as rank constraints or zero norm balls, and nonconvex regularizers such as ℓ^p with $p \in [0, 1)$, [28].

Another prominent example in distributed applications is the *"sharing"* problem [14]:

$$\underset{\boldsymbol{x}\in\mathbb{R}^{nN}}{\text{minimize}}\ \Phi(\boldsymbol{x}) \coloneqq \tfrac{1}{N}\textstyle\sum_{i=1}^{N} f_i(x_i) + g\Big(\sum_{i=1}^{N} x_i\Big), \tag{1.4}$$

where $f_i : \mathbb{R}^n \to \mathbb{R}$ are smooth functions and $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ is nonsmooth, and all are possibly nonconvex. The sharing problem is cast as in (1.1) by setting $G \coloneqq g \circ A$, where $A \coloneqq [\mathrm{I}_n \ \ldots \ \mathrm{I}_n] \in \mathbb{R}^{n\times nN}$ (I_r denotes the $r \times r$ identity matrix).

1.1 The main block-coordinate algorithm

While gradient evaluations are the building blocks of smooth minimization, a fundamental tool to deal with a nonsmooth lsc term $\psi : \mathbb{R}^r \to \overline{\mathbb{R}}$ is its *V-proximal mapping*

$$\operatorname{prox}_{\psi}^{V}(x) \coloneqq \underset{w\in\mathbb{R}^r}{\arg\min}\left\{\psi(w) + \tfrac{1}{2}\|w - x\|_V^2\right\}, \tag{1.5}$$

where V is a symmetric and positive definite matrix and $\|\cdot\|_V$ indicates the norm induced by the scalar product $(x, y) \mapsto \langle x, Vy\rangle$. It is common to take $V = t^{-1}\mathrm{I}_r$ as a multiple of the $r \times r$ identity matrix I_r, in which case the notation $\operatorname{prox}_{t\psi}$ is typically used and t is referred to as a stepsize. While this operator enjoys nice regularity properties when g is convex, such as (single valuedness and) Lipschitz continuity, for nonconvex g it may fail to be a well-defined function and rather has to be intended as a point-to-set mapping $\operatorname{prox}_{\psi}^{V} : \mathbb{R}^r \rightrightarrows \mathbb{R}^r$. Nevertheless, the value function associated to the minimization problem in the definition (1.5), namely the *Moreau envelope*

$$\psi^V(x) \coloneqq \inf_{w\in\mathbb{R}^r}\left\{\psi(w) + \tfrac{1}{2}\|w - x\|_V^2\right\}, \tag{1.6}$$

is a well-defined real-valued function, in fact locally Lipschitz continuous, that lower bounds ψ and shares with ψ infima and minimizers. The proximal mapping is available in closed form for many useful functions, some of which are widely used regularizers in machine learning; for instance, the proximal mapping of the ℓ^0 and ℓ^1 regularizers amount to hard and soft thresholding operators.

In many applications the cost to be minimized is structured as the sum of a smooth term h and a proximable (i.e., with easily computable proximal mapping) term ψ. In these cases, the *proximal gradient method* [25,3] constitutes a cornerstone iterative method that interleaves gradient descent steps on the smooth function and proximal operations on the nonsmooth function, resulting in iterations of the form $x^+ \in \operatorname{prox}_{\gamma\psi}(x - \gamma\nabla h(x))$ for some suitable stepsize γ.

Our proposed scheme to address problem (1.1) is a BC variant of proximal gradient, in the sense that only some coordinates are updated according to the proximal gradient rule, while the others are left unchanged. This concept is synopsized in Algorithm 1, which constitutes the general algorithm addressed in this paper.

Algorithm 1 General forward-backward block-coordinate scheme

REQUIRE $\boldsymbol{x}^0 \in \mathbb{R}^{\sum_i n_i}$, $\gamma_i \in (0, {}^{N}/_{L_{f_i}})$, $i \in [N]$

$\Gamma = \operatorname{blkdiag}(\gamma_1 \mathrm{I}_{n_1}, \ldots, \gamma_N \mathrm{I}_{n_N})$, $k = 0$

REPEAT until convergence

1: $\boldsymbol{z}^k \in \operatorname{prox}_G^{\Gamma^{-1}}(\boldsymbol{x}^k - \Gamma \nabla F(\boldsymbol{x}^k))$

2: select a set of indices $\mathcal{I}^{k+1} \subseteq [N]$

3: update $x_i^{k+1} = z_i^k$ for $i \in \mathcal{I}^{k+1}$ and $x_i^{k+1} = x_i^k$ for $i \notin \mathcal{I}^{k+1}$, $k \leftarrow k+1$

RETURN $\boldsymbol{z}^k$

Although seemingly wasteful, in many cases one can efficiently compute individual blocks without the need of full operations. In fact, the BC Algorithm 1 bridges the gap between a BC framework and a class of incremental methods where a global computation typically involving the full gradient is carried out incrementally via performing computations only for a subset of coordinates. Two such broad applications, problems (1.2) and (1.4), are discussed in the dedicated Sections 3 and 4, where among other things we show that Algorithm 1 leads to the well known Finito/MISO algorithm [21,36].

1.2 Contribution

1) To the best of our knowledge this is the first analysis of BC schemes with a non-separable nonsmooth term and in the fully nonconvex setting. While the original cost Φ cannot serve as a Lyapunov function, we show that the forward-backward envelope (FBE) [43,56] decreases surely, not only in expectation (Lemma 2.5).

2) This allows for a quite general convergence analysis for different sampling criteria. This paper in particular covers randomized strategies (Section 2.3) where at each iteration one or more coordinates are sampled with possibly time-varying probabilities, as well as essentially cyclic (and in particular cyclic and shuffled) strategies in case the nonsmooth term is convex (Section 2.4).

3) We exploit the Kurdyka-Łojasiewicz (KL) property to show global (as opposed to subsequential) and linear convergence when the sampling is essentially cyclic and the nonsmooth function is convex, without imposing convexity requirements on the smooth functions (Theorem 2.11).

4) As immediate byproducts of our analysis we obtain **(a)** an incremental algorithm for the sharing problem [14] that to the best of our knowledge is novel (Section 4), and **(b)** the Finito/MISO algorithm [21,36] leading to a much simpler and more general

analysis than available in the literature with new convergence results both for randomized sampling strategies in the fully nonconvex setting and for essentially cyclic samplings when the nonsmooth term is convex (Section 3).

1.3 Organization

In the next subsection we introduce the adopted notation. The core of the paper lies in the convergence analysis of Algorithm 1 detailed in Section 2: Section 2.1 introduces the FBE, fundamental tool of our methodology and lists some of its properties whose proofs are detailed in the dedicated Appendix A.1, followed by other ancillary results documented in Appendix A.2. The algorithmic analysis begins in Section 2.2 with a collection of facts that hold independently of the chosen sampling strategy, and later specializes to randomized and essentially cyclic samplings in the dedicated Sections 2.3 and 2.4. Sections 3 and 4 discuss two particular instances of the investigated algorithmic framework, namely (a generalization of) the Finito/MISO algorithm for finite sum minimization and an incremental scheme for the sharing problem, both for fully nonconvex and nonsmooth formulations. Convergence results are immediately inferred from those of the more general BC Algorithm 1. Section 5 concludes the paper.

1.4 Notation

With id we indicate the identity function $x \mapsto x$ defined on a suitable space, and with I the identity matrix of suitable size. For a symmetric and positive definite matrix V, we denote by $\|\cdot\|_V$ the norm induced by the scalar product $(x, y) \mapsto \langle x, Vy\rangle$, namely $\|x\|_V := \sqrt{\langle x, Vx\rangle}$. We denote by $\|\cdot\|$ the standard Euclidean norm. For a set E and a sequence $(x^k)_{k\in\mathbb{N}}$ we write $(x^k)_{k\in\mathbb{N}} \subseteq E$ to indicate that $x^k \in E$ for all $k \in \mathbb{N}$, and we say that $(x^k)_{k\in\mathbb{N}}$ is *summable* if $\sum_{k\in\mathbb{N}} \|x^k\|$ is finite. We say that $(x^k)_{k\in\mathbb{N}}$ converges at *Q-linear rate* (resp. *R-linear rate*) to a point x if there exist $c \in (0, 1)$ such that $\|x^{k+1} - x\| \le c\|x^k - x\|$ (resp. $\|x^k - x\| \le \rho c^k$ for some $\rho > 0$) holds for all $k \in \mathbb{N}$.

We use the notation $H : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ to indicate a point-to-set mapping $H : \mathbb{R}^n \to 2^{\mathbb{R}^m}$, where $2^{\mathbb{R}^m}$ is the power set of $\mathbb{R}^m$ (the set of all subsets of $\mathbb{R}^m$). The *graph* of H is the set $\operatorname{gph} H := \{(x, y) \in \mathbb{R}^n \times \mathbb{R}^m \mid y \in H(x)\}$. We say that H is *outer semicontinuous (osc)* if $\operatorname{gph} H$ is a closed subset of $\mathbb{R}^n \times \mathbb{R}^m$, and *locally bounded* if for every bounded $U \subset \mathbb{R}^n$ the set $\bigcup_{x\in U} H(x)$ is bounded.

The *domain* and *epigraph* of an extended-real-valued function $h : \mathbb{R}^n \to \overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$ are the sets respectively defined as $\operatorname{dom} h := \{x \in \mathbb{R}^n \mid h(x) < \infty\}$ and $\operatorname{epi} h := \{(x, \alpha) \in \mathbb{R}^n \times \mathbb{R} \mid h(x) \le \alpha\}$. Function h is said to be *proper* if $\operatorname{dom} h \neq \emptyset$, and *lower semicontinuous (lsc)* if $\operatorname{epi} h$ is a closed subset of $\mathbb{R}^{n+1}$. For $\alpha \in \mathbb{R}$, $\operatorname{lev}_{\le\alpha} h$ is the *α-sublevel set* of h, i.e., $\operatorname{lev}_{\le\alpha} h := \{x \in \mathbb{R}^n \mid h(x) \le \alpha\}$. We say that h is *level bounded* if $\operatorname{lev}_{\le\alpha} h$ is bounded for all $\alpha \in \mathbb{R}$. We denote by $\hat{\partial} h : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ the *regular subdifferential* of h, where

$$v \in \hat{\partial} h(\bar{x}) \quad \Leftrightarrow \quad \liminf_{\substack{x\to\bar{x}\\ x\neq\bar{x}}} \frac{h(x) - h(\bar{x}) - \langle v, x - \bar{x}\rangle}{\|x - \bar{x}\|} \ge 0.$$

A necessary condition for local minimality of x for h is $0 \in \hat{\partial} h(x)$, see [51, Th. 10.1]. The (limiting) *subdifferential* of h is $\partial h : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$, where $v \in \partial h(x)$ iff $x \in \operatorname{dom} h$ and there exists a sequence $(x^k, v^k)_{k\in\mathbb{N}} \subseteq \operatorname{gph} \hat{\partial} h$ such that $(x^k, h(x^k), v^k) \to (x, h(x), v)$ as $k \to \infty$.

The B-subdifferential (also known as *Bouligand* or *limiting Jacobian*) of a locally Lipschitz-continuous function $G : \mathbb{R}^n \to \mathbb{R}^m$ at $\bar{x} \in \mathbb{R}^n$ is the set-valued mapping $\partial_B G : \mathbb{R}^n \rightrightarrows \mathbb{R}^{m\times n}$ defined as

$$\partial_B G(\bar{x}) := \Big\{H \in \mathbb{R}^{m\times n} \mid \exists (x^k)_{k\in\mathbb{N}} \subset C_G \text{ with } x^k \to \bar{x}, JG(x^k) \to H\Big\},$$

where $C_G \subseteq \mathbb{R}^n$ denotes the (dense) set of points at which G is differentiable (in the classical sense) and JG denotes the Jacobian of G. If $G : \mathbb{R}^n \to \mathbb{R}^m$ is locally Lipschitz on $\mathbb{R}^n$, then $\partial_B G(x)$ is a nonempty and compact subset of $\mathbb{R}^{m\times n}$ matrices, and as a set-valued mapping it is osc at every $x \in \mathbb{R}^n$. The interested reader is referred to the textbooks [17,22,51] for the details.

2 Convergence analysis

We begin by observing that Assumption I is enough to guarantee the well definedness of the forward-backward operator in Algorithm 1, which for notational convenience will be henceforth denoted as $\mathrm{T}_\Gamma^{\mathrm{FB}}(\boldsymbol{x})$. Namely, $\mathrm{T}_\Gamma^{\mathrm{FB}} : \mathbb{R}^{\sum_i n_i} \rightrightarrows \mathbb{R}^{\sum_i n_i}$ is the point-to-set mapping

$$\begin{aligned}\mathrm{T}_\Gamma^{\mathrm{FB}}(\boldsymbol{x}) &:= \operatorname{prox}_G^{\Gamma^{-1}}(\boldsymbol{x} - \Gamma\nabla F(\boldsymbol{x})) \\ &= \operatorname*{arg\,min}_{\boldsymbol{w}\in\mathbb{R}^{\sum_i n_i}} \Big\{F(\boldsymbol{x}) + \langle \nabla F(\boldsymbol{x}), \boldsymbol{w} - \boldsymbol{x}\rangle + G(\boldsymbol{w}) + \tfrac{1}{2}\|\boldsymbol{w} - \boldsymbol{x}\|^2_{\Gamma^{-1}}\Big\}. \end{aligned} \tag{2.1}$$

Lemma 2.1. *Suppose that Assumption I holds, and let* $\Gamma := \operatorname{blkdiag}(\gamma_1 \mathrm{I}_{n_1}, \ldots, \gamma_N \mathrm{I}_{n_N})$ *with* $\gamma_i \in (0, {}^{N}/_{L_{f_i}})$, $i \in [N]$. *Then* $\operatorname{prox}_G^{\Gamma^{-1}}$ *and* $\mathrm{T}_\Gamma^{\mathrm{FB}}$ *are locally bounded, outer semicontinuous (osc), nonempty- and compact-valued mappings.*

Proof. See Appendix A.1. □

2.1 The forward-backward envelope

The fundamental challenge in the analysis of (1.1) is the fact that, without separability of G, descent on the cost function cannot be established even in expectation. Instead, we show that the *forward-backward envelope* (FBE) [43,56] can be used as Lyapunov function. This subsection formally introduces the FBE, here generalized to account for a matrix-valued stepsize parameter Γ, and lists some of its basic properties needed for the convergence analysis of Algorithm 1. Although easy adaptations of the similar results in [43,56,55], for the sake of self-containedness the proofs are detailed in the dedicated Appendix A.1.

Definition 2.2 (forward-backward envelope)**.** *In problem* (1.1)*, let f_i be differentiable functions, $i \in [N]$, and for $\gamma_1, \dots, \gamma_N > 0$ let $\Gamma = \operatorname{blkdiag}(\gamma_1 \mathrm{I}_{n_1}, \dots, \gamma_N \mathrm{I}_{n_N})$. The forward-backward envelope (FBE) associated to* (1.1) *with stepsize Γ is the function $\Phi_\Gamma^{\mathrm{FB}} : \mathbb{R}^{\sum_i n_i} \to [-\infty, \infty)$ defined as*

$$\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}) := \inf_{\boldsymbol{w} \in \mathbb{R}^{\sum_i n_i}} \left\{ F(\boldsymbol{x}) + \langle \nabla F(\boldsymbol{x}), \boldsymbol{w} - \boldsymbol{x} \rangle + G(\boldsymbol{w}) + \tfrac{1}{2}\|\boldsymbol{w} - \boldsymbol{x}\|_{\Gamma^{-1}}^2 \right\}. \tag{2.2a}$$

Definition 2.2 highlights an important symmetry between the Moreau envelope and the FBE: similarly to the relation between the Moreau envelope (1.6) and the proximal mapping (1.5), the FBE (2.2a) is the value function associated with the proximal gradient mapping (2.1). By replacing any minimizer $\boldsymbol{z} \in \mathrm{T}_\Gamma^{\mathrm{FB}}(\boldsymbol{x})$ in the right-hand side of (2.2a) one obtains yet another interesting interpretation of the FBE in terms of the Γ^{-1}-augmented Lagrangian associated to (1.1)

$$\mathscr{L}_{\Gamma^{-1}}(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{y}) := F(\boldsymbol{x}) + G(\boldsymbol{z}) + \langle \boldsymbol{y}, \boldsymbol{x} - \boldsymbol{z} \rangle + \tfrac{1}{2}\|\boldsymbol{x} - \boldsymbol{z}\|_{\Gamma^{-1}}^2,$$

namely,

$$\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}) = F(\boldsymbol{x}) + \langle \nabla F(\boldsymbol{x}), \boldsymbol{z} - \boldsymbol{x} \rangle + G(\boldsymbol{z}) + \tfrac{1}{2}\|\boldsymbol{z} - \boldsymbol{x}\|_{\Gamma^{-1}}^2 \tag{2.2b}$$

$$= \mathscr{L}_{\Gamma^{-1}}(\boldsymbol{x}, \boldsymbol{z}, -\nabla F(\boldsymbol{x})). \tag{2.2c}$$

Lastly, by rearranging the terms it can easily be seen that

$$\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}) = F(\boldsymbol{x}) - \tfrac{1}{2}\|\nabla F(\boldsymbol{x})\|_\Gamma^2 + G^{\Gamma^{-1}}(\boldsymbol{x} - \Gamma \nabla F(\boldsymbol{x})), \tag{2.2d}$$

hence in particular that the FBE inherits regularity properties of F, ∇F, and the Moreau envelope $G^{\Gamma^{-1}}$ (cf. (1.6)), some of which are summarized in the next result.

Lemma 2.3 (FBE: fundamental inequalities)**.** *Suppose that Assumption I is satisfied and let $\gamma_i \in (0, {}^{N}/_{L_{f_i}})$, $i \in [N]$. Then, the FBE $\Phi_\Gamma^{\mathrm{FB}}$ is a (real-valued and) locally Lipschitz-continuous function. Moreover, the following hold for any $\boldsymbol{x} \in \mathbb{R}^{\sum_i n_i}$:*

(i) $\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}) \le \Phi(\boldsymbol{x})$.

(ii) $\frac{1}{2}\|\boldsymbol{z} - \boldsymbol{x}\|_{\Gamma^{-1} - \Lambda_F}^2 \le \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}) - \Phi(\boldsymbol{z}) \le \frac{1}{2}\|\boldsymbol{z} - \boldsymbol{x}\|_{\Gamma^{-1} + \Lambda_F}^2$ *for any $\boldsymbol{z} \in \mathrm{T}_\Gamma^{\mathrm{FB}}(\boldsymbol{x})$, where* $\Lambda_F := \frac{1}{N} \operatorname{blkdiag}(L_{f_1} \mathrm{I}_{n_1}, \dots, L_{f_n} \mathrm{I}_{n_N})$.

Proof. See Appendix A.1. □

Another key property that the FBE shares with the Moreau envelope is that minimizing the extended-real valued function Φ is equivalent to minimizing the continuous function $\Phi_\Gamma^{\mathrm{FB}}$. Moreover, the former is level bounded iff so is the latter. This fact will be particularly useful for the analysis of Algorithm 1, as it will be shown in Lemma 2.5 that the FBE (surely) decreases along its iterates. As a consequence, despite the fact that the same does not hold for Φ (in fact, iterates may even be infeasible), coercivity of Φ is enough to guarantee boundedness of $(\boldsymbol{x}^k)_{k \in \mathbb{N}}$ and $(\boldsymbol{z}^k)_{k \in \mathbb{N}}$.

Lemma 2.4 (FBE: minimization equivalence)**.** *Suppose that Assumption I is satisfied and that $\gamma_i \in (0, {}^{N}/_{L_i})$, $i \in [N]$. Then the following hold:*

(i) $\min \Phi_\Gamma^{\mathrm{FB}} = \min \Phi$;

(ii) $\arg\min \Phi_\Gamma^{\mathrm{FB}} = \arg\min \Phi$;

(iii) $\Phi_\Gamma^{\mathrm{FB}}$ *is level bounded iff so is* Φ.

Proof. See Appendix A.1. □

We remark that the kinship of $\Phi_\Gamma^{\mathrm{FB}}$ and Φ extends also to local minimality; the interested reader is referred to [54, Th. 3.6] for details.

2.2 A sure descent lemma

We now proceed to the theoretical analysis of Algorithm 1. Clearly, some assumptions on the index selection criterion are needed in order to establish reasonable convergence results, for little can be guaranteed if, for instance, one of the indices is never selected. Nevertheless, for the sake of a general analysis it is instrumental to first investigate which properties hold independently of such criteria. After listing some of these facts in Lemma 2.5, in Sections 2.3 and 2.4 we will specialize the results to randomized and (essentially) cyclic sampling strategies.

Lemma 2.5 (sure descent)**.** *Suppose that Assumption I is satisfied. Then, the following hold for the iterates generated by Algorithm 1:*

(i) $\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{k+1}) \leq \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^k) - \sum_{i\in\mathcal{I}^{k+1}} \frac{\xi_i}{2\gamma_i}\|z_i^k - x_i^k\|^2$, *where* $\xi_i := \frac{N-\gamma_i L_{f_i}}{N}$, $i \in [N]$, *are strictly positive;*

(ii) $(\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^k))_{k\in\mathbb{N}}$ *monotonically decreases to a finite value* $\Phi_\star \geq \min \Phi$;

(iii) $\Phi_\Gamma^{\mathrm{FB}}$ *is constant (and equals* $\Phi_\star$ *as above) on the set of cluster points of* $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$;

(iv) the sequence $(\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2)_{k\in\mathbb{N}}$ *has finite sum (and in particular vanishes);*

(v) if Φ *is coercive, then* $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ *and* $(\boldsymbol{z}^k)_{k\in\mathbb{N}}$ *are bounded.*

Proof.

♠ 2.5*(i)* To ease notation, let $\Lambda_F := \frac{1}{N}\operatorname{blkdiag}(L_{f_1}\mathrm{I}_{n_1}, \dots, L_{f_n}\mathrm{I}_{n_N})$ and for $\boldsymbol{w} \in \mathbb{R}^{\sum_i n_i}$ let $w_{\mathcal{I}} \in \mathbb{R}^{\sum_{i\in\mathcal{I}} n_i}$ denote the slice $(w_i)_{i\in\mathcal{I}}$, and let $\Lambda_{F_{\mathcal{I}}}, \Gamma_{\mathcal{I}} \in \mathbb{R}^{\sum_{i\in\mathcal{I}} n_i \times \sum_{i\in\mathcal{I}} n_i}$ be defined accordingly. Start by observing that, since $\boldsymbol{z}^{k+1} \in \operatorname{prox}_G^{\Gamma^{-1}}(\boldsymbol{x}^{k+1} - \Gamma\nabla F(\boldsymbol{x}^{k+1}))$, from the proximal inequality on G it follows that

$$\begin{aligned} G(\boldsymbol{z}^{k+1}) - G(\boldsymbol{z}^k) &\leq \tfrac{1}{2}\|\boldsymbol{z}^k - \boldsymbol{x}^{k+1} + \Gamma\nabla F(\boldsymbol{x}^{k+1})\|^2_{\Gamma^{-1}} - \tfrac{1}{2}\|\boldsymbol{z}^{k+1} - \boldsymbol{x}^{k+1} + \Gamma\nabla F(\boldsymbol{x}^{k+1})\|^2_{\Gamma^{-1}} \\ &= \tfrac{1}{2}\|\boldsymbol{z}^k - \boldsymbol{x}^{k+1}\|^2_{\Gamma^{-1}} - \tfrac{1}{2}\|\boldsymbol{z}^{k+1} - \boldsymbol{x}^{k+1}\|^2_{\Gamma^{-1}} + \langle \nabla F(\boldsymbol{x}^{k+1}), \boldsymbol{z}^k - \boldsymbol{z}^{k+1}\rangle. \end{aligned} \tag{2.3}$$

We have

$$\begin{aligned} \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{k+1}) - \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^k) = F(\boldsymbol{x}^{k+1}) + \langle \nabla F(\boldsymbol{x}^{k+1}), \boldsymbol{z}^{k+1} - \boldsymbol{x}^{k+1}\rangle + G(\boldsymbol{z}^{k+1}) + \tfrac{1}{2}\|\boldsymbol{z}^{k+1} - \boldsymbol{x}^{k+1}\|^2_{\Gamma^{-1}} \\ - \Big(F(\boldsymbol{x}^k) + \langle \nabla F(\boldsymbol{x}^k), \boldsymbol{z}^k - \boldsymbol{x}^k\rangle + G(\boldsymbol{z}^k) + \tfrac{1}{2}\|\boldsymbol{z}^k - \boldsymbol{x}^k\|^2_{\Gamma^{-1}}\Big) \end{aligned}$$

apply the upper bound in (A.1) with $\boldsymbol{w} = \boldsymbol{x}^{k+1}$ and the proximal inequality (2.3)

$$\begin{aligned} &\leq \langle \nabla F(\boldsymbol{x}^k), \boldsymbol{x}^{k+1} - \boldsymbol{z}^k\rangle + \tfrac{1}{2}\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2_{\Lambda_F} + \langle \nabla F(\boldsymbol{x}^{k+1}), \boldsymbol{z}^k - \boldsymbol{x}^{k+1}\rangle \\ &\quad - \tfrac{1}{2}\|\boldsymbol{z}^k - \boldsymbol{x}^k\|^2_{\Gamma^{-1}} + \tfrac{1}{2}\|\boldsymbol{z}^k - \boldsymbol{x}^{k+1}\|^2_{\Gamma^{-1}}. \end{aligned}$$

To conclude, notice that the ℓ-th block of $\boldsymbol{x}^{k+1} - \boldsymbol{z}^k$ is zero if $\ell \in \mathcal{I}^{k+1}$, and the ℓ-th block of $\nabla F(\boldsymbol{x}^k) - \nabla F(\boldsymbol{x}^{k+1})$ is zero for $\ell \notin \mathcal{I}^{k+1}$ (due to separability of F). Hence, the scalar product vanishes. For similar reasons, one has $\|\boldsymbol{z}^k - \boldsymbol{x}^{k+1}\|^2_{\Gamma^{-1}} - \|\boldsymbol{z}^k - \boldsymbol{x}^k\|^2_{\Gamma^{-1}} = -\|z^k_{\mathcal{I}^{k+1}} - x^k_{\mathcal{I}^{k+1}}\|^2_{\Gamma^{-1}_{\mathcal{I}^{k+1}}}$ and $\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2_{\Lambda_F} = \|z^k_{\mathcal{I}^{k+1}} - x^k_{\mathcal{I}^{k+1}}\|^2_{\Lambda_{F_{\mathcal{I}^{k+1}}}}$, yielding the claimed expression.

♠ 2.5*(ii)* Monotonic decrease of $(\Phi^{\text{FB}}_\Gamma(\boldsymbol{x}^k))_{k\in\mathbb{N}}$ is a direct consequence of assertion 2.5*(i)*. This ensures that the sequence converges to some value $\Phi_\star$, bounded below by $\min \Phi$ in light of Lemma 2.4*(i)*.

♠ 2.5*(iii)* Directly follows from assertion 2.5*(ii)* together with the continuity of Φ^{FB}_Γ, see Lemma 2.3.

♠ 2.5*(iv)* Denoting $\xi_{\min} \coloneqq \min_{i\in[N]}\{\xi_i\}$ which is a strictly positive constant, it follows from assertion 2.5*(i)* that for each $k \in \mathbb{N}$ it holds that

$$\begin{aligned}\Phi^{\text{FB}}_\Gamma(\boldsymbol{x}^{k+1}) - \Phi^{\text{FB}}_\Gamma(\boldsymbol{x}^k) &\leq -\sum_{i\in\mathcal{I}^{k+1}} \tfrac{\xi_i}{2\gamma_i}\|z^k_i - x^k_i\|^2 \\ &\leq -\tfrac{\xi_{\min}}{2}\sum_{i\in\mathcal{I}^{k+1}} \gamma_i^{-1}\|z^k_i - x^k_i\|^2 \\ &= -\tfrac{\xi_{\min}}{2}\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2_{\Gamma^{-1}}.\end{aligned}$$

By summing for $k \in \mathbb{N}$ and using the positive definiteness of Γ^{-1} together with the fact that $\min \Phi^{\text{FB}}_\Gamma = \min \Phi > -\infty$ as ensured by Lemma 2.4*(i)* and Requirement I.A3, we obtain that $\sum_{k\in\mathbb{N}} \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2 < \infty$.

♠ 2.5*(v)* It follows from assertion 2.5*(ii)* that the entire sequence $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ is contained in the sublevel set $\left\{\boldsymbol{w} \mid \Phi^{\text{FB}}_\Gamma(\boldsymbol{w}) \leq \Phi^{\text{FB}}_\Gamma(\boldsymbol{x}^0)\right\}$, which is bounded provided that Φ is coercive as shown in Lemma 2.4*(iii)*. In turn, boundedness of $(\boldsymbol{z}^k)_{k\in\mathbb{N}}$ then follows from local boundedness of $\mathrm{T}^{\text{FB}}_\Gamma$, cf. Lemma 2.1. □

2.3 Randomized sampling

In this section we provide convergence results for Algorithm 1 where the index selection criterion complies with the following requirement.

Assumption II (randomized sampling requirements)**.** *There exist $p_1, \dots, p_N > 0$ such that, at any iteration and independently of the past, each $i \in [N]$ is sampled with probability at least p_i.*

Our notion of randomization is general enough to allow for time-varying probabilities and mini-batch selections. The role of parameters p_i in Assumption II is to prevent that an index is sampled with arbitrarily small probability. In more rigorous terms, $\mathcal{P}_k[i \in \mathcal{I}^{k+1}] \geq p_i$ shall hold for all $i \in [N]$, where $\mathcal{P}_k$ represents the probability conditional to the knowledge at iteration k. Notice that we do not require the p_i's to sum up to one, as multiple index selections are allowed, similar to the setting of [10,31] in the convex case.

Due to the possible nonconvexity of problem (1.1), unless additional assumptions are made not much can be said about convergence of the iterates to a unique

point. Nevertheless, the following result shows that any cluster point $\boldsymbol{x}^\star$ of sequences $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ and $(\boldsymbol{z}^k)_{k\in\mathbb{N}}$ generated by Algorithm 1 is a stationary point, in the sense that it satisfies the necessary condition for minimality $0 \in \hat{\partial}\Phi(\boldsymbol{x}^\star)$, see [51, Th. 10.1].

Theorem 2.6 (randomized sampling: subsequential convergence)**.** *Suppose that Assumptions I and II are satisfied. Then, the following hold almost surely for the iterates generated by Algorithm 1:*

(i) the sequence $(\|\boldsymbol{x}^k - \boldsymbol{z}^k\|^2)_{k\in\mathbb{N}}$ *has finite sum (and in particular vanishes);*
(ii) the sequence $(\Phi(\boldsymbol{z}^k))_{k\in\mathbb{N}}$ *converges to* $\Phi_\star$ *as in Lemma 2.5(ii);*
(iii) $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ *and* $(\boldsymbol{z}^k)_{k\in\mathbb{N}}$ *have the same cluster points, all stationary and on which* Φ *and* $\Phi_\Gamma^{\mathrm{FB}}$ *equal* $\Phi_\star$.

Proof. In what follows, $\mathbb{E}_k$ denotes the expectation conditional to the knowledge at iteration k.

♠ 2.6*(i)* Let $\xi_i := \frac{N-\gamma_i L_{f_i}}{N} > 0$, $i \in [N]$, be as in Lemma 2.5*(i)*. We have

$$\begin{aligned}
\mathbb{E}_k\Big[\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{k+1})\Big] &\overset{2.5(i)}{\leq} \mathbb{E}_k\left[\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^k) - \sum_{i\in\mathcal{I}^{k+1}} \tfrac{\xi_i}{2\gamma_i}\|z_i^k - x_i^k\|^2\right] \\
&= \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^k) - \sum_{\mathcal{I}\in\Omega} \mathcal{P}_k\Big[\mathcal{I}^{k+1} = \mathcal{I}\Big] \sum_{i\in\mathcal{I}} \tfrac{\xi_i}{2\gamma_i}\|z_i^k - x_i^k\|^2 \\
&= \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^k) - \sum_{i=1}^{N} \sum_{\mathcal{I}\in\Omega, \mathcal{I}\ni i} \mathcal{P}_k\Big[\mathcal{I}^{k+1} = \mathcal{I}\Big] \tfrac{\xi_i}{2\gamma_i}\|z_i^k - x_i^k\|^2 \\
&\leq \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^k) - \sum_{i=1}^{N} \tfrac{p_i\xi_i}{2\gamma_i}\|z_i^k - x_i^k\|^2, \qquad (2.4)
\end{aligned}$$

where $\Omega \subseteq 2^{[N]}$ is the sample space. Therefore,

$$\mathbb{E}_k\Big[\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{k+1})\Big] \leq \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^k) - \tfrac{\sigma}{2}\|\boldsymbol{x}^k - \boldsymbol{z}^k\|^2_{\Gamma^{-1}}, \quad \text{where } \sigma := \min_{i=1\dots N} p_i\xi_i > 0. \qquad (2.5)$$

The claim follows from the Robbins-Siegmund supermartingale theorem, see e.g., [50] or [7, Prop. 2].

♠ 2.6*(ii)* Observe that $\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^k) - \|\boldsymbol{z}^k - \boldsymbol{x}^k\|^2_{\Gamma^{-1}+\Lambda_F} \leq \Phi(\boldsymbol{z}^k) \leq \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^k) - \|\boldsymbol{z}^k - \boldsymbol{x}^k\|^2_{\Gamma^{-1}-\Lambda_F}$ holds (surely) for $k \in \mathbb{N}$ in light of Lemma 2.3*(ii)*. The claim then follows by invoking Lemma 2.5*(ii)* and assertion 2.6*(i)*.

♠ 2.6*(iii)* In the rest of the proof, for conciseness the "almost sure" nature of the results will be implied without mention. It follows from assertion 2.6*(i)* that a subsequence $(\boldsymbol{x}^k)_{k\in K}$ converges to some point $\boldsymbol{x}^\star$ iff so does the subsequence $(\boldsymbol{z}^k)_{k\in K}$. Since $\mathrm{T}_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^k) \ni \boldsymbol{z}^k$ and both $\boldsymbol{x}^k$ and $\boldsymbol{z}^k$ converge to $\boldsymbol{x}^\star$ as $K \ni k \to \infty$, the inclusion $0 \in \hat{\partial}\Phi(\boldsymbol{x}^\star)$ follows from Lemma A.1. Since the full sequences $(\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^k))_{k\in\mathbb{N}}$ and $(\Phi(\boldsymbol{z}^k))_{k\in\mathbb{N}}$ converge to the same value $\Phi_\star$ (cf. Lemma 2.5*(ii)* and assertion 2.6*(ii)*), due to continuity of $\Phi_\Gamma^{\mathrm{FB}}$ (Lemma 2.3) it holds that $\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^\star) = \Phi_\star$, and in turn the bounds in Lemma 2.3*(ii)* together with assertion 2.6*(i)* ensure that $\Phi(\boldsymbol{x}^\star) = \Phi_\star$ too. □

When G is convex and F is strongly convex (that is, each of the functions f_i is strongly convex), the FBE decreases Q-linearly in expectation along the iterates generated by the randomized BC Algorithm 1.

Theorem 2.7 (randomized sampling: linear convergence under strong convexity)**.** *Additionally to Assumptions I and II, suppose that G is convex and that each f_i is μ_{f_i}-strongly convex. Then, for all k the following hold for the iterates generated by Algorithm 1:*

$$\mathbb{E}_k\Big[\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{k+1}) - \min \Phi\Big] \leq (1-c)(\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^k) - \min \Phi) \tag{2.6a}$$

$$\mathbb{E}\Big[\Phi(\boldsymbol{z}^k) - \min \Phi\Big] \leq (\Phi(\boldsymbol{x}^0) - \min \Phi)(1-c)^k \tag{2.6b}$$

$$\tfrac{1}{2}\mathbb{E}\Big[\|\boldsymbol{z}^k - \boldsymbol{x}^\star\|_{\mu_F}^2\Big] \leq (\Phi(\boldsymbol{x}^0) - \min \Phi)(1-c)^k \tag{2.6c}$$

where $\boldsymbol{x}^\star \coloneqq \arg\min \Phi$, $\mu_F \coloneqq \frac{1}{N}\operatorname{blkdiag}(\mu_{f_1}\mathrm{I}_{n_1}, \ldots \mu_{f_n}\mathrm{I}_{n_N})$, and denoting $\xi_i = \frac{N-\gamma_i L_{f_i}}{N}$, $i \in [N]$,

$$c = \min_{i\in[N]}\left\{\frac{\xi_i p_i}{\gamma_i}\right\} \Big/ \max_{i\in[N]}\left\{\frac{N-\gamma_i\mu_{f_i}}{\gamma_i^2\mu_{f_i}}\right\}. \tag{2.7}$$

Moreover, by setting the stepsizes γ_i and minimum sampling probabilities p_i as

$$\gamma_i = \frac{N}{\mu_{f_i}}\Big(1 - \sqrt{1 - 1/\kappa_i}\Big) \quad \text{and} \quad p_i = \frac{\Big(\sqrt{\kappa_i} + \sqrt{\kappa_i - 1}\Big)^2}{\sum_{j=1}^N \Big(\sqrt{\kappa_j} + \sqrt{\kappa_j - 1}\Big)^2} \tag{2.8}$$

with $\kappa_i \coloneqq \frac{L_{f_i}}{\mu_{f_i}}$, $i \in [N]$, then the constant c in (2.6) can be tightened to

$$c = \frac{1}{\sum_{i=1}^N \big(\sqrt{\kappa_i} + \sqrt{\kappa_i - 1}\big)^2}. \tag{2.9}$$

Proof. The claimed Q-linear convergence rate (2.6a) with factor c as in (2.7) is obtained by combining the upper bound in Lemma A.2*(vi)* with (2.4). The R-linear rates in terms of the cost function and distance from the solution are obtained by repeated application of (2.6a) after taking (unconditional) expectation from both sides and using Lemma 2.3 and the lower bound in Lemma A.2*(vi)*.

To obtain the tighter estimate (2.9), observe that (2.4) with the choice

$$p_i \coloneqq \frac{1}{\gamma_i\mu_{f_i}}\frac{N-\gamma_i\mu_{f_i}}{N-\gamma_i L_{f_i}}\Big(\sum_j \frac{1}{\gamma_j\mu_{f_j}}\frac{N-\gamma_j\mu_{f_j}}{N-\gamma_j L_{f_j}}\Big)^{-1},$$

which equals the one in (2.8) with γ_i as prescribed, yields

$$\begin{aligned}\mathbb{E}_k\Big[\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{k+1})\Big] &\leq \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^k) - \Big(2N\textstyle\sum_j \frac{1}{\gamma_j\mu_j}\frac{N-\gamma_j\mu_j}{N-\gamma_j L_j}\Big)^{-1}\displaystyle\sum_{i=1}^N \frac{N-\gamma_i\mu_{f_i}}{\gamma_i^2\mu_{f_i}}\|z_i^k - x_i^k\|^2 \\ &= \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^k) - \Big(2N\textstyle\sum_j \frac{1}{\gamma_j\mu_j}\frac{N-\gamma_j\mu_j}{N-\gamma_j L_j}\Big)^{-1}\|\boldsymbol{z}^k - \boldsymbol{x}^k\|^2_{\Gamma^{-1}\mu_F^{-1}(\Gamma^{-1}-\mu_F)}.\end{aligned}$$

The assertion now follows by combining this with the upper bound in Lemma A.2*(vi)* and replacing the values of γ_i as proposed in (2.8). Notice that as κ_i's approach 1 the linear rate tends to $1 - 1/N$. □

2.4 Cyclic, shuffled and essentially cyclic samplings

In this section we analyze the convergence of the BC Algorithm 1 when a cyclic, shuffled cyclic or (more generally) an essentially cyclic sampling [58,57,27,16,63] is used. As formalized in the following standing assumption, an additional convexity requirement for the nonsmooth term G is needed.

Assumption III (essentially cyclic sampling requirements)**.** *In problem* (1.1)*, function G is convex. Moreover, there exists $T \geq 1$ such that in Algorithm 1 each index is selected at least once within any interval of T iterations.*

Note that having $T < N$ is possible because of our general sampling strategy where sets of indices can be sampled within the same iteration. For instance, $T = 1$ corresponds to $\mathcal{I}^{k+1} = [N]$ for all k, in which case Algorithm 1 would reduce to a (full) proximal gradient scheme.

Two notable special cases of single index selection rules are the cyclic and shuffled cyclic sampling strategies.

Shuffled cyclic sampling: corresponds to setting

$$\mathcal{I}^{k+1} = \{\pi_{\lfloor k/N \rfloor}(\mathrm{mod}(k, N) + 1)\} \quad \text{for all } k \in \mathbb{N}, \tag{2.10}$$

where $\pi_0, \pi_1, \ldots$ are permutations of the set of indices $[N]$ (chosen randomly or deterministically).

Cyclic sampling: corresponds to the case (2.10) with $\pi_{\lfloor k/N \rfloor} = \mathrm{id}$, i.e.,

$$\mathcal{I}^{k+1} = \{\mathrm{mod}(k, N) + 1\} \quad \text{for all } k \in \mathbb{N}. \tag{2.11}$$

Consistently with the deterministic nature of the essentially cyclic sampling, all the results of the previous section hold surely, as opposed to almost surely.

Theorem 2.8 (essentially cyclic sampling: subsequential convergence)**.** *Suppose that Assumptions I and III are satisfied. Then, all the assertions of Theorem 2.6 hold surely.*

Proof. We first establish an important descent inequality for $\Phi_\Gamma^{\mathrm{FB}}$ after every T iterations, cf. (2.18). Convexity of G, entailing $\mathrm{prox}_G^{\Gamma^{-1}}$ being Lipschitz continuous (cf. Lemma A.2*(i)*), allows the employment of techniques similar to those in [6, Lemma 3.3]. Since all indices are updated at least once every T iterations, one has that

$$t_\nu(i) \coloneqq \min\{t \in [T] \mid i \text{ is sampled at iteration } T\nu + t - 1\} \tag{2.12}$$

is well defined for each index $i \in [N]$ and $\nu \in \mathbb{N}$. Since i is sampled at iteration $T\nu + t_\nu(i) - 1$ and $x_i^{T\nu} = x_i^{T\nu+1} = \cdots = x_i^{T\nu+t_\nu(i)-1}$ by definition of $t_\nu(i)$, it holds that

$$\begin{aligned} x_i^{T\nu+t_\nu(i)} &= x_i^{T\nu+t_\nu(i)-1} + U_i^\top\left(\mathrm{T}_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{T\nu+t_\nu(i)-1}) - \boldsymbol{x}^{T\nu+t_\nu(i)-1}\right) \\ &= x_i^{T\nu+t_\nu(i)-1} + U_i^\top\left(\mathrm{T}_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{T\nu+t_\nu(i)-1}) - \boldsymbol{x}^{T\nu}\right), \end{aligned} \tag{2.13}$$

where $U_i \in \mathbb{R}^{(\sum_j n_j)\times n_i}$ denotes the i-th block column of the identity matrix so that for a vector $v \in \mathbb{R}^{n_i}$

$$U_i v = (0, \ldots, 0, \overset{i\text{-th}}{\overbrace{v}}, 0, \ldots, 0)^\top. \tag{2.14}$$

For all $t \in [T]$ the following holds

$$\begin{aligned} \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{T(\nu+1)}) - \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{T\nu}) &= \sum_{\tau=1}^{T} \left(\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{T\nu+\tau}) - \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{T\nu+\tau-1})\right) \\ &\leq \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{T\nu+t}) - \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{T\nu+t-1}) \\ &\leq -\tfrac{\xi_{\min}}{2}\|\boldsymbol{x}^{T\nu+t} - \boldsymbol{x}^{T\nu+t-1}\|_{\Gamma^{-1}}^2, \end{aligned} \tag{2.15}$$

where $\xi_i \coloneqq \frac{N-\gamma_i L_{f_i}}{N}$ as in Lemma 2.5*(i)*, $\xi_{\min} \coloneqq \min_{i\in[N]}\{\xi_i\}$, and the two inequalities follow from Lemma 2.5*(i)*. Moreover, the triangular inequality for $i \in [N]$ yields

$$\begin{aligned} \|\boldsymbol{x}^{T\nu+t_\nu(i)-1} - \boldsymbol{x}^{T\nu}\|_{\Gamma^{-1}} &\leq \sum_{\tau=1}^{t_\nu(i)-1} \|\boldsymbol{x}^{T\nu+\tau} - \boldsymbol{x}^{T\nu+\tau-1}\|_{\Gamma^{-1}} \\ &\leq \tfrac{T}{\sqrt{\xi_{\min}/2}}\left(\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{T\nu}) - \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{T(\nu+1)})\right)^{1/2}, \end{aligned} \tag{2.16}$$

where the second inequality follows from (2.15) together with the fact that $t_\nu(i) \leq T$. For all $i \in [N]$, from the triangular inequality and the $L_{\mathbf{T}}$-Lipschitz continuity of $\mathrm{T}_\Gamma^{\mathrm{FB}}$ (Lemma A.2*(iv)*) we have

$$\begin{aligned} \gamma_i^{-1/2}\|U_i^\top(\boldsymbol{x}^{T\nu} - \mathrm{T}_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{T\nu}))\| &\leq \gamma_i^{-1/2}\|U_i^\top(\boldsymbol{x}^{T\nu} - \mathrm{T}_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{T\nu+t_\nu(i)-1}))\| \\ &\quad + \gamma_i^{-1/2}\|U_i^\top(\mathrm{T}_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{T\nu+t_\nu(i)-1}) - \mathrm{T}_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{T\nu}))\| \\ &\overset{(2.13)}{\leq} \gamma_i^{-1/2}\|x_i^{T\nu+t_\nu(i)-1} - x_i^{T\nu+t_\nu(i)}\| \\ &\quad + \|\mathrm{T}_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{T\nu+t_\nu(i)-1}) - \mathrm{T}_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{T\nu})\|_{\Gamma^{-1}} \\ &\leq \|\boldsymbol{x}^{T\nu+t_\nu(i)-1} - \boldsymbol{x}^{T\nu+t_\nu(i)}\|_{\Gamma^{-1}} + L_{\mathbf{T}}\|\boldsymbol{x}^{T\nu+t_\nu(i)-1} - \boldsymbol{x}^{T\nu}\|_{\Gamma^{-1}} \\ &\overset{(2.15),\,(2.16)}{\leq} \tfrac{1+TL_{\mathbf{T}}}{\sqrt{\xi_{\min}/2}}\left(\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{T\nu}) - \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{T(\nu+1)})\right)^{1/2}. \end{aligned} \tag{2.17}$$

By squaring and summing over $i \in [N]$ we obtain

$$\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{T(\nu+1)}) - \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{T\nu}) \leq -\tfrac{\xi_{\min}}{2N(1+TL_{\mathbf{T}})^2}\|\boldsymbol{z}^{T\nu} - \boldsymbol{x}^{T\nu}\|_{\Gamma^{-1}}^2. \tag{2.18}$$

By telescoping the inequality and using the fact that $\min \Phi_\Gamma^{\mathrm{FB}} = \min \Phi$ as shown in Lemma 2.4*(i)*, we obtain that $(\|\boldsymbol{z}^{T\nu} - \boldsymbol{x}^{T\nu}\|_{\Gamma^{-1}}^2)_{\nu\in\mathbb{N}}$ has finite sum, and in particular vanishes. Clearly, by suitably shifting, for every $t \in [T]$ the same can be said for the sequence $(\|\boldsymbol{z}^{T\nu+t} - \boldsymbol{x}^{T\nu+t}\|_{\Gamma^{-1}}^2)_{\nu\in\mathbb{N}}$. The whole sequence $(\|\boldsymbol{z}^k - \boldsymbol{x}^k\|^2)_{k\in\mathbb{N}}$ is thus summable, and we may now infer the claim as done in the proof of Theorem 2.6. □

In the next theorem explicit linear convergence rates are derived under the additional strong convexity assumption for the smooth functions. The cyclic and shuffled cyclic cases are treated separately, as tighter bounds can be obtained by leveraging the fact that within cycles of N iterations every index is updated exactly once.

Theorem 2.9 (essentially cyclic sampling: linear convergence under strong convexity)**.** *Additionally to Assumptions I and III, suppose that each function f_i is μ_{f_i}-strongly convex. Then, denoting $\delta := \min_{i\in[N]}\left\{\frac{\gamma_i\mu_{f_i}}{N}\right\}$ and $\Delta := \max_{i\in[N]}\left\{\frac{\gamma_i L_{f_i}}{N}\right\}$, for all $\nu \in \mathbb{N}$ the following hold for the iterates generated by Algorithm 1:*

$$\Phi_\Gamma^{\text{FB}}(\boldsymbol{x}^{T(\nu+1)}) - \min\Phi \le (1-c)(\Phi_\Gamma^{\text{FB}}(\boldsymbol{x}^{T\nu}) - \min\Phi) \tag{2.19a}$$

$$\Phi(\boldsymbol{z}^{T\nu}) - \min\Phi \le (\Phi(\boldsymbol{x}^0) - \min\Phi)(1-c)^\nu \tag{2.19b}$$

$$\tfrac{1}{2}\|\boldsymbol{z}^{T\nu} - \boldsymbol{x}^\star\|^2_{\mu_F} \le (\Phi(\boldsymbol{x}^0) - \min\Phi)(1-c)^\nu \tag{2.19c}$$

where $\boldsymbol{x}^\star := \arg\min\Phi$, $\mu_F := \frac{1}{N}\operatorname{blkdiag}(\mu_{f_1}\mathrm{I}_{n_1}, \dots \mu_{f_n}\mathrm{I}_{n_N})$, and

$$c = \frac{\delta(1-\Delta)}{N(1+T(1-\delta))^2(1-\delta)}. \tag{2.20}$$

In the case of shuffled cyclic (2.10) or cyclic (2.11) sampling, the inequalities can be tightened by replacing T with N and with

$$c = \frac{\delta(1-\Delta)}{N(2-\delta)^2(1-\delta)}. \tag{2.21}$$

Proof.

♠ *The general essentially cyclic case.* Since $\mathrm{T}_\Gamma^{\text{FB}}$ is $L_{\mathbf{T}}$-Lipschitz continuous with $L_{\mathbf{T}} = 1-\delta$ as shown in Lemma A.2*(v)*, inequality (2.18) becomes

$$\Phi_\Gamma^{\text{FB}}(\boldsymbol{x}^{T(\nu+1)}) - \Phi_\Gamma^{\text{FB}}(\boldsymbol{x}^{T\nu}) \le -\tfrac{1-\Delta}{2N(1+T(1-\delta))^2}\|\boldsymbol{z}^{T\nu} - \boldsymbol{x}^{T\nu}\|^2_{\Gamma^{-1}}.$$

Moreover, it follows from Lemma A.2*(vi)* that

$$\Phi_\Gamma^{\text{FB}}(\boldsymbol{x}^{T\nu}) - \min\Phi \le \tfrac{1}{2}(\delta^{-1}-1)\|\boldsymbol{z}^{T\nu} - \boldsymbol{x}^{T\nu}\|^2_{\Gamma^{-1}}. \tag{2.22}$$

By combining the two inequalities the claimed Q-linear convergence (2.19a) with factor c as in (2.20) is obtained. In turn, the R-linear rates (2.19b) and (2.19c) are obtained by repeated application of (2.19a) and using Lemma 2.3 and Lemma A.2*(vi)*.

♠ *The shuffled cyclic case.* Let us now suppose that the sampling strategy follows a shuffled rule as in (2.10) with permutations $\pi_0, \pi_1, \dots$ (hence in the cyclic case $\pi_\nu = \mathrm{id}$ for all $\nu \in \mathbb{N}$). Let U_i be as in (2.14) and $\xi_{\min}$ as in the proof of Theorem 2.8. Observe that $t_\nu(i) = \pi_\nu^{-1}(i) \le N$ for $t_\nu(i)$ as defined in (2.12). For all $t \in [N]$

$$\begin{aligned}\Phi_\Gamma^{\text{FB}}(\boldsymbol{x}^{N(\nu+1)}) - \Phi_\Gamma^{\text{FB}}(\boldsymbol{x}^{N\nu}) &\le \Phi_\Gamma^{\text{FB}}(\boldsymbol{x}^{N\nu+t-1}) - \Phi_\Gamma^{\text{FB}}(\boldsymbol{x}^{N\nu}) \\ &\le -\tfrac{\xi_{\min}}{2}\sum_{\tau=1}^{t-1}\|\boldsymbol{x}^{N\nu+\tau} - \boldsymbol{x}^{N\nu+\tau-1}\|^2_{\Gamma^{-1}} \\ &= -\tfrac{\xi_{\min}}{2}\|\boldsymbol{x}^{N\nu+t-1} - \boldsymbol{x}^{N\nu}\|^2_{\Gamma^{-1}},\end{aligned} \tag{2.23}$$

where the equality follows from the fact that at every iteration a different coordinate is updated (and that Γ is diagonal), and the inequalities from Lemma 2.5*(i)*. Similarly, (2.15) holds with T replaced by N (despite the fact that T is not necessarily N, but is

rather bounded as $T \leq 2N-1$). By using (2.23) in place of (2.16), inequality (2.17) is tightened as follows

$$\gamma_i^{-1/2}\|U_i^\top(\boldsymbol{x}^{N\nu} - \mathrm{T}_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{N\nu}))\| \leq \tfrac{1+L_{\mathbf{T}}}{\sqrt{\xi_{\min}/2}}\Big(\varPhi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{N\nu}) - \varPhi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{N(\nu+1)})\Big)^{1/2}.$$

By squaring and summing for $i \in [N]$ we obtain

$$\varPhi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{N(\nu+1)}) - \varPhi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{N\nu}) \leq -\tfrac{\xi_{\min}}{2N(1+L_{\mathbf{T}})^2}\|\boldsymbol{z}^{N\nu} - \boldsymbol{x}^{N\nu}\|^2_{\Gamma^{-1}} = -\tfrac{1-\varDelta}{2N(1+L_{\mathbf{T}})^2}\|\boldsymbol{z}^{N\nu} - \boldsymbol{x}^{N\nu}\|^2_{\Gamma^{-1}},$$

where $L_{\mathbf{T}} = 1-\delta$ as discussed above. By combining this and (2.22) (with T replaced by N) the improved coefficient (2.21) is obtained. □

Note that if one sets $\gamma_i = \alpha N/L_{f_i}$ for some $\alpha \in (0,1)$, then $\delta = \alpha \min_{i\in[N]}\{\mu_{f_i}/L_{f_i}\}$ and $\varDelta = \alpha$. With this selection, as the condition number approaches 1 the rate in (2.21) tends to $1 - \frac{\alpha}{N(2-\alpha)^2}$.

2.5 Global and linear convergence with KL inequality

The convergence analyses of the randomized and essentially cyclic cases both rely on a descent property on the FBE that quantifies the progress in the minization of $\varPhi_\Gamma^{\mathrm{FB}}$ in terms of the squared forward-backward residual $\|\boldsymbol{x}-\boldsymbol{z}\|^2$. A subtle but important difference, however, is that the inequality (2.5) in the former case involves a conditional expectation, whereas (2.18) in the latter does not. The *sure* descent property occurring for essentially cyclic sampling strategies is the key for establishing global (as opposed to subsequential) convergence based on the Kurdyka-Łojasiewicz (KL) property [34,35,30]. A similar result is achieved in [63], which however considers the complementary case to problem (1.1) where the nonsmooth function G is assumed to be separable, and thus the cost function itself can serve as Lyapunov function.

Definition 2.10 (KL property with exponent θ)**.** *A proper lsc function $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ is said to have the* Kurdyka-Łojasiewicz *(KL) property with exponent $\theta \in (0,1)$ at $\bar{w} \in \operatorname{dom} h$ if there exist $\varepsilon, \eta, \varrho > 0$ such that*

$$\psi'(h(w) - h(\bar{w}))\operatorname{dist}(0, \partial h(w)) \geq 1$$

holds for all w such that $\|w-\bar{w}\| < \varepsilon$ and $h(\bar{w}) < h(w) < h(\bar{w})+\eta$, where $\psi(s) \coloneqq \varrho s^{1-\theta}$. We say that h satisfies the KL property with exponent θ (without mention of $\bar{w}$) if it satisfies the KL property with exponent θ at any $\bar{w} \in \operatorname{dom} \partial h$.

Semialgebraic functions comprise a wide class of functions that enjoy this property [12,11], which has been extensively exploited to provide convergence rates of optimization algorithms [1,2,3,13,24,42,32,62]. Based on this, in the next result we provide sufficient conditions ensuring global and R-linear convergence of Algorithm 1 with essentially cyclic sampling.

Theorem 2.11 (essentially cyclic sampling: global and linear convergence)**.** *Additionally to Assumptions I and III, suppose that Φ has the KL property with exponent $\theta \in (0,1)$ (as is the case when f_i and G are semialgebraic), and is coercive. Then, any sequences $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ and $(\boldsymbol{z}^k)_{k\in\mathbb{N}}$ generated by Algorithm 1 converge to (the same) stationary point $\boldsymbol{x}^\star$. Moreover, if $\theta \le 1/2$ then $(\|\boldsymbol{z}^k - \boldsymbol{x}^k\|)_{k\in\mathbb{N}}$, $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ and $(\boldsymbol{z}^k)_{k\in\mathbb{N}}$ converge at R-linear rate.*

Proof. Let $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ and $(\boldsymbol{z}^k)_{k\in\mathbb{N}}$ be sequences generated by Algorithm 1 with essentially cyclic sampling, and let $\Phi_\star$ be the limit of the sequence $(\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^k))_{k\in\mathbb{N}}$ as in Lemma 2.5*(ii)*. To avoid trivialities, we may assume that $\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^k) \gneqq \Phi_\star$ for all k, for otherwise the sequence $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ is asymptotically constant, and thus so is $(\boldsymbol{z}^k)_{k\in\mathbb{N}}$. Let Ω be the set of cluster points of $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$, which is compact and such that $\Phi_\Gamma^{\mathrm{FB}} \equiv \Phi_\star$ on Ω, as ensured by Theorem 2.8. It follows from Lemma A.3 and [1, Lem. 1(ii)] that $\Phi_\Gamma^{\mathrm{FB}}$ enjoys a *uniform* KL property on Ω; in particular, $\psi'(\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^k)-\Phi_\star)\operatorname{dist}(0,\partial\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^k)) \ge 1$ holds for all k large enough such that $\boldsymbol{x}^k$ is sufficiently close to Ω and $\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^k)$ is sufficiently close to $\Phi_\star$, where $\psi(s) = \varrho s^{1-\theta'}$ for some $\varrho > 0$ and $\theta' = \max\{\theta, 1/2\}$. Combined with Lemma A.2*(iii)*, for all k large enough we thus have

$$\psi'(\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^k) - \Phi_\star) \ge \frac{c}{\|\boldsymbol{x}^k - \boldsymbol{z}^k\|_{\Gamma^{-1}}}, \tag{2.24}$$

where $c := \frac{N\min_i\{\sqrt{\gamma_i}\}}{N+\max_i\{\gamma_i L_{f_i}\}} > 0$. Let $\Delta_k := \psi(\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^k) - \Phi_\star)$. By combining (2.24) and (2.18) we have that there exists a constant $c' > 0$ such that

$$\Delta_{(\nu+1)T} - \Delta_{\nu T} \le \psi'(\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{\nu T}) - \Phi_\star)\Big(\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{(\nu+1)T}) - \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{\nu T})\Big) \le -c'\|\boldsymbol{x}^{\nu T} - \boldsymbol{z}^{\nu T}\|_{\Gamma^{-1}} \tag{2.25}$$

holds for all $\nu \in \mathbb{N}$ large enough (the first inequality uses concavity of ψ). By summing over ν (sure) summability of the sequence $(\|\boldsymbol{x}^{\nu T} - \boldsymbol{z}^{\nu T}\|)_{\nu\in\mathbb{N}}$ is obtained. By suitably shifting, for every $t \in [T]$ the same can be said for the sequence $(\|\boldsymbol{z}^{T\nu+t} - \boldsymbol{x}^{T\nu+t}\|)_{\nu\in\mathbb{N}}$, and since T is finite we conclude that the whole sequence $(\|\boldsymbol{z}^k - \boldsymbol{x}^k\|)_{k\in\mathbb{N}}$ is summable. Since $\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\| \le \|\boldsymbol{z}^k - \boldsymbol{x}^k\|$ we conclude that $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ has finite length and is thus convergent (to a single point), and consequently so is $(\boldsymbol{z}^k)_{k\in\mathbb{N}}$.

Suppose now that $\theta \le 1/2$, so that $\psi(s) = \varrho\sqrt{s}$. Then,

$$\|\boldsymbol{x}^{\nu T} - \boldsymbol{z}^{\nu T}\|_{\Gamma^{-1}} \overset{(2.24)}{\ge} \tfrac{2c}{\varrho}\sqrt{\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{\nu T}) - \Phi_\star} = \tfrac{2c}{\varrho^2}\psi(\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}^{\nu T}) - \Phi_\star) = \tfrac{2c}{\varrho^2}\Delta_{\nu T}.$$

Combined with (2.25) it follows that $(\Delta_{\nu T})_{\nu\in\mathbb{N}}$ converges Q-linearly. By rearranging (2.25) as

$$c'\|\boldsymbol{x}^{\nu T} - \boldsymbol{z}^{\nu T}\|_{\Gamma^{-1}} \le \Delta_{\nu T} - \Delta_{(\nu+1)T} \le \Delta_{\nu T},$$

R-linear convergence of $(\|\boldsymbol{x}^{\nu T} - \boldsymbol{z}^{\nu T}\|)_{\nu\in\mathbb{N}}$ follows. By suitably shifting, for every $t \in [T]$ the same can be said for the sequence $(\|\boldsymbol{z}^{T\nu+t} - \boldsymbol{x}^{T\nu+t}\|)_{\nu\in\mathbb{N}}$, and since T is finite we conclude that the whole sequence $(\|\boldsymbol{z}^k - \boldsymbol{x}^k\|)_{k\in\mathbb{N}}$ converges R-linearly. On the other hand, since $\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\| \le \|\boldsymbol{z}^k - \boldsymbol{x}^k\|$, also $(\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|)_{k\in\mathbb{N}}$ converges R-linearly, hence so does $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$. By combining the two, we conclude that also $(\boldsymbol{z}^k)_{k\in\mathbb{N}}$ converges R-linearly. □

3 Nonconvex finite sum problems: the Finito/MISO algorithm

As mentioned in Section 1, if G is of the form (1.3) then problem (1.1) reduces to the finite sum minimization presented in (1.2). Most importantly, the proximal mapping of the original nonsmooth function G (in the larger space $\mathbb{R}^{nN}$) can be easily expressed in terms of that of the small function g (in the original space $\mathbb{R}^n$) in the reduced finite sum reformulation, as shown in the next lemma. We remark that, when g is convex, this result can also be deduced from [23, Lem. 5] through duality arguments.

Lemma 3.1. *Given $\gamma_i > 0$, $i \in [N]$, let $\Gamma := \operatorname{blkdiag}(\gamma_1 \mathrm{I}_n, \dots, \gamma_N \mathrm{I}_n)$ and $\hat{\gamma} := (\sum_{i=1}^N \gamma_i^{-1})^{-1}$. Then, for G as in (1.3) and any $\boldsymbol{u} \in \mathbb{R}^{nN}$*

$$\operatorname{prox}_G^{\Gamma^{-1}}(\boldsymbol{u}) = \left\{(\hat{v}, \dots, \hat{v}) \mid \hat{v} \in \operatorname{prox}_{\hat{\gamma} g}(\hat{u})\right\}, \quad \text{where } \hat{u} := \hat{\gamma} \textstyle\sum_{i=1}^N \gamma_i^{-1} u_i.$$

Proof. Observe first that for every $w \in \mathbb{R}^n$ one has

$$\begin{aligned} \textstyle\sum_i \gamma_i^{-1}\|w - u_i\|^2 &= \textstyle\sum_i \gamma_i^{-1}\|\hat{u} - u_i\|^2 + \sum_i \gamma_i^{-1}\|w - \hat{u}\|^2 + \overbrace{2\textstyle\sum_i \gamma_i^{-1}\langle \hat{u} - u_i, w - \hat{u}\rangle}^{=0} \\ &= \textstyle\sum_i \gamma_i^{-1}\|\hat{u} - u_i\|^2 + \hat{\gamma}^{-1}\|w - \hat{u}\|^2. \end{aligned} \tag{3.1}$$

Next, observe that since $\operatorname{dom} G \subseteq C$ (the consensus set),

$$\begin{aligned} \operatorname{prox}_G^{\Gamma^{-1}}(\boldsymbol{u}) &= \operatorname*{arg\,min}_{\boldsymbol{w} \in \mathbb{R}^{nN}} \left\{G(\boldsymbol{w}) + \textstyle\sum_{i=1}^N \frac{1}{2\gamma_i}\|w_i - u_i\|^2\right\} \\ &= \operatorname*{arg\,min}_{\boldsymbol{w} \in \mathbb{R}^{nN}} \left\{G(\boldsymbol{w}) + \textstyle\sum_{i=1}^N \frac{1}{2\gamma_i}\|w_i - u_i\|^2 \mid w_1 = \cdots = w_N\right\} \\ &= \operatorname*{arg\,min}_{(w,\dots,w)} \left\{g(w) + \textstyle\sum_{i=1}^N \frac{1}{2\gamma_i}\|w - u_i\|^2\right\} \\ &\overset{(3.1)}{=} \operatorname*{arg\,min}_{(w,\dots,w)} \left\{g(w) + \tfrac{1}{2\hat{\gamma}}\|w - \hat{u}\|^2\right\} = \left\{(\hat{v}, \dots, \hat{v}) \mid \hat{v} \in \operatorname{prox}_{\hat{\gamma} g}(\hat{u})\right\} \end{aligned}$$

as claimed. □

If all stepsizes are set to the same value γ, so that $\Gamma = \gamma \mathrm{I}_{nN}$, then the forward-backward step reduces to

$$\begin{aligned} \boldsymbol{z} \in \operatorname{prox}_G^{\Gamma^{-1}}(\boldsymbol{x} - \Gamma \nabla F(\boldsymbol{x})) \quad \Leftrightarrow \quad & \boldsymbol{z} = (\bar{z}, \dots, \bar{z}), \\ & \bar{z} \in \operatorname{prox}_{\gamma g/N}\Big(\tfrac{1}{N}\textstyle\sum_{j=1}^N (x_j - \frac{\gamma}{N}\nabla f_j(x_j))\Big). \end{aligned} \tag{3.2}$$

The argument of $\operatorname{prox}_{\gamma g/N}$ is the (unweighted) average of the forward operator. By applying Algorithm 1 with (3.2), Finito/MISO [21,36] is recovered. Differently from the existing convergence analyses, ours covers fully nonconvex and nonsmooth problems, more general sampling strategies and the possibility to select different stepsizes γ_i for each block, which can have a significant impact on the performance compared to the case where all stepsizes are equal. Moreover, to the best of our knowledge this is the first work that shows global convergence and linear rates even when the smooth functions are nonconvex. The resulting scheme is presented in Algorithm 2. We remark that the consensus formulation to recover Finito/MISO (although from a different umbrella algorithm) was also observed in [19] in the convex case. Moreover, the

Finito/MISO algorithm with cyclic sampling is also studied in [37] when $g \equiv 0$ and f_i are strongly convex functions; consistently with Assumption III, our analysis covers the more general essentially cyclic sampling even in the presence of a nonsmooth convex term g and allowing the smooth functions f_i to be nonconvex. Randomized Finito/MISO with $g \equiv 0$ is also studied in the recent work [46]; although their analysis is limited to a single stepsize, in the convex case it is allowed to be larger than our worst-case stepsize $\min_i \gamma_i$.

Algorithm 2 Nonconvex proximal Finito/MISO for problem (1.2)

REQUIRE $x^{\text{init}} \in \mathbb{R}^n$, $\gamma_i \in (0, N/L_{f_i})$, $i \in [N]$

$\hat{\gamma} := (\sum_{i=1}^N \gamma_i^{-1})^{-1}$, $s_i = x^{\text{init}} - \frac{\gamma_i}{N}\nabla f_i(x^{\text{init}})$ $i \in [N]$, $\hat{s} = \hat{\gamma}\sum_{i=1}^N \gamma_i^{-1} s_i$

REPEAT until convergence

1: select a set of indices $\mathcal{I} \subseteq [N]$
2: $z \in \operatorname{prox}_{\hat{\gamma} g}(\hat{s})$
3: **for** $i \in \mathcal{I}$ **do**
4: $\quad v \leftarrow z - \frac{\gamma_i}{N}\nabla f_i(z)$
5: $\quad$ update $\hat{s} \leftarrow \hat{s} + \frac{\hat{\gamma}}{\gamma_i}(v - s_i)$ and $s_i \leftarrow v$

RETURN z

The convergence results from Section 2 are immediately translated to this setting by noting that the bold variable $\boldsymbol{z}^k$ corresponds to $(z^k, \dots, z^k)$. Therefore, $\Phi(\boldsymbol{z}^k) = \varphi(z^k)$ where φ is the cost function for the finite sum problem.

Corollary 3.2 (subsequential convergence of Algorithm 2)**.** *In the finite sum problem (1.2) suppose that* $\arg\min \varphi$ *is nonempty, g is proper and lsc, and each f_i is L_{f_i}-Lipschitz differentiable, $i \in [N]$. Then, the following hold almost surely (resp. surely) for the sequence $(z^k)_{k\in\mathbb{N}}$ generated by Algorithm 2 with randomized sampling strategy as in Assumption II (resp. with any essentially cyclic sampling strategy and g convex as required in Assumption III):*

(i) the sequence $(\varphi(z^k))_{k\in\mathbb{N}}$ converges to a finite value $\varphi_\star \le \varphi(x^{\text{init}})$;
(ii) all cluster points of $(z^k)_{k\in\mathbb{N}}$ are stationary and on which φ equals $\varphi_\star$.

If, additionally, φ is coercive, then the following also hold:

(iii) $(z^k)_{k\in\mathbb{N}}$ is bounded (in fact, this holds surely for arbitrary sampling criteria).

Corollary 3.3 (linear convergence of Algorithm 2 under strong convexity)**.** *Additionally to the assumptions of Corollary 3.2, suppose that g is convex and that each f_i is μ_{f_i}-strongly convex. The following hold for the iterates generated by Algorithm 2:*
RANDOMIZED SAMPLING: *under Assumption II,*

$$\mathbb{E}\Big[\varphi(z^k) - \min\varphi\Big] \le (\varphi(x^{\text{init}}) - \min\varphi)(1-c)^k$$
$$\tfrac{1}{2}\mathbb{E}\Big[\|z^k - x^\star\|^2\Big] \le \frac{N(\varphi(x^{\text{init}}) - \min\varphi)}{\sum_i \mu_{f_i}}(1-c)^k$$

holds for all $k \in \mathbb{N}$, where c is as in (2.7) and $x^\star := \arg\min \varphi$. If the stepsizes γ_i and the sampling probabilities p_i are set as in Theorem 2.7, then the tighter constant c as in (2.9) is obtained.

SHUFFLED CYCLIC OR CYCLIC SAMPLING: *under either sampling strategy (2.10) or (2.11),*

$$\varphi(z^{\nu N}) - \min\varphi \le (\varphi(x^{\text{init}}) - \min\varphi)(1-c)^\nu$$
$$\tfrac{1}{2}\mathbb{E}\Big[\|z^{\nu N} - x^\star\|^2\Big] \le \frac{N(\varphi(x^{\text{init}}) - \min\varphi)}{\sum_i \mu_{f_i}}(1-c)^\nu$$

holds surely for all $\nu \in \mathbb{N}$, where c is as in (2.21).

The next result follows from Theorem 2.11 once the needed properties of Φ as in the umbrella formulation (1.1) are shown to hold.

Corollary 3.4 (global convergence of Algorithm 2)**.** *In the finite sum problem (1.2), suppose that φ has the KL property with exponent $\theta \in (0,1)$ (as is the case when f_i and g are semialgebraic) and is coercive, g is proper convex and lsc, and each f_i is L_{f_i}-Lipschitz differentiable, $i \in [N]$. Then the sequence $(z^k)_{k\in\mathbb{N}}$ generated by Algorithm 2 with any essentially cyclic sampling strategy as in Assumption III converges surely to a stationary point for φ. Moreover, if $\theta \le 1/2$ then it converges at R-linear rate.*

Proof. Function $\Phi = F + G$ with G as in (1.3) is clearly coercive and satisfies Assumption I. In order to invoke Theorem 2.11 is suffices to show that there exists a constant $c > 0$ such that

$$c\,\operatorname{dist}(0, \partial\Phi(\boldsymbol{x})) \ge \operatorname{dist}(0, \partial\varphi(x)) \quad \text{for all } x \in \mathbb{R}^n \text{ and } \boldsymbol{x} = (x, \dots, x), \tag{3.3}$$

as this will ensure that Φ enjoys the KL property at $\boldsymbol{x}^\star = (x^\star, \dots, x^\star)$ with the same desingularizing function (up to a positive scaling). Notice that for $x \in \mathbb{R}^n$ and $\boldsymbol{x} = (x, \dots, x)$, one has $\boldsymbol{v} \in \partial G(\boldsymbol{x})$ iff $\sum_{i=1}^N v_i \in \partial g(x)$. Since $\partial\Phi(\boldsymbol{x}) = \frac{1}{N}\times_{i=1}^N \nabla f_i(x_i) + \partial G(\boldsymbol{x})$ and $\partial\varphi(x) = \frac{1}{N}\sum_{i=1}^N \nabla f_i(x) + \partial g(x)$, see [51, Ex. 8.8(c) and Prop. 10.5], for $x \in \mathbb{R}^n$ and denoting $\boldsymbol{x} = (x, \dots, x)$ we have

$$\operatorname{dist}(0, \partial\varphi(x)) \le \inf_{\boldsymbol{v}\in\partial G(\boldsymbol{x})} \big\|\tfrac{1}{N}\textstyle\sum_{i=1}^N \nabla f_i(x) + \sum_{i=1}^N v_i\big\| \le c \inf_{\boldsymbol{u}\in\partial\Phi(\boldsymbol{x})} \|\boldsymbol{u}\|,$$

for some positive c, thus establishing inequality (3.3). □

4 Nonconvex sharing problem

In this section we consider the sharing problem (1.4). As discussed in Section 1, (1.4) fits into the problem framework (1.1) by simply letting $G := g \circ A$, where $A := [\mathrm{I}_n \ \dots \ \mathrm{I}_n] \in \mathbb{R}^{n\times nN}$. By arguing as in [5, Th. 6.15] it can be shown that, given a general matrix A with full row rank, the proximal mapping of $G = g \circ A$ is given by

$$\operatorname{prox}_G^{\Gamma^{-1}}(\boldsymbol{u}) = \boldsymbol{u} + \Gamma A^\top (A\Gamma A^\top)^{-1}\Big(\operatorname{prox}_g^{(A\Gamma A^\top)^{-1}}(A\boldsymbol{u}) - A\boldsymbol{u}\Big). \tag{4.1}$$

Since $A\Gamma A^\top = (\sum_{i=1}^N \gamma_i)\mathrm{I}_n$ for the sharing problem (1.4),

$$\boldsymbol{v} \in \operatorname{prox}_G^{\Gamma^{-1}}(\boldsymbol{u}) \iff \boldsymbol{v} = (u_1 + \gamma_1 w, \dots, u_N + \gamma_N w)$$

$$w \in \tilde{\gamma}^{-1}\big(\mathrm{prox}_{\tilde{\gamma}g}(\tilde{u}) - \tilde{u}\big), \quad \tilde{\gamma} := \textstyle\sum_{i=1}^N \gamma_i, \quad \tilde{u} := \sum_{i=1}^N u_i.$$

Consequently, the general BC Algorithm 1 when applied to the sharing problem (1.4) reduces to Algorithm 3.

Algorithm 3 Block-coordinate method for the nonconvex sharing problem (1.4)

REQUIRE $x_i^{\text{init}} \in \mathbb{R}^n$, $\gamma_i \in (0, {}^N\!/_{L_{f_i}})$, $i \in [N]$

$\tilde{\gamma} := \sum_{i=1}^N \gamma_i$, $s_i = x_i^{\text{init}} - \frac{\gamma_i}{N}\nabla f_i(x_i^{\text{init}})$ $i \in [N]$, $\tilde{s} = \sum_{i=1}^N s_i$

REPEAT until convergence

1: select a set of indices $\mathcal{I} \subseteq [N]$

2: $w \leftarrow \tilde{\gamma}^{-1}(\mathrm{prox}_{\tilde{\gamma}g}(\tilde{s}) - \tilde{s})$

3: **for** $i \in \mathcal{I}$ **do**

4: $\quad v_i \leftarrow s_i + \gamma_i w - \frac{\gamma_i}{N}\nabla f_i(s_i + \gamma_i w)$

5: $\quad$ update $\tilde{s} \leftarrow \tilde{s} + (v_i - s_i)$ and $s_i \leftarrow v_i$

RETURN $\boldsymbol{z} = (s_1 + \gamma_1 w, \dots, s_N + \gamma_N w)$ with $w \in \tilde{\gamma}^{-1}(\mathrm{prox}_{\tilde{\gamma}g}(\tilde{s}) - \tilde{s})$

Remark 4.1 (generalized sharing constraint)**.** Another notable instance of $G = g \circ A$ well suited for the BC framework of Algorithm 1 is when $g = \delta_{\{0\}}$ and $A = [A_1 \ \dots \ A_N]$, $A_i \in \mathbb{R}^{n \times n_i}$ such that A is full row rank. This models the generalized sharing problem

$$\underset{\boldsymbol{x} \in \mathbb{R}^{\sum_i n_i}}{\text{minimize}} \ \tfrac{1}{N}\textstyle\sum_{i=1}^N f_i(x_i) \quad \text{subject to } \sum_{i=1}^N A_i x_i = 0.$$

In this case (4.1) simplifies to

$$\big(\mathrm{prox}_G^{\Gamma^{-1}}(\boldsymbol{u})\big)_i = u_i - \gamma_i A_i^\top \mathcal{A}^{-1} \sum_{i=1}^N A_i u_i,$$

where $\mathcal{A} := A\Gamma A^\top$ can be factored offline and $\sum_{i=1}^N A_i x_i$ can be updated in an incremental fashion in the spirit of Algorithm 3. □

The convergence results for Algorithm 3 summarized below fall as special cases of those in Section 2.

Corollary 4.2 (convergence of Algorithm 3)**.** *In the sharing problem (1.4), suppose that* $\arg\min \Phi$ *is nonempty, g is proper and lsc, and each f_i is L_{f_i}-Lipschitz differentiable, $i \in [N]$. Consider the sequences $(w^k)_{k\in\mathbb{N}}$ and $(\boldsymbol{s}^k)_{k\in\mathbb{N}}$ generated by Algorithm 3 and let $(\boldsymbol{z}^k)_{k\in\mathbb{N}} = (s_1^k + \gamma_1 w^k, \dots, s_N^k + \gamma_N w^k)_{k\in\mathbb{N}}$. Then, the following hold almost surely (resp. surely) with randomized sampling strategy as in Assumption II (resp. with any essentially cyclic sampling strategy and g convex as required in Assumption III):*

(i) the sequence $(\Phi(\boldsymbol{z}^k))_{k\in\mathbb{N}}$ converges to a finite value $\Phi_\star \le \Phi(\boldsymbol{x}^{\text{init}})$;

(ii) all cluster points of $(\boldsymbol{z}^k)_{k\in\mathbb{N}}$ are stationary and on which Φ equals $\Phi_\star$.

If, additionally, Φ is coercive, then the following also hold:

(iii) $(\boldsymbol{z}^k)_{k\in\mathbb{N}}$ *is bounded (in fact, this holds surely for arbitrary sampling criteria).*

Corollary 4.3 (linear convergence of Algorithm 3 under strong convexity)**.** *Additionally to the assumptions of Corollary 4.2, suppose that* g *is convex and that each* f_i *is* μ_{f_i}*-strongly convex. The following hold:*
RANDOMIZED SAMPLING: *under Assumption II,*

$$\mathbb{E}\Big[\Phi(\boldsymbol{z}^k) - \min \Phi\Big] \leq (\Phi(\boldsymbol{x}^{\text{init}}) - \min \Phi)(1-c)^k$$
$$\tfrac{1}{2}\mathbb{E}\Big[\|\boldsymbol{z}^k - \boldsymbol{x}^\star\|^2_{\mu_F}\Big] \leq (\Phi(\boldsymbol{x}^{\text{init}}) - \min \Phi)(1-c)^k$$

holds for all $k \in \mathbb{N}$*, where* $\boldsymbol{x}^\star \coloneqq \arg\min \Phi$*,* $\mu_F \coloneqq \frac{1}{N}\operatorname{blkdiag}(\mu_{f_1}\mathrm{I}_{n_1}, \dots \mu_{f_n}\mathrm{I}_{n_N})$*, and* c *is as in* (2.7)*. If the stepsizes* γ_i *and the sampling probabilities* p_i *are set as in Theorem 2.7, then the tighter constant* c *as in* (2.9) *is obtained.*
SHUFFLED CYCLIC OR CYCLIC SAMPLING: *under either sampling strategy* (2.10) *or* (2.11)*,*

$$\Phi(\boldsymbol{z}^{N\nu}) - \min \Phi \leq (\Phi(\boldsymbol{x}^{\text{init}}) - \min \Phi)(1-c)^\nu$$
$$\tfrac{1}{2}\|\boldsymbol{z}^{N\nu} - \boldsymbol{x}^\star\|^2_{\mu_F} \leq (\Phi(\boldsymbol{x}^{\text{init}}) - \min \Phi)(1-c)^\nu$$

holds surely for all $\nu \in \mathbb{N}$*, where* c *is as in* (2.21)*.*

We conclude with an immediate consequence of Theorem 2.11 that shows that (strong) convexity is in fact not necessary for global or linear convergence to hold.

Corollary 4.4 (global and linear convergence of Algorithm 3)**.** *In problem* (1.4)*, suppose that* Φ *has the KL property with exponent* $\theta \in (0, 1)$ *(as is the case when* g *and* f_i *are semialgebraic) and is coercive,* g *is proper convex lsc, and each* f_i *is* L_{f_i}*-Lipschitz differentiable,* $i \in [N]$*. Then the sequence* $(\boldsymbol{z}^k)_{k\in\mathbb{N}}$ *as defined in Corollary 4.2 with any essentially cyclic sampling strategy as in Assumption III converges surely to a stationary point for* Φ*. Moreover, if* $\theta \leq 1/2$ *it converges with R-linear rate.*

5 Conclusions

We presented a general block-coordinate forward-backward algorithm for minimizing the sum of a separable smooth and a nonseparable nonsmooth function, both allowed to be nonconvex. The framework is general enough to encompass regularized finite sum minimization and sharing problems, and leads to (a generalization of) the Finito/MISO algorithm [21,36] with new convergence results and with another novel incremental-type algorithm. The forward-backward envelope is shown to be a particularly suitable Lyapunov function for establishing convergence: additionally to enjoying favorable continuity properties, *sure* descent (as opposed to in expectation) occurs along the iterates. Possible future developments include extending the framework to account for a nonseparable smooth term, for instance by "quantifying the strength of coupling" between blocks of variables as in [9, §7.5].

A The key tool: the forward-backward envelope

This appendix contains some proofs and auxiliary results omitted in the main body. We begin by observing that, since F and $-F$ are 1-smooth in the metric induced by $\Lambda_F \coloneqq \frac{1}{N}\operatorname{blkdiag}(L_{f_1}\mathrm{I}_{n_1}, \dots, L_{f_N}\mathrm{I}_{n_N})$, one has

$$F(\boldsymbol{x}) + \langle \nabla F(\boldsymbol{x}), \boldsymbol{w} - \boldsymbol{x}\rangle - \tfrac{1}{2}\|\boldsymbol{w} - \boldsymbol{x}\|^2_{\Lambda_F} \le F(\boldsymbol{w}) \le F(\boldsymbol{x}) + \langle \nabla F(\boldsymbol{x}), \boldsymbol{w} - \boldsymbol{x}\rangle + \tfrac{1}{2}\|\boldsymbol{w} - \boldsymbol{x}\|^2_{\Lambda_F} \tag{A.1}$$

for all $\boldsymbol{x}, \boldsymbol{w} \in \mathbb{R}^{\sum_i n_i}$, see [8, Prop. A.24]. Let us denote

$$\mathcal{M}_\Gamma(\boldsymbol{w}, \boldsymbol{x}) \coloneqq F(\boldsymbol{x}) + \langle \nabla F(\boldsymbol{x}), \boldsymbol{w} - \boldsymbol{x}\rangle + G(\boldsymbol{w}) + \tfrac{1}{2}\|\boldsymbol{w} - \boldsymbol{x}\|^2_{\Gamma^{-1}}$$

the quantity being minimized (with respect to $\boldsymbol{w}$) in the definition (2.2a) of the FBE. It follows from (A.1) that

$$\Phi(\boldsymbol{w}) + \tfrac{1}{2}\|\boldsymbol{w} - \boldsymbol{x}\|^2_{\Gamma^{-1}-\Lambda_F} \le \mathcal{M}_\Gamma(\boldsymbol{w}, \boldsymbol{x}) \le \Phi(\boldsymbol{w}) + \tfrac{1}{2}\|\boldsymbol{w} - \boldsymbol{x}\|^2_{\Gamma^{-1}+\Lambda_F} \tag{A.2}$$

holds for all $\boldsymbol{x}, \boldsymbol{w} \in \mathbb{R}^{\sum_i n_i}$. In particular, $\mathcal{M}_\Gamma$ is a *majorizing model* for Φ, in the sense that $\mathcal{M}_\Gamma(\boldsymbol{x}, \boldsymbol{x}) = \Phi(\boldsymbol{x})$ and $\mathcal{M}_\Gamma(\boldsymbol{w}, \boldsymbol{x}) \ge \Phi(\boldsymbol{w})$ for all $\boldsymbol{x}, \boldsymbol{w} \in \mathbb{R}^{\sum_i n_i}$. In fact, as explained in Section 2.1, while a Γ-forward-backward step $\boldsymbol{z} \in \mathrm{T}^{\mathrm{FB}}_\Gamma(\boldsymbol{x})$ amounts to evaluating a minimizer of $\mathcal{M}_\Gamma(\,\cdot\,, \boldsymbol{x})$, the FBE is defined instead as the minimization value, namely $\Phi^{\mathrm{FB}}_\Gamma(\boldsymbol{x}) = \mathcal{M}_\Gamma(\boldsymbol{z}, \boldsymbol{x})$ where $\boldsymbol{z}$ is any element of $\mathrm{T}^{\mathrm{FB}}_\Gamma(\boldsymbol{x})$.

A.1 Proofs of Section 2.1

Proof of Lemma 2.1. For $\boldsymbol{x}^\star \in \arg\min \Phi$ it follows from (A.1) that

$$\min \Phi \le F(\boldsymbol{x}) + G(\boldsymbol{x}) \le G(\boldsymbol{x}) + F(\boldsymbol{x}^\star) + \langle \nabla F(\boldsymbol{x}^\star), \boldsymbol{x} - \boldsymbol{x}^\star\rangle + \tfrac{1}{2}\|\boldsymbol{x}^\star - \boldsymbol{x}\|^2_{\Lambda_F}.$$

Therefore, G is lower bounded by a quadratic function with quadratic term $-\frac{1}{2}\|\cdot\|^2_{\Lambda_F}$, and thus is prox-bounded in the sense of [51, Def. 1.23]. The claim then follows from [51, Th. 1.25 and Ex. 5.23(b)] and the continuity of the forward mapping $\mathrm{id} - \Gamma\nabla F$. □

***Proof of Lemma 2.3** (FBE: fundamental inequalities).* Local Lipschitz continuity follows from (2.2d) in light of Lemma 2.1 and [51, Ex. 10.32].

♠ 2.3*(i)* Follows by replacing $\boldsymbol{w} = \boldsymbol{x}$ in (2.2a).

♠ 2.3*(ii)* Directly follows from (A.2) and the identity $\Phi^{\mathrm{FB}}_\Gamma(\boldsymbol{x}) = \mathcal{M}_\Gamma(\boldsymbol{z}, \boldsymbol{x})$ for $\boldsymbol{z} \in \mathrm{T}^{\mathrm{FB}}_\Gamma(\boldsymbol{x})$. □

***Proof of Lemma 2.4** (FBE: minimization equivalence).*

♠ 2.4*(i)* and 2.4*(ii)* It follows from Lemma 2.3*(i)* that $\inf \Phi^{\mathrm{FB}}_\Gamma \le \min \Phi$. Conversely, let $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ be such that $\Phi^{\mathrm{FB}}_\Gamma(\boldsymbol{x}^k) \to \inf \Phi^{\mathrm{FB}}_\Gamma$ as $k \to \infty$, and for each k let $\boldsymbol{z}^k \in \mathrm{T}^{\mathrm{FB}}_\Gamma(\boldsymbol{x}^k)$. It then follows from Lemmas 2.3*(i)* and 2.3*(ii)* that

$$\inf \Phi^{\mathrm{FB}}_\Gamma \le \min \Phi \le \liminf_{k\to\infty} \Phi(\boldsymbol{z}^k) \le \liminf_{k\to\infty} \Phi^{\mathrm{FB}}_\Gamma(\boldsymbol{x}^k) = \inf \Phi^{\mathrm{FB}}_\Gamma,$$

hence $\min \Phi = \inf \Phi^{\mathrm{FB}}_\Gamma$. Suppose now that $\boldsymbol{x} \in \arg\min \Phi$ (which exists by Assumption I); then it follows from Lemma 2.3*(ii)* that $\mathrm{T}^{\mathrm{FB}}_\Gamma(\boldsymbol{x}) = \{\boldsymbol{x}\}$ (for otherwise another element would belong to a lower level set of Φ). Combining with Lemma 2.3*(i)* with $\boldsymbol{z} = \boldsymbol{x}$ we then have

$$\min \Phi = \Phi(\boldsymbol{z}) \le \Phi^{\mathrm{FB}}_\Gamma(\boldsymbol{x}) \le \Phi(\boldsymbol{x}) = \min \Phi.$$

Since $\min \Phi = \inf \Phi^{\mathrm{FB}}_\Gamma$, we conclude that $\boldsymbol{x} \in \arg\min \Phi^{\mathrm{FB}}_\Gamma$, and that in particular $\inf \Phi^{\mathrm{FB}}_\Gamma = \min \Phi^{\mathrm{FB}}_\Gamma$. Conversely, suppose $\boldsymbol{x} \in \arg\min \Phi^{\mathrm{FB}}_\Gamma$ and let $\boldsymbol{z} \in \mathrm{T}^{\mathrm{FB}}_\Gamma(\boldsymbol{x})$. By combining Lemmas 2.3*(i)* and 2.3*(ii)* we have that $\boldsymbol{z} = \boldsymbol{x}$, that is, that $\mathrm{T}^{\mathrm{FB}}_\Gamma(\boldsymbol{x}) = \{\boldsymbol{x}\}$. It then follows from Lemma 2.3*(ii)* and assertion 2.4*(i)* that

$$\Phi(\boldsymbol{x}) = \Phi(\boldsymbol{z}) \le \Phi^{\mathrm{FB}}_\Gamma(\boldsymbol{x}) = \min \Phi^{\mathrm{FB}}_\Gamma = \min \Phi,$$

hence $\boldsymbol{x} \in \arg\min \Phi$.

♠ 2.4*(iii)* Due to Lemma 2.3*(i)*, if Φ_Γ^{FB} is level bounded clearly so is Φ. Conversely, suppose that Φ_Γ^{FB} is not level bounded. Then, there exist $\alpha \in \mathbb{R}$ and $(\boldsymbol{x}^k)_{k\in\mathbb{N}} \subseteq \operatorname{lev}_{\leq\alpha} \Phi_\Gamma^{\text{FB}}$ such that $\|\boldsymbol{x}^k\| \to \infty$ as $k \to \infty$. Let $\lambda = \min_i \left\{\gamma_i^{-1} - L_{f_i} N^{-1}\right\} > 0$, and for each $k \in \mathbb{N}$ let $\boldsymbol{z}^k \in \mathrm{T}_\Gamma^{\text{FB}}(\boldsymbol{x}^k)$. It then follows from Lemma 2.3*(ii)* that

$$\min \Phi \leq \Phi(\boldsymbol{z}^k) \leq \Phi_\Gamma^{\text{FB}}(\boldsymbol{x}^k) - \tfrac{\lambda}{2}\|\boldsymbol{x}^k - \boldsymbol{z}^k\|^2 \leq \alpha - \tfrac{\lambda}{2}\|\boldsymbol{x}^k - \boldsymbol{z}^k\|^2,$$

hence $(\boldsymbol{z}^k)_{k\in\mathbb{N}} \subseteq \operatorname{lev}_{\leq\alpha} \Phi$ and $\|\boldsymbol{x}^k - \boldsymbol{z}^k\|^2 \leq \frac{2}{\lambda}(\alpha - \min \Phi)$. Consequently, also the sequence $(\boldsymbol{z}^k)_{k\in\mathbb{N}} \subseteq \operatorname{lev}_{\leq\alpha} \Phi$ is unbounded, proving that Φ is not level bounded. □

A.2 Further results

This section contains a list of auxiliary results invoked in the main proofs of Section 2.

Lemma A.1. *Suppose that Assumption I holds, and let two sequences $(\boldsymbol{u}^k)_{k\in\mathbb{N}}$ and $(\boldsymbol{v}^k)_{k\in\mathbb{N}}$ satisfy $\boldsymbol{v}^k \in \mathrm{T}_\Gamma^{\text{FB}}(\boldsymbol{u}^k)$ for all k and be such that both converge to a point $\boldsymbol{u}^\star$ as $k \to \infty$. Then, $\boldsymbol{u}^\star \in \mathrm{T}_\Gamma^{\text{FB}}(\boldsymbol{u}^\star)$, and in particular $0 \in \hat{\partial}\Phi(\boldsymbol{u}^\star)$.*

Proof. Since ∇F is continuous, it holds that $\boldsymbol{u}^k - \Gamma\nabla F(\boldsymbol{u}^k) \to \boldsymbol{u}^\star - \Gamma\nabla F(\boldsymbol{u}^\star)$ as $k \to \infty$. From outer semicontinuity of $\operatorname{prox}_G^{\Gamma^{-1}}$ [51, Ex. 5.23(b)] it then follows that

$$\boldsymbol{u}^\star = \lim_{k\to\infty} \boldsymbol{v}^k \in \limsup_{k\to\infty} \operatorname{prox}_G^{\Gamma^{-1}}(\boldsymbol{u}^k - \Gamma\nabla F(\boldsymbol{u}^k)) \subseteq \operatorname{prox}_G^{\Gamma^{-1}}(\boldsymbol{u}^\star - \Gamma\nabla F(\boldsymbol{u}^\star)) = \mathrm{T}_\Gamma^{\text{FB}}(\boldsymbol{u}^\star),$$

where the limit superior is meant in the Painlevé-Kuratowski sense, cf. [51, Def. 4.1]. The optimality conditions defining $\operatorname{prox}_G^{\Gamma^{-1}}$ [51, Th. 10.1] then read

$$\begin{aligned} 0 &\in \hat{\partial}\Big(G + \tfrac{1}{2}\|\cdot - (\boldsymbol{u}^\star - \Gamma\nabla F(\boldsymbol{u}^\star))\|_{\Gamma^{-1}}^2\Big)(\boldsymbol{u}^\star) = \hat{\partial}G(\boldsymbol{u}^\star) + \Gamma^{-1}\Big(\boldsymbol{u}^\star - (\boldsymbol{u}^\star - \Gamma\nabla F(\boldsymbol{u}^\star))\Big) \\ &= \hat{\partial}G(\boldsymbol{u}^\star) + \nabla F(\boldsymbol{u}^\star) = \hat{\partial}\Phi(\boldsymbol{u}^\star), \end{aligned}$$

where the first and last equalities follow from [51, Ex. 8.8(c)]. □

Lemma A.2. *Suppose that Assumption I holds and that function G is convex. Then, the following hold:*

(i) $\operatorname{prox}_G^{\Gamma^{-1}}$ *is (single-valued and) firmly nonexpansive (FNE) in the metric $\|\cdot\|_{\Gamma^{-1}}$; namely,*

$$\|\operatorname{prox}_G^{\Gamma^{-1}}(\boldsymbol{u}) - \operatorname{prox}_G^{\Gamma^{-1}}(\boldsymbol{v})\|_{\Gamma^{-1}}^2 \leq \langle \operatorname{prox}_G^{\Gamma^{-1}}(\boldsymbol{u}) - \operatorname{prox}_G^{\Gamma^{-1}}(\boldsymbol{v}), \Gamma^{-1}(\boldsymbol{u} - \boldsymbol{v})\rangle \leq \|\boldsymbol{u} - \boldsymbol{v}\|_{\Gamma^{-1}}^2 \quad \forall \boldsymbol{u}, \boldsymbol{v};$$

(ii) the Moreau envelope $G^{\Gamma^{-1}}$ is differentiable with $\nabla G^{\Gamma^{-1}} = \Gamma^{-1}(\mathrm{id} - \operatorname{prox}_G^{\Gamma^{-1}})$;

(iii) for every $\boldsymbol{x} \in \mathbb{R}^{\sum_i n_i}$ it holds that $\operatorname{dist}(0, \partial\Phi_\Gamma^{\text{FB}}(\boldsymbol{x})) \leq \frac{N + \max_i\{\gamma_i L_{f_i}\}}{N \min_i\{\sqrt{\gamma_i}\}} \|\boldsymbol{x} - \mathrm{T}_\Gamma^{\text{FB}}(\boldsymbol{x})\|_{\Gamma^{-1}}$;

(iv) $\mathrm{T}_\Gamma^{\text{FB}}$ is $L_{\mathbf{T}}$-Lipschitz continuous in the metric $\|\cdot\|_{\Gamma^{-1}}$ for some $L_{\mathbf{T}} \geq 0$;

If in addition f_i is μ_{f_i}-strongly convex, $i \in [N]$, then the following hold:

(v) In A.2(iv), $L_{\mathbf{T}} \leq 1 - \delta$ for $\delta = \frac{1}{N}\min_{i\in[N]}\{\gamma_i\mu_{f_i}\}$;

(vi) For every $\boldsymbol{x} \in \mathbb{R}^{\sum_i n_i}$

$$\tfrac{1}{2}\|\boldsymbol{z} - \boldsymbol{x}^\star\|_{\mu_F}^2 \leq \Phi_\Gamma^{\text{FB}}(\boldsymbol{x}) - \min \Phi \leq \tfrac{1}{2}\|\boldsymbol{z} - \boldsymbol{x}\|^2_{\Gamma^{-2}\mu_F^{-1}(\mathrm{I} - \Gamma\mu_F)}$$

where $\boldsymbol{x}^\star := \arg\min \Phi$, $\mu_F := \frac{1}{N}\operatorname{blkdiag}(\mu_{f_1}\mathrm{I}_{n_1}, \ldots, \mu_{f_N}\mathrm{I}_{n_N})$, and $\boldsymbol{z} = \mathrm{T}_\Gamma^{\text{FB}}(\boldsymbol{x})$.

Proof.

♠ A.2*(i)* and A.2*(ii)* See [4, Prop.s 12.28 and 12.30].

♠ A.2*(iii)* Let $D \subseteq \mathbb{R}^{\sum_i n_i}$ be the set of points at which ∇F is differentiable. From the chain rule of differentiation applied to the expression (2.2d) and using assertion A.2*(ii)*, we have that $\Phi_\Gamma^{\mathrm{FB}}$ is differentiable on D with gradient

$$\nabla\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}) = [\mathrm{I} - \Gamma\nabla^2 F(\boldsymbol{x})]\Gamma^{-1}[\boldsymbol{x} - \mathrm{T}_\Gamma^{\mathrm{FB}}(\boldsymbol{x})] \quad \forall \boldsymbol{x} \in D.$$

Since D is dense in $\mathbb{R}^{\sum_i n_i}$ owing to Lipschitz continuity of ∇F, we may invoke [51, Th. 9.61] to infer that $\partial\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x})$ is nonempty for every $\boldsymbol{x} \in \mathbb{R}^{\sum_i n_i}$ and

$$\partial\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}) \supseteq \partial_B\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}) = [\mathrm{I} - \Gamma\partial_B\nabla F(\boldsymbol{x})]\Gamma^{-1}[\boldsymbol{x} - \mathrm{T}_\Gamma^{\mathrm{FB}}(\boldsymbol{x})] = [\Gamma^{-1} - \partial_B\nabla F(\boldsymbol{x})][\boldsymbol{x} - \mathrm{T}_\Gamma^{\mathrm{FB}}(\boldsymbol{x})],$$

where ∂_B denotes the (set-valued) Bouligand differential [22, §7.1]. The claim now follows by observing that $\partial_B\nabla F(\boldsymbol{x}) = \frac{1}{N}\,\mathrm{blkdiag}(\partial_B\nabla f_1(x_1), \dots, \partial_B\nabla f_N(x_N))$ and that each element of $\partial_B\nabla f_i(x_i)$ has norm bounded by L_{f_i}.

♠ A.2*(iv)* Lipschitz continuity follows from assertion A.2*(i)* together with the fact that Lipschitz continuity is preserved by composition.

♠ A.2*(v)* By [41, Thm 2.1.12] for all $x_i, y_i \in \mathbb{R}^{n_i}$

$$\langle \nabla f_i(x_i) - \nabla f_i(y_i), x_i - y_i\rangle \geq \tfrac{\mu_{f_i} L_{f_i}}{\mu_{f_i}+L_{f_i}}\|x_i - y_i\|^2 + \tfrac{1}{\mu_{f_i}+L_{f_i}}\|\nabla f_i(x_i) - \nabla f_i(y_i)\|^2. \tag{A.3}$$

For the forward operator we have

$$\begin{aligned}
&\|(\mathrm{id} - \tfrac{\gamma_i}{N}\nabla f_i)(x_i) - (\mathrm{id} - \tfrac{\gamma_i}{N}\nabla f_i)(y_i)\|^2 \\
&= \|x_i - y_i\|^2 + \tfrac{\gamma_i^2}{N^2}\|\nabla f_i(x_i) - \nabla f_i(y_i)\|^2 - \tfrac{2\gamma_i}{N}\langle x_i - y_i, \nabla f_i(x_i) - \nabla f_i(y_i)\rangle \\
&\overset{(A.3)}{\leq} \Big(1 - \tfrac{\gamma_i^2\mu_{f_i}L_{f_i}}{N^2}\Big)\|x_i - y_i\|^2 - \tfrac{\gamma_i}{N}\Big(2 - \tfrac{\gamma_i}{N}(\mu_{f_i} + L_{f_i})\Big)\langle \nabla f_i(x_i) - \nabla f_i(y_i), x_i - y_i\rangle \\
&\leq \left(1 - \tfrac{\gamma_i^2\mu_{f_i}L_{f_i}}{N^2}\right)\|x_i - y_i\|^2 - \tfrac{\gamma_i\mu_{f_i}}{N}\Big(2 - \tfrac{\gamma_i}{N}(\mu_{f_i} + L_{f_i})\Big)\|x_i - y_i\|^2 \\
&= \Big(1 - \tfrac{\gamma_i\mu_{f_i}}{N}\Big)^2\|x_i - y_i\|^2,
\end{aligned}$$

where strong convexity and the fact that $\gamma_i < {}^{N}/_{L_{f_i}} \leq {}^{2N}/_{(\mu_{f_i}+L_{f_i})}$ were used in the second inequality. Multiplying by γ_i^{-1} and summing over i shows that $\mathrm{id} - \Gamma\nabla F$ is $(1-\delta)$-contractive in the metric $\|\cdot\|_{\Gamma^{-1}}$, and so is $\mathrm{T}_\Gamma^{\mathrm{FB}} = \mathrm{prox}_G^{\Gamma^{-1}} \circ (\mathrm{id} - \Gamma\nabla F)$ as it follows from assertion A.2*(i)*.

♠ A.2*(vi)* By strong convexity, denoting $\Phi_\star := \min\Phi$, we have

$$\Phi_\star \leq \Phi(\boldsymbol{z}) - \tfrac{1}{2}\|\boldsymbol{z} - \boldsymbol{x}^\star\|_{\mu_F}^2 \leq \Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}) - \tfrac{1}{2}\|\boldsymbol{z} - \boldsymbol{x}^\star\|_{\mu_F}^2$$

where the second inequality follows from Lemma 2.3*(ii)*. This establishes the lower bound.
Since $\boldsymbol{z}$ is a minimizer in (2.2a), the necessary stationarity condition reads $\Gamma^{-1}(\boldsymbol{x} - \boldsymbol{z}) - \nabla F(\boldsymbol{x}) \in \partial G(\boldsymbol{z})$. Convexity of G then implies

$$G(\boldsymbol{x}^\star) \geq G(\boldsymbol{z}) + \langle \Gamma^{-1}(\boldsymbol{x} - \boldsymbol{z}) - \nabla F(\boldsymbol{x}), \boldsymbol{x}^\star - \boldsymbol{z}\rangle,$$

whereas from strong convexity of F we have

$$F(\boldsymbol{x}^\star) \geq F(\boldsymbol{x}) + \langle \nabla F(\boldsymbol{x}), \boldsymbol{x}^\star - \boldsymbol{x}\rangle + \tfrac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}^\star\|_{\mu_F}^2.$$

By combining these inequalities and (2.2b), we have

$$\begin{aligned}
\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}) - \Phi_\star &\leq \tfrac{1}{2}\|\boldsymbol{z} - \boldsymbol{x}\|_{\Gamma^{-1}}^2 - \tfrac{1}{2}\|\boldsymbol{x}^\star - \boldsymbol{x}\|_{\mu_F}^2 + \langle \Gamma^{-1}(\boldsymbol{z} - \boldsymbol{x}), \boldsymbol{x}^\star - \boldsymbol{z}\rangle \\
&= \tfrac{1}{2}\|\boldsymbol{z} - \boldsymbol{x}\|_{\Gamma^{-1}-\mu_F}^2 + \langle (\Gamma^{-1} - \mu_F)(\boldsymbol{z} - \boldsymbol{x}), \boldsymbol{x}^\star - \boldsymbol{z}\rangle - \tfrac{1}{2}\|\boldsymbol{x}^\star - \boldsymbol{z}\|_{\mu_F}^2.
\end{aligned}$$

Next, by using the inequality $\langle \boldsymbol{a}, \boldsymbol{b}\rangle \leq \frac{1}{2}\|\boldsymbol{a}\|_{\mu_F}^2 + \frac{1}{2}\|\boldsymbol{b}\|_{\mu_F^{-1}}^2$ to cancel out the last term, we obtain

$$\begin{aligned}
\Phi_\Gamma^{\mathrm{FB}}(\boldsymbol{x}) - \Phi_\star &\leq \tfrac{1}{2}\|\boldsymbol{z} - \boldsymbol{x}\|_{\Gamma^{-1}-\mu_F}^2 + \tfrac{1}{2}\|(\Gamma^{-1} - \mu_F)(\boldsymbol{x} - \boldsymbol{z})\|_{\mu_F^{-1}}^2 \\
&= \tfrac{1}{2}\|\boldsymbol{z} - \boldsymbol{x}\|_{\Gamma^{-2}\mu_F^{-1}(\mathrm{I}-\Gamma\mu_F)}^2,
\end{aligned}$$

where the last identity uses the fact that the matrices are diagonal. □

The next result recaps an important property that the FBE inherits from the cost function Φ that is instrumental for establishing global convergence and asymptotic linear rates for the BC Algorithm 1. The result falls as special case of [64, Th. 5.2] after observing that

$$\Phi_\Gamma^{\text{FB}}(\boldsymbol{x}) = \inf_{\boldsymbol{w}} \{\Phi(\boldsymbol{w}) + D_H(\boldsymbol{w}, \boldsymbol{x})\},$$

where $D_H(\boldsymbol{w}, \boldsymbol{x}) = H(\boldsymbol{w}) - H(\boldsymbol{x}) - \langle \nabla H(\boldsymbol{x}), \boldsymbol{w} - \boldsymbol{x} \rangle$ is the Bregman distance with kernel $H = \frac{1}{2}\|\cdot\|^2_{\Gamma^{-1}} - F$.

Lemma A.3 ([64, Th. 5.2])**.** *Suppose that Assumption I holds and for $\gamma_i \in (0, {}^{N}/_{L_{f_i}})$, $i \in [N]$, let $\Gamma = \operatorname{blkdiag}(\gamma_1 \mathrm{I}_{n_1}, \ldots, \gamma_N \mathrm{I}_{n_N})$. If Φ has the KL property with exponent $\theta \in (0, 1)$ (as is the case when f_i and G are semialgebraic), then so does Φ_Γ^{FB} with exponent* $\max\{{}^1/_2, \theta\}$.

References

1. Attouch, H., Bolte, J.: On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. Mathematical Programming **116**(1-2), 5–16 (2009)
2. Attouch, H., Bolte, J., Redont, P., Soubeyran, A.: Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality. Mathematics of Operations Research **35**(2), 438–457 (2010)
3. Attouch, H., Bolte, J., Svaiter, B.F.: Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. Mathematical Programming **137**(1), 91–129 (2013)
4. Bauschke, H.H., Combettes, P.L.: Convex analysis and monotone operator theory in Hilbert spaces. CMS Books in Mathematics. Springer (2017)
5. Beck, A.: First-Order Methods in Optimization. Society for Industrial and Applied Mathematics, Philadelphia, PA (2017)
6. Beck, A., Tetruashvili, L.: On the convergence of block coordinate descent type methods. SIAM journal on Optimization **23**(4), 2037–2060 (2013)
7. Bertsekas, D.P.: Incremental proximal methods for large scale convex optimization. Mathematical programming **129**(2), 163–195 (2011)
8. Bertsekas, D.P.: Nonlinear Programming. Athena Scientific (2016)
9. Bertsekas, D.P., Tsitsiklis, J.N.: Parallel and distributed computation: numerical methods, vol. 23. Prentice-Hall (1989)
10. Bianchi, P., Hachem, W., Iutzeler, F.: A coordinate descent primal-dual algorithm and application to distributed asynchronous optimization. IEEE Transactions on Automatic Control **61**(10), 2947–2957 (2016)
11. Bolte, J., Daniilidis, A., Lewis, A.: The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. SIAM Journal on Optimization **17**(4), 1205–1223 (2007)
12. Bolte, J., Daniilidis, A., Lewis, A., Shiota, M.: Clarke subgradients of stratifiable functions. SIAM Journal on Optimization **18**(2), 556–572 (2007)
13. Bolte, J., Sabach, S., Teboulle, M.: Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Mathematical Programming **146**(1-2), 459–494 (2014)
14. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends in Machine Learning **3**(1), 1–122 (2011)
15. Chouzenoux, E., Pesquet, J.C., Repetti, A.: A block coordinate variable metric forward–backward algorithm. Journal of Global Optimization **66**(3), 457–485 (2016)
16. Chow, Y.T., Wu, T., Yin, W.: Cyclic coordinate-update algorithms for fixed-point problems: Analysis and applications. SIAM Journal on Scientific Computing **39**(4), A1280–A1300 (2017)
17. Clarke, F.H.: Optimization and Nonsmooth Analysis. Society for Industrial and Applied Mathematics (1990). DOI 10.1137/1.9781611971309
18. Combettes, P.L., Pesquet, J.C.: Stochastic quasi-Fejér block-coordinate fixed point iterations with random sweeping. SIAM Journal on Optimization **25**(2), 1221–1248 (2015)
19. Davis, D.: Smart: The stochastic monotone aggregated root-finding algorithm. arXiv:1601.00698 (2016)

20. Defazio, A., Bach, F., Lacoste-Julien, S.: SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In: Advances in neural information processing systems, pp. 1646–1654 (2014)
21. Defazio, A., Domke, J.: Finito: A faster, permutable incremental gradient method for big data problems. In: International Conference on Machine Learning, pp. 1125–1133 (2014)
22. Facchinei, F., Pang, J.S.: Finite-dimensional variational inequalities and complementarity problems, vol. II. Springer (2003)
23. Fercoq, O., Bianchi, P.: A coordinate-descent primal-dual algorithm with large step size and possibly nonseparable functions. SIAM Journal on Optimization **29**(1), 100–134 (2019)
24. Frankel, P., Garrigos, G., Peypouquet, J.: Splitting methods with variable metric for Kurdyka-Łojasiewicz functions and general convergence rates. Journal of Optimization Theory and Applications **165**(3), 874–900 (2015)
25. Fukushima, M., Mine, H.: A generalized proximal point algorithm for certain non-convex minimization problems. International Journal of Systems Science **12**(8), 989–1000 (1981)
26. Hanzely, F., Mishchenko, K., Richtarik, P.: SEGA: Variance reduction via gradient sketching. In: Advances in Neural Information Processing Systems, pp. 2082–2093 (2018)
27. Hong, M., Wang, X., Razaviyayn, M., Luo, Z.Q.: Iteration complexity analysis of block coordinate descent methods. Mathematical Programming **163**(1-2), 85–114 (2017)
28. Hou, Y., Song, I., Min, H.K., Park, C.H.: Complexity-reduced scheme for feature extraction with linear discriminant analysis. IEEE transactions on neural networks and learning systems **23**(6), 1003–1009 (2012)
29. Iutzeler, F., Bianchi, P., Ciblat, P., Hachem, W.: Asynchronous distributed optimization using a randomized alternating direction method of multipliers. In: 52nd IEEE Conference on Decision and Control (CDC), pp. 3671–3676 (2013)
30. Kurdyka, K.: On gradients of functions definable in o-minimal structures. Annales de l'institut Fourier **48**(3), 769–783 (1998)
31. Latafat, P., Freris, N.M., Patrinos, P.: A new randomized block-coordinate primal-dual proximal algorithm for distributed optimization. IEEE Transactions on Automatic Control **64**(10), 4050–4065 (2019)
32. Li, G., Pong, T.K.: Douglas-Rachford splitting for nonconvex optimization with application to nonconvex feasibility problems. Mathematical Programming **159**(1), 371–401 (2016)
33. Lin, Q., Lu, Z., Xiao, L.: An accelerated randomized proximal coordinate gradient method and its application to regularized empirical risk minimization. SIAM Journal on Optimization **25**(4), 2244–2273 (2015)
34. Łojasiewicz, S.: Une propriété topologique des sous-ensembles analytiques réels. Les équations aux dérivées partielles pp. 87–89 (1963)
35. Łojasiewicz, S.: Sur la géométrie semi- et sous- analytique. Annales de l'institut Fourier **43**(5), 1575–1595 (1993)
36. Mairal, J.: Incremental majorization-minimization optimization with application to large-scale machine learning. SIAM Journal on Optimization **25**(2), 829–855 (2015)
37. Mokhtari, A., Gürbüzbalaban, M., Ribeiro, A.: Surpassing gradient descent provably: A cyclic incremental method with linear convergence rate. SIAM Journal on Optimization **28**(2), 1420–1447 (2018)
38. Necoara, I.: Random coordinate descent algorithms for multi-agent convex optimization over networks. IEEE Transactions on Automatic Control **58**(8), 2001–2012 (2013)
39. Necoara, I., Patrascu, A.: A random coordinate descent algorithm for optimization problems with composite objective function and linear coupled constraints. Computational Optimization and Applications **57**(2), 307–337 (2014)
40. Nesterov, Y.: Efficiency of coordinate descent methods on huge-scale optimization problems. SIAM Journal on Optimization **22**(2), 341–362 (2012)
41. Nesterov, Y.: Introductory lectures on convex optimization: A basic course, vol. 87. Springer Science & Business Media (2013)
42. Ochs, P., Chen, Y., Brox, T., Pock, T.: iPiano: Inertial proximal algorithm for nonconvex optimization. SIAM Journal on Imaging Sciences **7**(2), 1388–1419 (2014)
43. Patrinos, P., Bemporad, A.: Proximal Newton methods for convex composite optimization. In: 52nd IEEE Conference on Decision and Control, pp. 2358–2363 (2013)
44. Peng, Z., Xu, Y., Yan, M., Yin, W.: ARock: An algorithmic framework for asynchronous parallel coordinate updates. SIAM Journal on Scientific Computing **38**(5), A2851–A2879 (2016)

45. Pesquet, J.C., Repetti, A.: A class of randomized primal-dual algorithms for distributed optimization. Journal of Nonlinear and Convex Analysis **16**(12), 2453–2490 (2015)
46. Qian, X., Sailanbayev, A., Mishchenko, K., Richtárik, P.: MISO is making a comeback with better proofs and rates. arXiv:1906.01474 (2019)
47. Reddi, S.J., Hefny, A., Sra, S., Poczos, B., Smola, A.J.: Stochastic variance reduction for nonconvex optimization. In: International conference on machine learning, pp. 314–323 (2016)
48. Reddi, S.J., Sra, S., Poczos, B., Smola, A.J.: Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization. In: Advances in Neural Information Processing Systems, pp. 1145–1153 (2016)
49. Richtárik, P., Takáč, M.: Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. Mathematical Programming **144**(1-2), 1–38 (2014)
50. Robbins, H., Siegmund, D.: A convergence theorem for non negative almost supermartingales and some applications. In: Herbert Robbins Selected Papers, pp. 111–135. Springer (1985)
51. Rockafellar, R.T., Wets, R.J.B.: Variational analysis, vol. 317. Springer Science & Business Media (2011)
52. Schmidt, M., Le Roux, N., Bach, F.: Minimizing finite sums with the stochastic average gradient. Mathematical Programming **162**(1), 83–112 (2017)
53. Shalev-Shwartz, S., Zhang, T.: Stochastic dual coordinate ascent methods for regularized loss minimization. Journal of Machine Learning Research **14**(Feb), 567–599 (2013)
54. Themelis, A.: Proximal algorithms for structured nonconvex optimization. Ph.D. thesis, KU Leuven (2018)
55. Themelis, A., Ahookhosh, M., Patrinos, P.: On the acceleration of forward-backward splitting via an inexact Newton method. In: H.H. Bauschke, R.S. Burachik, D.R. Luke (eds.) Splitting Algorithms, Modern Operator Theory, and Applications, pp. 363–412. Springer International Publishing, Cham (2019)
56. Themelis, A., Stella, L., Patrinos, P.: Forward-backward envelope for the sum of two nonconvex functions: Further properties and nonmonotone linesearch algorithms. SIAM Journal on Optimization **28**(3), 2274–2303 (2018)
57. Tseng, P.: Convergence of a block coordinate descent method for nondifferentiable minimization. Journal of optimization theory and applications **109**(3), 475–494 (2001)
58. Tseng, P., Bertsekas, D.P.: Relaxation methods for problems with strictly convex separable costs and linear constraints. Mathematical Programming **38**(3), 303–321 (1987)
59. Tseng, P., Yun, S.: Block-coordinate gradient descent method for linearly constrained nonsmooth separable optimization. Journal of Optimization Theory and Applications **140**(3), 513–535 (2008)
60. Tseng, P., Yun, S.: A coordinate gradient descent method for nonsmooth separable minimization. Mathematical Programming **117**(1), 387–423 (2009)
61. Tseng, P., Yun, S.: A coordinate gradient descent method for linearly constrained smooth optimization and support vector machines training. Computational Optimization and Applications **47**(2), 179–206 (2010)
62. Xu, Y., Yin, W.: A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. SIAM Journal on Imaging Sciences **6**(3), 1758–1789 (2013)
63. Xu, Y., Yin, W.: A globally convergent algorithm for nonconvex optimization based on block coordinate update. Journal of Scientific Computing **72**(2), 700–734 (2017)
64. Yu, P., Li, G., Pong, T.K.: Deducing Kurdyka-Łojasiewicz exponent via inf-projection. arXiv:1902.03635 (2019)