Block-coordinate and incremental aggregated nonconvex proximal gradient methods: a unified view

Puya Latafat  
Andreas Themelis  
Panagiotis Patrinos

Department of Electrical Engineering  
ESAT STADIUS — KU Leuven  
Kasteelpark Arenberg 10, 3001 Leuven, Belgium

Abstract

In this paper we study block-coordinate proximal gradient algorithms for minimizing the sum of a separable smooth function and a (nonseparable) nonsmooth function, both of which are allowed to be nonconvex. Two prominent special cases are the regularized finite sum minimization and the sharing problem. The main tool in the analysis of the block-coordinate algorithm is the forward-backward envelope (FBE), which serves as a Lyapunov function. Favorable properties of the FBE under extra assumptions lead to an accelerated variant. Moreover, our analysis covers the popular Finito/MISO algorithm as a special case, allowing to address nonsmooth and nonconvex problems with rather general sampling strategies.

1 Introduction

This paper proposes block-coordinate (BC) proximal algorithms for problems of the form

$$\min_{x=(x_1,\ldots,x_N)\in\mathbb{R}^{nN}} \Phi(x) := F(x) + G(x), \quad \text{where} \quad F(x) := \frac{1}{N} \sum_{i=1}^{N} f_i(x_i),$$

where $f_i: \mathbb{R}^{n_i} \to \mathbb{R}$ are smooth possibly nonconvex functions, $i \in [N] := \{1,\ldots,N\}$, and $G: \mathbb{R}^{n_N} \to \mathbb{R}$ is possibly nonconvex, nonsmooth and extended-real valued ($\mathbb{R} := \mathbb{R} \cup \{\infty\}$ denotes the extended-real line). Unlike typical cases analyzed in the literature where $G$ is separable [39, 33, 11, 26, 23, 18, 6], 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) especially in the nonconvex setting is the fact that even in expectation the cost does not necessarily decrease along the trajectories.

Thanks to the nonconvexity and nonseparability of $G$, many machine learning problems can be formulated as in (1), a primary example being constrained and/or regularized finite sum problems [36, 14, 15, 24, 37, 31, 32]

$$\min_{x \in \mathbb{R}^n} \varphi(x) := \frac{1}{N} \sum_{i=1}^{N} f_i(x) + g(x),$$

where $f_i: \mathbb{R}^{n_i} \to \mathbb{R}$ are smooth functions and $g: \mathbb{R}^{n_N} \to \mathbb{R}$ is possibly nonsmooth, and everything here can be nonconvex. In fact, one way to cast (2) into the form of problem (1) is by setting

$$G(x) := \frac{1}{N} \sum_{i=1}^{N} g(x_i) + \delta_C(x),$$

where $C := \{x \in \mathbb{R}^{n_N} \mid x_1 = x_2 = \cdots = x_N\}$ is the consensus set, and $\delta_C$ is the indicator function of set $C$, namely $\delta_C(x) = 0$ for $x \in C$ and $\infty$ otherwise. Since the nonsmooth term $g$ is allowed to be nonconvex, formulation (2) can account for nonconvex constraints such as rank constraints or zero norm balls, and nonconvex regularizers such as $\ell^p$ with $p \in [0, 1)$, [19].

Preprint. Under review.
1.1 Contribution

1) To the best of our knowledge this is the first analysis of BC schemes with nonseparable nonsmooth terms and in the fully nonconvex setting. Although the original cost $\Phi$ cannot serve as Lyapunov function, we show that (a generalized version of) the forward-backward envelope (FBE) \cite{ levy2014exact, levy2015global} decreases surely, not only in expectation (Lemma 1).

2) This allows for a quite general convergence analysis for different sampling criteria. This paper in particular covers very general randomized strategies, where at each iteration 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.

3) When $G$ is convex and $F$ is twice continuously differentiable, the FBE is continuously differentiable. If, additionally, $F$ is (strongly) convex and quadratic, then the FBE is (strongly) convex and has Lipschitz-continuous gradient. Owing to these favorable properties, we propose a new BC Nesterov-type acceleration algorithm for minimizing the sum of a block-separable convex quadratic plus a nonsmooth convex function, whose analysis directly follows from existing work on smooth BC minimization \cite{beck2009fast}.

4) As a byproduct of our analysis, we obtain new convergence results for the Finito/MISO algorithm \cite{haddad2017finite, beck2015fast} both for randomized sampling strategies in the fully nonconvex setting and for essentially cyclic samplings when the nonsmooth term is convex. Furthermore, we extend the linear convergence analysis for strongly convex problems by allowing for a convex nonsmooth term $g$, and further waiving the “big-data assumptions” required in the original analysis.

2 The main block-coordinate algorithm

Let us start by formally stating the main requirements for our analysis.

**Assumption 1** (problem setting). In problem (1) the following hold:

A1 functions $f_i$ are $L_{f_i}$-smooth (Lipschitz differentiable with modulus $L_{f_i}$), $i \in [N]$;

A2 function $G$ is proper and lower semicontinuous (lsc);

A3 a solution exists: $\arg\min_{x} \Phi(x) \neq \emptyset$.

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

$$\text{prox}^V_\psi(x) := \arg\min_{w \in \mathbb{R}^r} \left\{ \psi(w) + \frac{1}{2} \|w - x\|^2_V \right\},$$

(4)

where $V$ is a symmetric and positive definite matrix and $\| \cdot \|_V$ indicates the norm induced by the scalar product $(x, y) \mapsto \langle x, V y \rangle$. It is common to take $V = t^{-1}I_r$ as a multiple of the $r \times r$ identity matrix $I_r$, in which case the notation $\text{prox}^V_\psi$ 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 $\text{prox}^V_\psi : \mathbb{R}^r \rightrightarrows \mathbb{R}^r$. Nevertheless, the value function associated to the minimization problem in the definition (4), namely the Moreau envelope

$$\psi^V(x) := \min_{w \in \mathbb{R}^r} \left\{ \psi(w) + \frac{1}{2} \|w - x\|^2_V \right\},$$

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

In many machine learning 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 $\psi$. In these cases, the proximal gradient method \cite{parikh2014proximal, beck2017fast} constitutes a cornerstone iterative method that interleaves gradient descent steps on the smooth function and proximal operations on the nonsmooth functions, resulting in iterations of the form $x^{t+1} = \text{prox}^V_\psi(x^t - \gamma h(x^t))$ for some suitable stepsize $\gamma$. 


Our proposed scheme to address problem (1) is a BC variant of the proximal gradient method, 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. Although seemingly wasteful, in many cases one can efficiently

| Algorithm 1 General block-coordinate scheme |
|--------------------------------------------|
| **Require** $x^0 \in \mathbb{R}^{\sum_{i=1}^{N} n_i}$, $\gamma_i \in (0, 1/L_i)$, $i \in [N]$ |
| **Repeat** for $k = 0, 1, \ldots$ |
| 1: $z^k_i \in \text{prox}_G^{-1}(x^k - \Gamma \nabla F(x^k))$ |
| 2: select a set of indices $I^{k+1} \subseteq [N]$ |
| 3: update $x_i^{k+1} = z_i^k$ for $i \in I^{k+1}$ and $x_i^{k+1} = x_i^k$ for $i \notin I^{k+1}$ |

compute individual blocks without the need of full operations. Two such broad applications are discussed in the dedicated Sections 3 and 4, where among other things we will show that Algorithm 1 leads to the well known Finito/MISO algorithm [15, 24].

### 2.1 Convergence analysis

This subsection is devoted to the theoretical analysis of the BC-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 1, in Sections 2.1.1 and 2.1.2 we will specialize the results to randomized and (essentially) cyclic sampling strategies.

The fundamental challenge in the analysis of (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) [29, 38] can be used as Lyapunov function. Similarly to the relation existing among the Moreau envelope and the proximal mapping, the FBE is the value function associated with the proximal gradient mapping. We formally define the FBE as follows where we use $\| \cdot \|_{l^{-1}}$, with $\Gamma$ a diagonal matrix containing the stepsizes associated with each block of coordinates.

**Definition 1** (block-FBE). For given scalars $\gamma_1, \ldots, \gamma_N > 0$ let $\Gamma = \text{blkdiag}(\gamma_1 I_{n_1}, \ldots, \gamma_N I_{n_N})$. The $\Gamma$-forward-backward envelope ($\Gamma$-FBE) associated to (1) is the function

$$
\Phi^{\Gamma}_{n}(x) := \inf_{w \in \mathbb{R}^{\sum_{i=1}^{N} n_i}} \left\{ F(x) + \langle \nabla F(x), w - x \rangle + G(w) + \frac{1}{2\gamma_i} \|w - x\|_{l^{-1}}^2 \right\}. 
$$

(5a)

Alternatively, letting $z$ be any element of $\text{prox}_{G}^{-1}(x - \Gamma \nabla F(x))$, we have

$$
\Phi^{\Gamma}_{n}(x) := F(x) + \langle \nabla F(x), z - x \rangle + G(z) + \frac{1}{2} \|z - x\|_{l^{-1}}^2. 
$$

(5b)

Next, in Lemma 1(ii) we establish the sure descent property that is instrumental to our analysis. Equipped with the descent property we also establish other useful properties for the BC update.

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

(i) $\Phi^{\Gamma}_{n}(x^{k+1}) \leq \Phi^{\Gamma}_{n}(x^k) - \sum_{i \in I^{k+1}} \xi_i^k \|z_i^k - x_i^k\|_{l^{-1}}^2$, where $\xi_i^k := \frac{N - \gamma_i L_i}{N}$, $i \in [N]$, are strictly positive;

(ii) the sequence $(\Phi^{\Gamma}_{n}(x^k))_{k \in \mathbb{N}}$ monotonically decreases to a finite value $\Phi_* \geq \min \Phi$;

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

(iv) if $\Phi$ is coercive, then $(x^k)_{k \in \mathbb{N}}$ and $(z^k)_{k \in \mathbb{N}}$ are bounded.

#### 2.1.1 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 \( p_1, \ldots, p_N > 0 \) such that, at any iteration and independently of the past, each \( i \in [N] \) is sampled with probability at least \( p_i \).

Differently from classical approaches that require i.i.d. probability spaces, 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, \( \mathcal{P}_i[i \in I^{k+1}] \geq p_i \) 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, 21] in the convex case.

Due to the possible nonconvexity of problem (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 accumulation point \( x^* \) of the generated sequences is a stationary point, in the sense that it satisfies the necessary condition for minimality \( 0 \in \partial \Phi(x^*) \), where \( \partial \) denotes the (regular) nonconvex subdifferential, see [35, Thm. 10.1]. Later, in Theorem 3 the mild additional requirements ensuring global convergence will also be given.

Theorem 2 (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:

\begin{enumerate}[(i)]
  \item the sequence \( (\|\mathbf{x}^k - \mathbf{z}^k\|^2)_{k \in \mathbb{N}} \) has finite sum (and in particular vanishes);
  \item the sequence \( (\Phi(\mathbf{z}^k))_{k \in \mathbb{N}} \) converges to \( \Phi_\star \) as in Lemma I(ii);
  \item \( (\mathbf{x}^k)_{k \in \mathbb{N}} \) and \( (\mathbf{z}^k)_{k \in \mathbb{N}} \) have same cluster points, all stationary and on which \( \Phi \) and \( \Phi^{\text{true}} \) equal \( \Phi_\star \).
\end{enumerate}

Semialgebraic functions comprise a wide class of functions that enjoy the so-called Kurdyka-Łojasiewicz (KL) property, an important tool that has been extensively exploited to provide convergence rates of optimization algorithms [2, 3, 4, 11, 16, 28]. In the next result we show that whenever \( F \) and \( G \) are semialgebraic the randomized BC-Algorithm 1 converges globally to a stationary point. The proof is largely inspired by the analysis in [22] for the Douglas-Rachford splitting algorithm in the nonconvex setting.

Theorem 3 (randomized sampling: global convergence). Additionally to Assumptions I and II, suppose that the cost function \( \Phi \) is coercive and that \( f_i \) and \( G \) are semialgebraic functions. Then, the sequences \( (\mathbf{x}^k)_{k \in \mathbb{N}} \) and \( (\mathbf{z}^k)_{k \in \mathbb{N}} \) generated by Algorithm 1 converge almost surely to (the same) stationary point \( \mathbf{x}_\star \).

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

Theorem 4 (randomized sampling: linear convergence). Additionally to Assumptions I and II, suppose that \( G \) is convex and that each \( f_i \) is \( \mu_i \)-strongly convex. Denote \( \kappa_i := \frac{L_i}{\mu_i} \), and let the stepsizes \( \gamma_i \) and minimum sampling probabilities \( p_i \) be set to

\begin{equation}
\gamma_i = \frac{N}{\mu_i} \left( 1 - \sqrt{1 - 1/\kappa_i} \right) \quad \text{and} \quad p_i = \frac{\left( \sqrt{\kappa_i} + \sqrt{\kappa_i - 1} \right)^2}{\sum_{j=1}^{N} \left( \sqrt{\kappa_j} + \sqrt{\kappa_j - 1} \right)^2}, \quad i \in [N].
\end{equation}

Then, for all \( k \) the following holds for the iterates generated by Algorithm 1:

\[ \mathbb{E}_\epsilon [\Phi^{\text{true}}_\Gamma(\mathbf{x}^{k+1}) - \min \Phi] \leq (1 - c) \left( \Phi^{\text{true}}_\Gamma(\mathbf{x}^k) - \min \Phi \right), \quad \text{where} \quad c = \frac{1}{\sum_{i=1}^{N} \left( \sqrt{\kappa_i} + \sqrt{\kappa_i - 1} \right)^2}.
\]

Notice that as \( \kappa_i \)'s approach 1 the linear rate tends to \( 1 - 1/N \). We also remark that although Theorem 4 prescribes (6), a \( Q \)-linear rate still holds with any \( p_i \) and \( \gamma_i \), with a more conservative coefficient

\begin{equation}
c = \min_{i \in [N]} \left( \frac{\xi_i}{\gamma_i} \right)/\max_{i \in [N]} \left( \frac{N - \gamma_i \mu_i}{p_i^2 \mu_i} \right),
\end{equation}

where \( \xi_i \) are as in Lemma 1(i). Moreover, using the above \( Q \)-linear rate in Theorem 4, by lower bounding the envelope as in Lemma A.2(ii), the following \( R \)-linear rate follows immediately for the
distance from the solution:
\[
\mathbb{E}[\|x^k - x^*\|_M^2] \leq 4\mathbb{E}[\Phi^\mu_I(x^k) - \min \Phi] \leq 4(1 - c)^k(\Phi^\mu_I(x^0) - \min \Phi),
\]
where \( M > 0 \) is as in Lemma A.2(iii).

2.1.2 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 [40, 39, 18, 12] 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), 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.

Two notable special cases are the cyclic and shuffled cyclic sampling strategies.

**Shuffled cyclic sampling**: corresponds to setting
\[
I^{k+1} = \{\pi_{(k)}[\text{mod}(k, N) + 1]\} \quad \text{for all} \quad k \in \mathbb{N}, \tag{8}
\]
where \( \pi_0, \pi_1, \ldots \) are permutations of the set of indices \([N]\) (chosen randomly or deterministically).

**Cyclic sampling**: corresponds to the case (8) with \( \pi_{(k)} = \text{id}, \) i.e.,
\[
I^{k+1} = \{\text{mod}(k, N) + 1\} \quad \text{for all} \quad k \in \mathbb{N}. \tag{9}
\]

We remark that in practice it has been observed that an effective sampling technique is to use random shuffling after each cycle [8, §2].

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

**Theorem 5** (essentially cyclic sampling: subsequential convergence). Suppose that Assumptions I and III are satisfied. Then, all the asserts of Theorem 2 hold surely.

**Theorem 6** (essentially cyclic sampling: global convergence). Additionally to Assumptions I and III, suppose that the cost function \( \Phi \) is coercive and that \( f_i \) and \( G \) are semialgebraic functions. Then, the sequences \( (x^k)_{k \in \mathbb{N}} \) and \( (\bar{x}^k)_{k \in \mathbb{N}} \) generated by Algorithm 1 converge to (the same) stationary point \( x^\ast \).

**Theorem 7** (essentially cyclic sampling: linear convergence). Additionally to Assumptions I and III, suppose that each function \( f_i \) is \( \mu_f \) strongly convex. Then, denoting \( \delta := \min_{v \in [N]} \left\{ \frac{2\mu_f}{N} \right\} \) and \( \Delta := \max_{v \in [N]} \left\{ \frac{2\mu_f}{N} \right\} \), for all \( v \in \mathbb{N} \) the following holds for the iterates generated by Algorithm 1:
\[
\Phi^\mu_I(x^{(v+1)}) - \min \Phi \leq (1 - c)(\Phi^\mu_I(x^{(v)}) - \min \Phi), \quad \text{where} \quad c = \frac{\delta(1 - \Delta)}{N(1 + T(1 - \delta))^2(1 - \delta)}.
\]

In the case of shuffled cyclic (8) or cyclic (9) sampling the following tighter bound holds
\[
\Phi^\mu_I(x^{(v+1)}) - \min \Phi \leq (1 - c)(\Phi^\mu_I(x^{(v)}) - \min \Phi), \quad \text{where} \quad c = \frac{\delta(1 - \Delta)}{N(2 - \delta)^2(1 - \delta)}. \tag{10}
\]

Note that if one sets \( \gamma_i = \alpha N/L_f \) for some \( \alpha \in (0, 1) \), then \( \delta = \alpha \min_{v \in [N]} \left\{ \frac{\mu_f}{L_f} \right\} \) and \( \Delta = \alpha \). With this selection, as the condition number approaches 1 the rate in (10) tends to \( 1 - \frac{\alpha}{N(2 - \alpha)} \).

As argued in the randomized case, the \( R \)-linear rate
\[
\mathbb{E}[\|x^N - x^\ast\|_M^2] \leq 4(1 - c)^k(\Phi^\mu_I(x^0) - \min \Phi)
\]
for the (shuffled) cyclic case with \( M \) as in Lemma A.2(iii) is easily deduced.
3 Nonconvex finite sum problems: the Finito/MISO algorithm

As mentioned in Section 1, if $G$ is of the form (3) then problem (1) reduces to the finite sum minimization presented in (2). Most importantly, the proximal mapping of the original nonsmooth function $G$ in (1) can be easily expressed in terms of that of the small function $g$ in the reduced finite sum reformulation (2). To see this, observe that for any $w, x_j \in \mathbb{R}^p$ and $\pi_i > 0$, $i \in [N]$, it holds that

$$\sum_{j=1}^{N} \pi_j \|x_j - w\|^2 = \sum_{j=1}^{N} \pi_j \|x_j - \hat{x}\|^2 + P\|\hat{x} - w\|^2$$

where $P := \sum_{j=1}^{N} \pi_j$ and $\hat{x} := \frac{1}{\gamma} \sum_{j=1}^{N} \pi_j x_j$. \hspace{1cm} (11)

Next, observe that since dom $G \subseteq C$ (the consensus set), necessarily there exists $\hat{w}$ such that $\text{prox}_G^{-1}(x_i) \ni \hat{w}$ for all $i \in [N]$. Thus, $\text{prox}_G^{-1}(x_i)$, is characterized as

$$\text{prox}_G^{-1}(x_i) \in \arg\min_{w \in \mathbb{R}^p} \left\{ g(w) + \frac{1}{\gamma} \sum_{i=1}^{N} \gamma_i \|w - x_i\|^2 \right\} = \text{arg\min}_{w \in \mathbb{R}^p} \left\{ g(w) + \frac{1}{\gamma} \sum_{i=1}^{N} \pi_i^{-1} \|w - \hat{x}\|^2 \right\} = \text{prox}_g(\hat{x})$$

where the first equality comes from (11) with $\pi_j = \gamma_j^{-1}$, $\hat{y} := (\sum_{i=1}^{N} \gamma_i^{-1})^{-1}$ and $\hat{x} := \gamma \sum_{i=1}^{N} \gamma_i^{-1} x_i$.

If all stepizes are set to a same value $\gamma$, so that $\Gamma = \gamma I$, then $\hat{x} = \frac{1}{\gamma} \sum_{i=1}^{N} x_i$ is the (unweighted) average of vector $x$, and the forward-backward step reduces to

$$\left(\text{prox}_G^{-1}(x - \Gamma \nabla F(x))\right)_i = \text{prox}_g\left(\frac{1}{\gamma} \sum_{j=1}^{N} (x_j - \gamma \nabla f_j(x_j))\right).$$

(12)

Apparently, the BC-Algorithm 1 applied to the finite sum problem (2) reduces to the Finito/MISO algorithm [15, 24]. Here, however, we make no big data assumptions and fully support nonconvex and nonsmooth problems, more general sampling strategies and the possibility to select different stepsizes $\gamma_i$ for each block, which can have a significant impact on the performance compared to the case where all stepsizes are equal. The resulting scheme is presented in Algorithm 2. Being a special case of the general BC-Algorithm 1, the convergence results of Section 2 can be directly imported. We remark that the consensus formulation to recover Finito/MISO (although from a different umbrella algorithm) was also observed in [13] in the convex case. Moreover, the Finito/MISO algorithm with cyclic sampling is also studied in [25] 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.

Algorithm 2 Nonconvex proximal Finito/MISO for problem (2) under Assumption I

REQUIRE $x^0 \in \mathbb{R}^p$, $\gamma_i \in (0, N/l_i)$, $i \in [N]

\hat{y} := (\sum_{i=1}^{N} \gamma_i^{-1})^{-1}, \quad p_i = x^0 - \frac{\gamma}{N} \nabla f_i(x^0), \quad i \in [N], \quad \hat{p} = \hat{y} \sum_{i=1}^{N} \gamma_i^{-1} p_i

\text{repeat for } k = 0, 1, \ldots

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

2: $z^k \in \text{prox}_g(\hat{p})$

3: for $i \in I^{k+1}$ do

4: $v \leftarrow z^k - \frac{\gamma}{N} \nabla f_i(z^k)$

5: update $\hat{p} \leftarrow \hat{p} + \frac{\gamma}{N} (v - p_i)$ and $p_i \leftarrow v$

The convergence results from Section 2.1 are immediately translated to this setting by noting that the bold variable $z^k \in \mathbb{R}^{m_i}$ corresponds to $(z^k, \ldots, z^k)$. Therefore, $\Phi(z^k) = \varphi(z^k)$ where $\varphi$ is the cost function for the finite sum problem.

Corollary 8 (convergence of nonconvex proximal Finito/MISO). In the finite sum problem (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 iterates 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 $\varphi_* \leq \varphi(x^0)$;

(ii) all cluster points of the sequence $(z^k)_{k \in \mathbb{N}}$ are stationary and on which $\varphi$ equals $\varphi_*$.

If, additionally, $\varphi$ is coercive, then the following also hold:
Corollary 8 (linear convergence of strongly convex proximal Finito/MISO). Additionally to the assumptions of Corollary 8, suppose that $g$ is convex and that each $f_i$ is $\mu_{f_i}$-strongly convex. The following hold for the iterates generated by Algorithm 2:

**Randomized sampling:** under Assumption II, $\mathbb{E} \left[ \varphi(z^k) - \min \varphi \right] \leq (1 - c)^k (\varphi(x^0) - \min \varphi)$ holds for all $k \in \mathbb{N}$, where $c$ is as in (7). If the stepizes $\gamma_i$ and the sampling probabilities $p_i$ are set as in Theorem 4, then the tighter constant $c$ as therein defined is obtained.

**Shuffled cyclic or cyclic sampling:** under either sampling strategy (8) or (9), $\varphi(z^N) - \min \varphi \leq (1 - c)^N (\varphi(x^0) - \min \varphi)$ holds surely for all $v \in \mathbb{N}$, where $c$ is as in (10).

Let the Lyapunov function $\Phi^n$ be as in Definition 1 with $G$ as in (3). The $R$-linear rates in terms of the cost function are obtained from Theorems 4 and 7 by simply using the inequalities $\Phi^n(x^0) \leq \Phi(x^0) = \varphi(x^0)$ and $\varphi(z^k) = \Phi(z^k) \leq \Phi^k(x^0)$, see Lemmas A.2(i) and A.2(ii), where bold variable $x^k \in \mathbb{R}^N$ is the main iterate as it appears in Algorithm 1, although hidden for computational and storage efficacy in Algorithm 2. Similarly, one can obtain rates in terms of the distance from the solution by lower bounding the envelope as in Lemma A.2(iii). Linear rate for the essentially cyclic case is also readily available as in Theorem 7.

4 Nonconvex sharing problem

In this section we consider another important immediate consequence of our analysis. Consider the following problem

$$
\min_{x \in \mathbb{R}^n} \frac{1}{N} \sum_{i=1}^{N} f_i(x_i) \quad \text{subject to} \quad \sum_{i=1}^{N} x_i = 0.
$$

The constraint in (13) models “sharing” problems such as resource allocation, and consists of the linear space orthogonal to the consensus set of (3). Clearly, (13) fits into the problem framework (1) by simply letting $G := \delta_S$, where $S := \{ x \in \mathbb{R}^{nN} \mid \sum_{i=1}^{N} x_i = 0 \} = \ker A$ with $A := [I_n \cdots I_n] \in \mathbb{R}^{n \times nN}$. Clearly, $\operatorname{prox}_{G}^{-1}$ is the projection on $\ker A$ in the metric $\| \cdot \|_{\operatorname{F}^{-1}}$, easily seen to equal

$$
\operatorname{prox}_{G}^{-1}(x) = x - \Gamma A \Gamma^\top \left( \Gamma \Gamma^\top \right)^{-1} A x = x - \Gamma \left( \tilde{y} + \gamma \sum_{i=1}^{N} p_i x_i \right),
$$

where $\tilde{y} := (\sum_{i=1}^{N} \gamma_i)^{-1}$ and $\tilde{x} := \tilde{y} \sum_{i=1}^{N} x_i$.

This formula in the general BC-Algorith 1 yields Algorithm 3 for the sharing problem (13). We remark that the convergence results for Algorithm 3 are as in Section 2.1 with $G$ replaced with $\delta_S$, covering the randomized and (essentially) cyclic implementations. Note that here in the notation of Section 2.1, the proximal gradient point $z = (z_1, \ldots, z_N)$ in Algorithm 3 is given by $z_i = p_i - \gamma_i \tilde{x}, i \in [N]$.

**Algorithm 3** Block-coordinate method for nonconvex sharing problem (13)

**Require** $x_0^i \in \mathbb{R}^n$, $\gamma_i \in [0, N/\mu_i)$, $i \in [N]$

$\tilde{y} := (\sum_{i=1}^{N} \gamma_i)^{-1}$, $p_i = x_0^i - \frac{x_i}{\mu_i} \nabla f_i(x_0^i)$ $i \in [N]$, $\tilde{p} = \tilde{y} \sum_{i=1}^{N} p_i$

**Repeat** for $k = 0, 1, \ldots$

1: select a set of indices $\mathcal{I}^{k+1} \subseteq [N]$
2: $w \leftarrow \tilde{p}$
3: for $i \in \mathcal{I}^{k+1}$ do
4: $v_i \leftarrow p_i - \gamma_i w - \frac{x_i}{\mu_i} \nabla f_i(p_i - \gamma_i w)$
5: update $\tilde{p} \leftarrow \tilde{p} + \tilde{y}(v_i - p_i)$ and $p_i \leftarrow v_i$

This framework can be extended to more general constraints of the form $Ax = 0$ for some full row rank matrix $A$, which require factoring once offline the matrix $\Gamma A \Gamma^\top$, needed for computing $\operatorname{prox}_{G}^{-1}$.

However, this may be expensive for general $A$, especially if one tunes the stepizes $\gamma_i$ during the
iterations in which case the whole matrix has to be refactored, unless a unique scalar stepsize $\gamma$ is used or matrix $A$ has a special structure such as the one considered in problem (13). One notable instance is when the constraint $\sum_{i=1}^{n} A_i x_i = 0$ is considered in (13) for some $A_i \in \mathbb{R}^{n \times n}$, in which case $A = [A_1 \cdots A_N]$ and one has $(A^T A)^{-1} = (\sum_i A_i A_i^T)^{-1}$.

5 Accelerated block-coordinate proximal gradient

The work [1] introduced a coordinate descent method for smooth convex minimization, in which each coordinate is randomly sampled according to an ad hoc probability distribution that provably leads to a remarkable speed up with respect to uniform sampling strategies. The unified analysis of BC-algorithms and the analytical tool introduced in this paper, the forward backward envelope function, allow the extension of this approach to nonsmooth convex minimization of the form (1), where functions $f_i$ are convex quadratic and $G$ is convex but possibly nonsmooth:

Assumption IV (requirements for the fast BC-Algorithm 4). In problem (1), $G : \mathbb{R}^{\sum_{i=1}^{n_i}} \rightarrow \mathbb{R}$ is proper convex and lsc, and $f_i(x_i) := \frac{1}{2} x_i^T H_i x_i + q_i^T x_i$ is convex quadratic, with $L_{f,i} := \lambda_{\max}(H_i)$ and $\mu_f := \lambda_{\min}(H_i) \geq 0, i \in [N]$.

Let $U_i \in \mathbb{R}^{\sum_{i=1}^{n_i} n \times n}$ denote the $i$-th block column of the identity matrix so that for a vector $v \in \mathbb{R}^n$

$$U_i v = (0, \ldots, 0, v^{i-th}, 0, \ldots, 0).$$  \hfill (14)

The accelerated BC scheme based on [1] (for both strongly convex and convex cases) is given in Algorithm 4. Similarly to the approach of [30] where an accelerated Douglas-Rachford algorithm is proposed, in order to derive Algorithm 4 we consider the scaled problem $\text{minimize}_x \tilde{\Phi}_I^n(x)$ where $\tilde{\Phi}_I^n := \Phi_I^n \circ Q^{-1/2}$, and $Q$ is the symmetric positive definite matrix

$$Q := \blkdiag(Q_1, \ldots, Q_N) > 0 \quad \text{with} \quad Q_i := \gamma_i^{-1} - \frac{1}{N} H_i \in \mathbb{R}^{n \times n}, \quad i \in [N].$$  \hfill (15)

As detailed in Lemma A.6, whenever Assumption IV is satisfied $\tilde{\Phi}_I^n$ is a convex Lipschitz-differentiable function, and its gradient is given by $\nabla \tilde{\Phi}_I^n(x) = \frac{Q^{1/2}}{Q^{-1/2}} (x - \text{prox}_{\tilde{\Phi}_I^n}(x - \nabla F(x)))$ where $x = Q^{1/2} \tilde{x}$. Note that, based on Lemma A.6, $\tilde{\Phi}_I^n$ is 1-smooth along the $i$-th block (in the notation of [1], $L_i = 1, S_i = N$, and $p_i = 1/n$). Hence the parameters of the algorithm simplify substantially in uniform sampling. Moreover, when functions $f_i$ are $\mu_{f,i}$-strongly convex, by Lemma A.6 $\tilde{\Phi}_I^n$ is $\sigma$-strongly convex with $\sigma = \frac{1}{\min_{i \in [N]} \gamma_i \mu_f}$.

Algorithm 4 is obtained by applying the fast BC to this problem and scaling the variables by $Q^{-1/2}$. Specifically, the update rule as in [1] reads

$$\begin{align*}
\tilde{x}^+ &:= \tau \tilde{w} + (1 - \tau) \tilde{y}
\tilde{y}^+ &:= \tilde{x} - U_i U_i^T \tilde{\Phi}_I^n(\tilde{x}^+) = \tilde{x} - U_i Q_i^{1/2} (x_i^+ - z_i^+)
\tilde{w}^+ &:= \frac{1}{1 + \sigma \rho} (\tilde{w} + \eta \sigma \tilde{x}^+ - N \eta U_i U_i^T \nabla \tilde{\Phi}_I^n(\tilde{x}^+)) = \frac{1}{1 + \sigma \rho} (\tilde{w} + \eta x^+ - N \eta U_i Q_i^{1/2} (x_i^+ - z_i^+))
\end{align*}$$

where $z^+ = \text{prox}_{\tilde{G}}(x^+ - \nabla F(x^+))$. Since $Q^{-1/2} U_i Q_i^{1/2} = U_i$, premultiplying by $Q^{-1/2}$ yields

$$\begin{align*}
x^+ &:= \tau z + (1 - \tau) y
z^+ &:= \text{prox}_{\tilde{G}}(x^+ - \nabla F(x^+))
y^+ &:= x - U_i^T (z_i^+ - x_i^+)
w^+ &:= \frac{1}{1 + \sigma \rho} (w + \eta \sigma x^+ + N \eta U_i (z_i^+ - x_i^+)).
\end{align*}$$

For computational efficiency, vectors $\nabla F(x^+)$ and $\nabla F(w^+)$ are stored in variables $x^+$ and $w^+$ and updated recursively using the fact that gradients are affine, in such a way that each iteration requires only the evaluation of the sampled gradient (see Step 5). For similar reasons, in Algorithm 4 the iterates start with the $x$-update rather than the $x$-update as in [1]. Moreover, in the same spirit of Algorithm 1 this accelerated variant can be implemented efficiently whenever the individual blocks of $z^+$ can be computed efficiently, similarly to the cases discussed in Sections 3 and 4.

The convergence rate results follow directly from those of [1] with parameters $L_i = 1$ and $S_i = N$ as described above.
Lemma A.2 satisfy Assumption IV. are satisfied. is satisfied and given Algorithm 4) Then, the iterates generated by $M$ where 

This section lists some useful properties of the (convergence rates of Theorem 10

Suppose that $\gamma_i \in (0, \eta_i/L_i), i \in [N], \sigma := \frac{1}{n} \min_{i \in [N]} \gamma_i/\mu_i$.

1: if $\sigma = 0$, then $\eta = 1/n$ otherwise set $\tau = \frac{2}{1 + \sqrt{1 + 4\sigma/\eta}}$, $\eta = \frac{1}{\tau N}$ end if

2: $w^0 = x^0$, $(\nu^0, r^0) = (\nabla F(x^0), \nabla F(x^0))$, $z^0 = \prox_G^{-(1)}(x^0 - r^0)$

3: for $k = 0, 1, \ldots$

4: sample $i \in [N]$ uniformly

5: $y^{k+1} \leftarrow x^{k} + U_i(x^{k} - x_i^{k})$, $d \leftarrow \frac{2}{\eta} \nabla f_i(x_i^{k}) - r_i^{k}$

6: $v^{k+1} = \frac{1}{\tau \sigma} (v_i^{k} + \eta \sigma r_i^{k} + N \eta U_i d_i)$, $w^{k+1} = \frac{1}{\tau \sigma} (w_i^{k} + \eta \sigma x^{k} + N \eta U_i (x_i^{k} - x_i^{k}))$

7: if $\sigma = 0$, then $\eta \leftarrow \frac{1}{2} \sqrt{\frac{1}{\tau N}}$, $\tau \leftarrow \frac{1}{x^2 \tau^2}$; end if

8: $x^{k+1} = \tau w^{k+1} + (1 - \tau) y^{k+1}$, $r^{k+1} = \tau v^{k+1} + (1 - \tau) (r^{k} + U_i d_i)$

9: $z^{k+1} = \prox_G^{-1}(x^{k+1} - r^{k+1})$

Algorithm 4 Accelerated block-coordinate proximal gradient for problem (1) under Assumption IV

Theorem 10 (convergence rates of Algorithm 4). Suppose that Assumptions I and IV are satisfied. Then, the iterates generated by Algorithm 4 satisfy

$$\mathbb{E}\left[\Phi^m_T(y^k) - \min \Phi\right] \leq \frac{2N^2\|x^0 - x^*\|_Q^2}{(k + 1)^2},$$

where $Q$ is as in (15). Moreover, in the strongly convex case ($\sigma = \frac{1}{n} \min_{i \in [N]} \gamma_i/\mu_i > 0$)

$$\mathbb{E}\left[\Phi^m_T(y^k) - \min \Phi\right] \leq O(1)(1 - c)^k \left(\Phi^m_T(x^0) - \min \Phi\right) \text{ where } c = \left(\frac{1}{2} + \frac{1}{\sqrt{4 + \frac{4}{N^2}}}\right)^{-1}.$$

Note that in the strongly convex case it follows from Lemma A.2(iii) that the distance from the solution decreases R-linearly as

$$\mathbb{E}\left[\|y^k - x^*\|_{1d}^2\right] \leq O(1)(1 - c)^k \left(\Phi^m_T(x^0) - \min \Phi\right),$$

where $M$ is as in Lemma A.2(iii).

Appendix

A The key tool: forward-backward envelope

This section lists some useful properties of the $\Gamma$-FBE. We start by observing that $\Phi^m_T$ can equivalently be expressed as

$$\Phi^m_T(x) = F(x) - \frac{1}{2}\|\nabla F(x)\|_{1d}^2 + G^{-1}(x - \nabla F(x)).$$

In what follows, we use the shorthand notation $T^m_T$ to indicate the point-to-set mapping

$$T^m_T(x) := \prox_{\Gamma}^{-1}(x - \nabla F(x)).$$

Since $F$ and $-F$ are 1-smooth in the metric induced by $\Lambda_F := \frac{1}{N} \blkdiag(L_1, L_2, \ldots, L_N)$, one has

$$F(x) + \langle \nabla F(x), w - x \rangle - \frac{1}{2}\|w - x\|_{\Lambda}^2 \leq F(w) \leq F(x) + \langle \nabla F(x), w - x \rangle + \frac{1}{2}\|w - x\|_{\Lambda}^2$$

for all $x, w \in \mathbb{R}^{\Sigma_n^m}$, see [9, Prop. A.24]. As shown in the next result, the upper bound guarantees that the forward-backward operator $T^m_T$ is a well-defined (set-valued) mapping whenever the stepsizes $\gamma_i$ are selected as prescribed in Algorithm 1.

Lemma A.1. Suppose that Assumption I is satisfied and given $\gamma_i \in (0, \eta_i/L_i), i \in [N]$, let $\Gamma := \blkdiag(\gamma_1 I_1, \ldots, \gamma_N I_N)$. Then $\prox_{\Gamma}^{-1}$ and $T^m_T$ are locally bounded, outer semicontinuous (osc), nonempty- and compact-valued mappings.
Proof. For \( x^* \in \text{arg min} \Phi \) it follows from (17) that
\[
\min \Phi \leq F(x) + G(x) \leq G(x) + F(x^*) + \langle \nabla F(x^*), x - x^* \rangle + \frac{1}{2}\|x^* - x\|_\nabla^2G.
\]
Therefore, \( G \) is lower bounded by a quadratic function with quadratic term \(-\frac{1}{2}\|\cdot\|_\nabla^2G\|^2\), and thus is prox-bounded in the sense of [35, Def. 1.23]. The claim then follows from [35, Thm. 1.25 and Ex. 5.23(b)] and the continuity of the forward mapping \( \text{id} - \Gamma\nabla F \).

Let us denote
\[
M_f(w, x) := F(x) + \langle \nabla F(x), w - x \rangle + G(w) + \frac{1}{2}\|w - x\|_{\Gamma^{-1}A_f}^2,
\]
the quantity being minimized (with respect to \( w \)) in the definition (5a) of the \( \Gamma \)-FBE. It follows from (17) that
\[
\Phi(w) + \frac{1}{2}\|w - x\|_{\Gamma^{-1}A_f}^2 \leq M_f(w, x) \leq \Phi(w) + \frac{1}{2}\|w - x\|_{\Gamma^{-1}+A_f}^2
\]
holds for all \( x, w \in \mathbb{R}^{\mathbb{R}} \). In particular, \( M_f \) is a majorizing model for \( \Phi \), in the sense that \( M_f(x, x) = \Phi(x) \) and \( M_f(w, x) \geq \Phi(w) \) for all \( x, w \in \mathbb{R}^{\mathbb{R}} \). In fact, while a \( \Gamma \)-forward-backward step \( \gamma \in T_f^\Gamma(x) \) amounts to evaluating a minimizer of \( M_f(-x, \cdot) \), the \( \Gamma \)-FBE is defined instead as the minimization value, namely \( \Phi^\Gamma_f(x) = M_f(z, x) \) where \( z \) is any element of \( T_f^\Gamma(x) \).

**Lemma A.2** (\( \Gamma \)-FBE: fundamental inequalities). Suppose that Assumption I is satisfied and let \( \gamma_i \in (0, \gamma_i]) \), \( i \in [N] \). Then, the \( \Gamma \)-FBE \( \Phi^\Gamma_f \) is a (real-valued and) locally Lipschitz-continuous function. Moreover, the following hold for any \( x \in \mathbb{R}^{\mathbb{R}} \):

(i) \( \Phi^\Gamma_f(x) \leq \Phi(x) \).

(ii) \( \frac{1}{2}\|z - x\|_{\Gamma^{-1}A_f}^2 \leq \Phi^\Gamma_f(x) - \Phi(z) \leq \frac{1}{2}\|z - x\|_{\Gamma^{-1}+A_f}^2 \) for any \( z \in T_f^\Gamma(x) \).

(iii) If in addition each \( f_i \) is \( \mu_i \)-strongly convex and \( G \) is convex, then denoting \( m_i := \min \left\{ \frac{\mu_i}{\gamma_i}, \frac{\xi_i}{\gamma_i} \right\} \) and \( M := \text{blkdiag}(m_1I_{n_1}, \ldots, m_NI_{n_N}) \) with \( \xi_i = \frac{N^{-\gamma_i \mu_i}}{N} \) as in Lemma 1(i), it holds that
\[
\frac{1}{2}\|x - x^*\|_M^2 \leq \Phi^\Gamma_f(x) - \min \Phi,
\]
where \( x^* := \text{arg min} \Phi \).

**Proof.** Local Lipschitz continuity follows from (16) in light of Lemma A.1 and [35, Ex. 10.32].

- **A.2(i)** Follows by replacing \( w = x \) in (5a).
- **A.2(ii)** Directly follows from (18) and the identity \( \Phi^\Gamma_f(z) = M_f(z, x) \) for \( z \in T_f^\Gamma(x) \).
- **A.2(iii)** By strong convexity, denoting \( \Phi_* := \min \Phi \), we have
\[
\Phi_* \leq \Phi(z) - \frac{1}{2}\|z - x^*\|_\mu^2 \leq \Phi^\Gamma_f(x) - \frac{1}{2}\|x - z\|_{\Gamma^{-1}A_f}^2 \leq \Phi^\Gamma_f(x) - \frac{1}{2}\|x - z\|_{A_f}^2 \leq \min \Phi,
\]
where \( \mu_F := \frac{1}{4} \text{blkdiag}(\mu_{i_1}I_{n_1}, \ldots, \mu_{i_N}I_{n_N}) \). The claim follows by using the elementary inequality \( \frac{1}{2}\|a + b\|_M \leq \|a\|_M + \|b\|_M \) with \( a = x - z \) and \( b = z - x^* \).

**Lemma A.3** (\( \Gamma \)-FBE: minimization equivalence). Suppose that Assumption I is satisfied and that \( \gamma_i \in (0, \gamma_i]) \), \( i \in [N] \). Then the following hold:

(i) \( \min \Phi^\Gamma_f = \min \Phi \);

(ii) \( \text{arg min} \Phi^\Gamma_f = \text{arg min} \Phi \);

(iii) \( \Phi^\Gamma_f \) is level bounded iff so is \( \Phi \).

**Proof.**
• A.3(i) and A.3(ii) It follows from Lemma A.2(i) that \( \inf \Phi^p_T \leq \min \Phi \). Conversely, let \((x^k)_{k \in \mathbb{N}}\) be such that \( \Phi^p_T(x^k) \to \inf \Phi^p_T \) as \( k \to \infty \), and for each \( k \) let \( z^k \in T^p_T(x^k) \). It then follows from Lemmas A.2(i) and A.2(ii) that \[
\inf \Phi^p_T \leq \min \Phi \leq \liminf_{k \to \infty} \Phi(z^k) \leq \liminf_{k \to \infty} \Phi^p_T(x^k) = \inf \Phi^p_T,
\]
hence \( \min \Phi = \inf \Phi^p_T \). Suppose now that \( x \in \arg \min \Phi \) (which exists by Assumption I); then it follows from Lemma A.2(ii) that \( T^p_T(x) = \{x\} \) (for otherwise another element would belong to a lower level set of \( \Phi \)). Combining with Lemma A.2(i) with \( z = x \) we then have \[
\min \Phi = \Phi(z) \leq \Phi^p_T(x) \leq \Phi(x) = \min \Phi.
\]

Since \( \min \Phi = \inf \Phi^p_T \), we conclude that \( x \in \arg \min \Phi^p_T \), and that in particular \( \inf \Phi^p_T = \min \Phi^p_T \). Conversely, suppose \( x \in \arg \min \Phi^p_T \), and let \( z \in T^p_T(x) \). By combining Lemmas A.2(i) and A.2(ii) we have that \( z = x \), that is, that \( T^p_T(x) = \{x\} \). It then follows from Lemma A.2(ii) and assert A.3(i) that \[
\Phi(x) = \Phi(z) \leq \Phi^p_T(x) = \min \Phi^p_T = \min \Phi,
\]
hence \( x \in \arg \min \Phi \).

• A.3(iii) Due to Lemma A.2(ii), if \( \Phi^p_T \) is level bounded clearly so is \( \Phi \). Conversely, suppose that \( \Phi^p_T \) is not level bounded. Then, there exist \( \alpha \in \mathbb{R} \) and \((x^k)_{k \in \mathbb{N}} \subseteq \text{lev}_{\leq 0} \Phi^p_T \) such that \( \|x^k\| \to \infty \) as \( k \to \infty \). Let \( \lambda = \min \left\{ \gamma^{-1}_{T} - L_{T} N_{1}^{-1} \right\} > 0 \), and for each \( k \in \mathbb{N} \) let \( z^k \in T^p_T(x^k) \). It then follows from Lemma A.2(ii) that \[
\min \Phi \leq \Phi(z^k) \leq \Phi^p_T(x^k) - \frac{\lambda}{2} \|x^k - z^k\| \leq \alpha - \frac{\lambda}{2} \|x^k - z^k\|^2,
\]
hence \((z^k)_{k \in \mathbb{N}} \subseteq \text{lev}_{\leq 0} \Phi \) and \( \|x^k - z^k\|^2 \leq 2 \gamma \alpha - \min \Phi \). Consequently, also the sequence \((z^k)_{k \in \mathbb{N}} \subseteq \text{lev}_{\leq 0} \Phi \) is unbounded, proving that \( \Phi \) is not level bounded.

We next investigate some properties of the proximal mapping and the Moreau envelope that will be used in the sequel.

**Lemma A.4.** Suppose that Assumption I holds, and let two sequences \((u^k)_{k \in \mathbb{N}}\) and \((v^k)_{k \in \mathbb{N}}\) satisfy \( v^k \in T^p_T(u^k) \) for all \( k \) and be such that both converge to a point \( u^* \) as \( k \to \infty \). Then, \( u^* \in T^p_T(u^*) \), and in particular \( 0 \in \partial \Phi(u^*) \).

**Proof.** Since \( \nabla F \) is continuous, it holds that \( u^k - \Gamma \nabla F(u^k) \to u^* - \Gamma \nabla F(u^*) \) as \( k \to \infty \). From outer semicontinuity of \( \text{prox}^{G^{-1}}_{\Gamma} \) [35, Ex. 5.23(b)] it then follows that \[
u^* = \lim_{k \to \infty} v^k \in \lim \sup_{k \to \infty} \text{prox}^{G^{-1}}_{\Gamma}(u^k - \Gamma \nabla F(u^k)) \subseteq \text{prox}^{G^{-1}}_{\Gamma}(u^* - \Gamma \nabla F(u^*)) = T^p_T(u^*),
\]
where the limit superior is meant in the Painlevé-Kuratowski sense, cf. [35, Def. 4.1]. The optimality conditions defining \( \text{prox}^{G^{-1}}_{\Gamma} \) [35, Thm. 10.1] then read \[
0 \in \partial \left( G + \frac{1}{2} \| \cdot - (u^* - \Gamma \nabla F(u^*)) \|^2_{T^{-1}} \right)(u^*) = \partial G(u^*) + \Gamma^{-1} \left( u^* - (u^* - \Gamma \nabla F(u^*)) \right) = \partial \Phi(u^*) + \nabla F(u^*) = \partial \Phi(u^*),
\]
where the first and last equalities follow from [35, Ex. 8.8(c)].

**Lemma A.5.** Suppose that Assumption I holds and that function \( G \) is convex. Then, the following hold:

(i) \( \text{prox}^{G^{-1}}_{\Gamma} \) is (single-valued and) firmly nonexpansive (FNE) in the metric \( \| \cdot \|_{\Gamma^{-1}} \); namely, \[
\| \text{prox}^{G^{-1}}_{\Gamma}(u) - \text{prox}^{G^{-1}}_{\Gamma}(v) \|^2_{\Gamma^{-1}} \leq \langle \text{prox}^{G^{-1}}_{\Gamma}(u) - \text{prox}^{G^{-1}}_{\Gamma}(v), \Gamma^{-1}(u - v) \rangle \leq \|u - v\|^2_{\Gamma^{-1}}, \quad \forall u, v;
\]

(ii) the Moreau envelope \( G^{\Gamma^{-1}} \) is differentiable with \( \nabla G^{\Gamma^{-1}} = \Gamma^{-1}(\text{id} - \text{prox}^{G^{-1}}_{\Gamma}) \);

(iii) \( T^p_T \) is \( L_T \)-Lipschitz continuous in the metric \( \| \cdot \|_{\Gamma^{-1}} \) for some \( L_T \geq 0 \); if in addition \( f_i \) is \( \mu_i \)-strongly convex, \( i \in \mathbb{N} \), then \( L_T \leq 1 - \delta \) for \( \delta = \frac{1}{N} \min_{i \in \mathbb{N}} \{ \gamma_i \mu_i \} \).
Proof.

\* **A.5(i)** See [5, Prop. 12.28].

\* **A.5(ii)** See [5, Prop. 12.30].

\* **A.5(iii)** Lipschitz continuity follows from assert A.5(i) together with the fact that Lipschitz continuity is preserved by composition. Suppose now that \( f_i \) is \( \mu_{f_i} \)-strongly convex, \( i \in [N] \). By [27, Thm 2.1.12] for all \( x_i, y_i \in \mathbb{R}^n \)

\[
\langle \nabla f_i(x_i) - \nabla f_i(y_i), x_i - y_i \rangle \geq \frac{\mu_{f_i} L_{f_i}}{\mu_{f_i} + L_{f_i}} \|x_i - y_i\|^2 + \frac{1}{\mu_{f_i} + L_{f_i}} \|\nabla f_i(x_i) - \nabla f_i(y_i)\|^2.
\]

(19)

For the forward operator we have

\[
\| (I - \frac{2}{N} \nabla f_i)(x_i) - (I - \frac{2}{N} \nabla f_i)(y_i) \|^2 = \|x_i - y_i\|^2 + \frac{2}{N} \|\nabla f_i(x_i) - \nabla f_i(y_i)\|^2 - \frac{2}{N} \langle x_i - y_i, \nabla f_i(x_i) - \nabla f_i(y_i) \rangle \\
\leq \left( 1 - \frac{\gamma_i L_{f_i}}{N} \right) \|x_i - y_i\|^2 - \frac{1}{N} \left( 2 - \frac{\gamma_i L_{f_i}}{N} \right) \|\nabla f_i(x_i) - \nabla f_i(y_i)\|^2 \\
\leq \left( 1 - \frac{\gamma_i L_{f_i}}{N} \right) \|x_i - y_i\|^2 - \frac{\gamma_i L_{f_i}}{N} \left( 2 - \frac{\gamma_i L_{f_i}}{N} \right) \|\nabla f_i(x_i) - \nabla f_i(y_i)\|^2 \\
= \left( 1 - \frac{\gamma_i L_{f_i}}{N} \right) \|x_i - y_i\|^2,
\]

where strong convexity and the fact that \( \gamma_i < \frac{N}{L_{f_i}} \leq 2N/\mu_{f_i} + L_{f_i} \) was used in the second inequality. Multiplying by \( \gamma_i^{-1} \) and summing over \( i \) shows that the forward operator \( \text{id} - \Gamma F \) is \((1-\delta)\)-contractive in the metric \( \| \cdot \|_{\tilde{L}_1} \), and so is \( T_i^\mu = \text{prox}_{\tilde{L}_1^\mu} \circ (\text{id} - \Gamma F) \) as it follows from assert A.5.5(i).

**Lemma A.6** (\( \Gamma \)-FBE: convexity and block-smoothness). Suppose that Assumptions I and IV are satisfied, and consider the notation introduced therein. Let \( \gamma_i \in (0, \frac{N}{L_{f_i}}) \) be defined. Define \( Q := \gamma_i^{-1} \Gamma - \frac{1}{N} H_i \in \mathbb{R}^{n \times n} \), \( \mathcal{Q} := \text{blkdiag}(Q_1, \ldots, Q_N) \), and \( H := \frac{1}{N} \text{blkdiag}(H_1, \ldots, H_N) \). Then, \( \Phi_{f_{\mathcal{Q}}} : \Phi_{f_{\mathcal{Q}}} = \Phi_{f_{\mathcal{Q}}} \circ \Gamma^{-1/2} \Gamma_{Q} \) is convex and smooth with \( \nabla \Phi_{f_{\mathcal{Q}}}^\mu(\tilde{x}) = \Gamma_{Q}^{-1/2}(x - \Gamma_{Q} T_{\mu_{f_{\mathcal{Q}}}}(x)) \) where \( x = \Gamma_{Q}^{-1/2} \tilde{x} \). In fact, for any \( \tilde{x}, \tilde{x}' \in \mathbb{R}^{\sum_{i} n_i} \) it holds that

\[
0 \leq \langle \nabla \Phi_{f_{\mathcal{Q}}}^\mu(\tilde{x}') - \nabla \Phi_{f_{\mathcal{Q}}}^\mu(\tilde{x}), \tilde{x}' - \tilde{x} \rangle \leq \|\tilde{x}' - \tilde{x}\|^2.
\]

(20)

In particular, function \( \Phi_{f_{\mathcal{Q}}}^\mu \) is \( 1 \)-smooth along each block \( i \in [N] \). If, additionally, all functions \( f_i \) are strongly convex, then \( \Phi_{f_{\mathcal{Q}}}^\mu \) is \( \sigma \)-strongly convex with \( \sigma := \frac{1}{N} \min_{i \in [N]} \{ \gamma_i \mu_{f_i} \} \).

Proof. Since \( \gamma_i < N/L_{f_i} \), \( Q \) is positive definite. We begin by showing that for any \( x, x' \in \mathbb{R}^{\sum_{i} n_i} \) it holds that

\[
0 \leq \|x' - x\|^2 Q - \|Q(x' - x)\|^2 \leq \langle \nabla \Phi_{f_{\mathcal{Q}}}^\mu(x' - x), x' - x \rangle \leq \|x' - x\|^2 Q_{\mathcal{Q}}.
\]

(21)

It follows from Lemma A.5(ii), the chain rule of differentiation applied to (16), and the continuous differentiability of \( F \) that \( \Phi_{f_{\mathcal{Q}}}^\mu \) is continuously differentiable with \( \nabla \Phi_{f_{\mathcal{Q}}}^\mu(x) = Q(x - x') \). For \( z^i := T_{\mu_{f_i}}(x) \) and \( z'^i := T_{\mu_{f_i}}(x') \) it holds that

\[
\langle \nabla \Phi_{f_{\mathcal{Q}}}^\mu(x'), \nabla \Phi_{f_{\mathcal{Q}}}^\mu(x), x' - x \rangle = \langle \Phi_{f_{\mathcal{Q}}}^\mu(x'), x' - x \rangle = \|x' - x\|^2 Q_{\mathcal{Q}} - \langle z^i - z'^i, x' - x \rangle.
\]

(22)

In order to bound the last scalar product, observe that

\[
0 \leq \langle \Gamma^{-1}(z^i - z'^i), x' - \Gamma \nabla F(x') \rangle = \langle (x' - \Gamma \nabla F(x')) - (x - \Gamma \nabla F(x)), x' - x \rangle \leq \|x' - x\|^2 Q_{\mathcal{Q}}.
\]

as it follows from Lemma A.5(i). Since \( \text{id} - \Gamma F = \Gamma Q \cdot - \Gamma q \) (with \( q := (\frac{1}{\mu_{f_1}}, \ldots, \frac{1}{\mu_{f_N}}) \)), the above inequality simplifies to

\[
0 \leq \langle z^i - z'^i, Q(x' - x) \rangle \leq \|\Phi_{Q}(x' - x)\|^2.
\]

which combined with (22) results in the claimed (21). If additionally \( \mu_{f_i} > 0 \) for all \( i \), then \( \Phi_{f_{\mathcal{Q}}}^\mu \) is \( 1 \)-strongly convex in the metric \( \| \cdot \|_{Q_{\mathcal{Q}}} \) (by observing that \( Q - Q_{\mathcal{Q}} < 0 \)). The result in (20) follows by using (21) with the change of variables \( x = \Gamma_{Q}^{-1/2} \tilde{x}, x' = \Gamma_{Q}^{-1/2} \tilde{x}' \) and noting that \( \nabla \Phi_{f_{\mathcal{Q}}}^\mu(x) = \Gamma_{Q}^{-1/2} \nabla \Phi_{f_{\mathcal{Q}}}^\mu(x) \). Since \( \Gamma \) is block-wise a multiple of identity it commutes with any block-diagonal matrix. Therefore, when \( f_i \) are strongly convex, using the lower bound in (21) and the above change of variable we obtain that \( \Phi_{f_{\mathcal{Q}}}^\mu \) is strongly convex in the metric \( \| \cdot \|_{Q_{\mathcal{Q}}} \). The result follows by noting that \( 1 - \Gamma_{Q} \Gamma = \Gamma H \). \( \square \)
B Proofs

Proof of Lemma 1 (sure descent)

\[\text{\textbullet 1(i) To ease notation, for } w \in \mathbb{R}^{\sum n_i} \text{ let } w_I \in \mathbb{R}^{\sum n_i} \text{ denote the slice } (w_i)_{i \in I}, \text{ and let } \Lambda_{F_I}, \Gamma_I \in \mathbb{R}^{\sum n_i \times \sum n_i} \text{ be defined similarly. Start by observing that, since } z^+ \in \text{prox}_G^{\Gamma^{-1}}(x^+ - \nabla F(x^+)), \text{ from the proximal inequality on } G \text{ it follows that}
\]
\[
G(z^+) - G(z) \leq \frac{1}{2} \|z - x^+ + \nabla F(x^+)\|^2_{\Gamma_I^{-1}} - \frac{1}{2} \|z^+ - x^+ + \nabla F(x^+)\|^2_{\Gamma_I^{-1}} \leq \frac{1}{2} \|z - x^+\|^2_{\Gamma_I^{-1}} - \frac{1}{2} \|z^+ - x^+\|^2_{\Gamma_I^{-1}} + \langle \nabla F(x^+), z - z^+ \rangle. \quad (23)
\]
\[\text{We have}
\]
\[
\Phi^{\Gamma}_I(x^+) - \Phi^{\Gamma}_I(x) = F(x^+) + \langle \nabla F(x^+), z^+ - x^+ \rangle + G(z^+) + \frac{1}{2} \|z^+ - x^+\|^2_{\Gamma_I^{-1}} 
\]
\[
- \left( F(x) + \langle \nabla F(x), z - x \rangle + G(z) + \frac{1}{2} \|z - x\|^2_{\Gamma_I^{-1}} \right)
\]
\[\text{apply the upper bound in (17) with } w = x^+ \text{ and the proximal inequality (23) }
\]
\[
\leq \langle \nabla F(x), x^+ - z \rangle + \frac{1}{2} \|x^+ - x\|^2_{\Gamma_I}, \quad \|x^+ - x\|^2_{\Gamma_I}, \quad \|z - x\|^2_{\Gamma_I^{-1}}, \quad \|z - x^+\|^2_{\Gamma_I^{-1}}.
\]
\[\text{To conclude, notice that the } \ell\text{-th block of } \nabla F(x) - \nabla F(x^+) \text{ is zero for } \ell \notin I, \text{ and that the } \ell\text{-th block of } x^+ - z \text{ is zero iff } \ell \in I. \text{ Hence, the scalar product vanishes. For similar reasons, one has}
\]
\[
\|z - x^+\|^2_{\Gamma_I^{-1}}, \quad \|z - x\|^2_{\Gamma_I^{-1}}, \quad \|x^+ - x\|^2_{\Gamma_I}, \quad \|z - x^+\|^2_{\Gamma_I^{-1}}.
\]
\[\text{yielding the claimed expression.}
\]
\[\text{\textbullet 1(ii) Monotonic decrease of } (\Phi^{\Gamma}_I(x^+))_{k \in \mathbb{N}} \text{ is a direct consequence of assert 1(i). This ensures that the sequence converges to some value } \Phi^*_I, \text{ bounded below by } \min \Phi \text{ in light of Lemma A.3(i).}
\]
\[\text{\textbullet 1(iii) Denoting } \xi_{\min} := \min_{i \in [N]} \{ \xi_i \} \text{ which is a strictly positive constant, it follows from assert 1(i) that for each } k \in \mathbb{N} \text{ it holds that}
\]
\[
\Phi_I^{\Gamma}(x^{k+1}) - \Phi_I^{\Gamma}(x^k) \leq - \sum_{i \in I \setminus \iota^I} \frac{\xi_i}{2\gamma_i} \|z^i_k - x^i_k\|^2 \leq - \frac{\sum_{i \in I \setminus \iota^I} \gamma_i^{-1} \|z^i_k - x^i_k\|^2}{2} = - \frac{\sum_{i \in I \setminus \iota^I} \|x^{k+1} - x^i_k\|^2_{\Gamma_I^{-1}}}{2}. \quad (24)
\]
\[\text{By summing for } k \in \mathbb{N} \text{ and using the positive definiteness of } \Gamma^{-1} \text{ together with the fact that } \min \Phi^{\Gamma}_I = \min \Phi > \infty \text{ as ensured by Lemma A.3(i) and Assumption I}\_3, \text{ we obtain that}
\]
\[
\sum_{k \in \mathbb{N}} \|x^{k+1} - x^k\|^2_{\Gamma_I} < \infty.
\]
\[\text{\textbullet 1(iv) It follows from assert 1(ii) that every } x^k \text{ belongs to the sublevel set } \{ w | \Phi^{\Gamma}_I(w) \leq \Phi^{\Gamma}_I(x^0) \}, \text{ which is bounded provided that } \Phi \text{ is coercive as shown in Lemma A.3(iii). In turn, boundedness of } (z^k), k \in \mathbb{N}, \text{ then follows from local boundedness of } T^{\Gamma}_I, \text{ cf. Lemma A.1.}
\]

Proof of Theorem 2 (randomized sampling: subsequential convergence)

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

\[\text{\textbullet 2(i) Let } \xi_i := \frac{\sum_{i \in \mathbb{N}} \xi_i}{\mathbb{N}} > 0, \ i \in [N], \text{ be as in Lemma 1(i). We have}
\]
\[
\mathbb{E}_k \left[ \Phi^{\Gamma}_I(x^{k+1}) \right] \leq \mathbb{E}_k \left[ \Phi^{\Gamma}_I(x^k) - \sum_{i \in \mathbb{N}} \frac{\xi_i}{2\gamma_i} \|z^i_k - x^i_k\|^2 \right] = \Phi^{\Gamma}_I(x^k) - \sum_{i \in \mathbb{N}} \mathbb{E}_k \left[ \left( I^{k+1} = I \right) \right] \frac{\xi_i}{2\gamma_i} \|z^i_k - x^i_k\|^2 \leq \Phi^{\Gamma}_I(x^k) - \sum_{i \in \mathbb{N}} \frac{\xi_i}{2\gamma_i} \|z^i_k - x^i_k\|^2, \quad (25)
\]
\[\text{where } \Omega \subseteq 2^{[N]} \text{ is the sample space (} 2^{[N]} \text{ denotes the power set of } [N]). \text{ Therefore,}
\]
\[
\mathbb{E}_k \left[ \Phi^{\Gamma}_I(x^{k+1}) \right] \leq \Phi^{\Gamma}_I(x^k) - \sigma \|x^k - x^{k-1}\|^2, \quad \text{where } \sigma := \min_{i \in [N]} p_i \xi_i > 0. \quad (26)
\]
\[\text{The claim follows from the Robbins Siegmund supermartingale theorem, see e.g., [34] or [7, Prop. 2].}
\]
\( \ast \) 2(ii) Observe that \( \Phi^p_k(x^i) - \|z^i - x^i\|_{L^{-1} - \gamma_i}^2 \leq \Phi(x^i) - \|z^i - x^i\|_{L^{-1} - \gamma_i}^2 \) holds (surely) for \( k \in \mathbb{N} \) in light of Lemma A.2(ii). The claim then follows by invoking Lemma 1(ii) and asserting 2(i).

\( \ast \) 2(iii) In the rest of the proof, for conciseness the “almost sure” nature of the results will be implied without mention. It follows from assert 2(i) that a subsequence \((x^i)_{k \in \mathbb{N}}\) converges to some point \(x^*\) iff so does the subsequence \((z^i)_{k \in \mathbb{N}}\). Since \(T^p_k(x^i) \ni z^i\) and both \(x^i\) and \(z^i\) converge to \(x^*\) as \(K \ni k \to \infty\), the inclusion \(0 \in \partial \Phi(x^*)\) follows from Lemma A.4. Since the full sequences \((\Phi^p_k(x^i))_{k \in \mathbb{N}}\) and \((\Phi(z^i))_{k \in \mathbb{N}}\) converge to the same value \(\Phi_*\) (cf., Lemma 1(i) and assert 2(ii)), due to continuity of \(\Phi^p_k\) (Lemma A.2) it holds that \(\Phi^p_k(x^*) = \Phi_*\), and in turn the bounds in Lemma A.2(ii) together with assert 2(ii) ensure that \(\Phi(x^*) = \Phi_*\) too.

**Proof of Theorem 3 (randomized sampling: global convergence)**

Let \(L(x, z, y) := F(x) + G(z) + \langle y, x - z \rangle + \frac{1}{\mu}(\|x - z\|_{L}^2)\) be a \(\Gamma^{-1}\)-augmented Lagrangian associated to problem (1), and let \(L_k := L(x^k, z^k, -\nabla F(x^k))\) and similarly \(\partial L_k := \partial L(x^k, z^k, -\nabla F(x^k))\). Note that \(\Phi^p_k(x^k) = L_k\); to avoid trivialities, we may thus assume that \(L_k \ni \Phi_*\) for all \(k\), for otherwise the sequence \((x^k)_{k \in \mathbb{N}}\) is asymptotically constant, cf. (24). Let \(\Omega\) be the set of accumulation points of \((x^k)_{k \in \mathbb{N}}\) and \((z^k)_{k \in \mathbb{N}}\), which is compact and such that \(\Phi^p_k = \Phi_*\) on \(\Omega\) for some \(\Phi_* \in \mathbb{R}\). Since \(F\) and \(G\) are semialgebraic, known properties of semialgebraic functions (see e.g., [20, §83.1]) ensure that \(L\) is semialgebraic, and as such it possesses the KL property on \(\Omega\), see [11, Thm. 3 and Lem. 6]. Namely, there exists a continuous increasing concave function \(\psi : [0, \infty) \to [0, \infty)\) which is differentiable on \((0, \infty)\) and with \(\psi(0) = 0\), such that \(\psi'(L_k - \Phi_*) \geq 1\) for all \(k\) large enough such that \(x^k\) and \(z^k\) are sufficiently close to \(\Omega\) and \(L_k\) is sufficiently close to \(\Phi_*\). Notice that \(\partial L_k \ni (\Gamma^{-1}(x^k) - z^k), 0, x^k - z^k\), which implies that

\[
dist(0, \partial L_k) \leq \psi_{\min} + \psi_{\max} ||x^k - z^k||_{L^{-1}} - 1, \tag{27}
\]

where \(\psi_{\min} := \min_{y \in \mathbb{N}} \{\gamma\}\) and \(\psi_{\max} := \max_{y \in \mathbb{N}} \{\gamma\}\). Denoting \(\Delta_k := \psi(L_k - \Phi_*)\), we have

\[
\mathbb{E}[\Delta_{k+1}] - \Delta_k \leq \psi(L_k - \Phi_*)\mathbb{E}[\Delta_{k+1} - L_k] \leq - \sigma||x^k - z^k||_{L^{-1}}^2 - \frac{2}{\psi_{\max} + \psi_{\min}} ||x^k - z^k||_{L^{-1}} - 1. \tag{28}
\]

The first inequality uses concavity of \(\psi\), and the second one the KL property and the expected sufficient decrease (26) with \(\sigma > 0\) as therein defined. By virtue of the Robbins Siegmund supermartingale theorem, see e.g., [34] or [7, Prop. 2], we conclude that \(||x^k - z^k||_{L^{-1}}\) is summable a.s., hence so is \(||x^k - z^k||\). Since \(||x^k - z^k|| \leq ||x^k - z^k||\) we conclude that, almost surely, \((x^k)_{k \in \mathbb{N}}\) has finite length and is thus convergent (to a single point), and consequently so is \((z^k)_{k \in \mathbb{N}}\).

**Proof of Theorem 4 (randomized sampling: linear convergence)**

Convexity of \(G\) and the optimality conditions for \(z = T^p_i(x)\) imply that \(\Gamma^{-1}(x - z) - \nabla F(x) \in \partial G(z)\), hence \(G(x^*) \geq G(z) + \langle \Gamma^{-1}(x - z) - \nabla F(x), x^* - z \rangle\) for \(x^* := \mathbf{arg \ min} \Phi\). Denoting \(\mu_F := \frac{1}{L} \mathbf{blkdiag}(\mu_l, \ldots, \mu_{l_d})\), we have \(F(x^*) \geq F(x) + \langle \nabla F(x), x^* - x \rangle + \frac{1}{\mu}(x^*-x)^2\). By combining these two inequalities into (5b), and denoting \(\Phi_* := \mathbf{min} \Phi = \mathbf{min} \Phi^p\), we have

\[
\Phi^p_k(x) - \Phi_* \leq \frac{1}{2}||x - z||_{\mu}^2 + \frac{1}{2}\|x^* - x\|^2 + \langle \Gamma^{-1}(x - z), x^* - z \rangle
\]

Next, by using the inequality \((a, b) \leq \frac{1}{2}\|a\|_{\mu_i} + \frac{1}{2}\|b\|_{\mu_j}^2\), to cancel out the last term, we obtain

\[
\Phi^p_k(x) - \Phi_* \leq \frac{1}{2}||x - z||_{\mu_i}^2 + \frac{1}{2}\|(\Gamma^{-1} - \mu_F)(x - z)\|_{\mu_i}^2 \leq \frac{1}{2}||x - z||_{\mu_i}^2(1 - \mu_i)^{1} \tag{29}
\]

where the last identity uses the fact that the matrices are diagonal. Now, observe that (25) with the choice \(p_i := \frac{1}{\mu_i} \frac{N - \gamma_i \mu_i}{N - \gamma_i L_i} \), which equals the one in (6) with \(\gamma_i\) as prescribed, yields

\[
\mathbb{E}_k[\Phi^p_k(x^*) - \Phi_*] \leq \Phi^p_k(x) - \Phi_* - \left(2N \sum_j \frac{N - \gamma_i \mu_i}{N - \gamma_i L_i} \right)^{-1} \sum_i \frac{N - \gamma_i \mu_i}{N - \gamma_i L_i} \|x_i - x\|^2
\]

\[
= \Phi^p_k(x) - \Phi_* - \left(2N \sum_j \frac{N - \gamma_i \mu_i}{N - \gamma_i L_i} \right)^{-1} \|x - x\|^2(1 - \mu_i)^{-1} \tag{29}
\]

14
Theorem 5, using the su

Lemma A.5 (essentially cyclic sampling: subsequential convergence)

We first establish an important descent inequality for the envelope after every T iterations, cf. (35). Convexity of G, entailing \( \text{prox}^{x^+} \) being Lipschitz continuous (cf. Lemma A.5(ii)), 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

\[
t_v(i) := \min \{ t \in [T] \mid i \text{ is sampled at iteration } T \nu + t - 1 \}
\]

is well defined for each index \( i \in [N] \) and \( \nu \in \mathbb{N} \). For \( i \in [N] \), since \( i \) is sampled at iteration \( T \nu + t_v(i) - 1 \), it holds that

\[
x^{T\nu+t_v(i)} = x^{T\nu+t_v(i)-1} + U_i U_i^T (T_{\Gamma}^n(x^{T\nu+t_v(i)-1}) - x^{T\nu+t_v(i)-1}).
\]

For all \( t \in [T] \) the following holds

\[
\Phi^n_{t}(x^{T(t+1)}) - \Phi^n_{t}(x^{T(t)}) = \sum_{r=1}^{T} (\Phi^n_{t}(x^{T(t)}) - \Phi^n_{t}(x^{T(t-1)})) \leq \Phi^n_{t}(x^{T(t)}) - \Phi^n_{t}(x^{T(t-1)})
\]

\[
\leq - \frac{\lambda_{min}(\Gamma_{-})}{2} \| x^{T\nu+t_v(i)} - x^{T\nu+t_v(i)-1} \|_{l-1}^2
\]

(32)

where \( \xi_{i} := \frac{N-\gamma L_{t}}{2} \) are as in Lemma 1(i), \( \xi_{min} := \min_{\nu \in \mathbb{N}} \{ \xi_{i} \} \), and the two inequalities follow from Lemma 1(i). Moreover, using triangular inequality for \( i \in [N] \) yields

\[
\| x^{T\nu+t_v(i)} - x^{T\nu} \|_{l-1} \leq \sum_{r=1}^{T} \| x^{T\nu+t_v(i-1)} - x^{T\nu+t_v(i-1)} \|_{l-1} \leq \frac{T}{\lambda_{max}(\Gamma_{-})} (\Phi^n_{T}(x^{T\nu}) - \Phi^n_{T}(x^{T\nu+1}))^{1/2}
\]

(33)

where the second inequality follows from (32) together with the fact that \( t_v(i) \leq T \). For all \( i \in [N] \), from the triangular inequality and the \( L_{T} \)-Lipschitz continuity of \( T_{\Gamma}^{n} \) (Lemma A.5(iii)) we have

\[
gamma\| U_{t} (x^{T\nu} - T_{\Gamma}^{n}(x^{T\nu})) \| \leq \gamma_{i}^{-1/2} \| U_{t}(x^{T\nu} - T_{\Gamma}^{n}(x^{T\nu+1})) \| + \gamma_{i}^{-1/2} \| U_{t}(T_{\Gamma}^{n}(x^{T\nu+1}) - T_{\Gamma}^{n}(x^{T\nu})) \|
\]

\[
\leq \| (x^{T\nu+1}) - x^{T\nu} \|_{l-1} + \| T_{\Gamma}^{n}(x^{T\nu+1}) - T_{\Gamma}^{n}(x^{T\nu}) \|_{l-1}
\]

\[
\leq \| x^{T\nu+1} - x^{T\nu} \|_{l-1} + \| x^{T\nu+1} - x^{T\nu} \|_{l-1} + L_{T} \| x^{T\nu+1} - x^{T\nu} \|_{l-1}
\]

(32), (33)

\[
\leq \frac{2^T L_{T}}{\gamma_{min}/2} (\Phi^n_{T}(x^{T\nu}) - \Phi^n_{T}(x^{T\nu+1}))^{1/2}
\]

(34)

The second inequality follows from (31) together with the fact that \( x^{T\nu} = x^{T\nu+i} = \cdots = x^{T\nu+t_v(i-1)} \) by definition of \( t_v(i) \). By squaring and summing over \( i \in [N] \), we obtain

\[
\Phi^n_{T}(x^{T\nu+1}) - \Phi^n_{T}(x^{T\nu}) \leq - \frac{\lambda_{min}(\Gamma_{-})}{2} \| x^{T\nu+1} - x^{T\nu} \|_{l-1}^2
\]

(35)

By telescoping the inequality and using the fact that \( \min \Phi^n_{T} = \min \Phi > -\infty \) (Lemma A.3(ii)), we obtain that \( (\| z^{T\nu} - x^{T\nu} \|_{l-1}^2) \in \mathbb{N} \) has finite sum, and in particular vanishes. Clearly, by suitably shifting, for every \( t \in [T] \) the same can be said for the sequence \( (\| z^{T\nu+i} - x^{T\nu+i} \|_{l-1}^2) \in \mathbb{N} \). The whole sequence \( (\| z_{k} - x_{k} \|_{l-1}^2) \in \mathbb{N} \) is thus summable, and in particular \( (x_{k}) \in \mathbb{N} \) and \( (z_{k}) \in \mathbb{N} \) have the same cluster points.

Let \( (x_{k}) \in \mathbb{K} \) be a subsequence converging to a point \( x^{*} \). Then, since \( T_{\Gamma}^{n}(x_{k}) \in z_{k} \) and \( z_{k} \) also converges to \( x^{*} \) as \( K \to \infty \), the inclusion \( 0 \in \partial \Phi(x^{*}) \) follows from Lemma A.4.

Proof of Theorem 6 (essentially cyclic sampling: global convergence)

The proof can pattern the arguments of the proof of Theorem 2, using the sufficient decrease established in (35), to obtain the following deterministic variant of (28):

\[
\Delta_{v+1} \leq \Delta_{v} - \frac{\omega_{T}}{2V_{max}(\gamma_{min})} \| x^{T\nu+1} - x^{T\nu} \|_{l-1} \quad \text{for all } v \in \mathbb{N},
\]
where $\Delta_t$, $\gamma_{\min}$ and $\gamma_{\max}$ are as in (28) and $\sigma'=\frac{\xi_{\min}}{N(1+T)\xi}$ with $\xi_{\min}$ and $L_T$ as in the proof of Theorem 5. By summing over $\nu \in \mathbb{N}$ (sure) summability of the sequence $(\|x^{T+\nu} - x^{T}\|_{\nu} \in \mathbb{N})$ is obtained. By suitably shifting, for every $t \in [T]$ the same can be said for the sequence $(\|z^{T+\nu} - x^{T+\nu}\|_{\nu} \in \mathbb{N})$, and since $T$ is finite we conclude that the whole sequence $(\|z^k - x^k\|_{k} \in \mathbb{N})$ is summable. We may invoke the final part of the proof of Theorem 3 to obtain the claim.

**Proof of Theorem 7 (essentially cyclic sampling: linear convergence)**

Here and in what follows, $\Delta$ and $\delta$ are as in the statement of Theorem 7.

- Since $T^n$ is $L_T$-Lipschitz continuous with $L_T = 1-\delta$ as shown in Lemma A.5(iii), inequality (35) can be tightened to

$$\Phi^m_t(x^{T(v+1)}) - \Phi^m_t(x^{Tv}) \leq -\frac{1-\Delta}{2(1+T(1-\delta))}\|\xi^{Tv} - x^{Tv}\|_{T-1}^2.$$  

Moreover, it follows from (29) that

$$\Phi^m_t(x^{Tv}) - \Phi_* \leq \frac{1}{2}(\delta^{-1} - 1)|\|z^{Tv} - x^{Tv}\|_{T-1}^2.$$

By combining the two inequalities the claimed linear bound for the essentially cyclic sampling is obtained.

- Let us now suppose that the sampling strategy follows a shuffled rule as in (8) with permutations $\pi_0, \pi_1, \ldots$ (hence in the cyclic case $\pi_{\nu} = \text{id}$ for all $\nu \in \mathbb{N}$). In this case the bound for the essentially cyclic case obtained in (35) can be tightened as follows. Let $U_i$ be as in (14) and $\xi_{\min}$ as in the proof of Theorem 5. Observe that $t_{U_i} = \pi_{i+1}^{-1}(i) \leq N$ for $t_{U_i}$ as defined in (30). For all $t \in [N]$

$$\Phi^m_t(x^{N(v+1)}) - \Phi^m_t(x^{Nv}) \leq \Phi^m_t(x^{Nv+1}) - \Phi^m_t(x^{Nv}) \leq -\frac{\xi_{\min}}{2} \sum_{t=1}^{v-1} \|x^{Nv+t} - x^{Nv+t-1}\|_{T-1}^2,$$

where the equality follows from the fact that at every iteration a different coordinate is updated (and that $\Gamma$ is diagonal), and the inequalities from Lemma 1(i). Similarly, (32) 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 (37) in place of (33), inequality (34) is tightened as follows

$$\gamma_{i/2}||U_i'(x^{Nv} - T^n_t(x^{Nv}))|| \leq \frac{1+L_T}{\sqrt{\xi_{\min}}/2} \left(\Phi^m_t(x^{Nv}) - \Phi^m_t(x^{N(v+1)})\right)^{1/2}.$$  

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

$$\Phi^m_t(x^{N(v+1)}) - \Phi^m_t(x^{Nv}) \leq \frac{\xi_{\min}}{2(1+L_t)}\|\xi^{Nv} - x^{Nv}\|_{T-1}^2 = -\frac{1-\Delta}{2(1+T(1-\delta))}\|\xi^{Nv} - x^{Nv}\|_{T-1}^2,$$

where $L_T = 1-\delta$ as discussed above. Moreover, it follows from (29) that

$$\Phi^m_t(x^{Nv}) - \Phi_* \leq \frac{1}{2}(\delta^{-1} - 1)|\|z^{Nv} - x^{Nv}\|_{T-1}^2.$$

By combining the two inequalities the claimed rate (10) is obtained.

**References**

[1] Allen-Zhu, Z., Qu, Z., Richtárik, P., Yuan, Y.: Even faster accelerated coordinate descent using non-uniform sampling. In: International Conference on Machine Learning, pp. 1110–1119 (2016)

[2] 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). DOI 10.1007/s10107-007-0133-5

[3] Attouch, H., Bolte, J., Redont, P., Soubeyran, A.: Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Lojasiewicz inequality. Mathematics of Operations Research 35(2), 438–457 (2008)
[4] 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)

[5] Bauschke, H.H., Combettes, P.L.: Convex analysis and monotone operator theory in Hilbert spaces. CMS Books in Mathematics. Springer (2017). DOI 10.1007/978-3-319-48311-5

[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.: Convex Optimization Algorithms. Athena Scientific, Belmont, Mass (2015)

[9] Bertsekas, D.P.: Nonlinear Programming. Athena Scientific (2016)

[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., Sabach, S., Teboulle, M.: Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Mathematical Programming 146(1-2), 459–494 (2014). DOI 10.1007/s10107-013-0701-9. URL http://dx.doi.org/10.1007/s10107-013-0701-9

[12] Chow, Y., 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)

[13] Davis, D.: Smart: The stochastic monotone aggregated root-finding algorithm. arXiv preprint arXiv:1601.00698 (2016)

[14] 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)

[15] 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)

[16] 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). DOI 10.1007/s10957-014-0642-3

[17] 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)

[18] 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)

[19] Hou, Y., Song, L., 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)

[20] Ioffe, A.: Variational Analysis of Regular Mappings: Theory and Applications. Springer Monographs in Mathematics. Springer International Publishing (2017)

[21] Latafat, P., Freris, N.M., Patrinos, P.: A new randomized block-coordinate primal-dual proximal algorithm for distributed optimization. IEEE Transactions on Automatic Control pp. 1–1 (2019). DOI 10.1109/TAC.2019.2906924

[22] Li, G., Pong, T.K.: Douglas-Rachford splitting for nonconvex optimization with application to nonconvex feasibility problems. Mathematical Programming 159(1), 371–401 (2016)

[23] 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)

[24] Mairal, J.: Incremental majorization-minimization optimization with application to large-scale machine learning. SIAM Journal on Optimization 25(2), 829–855 (2015)

[25] 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)
[26] Nesterov, Y.: Efficiency of coordinate descent methods on huge-scale optimization problems. SIAM Journal on Optimization 22(2), 341–362 (2012)

[27] Nesterov, Y.: Introductory lectures on convex optimization: A basic course, vol. 87. Springer Science & Business Media (2013)

[28] 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). DOI 10.1137/130942954

[29] Patrinos, P., Bemporad, A.: Proximal newton methods for convex composite optimization. In: 52nd IEEE Conference on Decision and Control, pp. 2358–2363 (2013)

[30] Patrinos, P., Stella, L., Bemporad, A.: Douglas-rachford splitting: Complexity estimates and accelerated variants. In: 53rd IEEE Conference on Decision and Control, pp. 4234–4239 (2014). DOI 10.1109/CDC.2014.7040049

[31] Reddi, S.J., Hefny, A., Sra, S., Poczos, B., Smola, A.: Stochastic variance reduction for nonconvex optimization. In: International conference on machine learning, pp. 314–323 (2016)

[32] 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)

[33] 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)

[34] 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)

[35] Rockafellar, R.T., Wets, R.J.B.: Variational analysis, vol. 317. Springer Science & Business Media (2011)

[36] Schmidt, M., Le Roux, N., Bach, F.: Minimizing finite sums with the stochastic average gradient. Mathematical Programming 162(1), 83–112 (2017)

[37] Shalev-Shwartz, S., Zhang, T.: Stochastic dual coordinate ascent methods for regularized loss minimization. Journal of Machine Learning Research 14(Feb), 367–399 (2013)

[38] 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)

[39] Tseng, P.: Convergence of a block coordinate descent method for nondifferentiable minimization. Journal of optimization theory and applications 109(3), 475–494 (2001)

[40] Tseng, P., Bertsekas, D.P.: Relaxation methods for problems with strictly convex separable costs and linear constraints. Mathematical Programming 38(3), 303–321 (1987)