Bregman Finito/MISO for Nonconvex Regularized Finite Sum Minimization without Lipschitz Gradient Continuity

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

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

Ahookhosh, Masoud
Department of Mathematics and Computer Science, University of Antwerp

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

出版情報：SIAM Journal on Optimization. 32 (3), pp.2230-2262, 2022-09-13. Society for Industrial and Applied Mathematics
バージョン：

SIAM J. OPTIM.
Vol. 32, No. 3, pp. 2230–2262

BREGMAN FINITO/MISO FOR NONCONVEX REGULARIZED FINITE SUM MINIMIZATION WITHOUT LIPSCHITZ GRADIENT CONTINUITY*

PUYA LATAFAT†, ANDREAS THEMELIS‡, MASOUD AHOOKHOSH§, AND PANAGIOTIS PATRINOS†

Abstract. We introduce two algorithms for nonconvex regularized finite sum minimization, where typical Lipschitz differentiability assumptions are relaxed to the notion of relative smoothness. The first one is a Bregman extension of Finito/MISO [A. Defazio and J. Domke, *Proc. Mach. Learn. Res. (PMLR)*, 32 (2014), pp. 1125–1133; J. Mairal, *SIAM J. Optim.*, 25 (2015), pp. 829–855], studied for fully nonconvex problems when the sampling is randomized, or under convexity of the nonsmooth term when it is essentially cyclic. The second algorithm is a low-memory variant, in the spirit of SVRG [R. Johnson and T. Zhang, Advances in *Neural Information Processing Systems* 26, Curran Associates, Red Hook, NY, 2013, pp. 315–323] and SARAH [L. M. Nguyen et al., *Proc. Mach. Learn. Res. (PMLR)*, 70 (2017), pp. 2613–2621], that also allows for fully nonconvex formulations. Our analysis is made remarkably simple by employing a Bregman–Moreau envelope as the Lyapunov function. In the randomized case, linear convergence is established when the cost function is strongly convex, yet with no convexity requirements on the individual functions in the sum. For the essentially cyclic and low-memory variants, global and linear convergence results are established when the cost function satisfies the Kurdyka–Łojasiewicz property.

Key words. nonsmooth nonconvex optimization, incremental aggregated algorithms, Bregman–Moreau envelope, KL inequality

MSC codes. 90C06, 90C25, 90C26, 49J52, 49J53

DOI. 10.1137/21M140376X

1. Introduction.

We consider the following regularized finite sum minimization

$$\text{(P)} \qquad \text{minimize}_{x\in\mathbb{R}^n}\ \varphi(x) \coloneqq \tfrac{1}{N}\textstyle\sum_{i=1}^N f_i(x) + g(x) \quad \text{subject to}\, x \in \overline{C},$$

where $\overline{C}$ denotes the closure of $C \coloneqq \bigcap_{i=1}^N \operatorname{int}\operatorname{dom} h_i$ for some convex functions h_i, $i \in [N] \coloneqq \{1,\dots,N\}$. Our goal in this paper is to study such problems without imposing convexity assumptions on f_i and g, and in a setting where f_i are differentiable but their gradients need not be Lipschitz continuous. Our full setting is formalized in Assumption I.

To relax the Lipschitz differentiability assumption, we adopt the notion of smoothness relative to a distance-generating function [7], and following [40] we will use the

*Received by the editors March 9, 2021; accepted for publication (in revised form) June 13, 2022; published electronically September 13, 2022.
https://doi.org/10.1137/21M140376X
Funding: This work was supported by Research Foundation Flanders (FWO) Ph.D. grant 1196820N and research projects G0A0920N, G086518N, and G086318N; by Research Council KU Leuven C1 project C14/18/068; by Fonds de la Recherche Scientifique - FNRS and the Fonds Wetenschappelijk Onderzoek - Vlaanderen under EOS project 30468160 (SeLMA); and by JSPS KAKENHI grant JP21K17710.
†Department of Electrical Engineering (ESAT-STADIUS) – KU Leuven, 3001 Leuven, Belgium (puya.latafat@kuleuven.be, panos.patrinos@esat.kuleuven.be).
‡Faculty of Information Science and Electrical Engineering (ISEE) – Kyushu University, 744 Motooka, Nishi-ku 819-0395, Fukuoka, Japan (andreas.themelis@ees.kyushu-u.ac.jp).
§Department of Mathematics and Computer Science, University of Antwerp, B-2020 Antwerp, Belgium (masoud.ahookhosh@uantwerp.be).

Algorithm 1 Bregman Finito/MISO for the regularized finite sum minimization (P).

REQUIRE Legendre kernels h_i such that f_i is L_{f_i}-smooth relative to h_i
stepsizes $\gamma_i \in (0, {}^N/_{L_{f_i}})$
initial point $x^{\text{init}} \in C := \bigcap_{i=1}^N \operatorname{int}\operatorname{dom} h_i$

INITIALIZE table $\boldsymbol{s}^0 = (s_1^0, \dots, s_N^0) \in \mathbb{R}^{nN}$ of vectors $s_i^0 = \frac{1}{\gamma_i}\nabla h_i(x^{\text{init}}) - \frac{1}{N}\nabla f_i(x^{\text{init}})$
$\mathbb{R}^n$-vector $\tilde{s}^0 = \sum_{i=1}^N s_i^0$

REPEAT for $k = 0, 1, \dots$ until convergence

1: Compute $z^k \in \arg\min_{w\in\mathbb{R}^n} \left\{ g(w) + \sum_{i=1}^N \frac{1}{\gamma_i} h_i(w) - \langle \tilde{s}^k, w\rangle \right\}$

2: Select a subset of indices $\mathcal{J}^{k+1} \subseteq [N] := \{1, \dots, N\}$ and update the table $\boldsymbol{s}^{k+1}$ as follows:

$$s_i^{k+1} = \begin{cases} \frac{1}{\gamma_i}\nabla h_i(z^k) - \frac{1}{N}\nabla f_i(z^k) & \text{if } i \in \mathcal{J}^{k+1}, \\ s_i^k & \text{otherwise} \end{cases}$$

3: Update the vector $\tilde{s}^{k+1} = \tilde{s}^k + \sum_{i\in\mathcal{J}^{k+1}} (s_i^{k+1} - s_i^k)$

RETURN z^k

terminology of *relative smoothness*. Despite the lack of Lipschitz differentiability, in many applications the involved functions satisfy a descent property where the usual quadratic upper bound is replaced by a Bregman distance (cf. Fact 2.5(i) and Definition 2.1). Owing to this property, Bregman extensions for many classical schemes have been proposed [7, 40, 6, 59, 49, 1].

In the setting of finite sum minimization, the incremental aggregated algorithm PLIAG was proposed recently [70] as a Bregman variant of the incremental aggregated gradient method [15, 16, 65]. The analysis of PLIAG is limited to the convex case and requires restrictive assumptions for the Bregman kernel [70, Thm. 1, Assump. 8]. Stochastic mirror descent (SMD) is another relevant algorithm which can tackle more general stochastic optimization problems. SMD may be viewed as a Bregman extension of the stochastic (sub)gradient method and has long been studied [46, 61, 11, 45]. More recently, [32] studied SMD for convex and relatively smooth formulations, and (sub)gradient versions have been analyzed under relative continuity in a convex setting [39], as well as relative weak convexity [71, 25].

Motivated by these recent works, we propose a Bregman extension of the popular Finito/MISO algorithm [28, 42] in a fully nonconvex setting and with very general sampling strategies that will be made precise shortly after. In a nutshell, our analysis revolves around the fact that, regardless of the index selection strategy, the function $\mathscr{L} : \mathbb{R}^n \times \mathbb{R}^{nN} \to \overline{\mathbb{R}}$ defined as

$$\mathscr{L}(z, \boldsymbol{s}) := \varphi(z) + \sum_{i=1}^N \mathrm{D}_{\hat{h}_i^*}\big(s_i, \nabla\hat{h}_i(z)\big), \tag{1.1}$$

where $\hat{h}_i^*$ denotes the convex conjugate of $\hat{h}_i := {}^{h_i}/_{\gamma_i} - {}^{f_i}/_N$, monotonically decreases along the iterates $(z^k, \boldsymbol{s}^k)_{k\in\mathbb{N}}$ generated by Algorithm 1 (see Assumption I for the requirements on h_i, f_i). Our methodology leverages an interpretation of Finito/MISO as a block-coordinate algorithm that was observed in [37] in the Euclidean setting. In fact, the analysis is here further simplified after noticing that the smooth function can be "hidden" in the distance-generating function, resulting in a Lyapunov function $\mathscr{L}$ that can be expressed as a *Bregman–Moreau envelope* (cf. Lemma 3.2).

Algorithm 2 Low-memory Bregman Finito/MISO.

REQUIRE Legendre kernels h_i such that f_i is L_{f_i}-smooth relative to h_i
stepsizes $\gamma_i \in (0, {}^{N}/_{L_{f_i}})$
initial point $x^{\text{init}} \in C := \bigcap_{i=1}^N \operatorname{int} \operatorname{dom} h_i$

INITIALIZE $\mathbb{R}^n$-vector $\tilde{s}^0 = \sum_{i=1}^N \left[\frac{1}{\gamma_i}\nabla h_i(x^{\text{init}}) - \frac{1}{N}\nabla f_i(x^{\text{init}})\right]$
set of selectable indices $\mathcal{K}^0 = \emptyset$ ▷ set to $\emptyset$ so as to start with a full update

REPEAT FOR $k = 0, 1, \ldots$ until convergence

1: $z^k \in \arg\min_{w\in\mathbb{R}^n} \left\{ g(w) + \sum_{i=1}^N \frac{1}{\gamma_i} h_i(w) - \langle \tilde{s}^k, w\rangle \right\}$
2: IF $\mathcal{K}^k = \emptyset$ THEN ▷ No index left to be sampled: full update
3: $\mathcal{I}^{k+1} = \mathcal{K}^{k+1} = [N]$ ▷ activate all indices and reset the selectable indices
4: $\tilde{z}^k = z^k$ ▷ store the full update z^k
5: $\tilde{s}^{k+1} = \sum_{i=1}^N \left[\frac{1}{\gamma_i}\nabla h_i(z^k) - \frac{1}{N}\sum_{i=1}^N \nabla f_i(z^k)\right]$
6: ELSE
7: select a nonempty subset $\mathcal{I}^{k+1} \subseteq \mathcal{K}^k$ ▷ select among the indices not yet sampled
8: $\mathcal{K}^{k+1} = \mathcal{K}^k \setminus \mathcal{I}^{k+1}$ ▷ update the set of selectable indices
9: $\tilde{z}^k = \tilde{z}^{k-1}$
10: $\tilde{s}^{k+1} = \tilde{s}^k + \sum_{i\in\mathcal{I}^{k+1}} \left[\left(\frac{1}{\gamma_i}\nabla h_i(z^k) - \frac{1}{N}\nabla f_i(z^k)\right) - \left(\frac{1}{\gamma_i}\nabla h_i(\tilde{z}^k) - \frac{1}{N}\nabla f_i(\tilde{z}^k)\right)\right]$

RETURN $\tilde{z}^k$

We cover a wide range of sampling strategies for the index set $\mathcal{I}^{k+1}$ at step 2, which we can summarize into the following two umbrella categories:

$(\mathcal{S}_1)$ Randomized rule: $\exists p_1, \ldots, p_N > 0 : \quad \mathcal{P}_k\left[i \in \mathcal{I}^{k+1}\right] = p \quad \forall k \in \mathbb{N}, i \in [N].$

$(\mathcal{S}_2)$
Essentially cyclic rule: $\exists T > 0 : \quad \bigcup_{t=1}^T \mathcal{I}^{k+t} = [N] \quad \forall k \in \mathbb{N}.$

The randomized setting $(\mathcal{S}_1)$, in which $\mathcal{P}_k$ denotes the probability conditional to the knowledge at iteration k, covers, for instance, a minibatch strategy of size b. Another notable case is when each index i is selected at random with probability p_i independently of other indices.

The essentially cyclic rule $(\mathcal{S}_2)$ is also very general and has been considered by many authors [62, 60, 33, 24, 67]. Two notable special cases of single index selection rules complying with $(\mathcal{S}_2)$ are the cyclic and shuffled cyclic sampling strategies:

$(\mathcal{S}_2^{\text{SHUF}})$ Shuffled cyclic rule: $\mathcal{I}^{k+1} = \left\{\pi_{\lfloor k/N\rfloor}\left(\operatorname{mod}(k, N) + 1\right)\right\},$
where $\pi_0, \pi_1, \ldots$ are permutations of the set of indices $[N]$ (chosen randomly or deterministically). When $\pi_{\lfloor k/N\rfloor} = \operatorname{id}$ one recovers the (plain) cyclic sampling rule

$(\mathcal{S}_2^{\text{CYCL}})$ Cyclic rule: $\mathcal{I}^{k+1} = \{\operatorname{mod}(k, N) + 1\}.$

We remark that, in the cyclic case, our algorithm generalizes DIAG [44] for smooth strongly convex problems, which itself may be seen as a cyclic variant of Finito/MISO.

1.1. Low-memory variant. One iteration of Algorithm 1 involves the computation of z^k at step 1 and that of the gradients $\nabla({}^{h_i}/_{\gamma_i} - {}^{f_i}/_{N})$, $i \in \mathcal{I}^{k+1}$, at step 3. Consequently, the overall complexity of each iteration is independent of the number N of functions appearing in problem (P), and is instead proportional to the number

of sampled indices, which the user is allowed to upper bound to any integer between 1 and N. As is the case for all incremental gradient methods, the low iteration cost comes at the price of having to store in memory a table $\boldsymbol{s}^k$ of N many $\mathbb{R}^n$ vectors, which can become problematic when N grows large. Other incremental algorithms for convex optimization such as IAG [15, 16, 65], IUG [64], SAG [54], and SAGA [27], can considerably reduce memory allocation from $O(nN)$ to $O(n)$ in applications such as logistic regression and lasso where the gradients ∇f_i can be expressed as scaled versions of the data vectors. Despite the favorable performance of the Finito/MISO algorithm on such problems as observed in [27], this memory reduction trick cannot be employed due to the fact that the vectors s_i stored in the table depend not only on the gradients, but also on the vectors $\nabla h_i(z^k)$. Nevertheless, inspired by the popular stochastic methods SVRG [34, 66] and SARAH [48], by suitably interleaving incremental and full gradient evaluations it is possible to completely waive the need of a memory table and match the $O(n)$ storage requirement.

In a nutshell, after a full update — which in Algorithm 1 corresponds to selecting $\mathcal{I}^{k+1} = [N]$ — all vectors s_i^{k+1} in the table only depend on variable z^k computed at step 1, until i is sampled again. As long as full gradient updates are frequent enough so that no index is sampled twice in-between, it thus suffices to keep track of $z^k \in \mathbb{R}^n$ instead of the table $\boldsymbol{s}^k \in \mathbb{R}^{nN}$. The variant is detailed in Algorithm 2, in which $\mathcal{K}^k \subseteq [N]$ keeps track of the indices that have not yet been sampled between full gradient updates (and is thus reset whenever such full steps occur; cf. step 3). Vector $\tilde{z}^k \in \mathbb{R}^n$ is equal to z^k corresponding to the latest full gradient update (cf. step 4) and acts as a low-memory surrogate of the table $\boldsymbol{s}^k$ of Algorithm 1. Similarly to SVRG and SARAH, this reduction in the storage requirements comes at the cost of an extra gradient evaluation per sampled index (cf. step 10).

Since full gradient updates correspond to selecting all indices, Algorithm 2 may be viewed as Algorithm 1 with an essentially cyclic sampling rule of period N, a claim that will be formally shown in Lemma 4.12. In fact, not only does it naturally inherit all the convergence results, but its particular sampling strategy also allows us to waive convexity requirements on g that are necessary for more general essentially cyclic rules.

1.2. Contributions. As a means to informally summarize the content of the paper, in Table 1 we synopsize the convergence results of the two algorithms.

1. To the best of our knowledge, this is the first analysis of an incremental aggregated method in a fully nonconvex setting and without Lipschitz differentiability assumptions. Our analysis, surprisingly simple and yet covering randomized and essentially cyclic samplings altogether, relies on a sure descent property on the Bregman–Moreau envelope (cf. Lemma 4.2).

2. We propose a novel low-memory variant of the (Bregman) Finito/MISO algorithm, that in the spirit of SVRG [34, 66] and SARAH [48] alternates between incremental steps and a full proximal gradient step. It is highly interesting even in the Euclidean case, as it can accommodate fully nonconvex formulations while maintaining an $O(n)$ memory requirement.

3. Linear convergence of Algorithm 1 in the randomized case is established when the cost function φ is strongly convex, yet with no convexity requirement on f_i or g. To the best of our knownledge, this is a novelty even in the Euclidean case, for all available results are bound to strong convexity of each term f_i in the sum; see, e.g., [28, 42, 44, 37, 50]. This type of assumption has also been considered in [2] for the case of SVRG. Although in practice it is hardly ever the case that f is strongly

TABLE 1

Summary of the convergence results for Algorithm 1 *with randomized rule* ($\mathcal{S}_1$) *and essentially cyclic rule* ($\mathcal{S}_2$)*, and for the low-memory variant of Algorithm* 2 (LM)*. Claims are either* sure*, almost sure* (a.s.)*, or in expectation* ($\mathbb{E}$)*.* Other abbreviations: loc*: locally;* cvx*: convex;* str*: strongly;* smooth*: Lipschitz differentiable;* ω*: set of limit points;* $\mathrm{KL}_{1/2}$*: Kurdyka–Łojasiewicz property with exponent* $1/2$*.*

	Sampling	Requirements (on top of Assumption I)	Property	Reference
z^k bounded	any	$\varphi + \delta_{\overline{C}}$ level bounded	sure	Lemma 4.2(iv)
$\varphi(z^k)$ convergent	$\mathcal{S}_1$		a.s.	Theorem 4.6(ii)
	$\mathcal{S}_2$	$C = \mathbb{R}^n$; g cvx; h_i loc str cvx smooth; φ level bounded[a]	sure	Theorem 4.9
	LM			Theorem 4.13
$\omega(z^k)$ stationary	$\mathcal{S}_1$	either $C = \mathbb{R}^n$	a.s.	Theorem 4.6(iv)
		or $\operatorname{dom} h_i$ closed; φ cvx		Theorem 4.6(vi)
	$\mathcal{S}_2$	$C = \mathbb{R}^n$; g cvx; h_i loc str cvx smooth; φ level bounded[a]	sure	Theorem 4.9
	LM	$C = \mathbb{R}^n$		Theorem 4.13
z^k convergent	$\mathcal{S}_1$	either $C = \mathbb{R}^n$; φ cvx or Assumption II; $\varphi + \delta_{\overline{C}}$ cvx level bounded	a.s.	Theorem 4.6(vii)
	$\mathcal{S}_2$	Assumption III; g cvx	sure	Theorem 4.11(i)
	LM	Assumption III		Theorem 4.14(i)
$\varphi(z^k)$ and z^k linearly convergent	$\mathcal{S}_1$	$C = \mathbb{R}^n$; φ str cvx; h_i loc smooth	$\mathbb{E}$	Theorem 4.7
	$\mathcal{S}_2$	Assumption III; φ $\mathrm{KL}_{1/2}$; g cvx	sure	Theorem 4.11(iii)
	LM	Assumption III; φ $\mathrm{KL}_{1/2}$		Theorem 4.14(iii)

[a] Level boundedness is not necessary if the h_i are *globally* smooth and strongly convex (cf. Theorem 4.9).

convex without each f_i also being (strongly) convex, our analysis being agnostic to the decomposition leads to tighter and more general results.

4. We leverage the Kurdyka–Łojasiewicz (KL) property to establish global (as apposed to subsequential) convergence as well as linear convergence, for Algorithm 1 with (essentially) cyclic sampling and for the low-memory Algorithm 2.

1.3. Organization. We conclude this section by introducing some notational conventions. The problem setting is formally described in section 2 together with a list of related definitions and known facts involving Bregman distances, relative smoothness, and proximal mapping. Section 3 offers an alternative interpretation of Algorithm 1 as the block-coordinate Bregman proximal point Algorithm 3, which majorly simplifies the analysis, addressed in section 4. Some auxiliary results are deferred to Appendix A. Section 5 applies the proposed algorithms to sparse phase retrieval problems, and section 6 concludes the paper.

1.4. Notation. The set of real and extended-real numbers are $\mathbb{R} := (-\infty, \infty)$ and $\overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$, while the positive and strictly positive reals are $\mathbb{R}_+ := [0, \infty)$ and $\mathbb{R}_{++} := (0, \infty)$. With id we indicate the identity function $x \mapsto x$ defined on a suitable space. We denote by $\langle \cdot\,, \cdot \rangle$ and $\|\cdot\|$ the standard Euclidean inner product and the induced norm. For a vector $\boldsymbol{w} = (w_1, \ldots, w_r) \in \mathbb{R}^{\sum_i n_i}$, $w_i \in \mathbb{R}^{n_i}$ is used to denote its ith block coordinate. $\operatorname{int} E$ and $\operatorname{bdry} E$, respectively, denote the interior and boundary of a set E, and for 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 $k \in \mathbb{N}$. We say that $(x^k)_{k\in\mathbb{N}}$ converges at *Q-linear rate* (resp., *R-linear rate*) to a point x if there exists $c \in (0,1)$ 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 mapping from each point $x \in \mathbb{R}^n$ to a subset $H(x)$ of $\mathbb{R}^m$. The set $\operatorname{gph} H := \{(x, y) \in \mathbb{R}^n \times \mathbb{R}^m \mid y \in H(x)\}$ is the *graph* of H. 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 $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ are the sets $\operatorname{dom} h := \{x \in \mathbb{R}^n \mid h(x) < \infty\}$ 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 $\mathbb{R}^{n+1}$. We say that h is *level bounded* if its α-sublevel set $\operatorname{lev}_{\le\alpha} h := \{x \in \mathbb{R}^n \mid h(x) \le \alpha\}$ is bounded for all $\alpha \in \mathbb{R}$. The *conjugate* of h is defined by $h^*(y) := \sup_{x\in\mathbb{R}^n} \{\langle y, x\rangle - h(x)\}$. The indicator function of a set $E \subseteq \mathbb{R}^n$ is denoted by δ_E, namely, $\delta_E(x) = 0$ if $x \in E$ and ∞ otherwise.

We denote by $\hat{\partial} h : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ the *regular subdifferential* of h, where

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

A necessary condition for local minimality of x for h is $0 \in \hat{\partial} h(x)$; see [53, Thm. 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 $k \to \infty$. Finally, the set of r times continuously differentiable functions from X to $\mathbb{R}$ is denoted by $\mathscr{C}^r(X)$.

2. Problem setting and preliminaries.

Throughout this paper, problem (P) is studied under the following assumptions.

Assumption I (basic requirements). In problem (P),

A1. $f_i : \mathbb{R}^n \to \overline{\mathbb{R}}$ are L_{f_i}-smooth relative to Legendre kernels h_i (Definitions 2.2 and 2.4);

A2. $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ is proper and lsc;

A3. a solution exists: $\arg\min \{\varphi(x) \mid x \in \overline{C}\} \neq \emptyset$;

A4. for given $\gamma_i \in (0, {}^{N}/_{L_{f_i}})$, $i \in [N]$, it holds that for any $s \in \mathbb{R}^n$

$$T(s) := \operatorname*{arg\,min}_{w\in\mathbb{R}^n} \Big\{ g(w) + \textstyle\sum_{i=1}^N \frac{1}{\gamma_i} h_i(w) - \langle s, w\rangle \Big\} \subseteq C. \tag{2.1}$$

As it will become clear in section 3, the subproblem (2.1) is in fact a reformulation of a (Bregman) proximal mapping. Assumption I.A4 excludes boundary points from $\operatorname{range} T$. This is a standard assumption that usually holds in practice [20, 56], e.g., when g is convex or when the intersection of $\operatorname{dom} h_i$, $i \in [N]$, is an open set.

Definition 2.1 (Bregman distance). *For a convex function $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ that is continuously differentiable on* $\operatorname{int}\operatorname{dom} h \neq \emptyset$, *the* Bregman distance $\mathrm{D}_h : \mathbb{R}^n \times \mathbb{R}^n \to \overline{\mathbb{R}}$ *is defined as*

$$\mathrm{D}_h(x, y) := \begin{cases} h(x) - h(y) - \langle \nabla h(y), x - y\rangle & \text{if } y \in \operatorname{int}\operatorname{dom} h, \\ \infty & \text{otherwise.} \end{cases} \tag{2.2}$$

Function h will be referred to as a distance-generating function.

Definition 2.2 (Legendre kernel). *A proper, lsc, and strictly convex function $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ with* $\operatorname{int}\operatorname{dom} h \neq \emptyset$ *and such that $h \in \mathscr{C}^1(\operatorname{int}\operatorname{dom} h)$ is said to be a Legendre kernel if it is* (i) 1-coercive, *i.e., such that* $\lim_{\|x\|\to\infty} {}^{h(x)}/_{\|x\|} = \infty$, *and* (ii)

essentially smooth, *i.e., if* $\|\nabla h(x_k)\| \to \infty$ *for every sequence* $(x_k)_{k\in\mathbb{N}} \subseteq \operatorname{int}\operatorname{dom} h$ *converging to a boundary point of* $\operatorname{dom} h$.

Fact 2.3. Let $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a Legendre kernel, $x \in \mathbb{R}^n$, and $y, z \in \operatorname{int}\operatorname{dom} h$. Then we have the following:

(i) $h^* \in \mathscr{C}^1(\mathbb{R}^n)$ is strictly convex and $\nabla h^{-1} = \nabla h^*$ [52, Thm. 26.5 and Cor. 13.3.1].

(ii) $\mathrm{D}_h(x, z) = \mathrm{D}_h(x, y) + \mathrm{D}_h(y, z) + \langle x - y, \nabla h(y) - \nabla h(z)\rangle$ [22, Lem. 3.1].

(iii) $\mathrm{D}_h(y, z) = \mathrm{D}_{h^*}(\nabla h(z), \nabla h(y))$ [8, Thm. 3.7(v)].

(iv) $\mathrm{D}_h(\,\cdot\,, z)$ and $\mathrm{D}_h(z, \,\cdot\,)$ are level bounded [9, Lem. 7.3(v)–(viii)].

(v) If $\operatorname{dom} h$ is closed and $\mathrm{D}_h(x^k, y^k) \to 0$ for some $x^k \in \operatorname{dom} h$ and $y^k \in \operatorname{int}\operatorname{dom} h$, then $(x^k)_{k\in\mathbb{N}}$ converges to a point x iff so does $(y^k)_{k\in\mathbb{N}}$ [56, Thm. 2.4].

Moreover, for any convex set $\mathcal{U} \subseteq \operatorname{int}\operatorname{dom} h$ and $u, v \in \mathcal{U}$ the following hold:

(vi) If h is $\mu_{h,\mathcal{U}}$-strongly convex on $\mathcal{U}$, then

$$\frac{\mu_{h,\mathcal{U}}}{2}\|v - u\|^2 \le \mathrm{D}_h(v, u) \le \frac{1}{2\mu_{h,\mathcal{U}}}\|\nabla h(v) - \nabla h(u)\|^2.$$

(vii) If ∇h is $\ell_{h,\mathcal{U}}$-Lipschitz on $\mathcal{U}$, then

$$\frac{1}{2\ell_{h,\mathcal{U}}}\|\nabla h(v) - \nabla h(u)\|^2 \le \mathrm{D}_h(v, u) \le \frac{\ell_{h,\mathcal{U}}}{2}\|v - u\|^2.$$

Definition 2.4 (relative smoothness [7]). *We say that a proper, lsc function* $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ *is* smooth relative to *a Legendre kernel* $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ *if* $\operatorname{dom} f \supseteq \operatorname{dom} h$, *and there exists* $L_f \ge 0$ *such that* $L_f h \pm f$ *are convex functions on* $\operatorname{int}\operatorname{dom} h$. *We will simply say that* f *is* L_f*-smooth relative to* h *to make the modulus* L_f *explicit.*

Fact 2.5. Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be L_f-smooth relative to a Legendre kernel $h : \mathbb{R}^n \to \overline{\mathbb{R}}$. Then, $f \in \mathscr{C}^1(\operatorname{int}\operatorname{dom} h)$ and the following hold:

(i) $\big|f(y) - f(x) - \langle \nabla f(x), y - x\rangle\big| \le L_f\,\mathrm{D}_h(y, x)$ for all $x, y \in \operatorname{int}\operatorname{dom} h$.

(ii) $-L_f\nabla^2 h \preceq \nabla^2 f \preceq L_f\nabla^2 h$ on $\operatorname{int}\operatorname{dom} h$, provided that $f, h \in \mathscr{C}^2(\operatorname{int}\operatorname{dom} h)$.

(iii) If ∇h is Lipschitz continuous with modulus $\ell_{h,\mathcal{U}}$ on a convex set $\mathcal{U}$, then so is ∇f with modulus $\ell_{f,\mathcal{U}} = L_f\ell_{h,\mathcal{U}}$ [1, Prop. 2.5(ii)].

Relative to a Legendre kernel $h : \mathbb{R}^n \to \overline{\mathbb{R}}$, the *Bregman proximal mapping* of ψ is the set-valued map $\operatorname{prox}_\psi^h : \operatorname{int}\operatorname{dom} h \rightrightarrows \mathbb{R}^n$ given by

$$\operatorname{prox}_\psi^h(x) \coloneqq \operatorname*{arg\,min}_{z\in\mathbb{R}^n} \{\psi(z) + \mathrm{D}_h(z, x)\}, \tag{2.3}$$

and the corresponding Bregman–Moreau envelope is $\psi^h : \mathbb{R}^n \to [-\infty, \infty]$ defined as

$$\psi^h(x) \coloneqq \inf_{z\in\mathbb{R}^n} \{\psi(z) + \mathrm{D}_h(z, x)\}. \tag{2.4}$$

Fact 2.6 (regularity properties of $\operatorname{prox}_\psi^h$ and ψ^h [35]). The following hold for a Legendre kernel $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ and a proper, lsc, lower bounded function $\psi : \mathbb{R}^n \to \overline{\mathbb{R}}$:

(i) $\operatorname{prox}_\psi^h$ is locally bounded, compact-valued, and osc on $\operatorname{int}\operatorname{dom} h$.

(ii) ψ^h is real-valued and continuous on $\operatorname{int}\operatorname{dom} h$; in fact, it is locally Lipschitz if so is ∇h.

Fact 2.7 (relation between ψ and ψ^h). Let h be a Legendre kernel and $\psi : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper, lsc, and lower bounded on $\operatorname{dom} h$. Then, for every $x \in \operatorname{int}\operatorname{dom} h$, $y \in \operatorname{dom} h$, and $\bar{x} \in \operatorname{prox}_\psi^h(x)$

(i) $\psi^h(x) \overset{(def)}{=} \psi(\bar{x}) + \mathrm{D}_h(\bar{x}, x) \leq \psi(y) + \mathrm{D}_h(y, x)$, and in particular $\psi^h(x) \leq \psi(x)$.

(ii) if ψ is convex, then $\psi^h(x) \leq \psi(y) + \mathrm{D}_h(y, x) - \mathrm{D}_h(y, \bar{x})$ [59, Lem. 3.1].

Moreover, if $\operatorname{range} \operatorname{prox}_\psi^h \subseteq \operatorname{int} \operatorname{dom} h$, then the following also hold [1, Prop.3.3]:

(iii) $\inf_{\overline{\operatorname{dom} h}} \psi \leq \inf_{\operatorname{int} \operatorname{dom} h} \psi = \inf \psi^h$ and $\arg\min \psi^h = \arg\min_{\operatorname{int} \operatorname{dom} h} \psi$.

(iv) $\psi + \delta_{\overline{\operatorname{dom} h}}$ is level bounded iff so is ψ^h.

3. A block-coordinate interpretation. By introducing N copies of x, problem (P) can equivalently be written as

$$\underset{\boldsymbol{x}=(x_1,\ldots,x_N)\in\mathbb{R}^{nN}}{\text{minimize}} \ \Phi(\boldsymbol{x}) = \overbrace{\tfrac{1}{N}\textstyle\sum_{i=1}^N f_i(x_i)}^{F(\boldsymbol{x})} + \overbrace{\tfrac{1}{N}\textstyle\sum_{i=1}^N g(x_i) + \delta_\Delta(\boldsymbol{x})}^{G(\boldsymbol{x})} \quad \text{subject to } \boldsymbol{x} \in \overline{C} \times \cdots \times \overline{C}, \tag{3.1}$$

where $\Delta := \left\{\boldsymbol{x} = (x_1, \ldots, x_N) \in \mathbb{R}^{nN} \mid x_1 = x_2 = \cdots = x_N\right\}$ is the consensus set. The equivalence between (3.1) and the original problem (P) is formally established in Lemma A.1. Note that Assumption I.A1 implies that F as in (3.1) is smooth with respect to the Legendre kernel

$$H : \mathbb{R}^{nN} \to \overline{\mathbb{R}} \quad \text{defined as} \quad H(\boldsymbol{x}) = \textstyle\sum_{i=1}^N h_i(x_i), \tag{3.2}$$

making Bregman forward-backward iterations

$$\boldsymbol{x}^+ \in \arg\min\{\langle \nabla F(\boldsymbol{x}), \cdot \rangle + G(\cdot) + \tfrac{1}{\gamma}\mathrm{D}_H(\cdot, \boldsymbol{x})\}$$

for some stepsize $\gamma > 0$ a suitable option to address problem (3.1). In fact, it can be easily verified that $L_F = \frac{1}{N}\max_{i=1\ldots N} L_{f_i}$ is a smoothness modulus of F relative to H, indicating that fixed point iterations $\boldsymbol{x} \leftarrow \boldsymbol{x}^+$ under Assumption I converge (in some sense to be made precise) to a stationary point of the problem whenever $\gamma \in (0, {}^1/_{L_F})$. Notice that a higher degree of flexibility can be granted by considering an N-tuple of individual stepsizes $\Gamma = (\gamma_1, \ldots, \gamma_N)$, giving rise to the forward-backward operator $\mathrm{T}_\Gamma^{F,G} : \mathbb{R}^{nN} \rightrightarrows \mathbb{R}^{nN}$ in the Bregman metric $(\boldsymbol{z}, \boldsymbol{x}) \mapsto \sum_{i=1}^N \frac{1}{\gamma_i}\mathrm{D}_{h_i}(z_i, x_i)$, namely,

$$\mathrm{T}_\Gamma^{F,G}(\boldsymbol{x}) := \underset{\boldsymbol{z}\in\mathbb{R}^{nN}}{\arg\min}\left\{F(\boldsymbol{x}) + \langle \nabla F(\boldsymbol{x}), \boldsymbol{z} - \boldsymbol{x}\rangle + G(\boldsymbol{z}) + \textstyle\sum_{i=1}^N \frac{1}{\gamma_i}\mathrm{D}_{h_i}(z_i, x_i)\right\}. \tag{3.3}$$

This intuition is validated in the next result, which asserts that whenever the stepsizes γ_i are selected as in Algorithm 1 the operator $\mathrm{T}_\Gamma^{F,G}$ coincides with a proximal mapping on a suitable Legendre kernel function $\hat{H}$. This observation leads to a much simpler analysis of Algorithm 1, which will be shown to be a block-coordinate variant of a Bregman proximal point method.

LEMMA 3.1. *Suppose that Assumption* I.A1 *holds and let* $\gamma_i \in (0, {}^N/_{L_{f_i}})$ *be selected as in Algorithm* 1. *Then,* $\hat{h}_i := \frac{1}{\gamma_i}h_i - \frac{1}{N}f_i$ *(with the convention* $\infty - \infty = \infty$*) is a Legendre kernel with* $\operatorname{dom}\hat{h}_i = \operatorname{dom} h_i$, $i \in [N]$, *and thus so is the function*

$$\hat{H} : \mathbb{R}^{nN} \to \overline{\mathbb{R}} \quad \textit{defined as} \quad \hat{H}(\boldsymbol{x}) = \textstyle\sum_{i=1}^N \hat{h}_i(x_i). \tag{3.4}$$

Moreover, for any $(\boldsymbol{z}, \boldsymbol{x}) \in \mathbb{R}^{nN} \times \mathbb{R}^{nN}$ *it holds that*

$$\Phi(\boldsymbol{z}) + \mathrm{D}_{\hat{H}}(\boldsymbol{z}, \boldsymbol{x}) = F(\boldsymbol{x}) + \langle \nabla F(\boldsymbol{x}), \boldsymbol{z} - \boldsymbol{x}\rangle + G(\boldsymbol{z}) + \textstyle\sum_{i=1}^N \frac{1}{\gamma_i}\mathrm{D}_{h_i}(z_i, x_i), \tag{3.5}$$

and in particular the forward-backward operator (3.3) *satisfies*

$$\mathrm{T}_\Gamma^{F,G}(\boldsymbol{x}) = \operatorname{prox}_\Phi^{\hat{H}}(\boldsymbol{x}). \tag{3.6}$$

When Assumption I *is satisfied, then the following also hold:*

(i) $\mathrm{D}_{\hat{H}}(\boldsymbol{z},\boldsymbol{x}) \geq \sum_{i=1}^N (\frac{1}{\gamma_i} - \frac{L_{f_i}}{N})\,\mathrm{D}_{h_i}(z_i,x_i)$.

(ii) $\operatorname{prox}_\Phi^{\hat{H}}(\boldsymbol{x}) = \Big\{(z,\cdots,z) \mid z \in T(\sum_{i=1}^N \nabla\hat{h}_i(x_i))\Big\}$ *with* T *as in* (2.1), *is a nonempty and compact subset of* $C\times\cdots\times C$ *for any* $\boldsymbol{x} \in \operatorname{int}\operatorname{dom} h_1 \times\cdots\times \operatorname{int}\operatorname{dom} h_N$.

(iii) *If* $\boldsymbol{z} \in \operatorname{prox}_\Phi^{\hat{H}}(\boldsymbol{x})$, *then* $\nabla\hat{H}(\boldsymbol{x}) - \nabla\hat{H}(\boldsymbol{z}) \in \hat{\partial}\Phi(\boldsymbol{z})$; *the converse also holds when* Φ *is convex.*

(iv) *If* ∇h_i *is* $\ell_{h_i,\mathcal{U}_i}$*-Lipschitz on a convex set* $\mathcal{U}_i \subseteq \operatorname{int}\operatorname{dom} h_i$, *then* $\nabla\hat{h}_i$ *is* $\ell_{\hat{h}_i,\mathcal{U}_i}$*-Lipschitz on* $\mathcal{U}_i$ *with* $\ell_{\hat{h}_i,\mathcal{U}_i} \leq \big(\frac{1}{\gamma_i} + \frac{L_{f_i}}{N}\big)\ell_{h_i,\mathcal{U}_i}$. *If, in addition,* $f_i - \frac{\mu_{f_i,\mathcal{U}_i}}{2}\|\cdot\|^2$ *is convex on* $\mathcal{U}_i$ *for some* $\mu_{f_i,\mathcal{U}_i} \in \mathbb{R}$, *then* $\ell_{\hat{h}_i} \leq \frac{\ell_{h_i,\mathcal{U}_i}}{\gamma_i} - \frac{\mu_{f_i,\mathcal{U}_i}}{N}$.

(v) *If* h_i *is* $\mu_{h_i,\mathcal{U}_i}$*-strongly convex on a convex set* $\mathcal{U}_i \subseteq \operatorname{dom} h_i$, *then* $\hat{h}_i$ *is* $\mu_{\hat{h}_i,\mathcal{U}_i}$*-strongly convex on* $\mathcal{U}_i$ *with* $\mu_{\hat{h}_i,\mathcal{U}_i} \geq \big(\frac{1}{\gamma_i} - \frac{L_{f_i}}{N}\big)\mu_{h_i,\mathcal{U}_i}$.

Proof. The claims on $\hat{h}_i$ are shown in [1, Thm. 4.1], and (3.5) and (3.6) then easily follow.

3.1(i) This is an immediate consenquence of Fact 2.5(i).

3.1(ii) Let $\boldsymbol{x}$ be as in the statement, and observe that $\boldsymbol{x} \in \operatorname{int}\operatorname{dom}\hat{H}$; nonemptyness and compactness of $\operatorname{prox}_\Phi^{\hat{H}}$ then follow from Fact 2.6(i). Let now $\boldsymbol{u} \in \operatorname{prox}_\Phi^{\hat{H}}(\boldsymbol{x})$ be fixed, and note that the consensus constraint encoded in Φ ensures that $u_i = u_j$ for all $i,j \in [N]$. Thus,

$$\begin{aligned}
u_i &= \operatorname*{arg\,min}_{w\in\mathbb{R}^n} \Big\{\Phi(w,\dots,w) + \hat{H}(w,\dots,w) - \langle\nabla\hat{H}(\boldsymbol{x}),(w,\dots,w)\rangle\Big\}\\
&= \operatorname*{arg\,min}_{w\in\mathbb{R}^n} \Big\{\tfrac{1}{N}\textstyle\sum_{i=1}^N f_i(w) + g(w) + \sum_{i=1}^N\big(\hat{h}_i(w) - \langle\nabla\hat{h}_i(x_i),w\rangle\big)\Big\}\\
&= \operatorname*{arg\,min}_{w\in\mathbb{R}^n} \Big\{g(w) + \textstyle\sum_{i=1}^N \frac{1}{\gamma_i}h_i(w) - \langle\sum_{i=1}^N\nabla\hat{h}_i(x_i),w\rangle\Big\} \overset{(2.1)}{=} T\big(\textstyle\sum_{i=1}^N\nabla\hat{h}_i(x_i)\big) \subseteq C,
\end{aligned}$$

where the inclusion follows from Assumption I.A4.

3.1(iii) Observe first that necessarily $\boldsymbol{x} \in \operatorname{int}\operatorname{dom} h_i \times\cdots\times\operatorname{int}\operatorname{dom} h_N$, for otherwise no such $\boldsymbol{z}$ exists. Moreover, from assertion 3.1(iii) it follows that $\boldsymbol{z}$ also belongs to such an open set, onto which $\hat{H}$ is continuously differentiable. The claim then follows from the necessary condition for optimality of $\boldsymbol{z}$ in the minimization problem (2.4) — which is also sufficient when Φ is convex, for so is $\Phi + \mathrm{D}_{\hat{H}}(\,\cdot\,,\boldsymbol{x})$ in this case — having

$$0 \in \hat{\partial}[\Phi + \mathrm{D}_{\hat{H}}(\,\cdot\,,\boldsymbol{x})](\boldsymbol{z}) = \hat{\partial}[\Phi + \hat{H} - \langle\nabla\hat{H}(\boldsymbol{x}),\,\cdot\,\rangle](\boldsymbol{z}) = \hat{\partial}\Phi(\boldsymbol{z}) + \nabla\hat{H}(\boldsymbol{z}) - \nabla\hat{H}(\boldsymbol{x}).$$

The last equality follows from [53, Ex. 8.8(c)], owing to the smoothness of $\hat{H}$ at $\boldsymbol{z}$.

3.1(iv) and 3.1(v) Observe that

$$\tag{3.7}
\tfrac{N-\gamma_iL_{f_i}}{N\gamma_i}h_i \preceq \tfrac{N-\gamma_iL_{f_i}}{N\gamma_i}h_i + \overbrace{\tfrac{1}{N}(L_{f_i}h_i - f_i)}^{\text{convex}} = \hat{h}_i = \tfrac{N+\gamma_iL_{f_i}}{N\gamma_i}h_i - \overbrace{\tfrac{1}{N}(L_{f_i}h_i + f_i)}^{\text{convex}} \preceq \tfrac{N+\gamma_iL_{f_i}}{N\gamma_i}h_i,$$

where for notational convenience we used the partial ordering "$\preceq$," defined as $\alpha \preceq \beta$ iff $\beta - \alpha$ is convex. The claimed moduli $\ell_{\hat{h}_i,\mathcal{U}_i} \leq \big(\frac{1}{\gamma_i} + \frac{L_{f_i}}{N}\big)\ell_{h_i,\mathcal{U}_i}$ and $\mu_{\hat{h}_i,\mathcal{U}_i} \geq \big(\frac{1}{\gamma_i} - \frac{L_{f_i}}{N}\big)\mu_{h_i,\mathcal{U}_i}$ are thus readily inferred. In case f_i is $\mu_{f_i,\mathcal{U}_i}$-strongly convex on $\mathcal{U}_i$, we may write

Algorithm 3 Block-coordinate proximal point formulation of Algorithm 1.

REQUIRE Legendre kernels h_i such that f_i is L_{f_i}-smooth relative to h_i
stepsizes $\gamma_i \in (0, {}^{N}/_{L_{f_i}})$
initial point $x^{\text{init}} \in \cap_{i=1}^N \operatorname{int}\operatorname{dom} h_i = C$

DENOTE $\boldsymbol{x}^0 = (x^{\text{init}}, \dots, x^{\text{init}})$, $\hat{h}_j := \frac{1}{\gamma_i} h_j - \frac{1}{N} f_j$, $\hat{H}(\boldsymbol{x}) := \sum_{i=1}^N \hat{h}_i(x_i)$

REPEAT for $k = 0, 1, \dots$ until convergence

1: $\boldsymbol{u}^k \in \arg\min_{\boldsymbol{w}\in\mathbb{R}^{nN}} \left\{\Phi(\boldsymbol{w}) + \hat{H}(\boldsymbol{w}) - \langle \nabla\hat{H}(\boldsymbol{x}^k), \boldsymbol{w}\rangle\right\} = \arg\min_{\boldsymbol{w}\in\mathbb{R}^{nN}} \{\Phi(\boldsymbol{w}) + \mathrm{D}_{\hat{H}}(\boldsymbol{w}, \boldsymbol{x}^k)\}$

2: Select a subset of indices $\mathcal{J}^{k+1} \subseteq [N]$

3: $x^{k+1}_{\mathcal{J}^{k+1}} = u^k_{\mathcal{J}^{k+1}}$ and $x^{k+1}_{[N]\setminus\mathcal{J}^{k+1}} = x^k_{[N]\setminus\mathcal{J}^{k+1}}$

$$\hat{h}_i = \tfrac{h_i}{\gamma_i} - \tfrac{\mu_{f_i,\mathcal{U}_i}}{2N}\|\cdot\|^2 - \tfrac{1}{N}\overbrace{(f_i - \tfrac{\mu_{f_i,\mathcal{U}_i}}{2}\|\cdot\|^2)}^{\text{convex}} \preceq \tfrac{h_i}{\gamma_i} - \tfrac{\mu_{f_i,\mathcal{U}_i}}{2N}\|\cdot\|^2$$

to obtain the tighter bound $\ell_{\hat{h}_i,\mathcal{U}_i} \le \frac{\ell_{h_i,\mathcal{U}_i}}{\gamma_i} - \frac{\mu_{f_i,\mathcal{U}_i}}{N}$. ∎

3.1. Block-coordinate proximal point reformulation of Algorithm 1. Algorithm 3 presents a block coordinate (BC) proximal point algorithm with the distance generating function $\hat{H}$. Note that in a departure from most of the existing literature on BC proximal methods that consider *separable* nonsmooth terms (see, e.g., [63, 47, 12, 19, 31]), here the nonsmooth function G in (3.1) is nonseparable. It is shown in the next lemma that this conceptual algorithm is equivalent to the Bregman Finito/MISO Algorithm 1.

LEMMA 3.2 (equivalence of Algorithms 1 and 3). *As long as the same initialization parameters are chosen in the two algorithms, to any sequence* $(\boldsymbol{s}^k, \tilde{s}^k, z^k, \mathcal{I}^{k+1})_{k\in\mathbb{N}}$ *generated by Algorithm* 1 *there corresponds a sequence* $(\boldsymbol{x}^k, \boldsymbol{u}^k, \mathcal{J}^{k+1})_{k\in\mathbb{N}}$ *generated by Algorithm* 3 *(and vice versa) satisfying the following identities for all* $k \in \mathbb{N}$ *and* $i \in [N]$*:*

(i) $\mathcal{I}^{k+1} = \mathcal{J}^{k+1}$.

(ii) $(z^k, \dots, z^k) = \boldsymbol{u}^k$.

(iii) $s_i^k = \frac{1}{\gamma_i}\nabla h_i(x_i^k) - \frac{1}{N}\nabla f_i(x_i^k)$ *(or, equivalently,* $x_i^k = \nabla\hat{h}_i^*(s_i^k)$*).*

(iv) $\tilde{s}^k = \sum_{i=1}^N \nabla\hat{h}_i(x_i^k)$.

(v) $\varphi(z^k) = \Phi(\boldsymbol{u}^k) = \Phi^{\hat{H}}(\boldsymbol{x}^k) - \mathrm{D}_{\hat{H}}(\boldsymbol{u}^k, \boldsymbol{x}^k)$.

(vi) $\Phi^{\hat{H}}(\boldsymbol{x}^k) = \mathcal{L}(z^k, \boldsymbol{s}^k)$*, where* $\mathcal{L}$ *is as in* (1.1).

Proof. Let the index sets $\mathcal{I}^{k+1}$ and $\mathcal{J}^{k+1}$ be chosen identically, $k \in \mathbb{N}$. It follows from Lemma 3.1(ii) that $u_i^k = u_j^k$ for all $k \in \mathbb{N}$ and $i, j \in [N]$ with

$$u_i^k = \arg\min_{w\in\mathbb{R}^n} \left\{ g(w) + \sum_{i=1}^N \tfrac{1}{\gamma_i} h_i(w) - \langle \overbrace{\textstyle\sum_{i=1}^N \nabla\hat{h}_i(x_i^k)}^{:=v^k}, w\rangle \right\}. \tag{3.8}$$

We now proceed by induction to show assertions 3.2(ii), 3.2(iii), and 3.2(iv). Note that the latter amounts to showing that v^k as defined in (3.8) coincides with $\tilde{s}^k$; by comparing (3.8) and the expression of z^k in step 1, the claimed correspondence of $\boldsymbol{u}^k$ and z^k as in assertion 3.2(ii) is then also obtained and, in turn, so is the identity in 3.2(v).

For $k = 0$ assertions 3.2(iii) and 3.2(iv) hold because of the initialization of $\tilde{s}^0$ in Algorithm 1 and of $\boldsymbol{x}^0$ in Algorithm 3; in turn, as motivated above, the base case for assertion 3.2(ii) also holds. Suppose now that the three assertions hold for some $k \geq 0$; then,

$$\begin{aligned} v^{k+1} = \textstyle\sum_{i=1}^N \nabla \hat{h}_i(x_i^{k+1}) &= \sum_{i\in\mathcal{I}^{k+1}} \nabla \hat{h}_i(u_i^k) + v^k - \sum_{i\in\mathcal{I}^{k+1}} \nabla \hat{h}_i(x_i^k) \\ \text{(induction)} &= \sum_{i\in\mathcal{I}^{k+1}} \nabla \hat{h}_i(z^k) + \tilde{s}^k - \sum_{i\in\mathcal{I}^{k+1}} s_i^k \\ &= \sum_{i\in\mathcal{I}^{k+1}} s_i^{k+1} + \tilde{s}^k - \sum_{i\in\mathcal{I}^{k+1}} s_i^k = \tilde{s}^{k+1}, \end{aligned}$$

where the last two equalities are due to steps 2 and 3. Therefore, $v^{k+1} = \tilde{s}^{k+1}$ and thus $\boldsymbol{u}^{k+1} = (z^{k+1}, \dots, z^{k+1})$. It remains to show that $s_i^{k+1} = \frac{1}{\gamma_i}\nabla h_i(x_i^{k+1}) - \frac{1}{N}\nabla f_i(x_i^{k+1})$. For $i \in \mathcal{I}^{k+1}$ this holds because of the update rule at step 3 and the fact that $x_i^{k+1} = u_i^k = z^k$ owing to step 3. For $i \notin \mathcal{I}^{k+1}$ this holds because $(x_i^{k+1}, s_i^{k+1}) = (x_i^k, s_i^k)$. Finally,

$$\Phi^{\hat{H}}(\boldsymbol{x}^k) \overset{(def)}{=} \Phi(\boldsymbol{u}^k) + \mathrm{D}_{\hat{H}}(\boldsymbol{u}^k, \boldsymbol{x}^k) \overset{3.2(\mathrm{v})}{=} \varphi(z^k) + \sum_{i=1}^N \mathrm{D}_{\hat{h}_i}(z^k, x_i^k) \overset{3.2(\mathrm{iii})}{=} \varphi(z^k) + \sum_{i=1}^N \mathrm{D}_{\hat{h}_i}(z^k, \nabla \hat{h}_i^*(s_i^k)),$$

and the last term is $\mathscr{L}(z^k, \boldsymbol{s}^k)$ (cf. Facts 2.3(i) and 2.3(iii)), yielding assertion 3.2(vi). ∎

4. Convergence analysis.

The BC interpretation of Algorithm 1 presented in section 3 plays a crucial role in the proposed methodology, and leads to a remarkably simple convergence analysis. In fact, many key facts can be established without confining the discussion to a particular sampling strategy. These preliminary results are presented in the next subsection and will be extensively referred to in the subsequent subsections that are instead devoted to a specific sampling strategy.

4.1. General sampling results.

Unlike classical analyses of BC proximal methods that employ the cost as a Lyapunov function (see, e.g., [10, sect. 11]), here, the nonseparability of G precludes this possibility. To address this challenge, we instead employ the Bregman–Moreau envelope equipped with the distance generating function $\hat{H}$ (see (3.4)). Before showing its Lyapunov-type behavior for Algorithm 3, we list some of its properties and its relation with the original problem. The proof is a simple consequence of Facts 2.6(ii) and 2.7 and the fact that $\hat{H}$ is a Legendre kernel with $\operatorname{dom}\hat{H} = \operatorname{dom} h_1 \times \cdots \times \operatorname{dom} h_N$ (cf. Lemma 3.1).

LEMMA 4.1 (connections between $\varphi + \delta_{\overline{C}}$ and $\Phi^{\hat{H}}$). *If Assumption* I *holds, then*

(i) $\Phi^{\hat{H}}$ *is continuous on* $\operatorname{dom}\Phi^{\hat{H}} = \operatorname{int}\operatorname{dom} h_1 \times \cdots \times \operatorname{int}\operatorname{dom} h_N$, *in fact, locally Lipschitz if* ∇h_i *is on* $\operatorname{int}\operatorname{dom} h_i$, $i \in [N]$.

(ii) $\min_{\overline{C}} \varphi \leq \inf_C \varphi = \inf \Phi^{\hat{H}}$ *and* $\arg\min \Phi^{\hat{H}} = \{(x^\star, \dots, x^\star) \mid x^\star \in \arg\min_C \varphi\}$.

(iii) $\Phi^{\hat{H}}$ *is level bounded iff so is* $\varphi + \delta_{\overline{C}}$.

LEMMA 4.2 (sure descent). *Suppose that Assumption* I *holds, and consider the iterates generated by Algorithm* 3. *Then,* $\boldsymbol{u}^k = (u^k, \dots, u^k)$ *for some* $u^k \in C$ *and* $\boldsymbol{x}^k \in C \times \cdots \times C \subseteq \operatorname{int}\operatorname{dom}\hat{H}$ *for every* $k \in \mathbb{N}$, *and the algorithm is thus well defined. Moreover, the following hold:*

(i) $\Phi^{\hat{H}}(\boldsymbol{x}^{k+1}) \leq \Phi^{\hat{H}}(\boldsymbol{x}^k) - \mathrm{D}_{\hat{H}}(\boldsymbol{x}^{k+1}, \boldsymbol{x}^k) = \Phi^{\hat{H}}(\boldsymbol{x}^k) - \sum_{i\in\mathcal{I}^{k+1}} \mathrm{D}_{\hat{h}_i}(u^k, x_i^k)$ *for every* $k \in \mathbb{N}$; *when* Φ *is convex (i.e., when so is* φ*), then the inequality can be strengthened to* $\Phi^{\hat{H}}(\boldsymbol{x}^{k+1}) \leq \Phi^{\hat{H}}(\boldsymbol{x}^k) - \mathrm{D}_{\hat{H}}(\boldsymbol{x}^{k+1}, \boldsymbol{x}^k) - \mathrm{D}_{\hat{H}}(\boldsymbol{u}^k, \boldsymbol{u}^{k+1})$.

(ii) $(\Phi^{\hat{H}}(\boldsymbol{x}^k))_{k\in\mathbb{N}}$ *monotonically decreases to a finite value* $\varphi_\star \geq \inf_C \varphi \geq \min_{\overline{C}} \varphi$.

(iii) *The sequence* $(\mathrm{D}_{\hat{H}}(\boldsymbol{x}^{k+1}, \boldsymbol{x}^k))_{k\in\mathbb{N}}$ *has finite sum (and in particular vanishes); the same also holds for* $(\mathrm{D}_{\hat{H}}(\boldsymbol{u}^k, \boldsymbol{u}^{k+1}))_{k\in\mathbb{N}}$ *when* Φ *is convex (i.e., when so is* φ*).*

(iv) *If* $\varphi + \delta_{\overline{C}}$ *is level bounded, then* $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ *and* $(\boldsymbol{u}^k)_{k\in\mathbb{N}}$ *are bounded.*

(v) *If* $\operatorname{dom} h_i$ *is closed, a subsequence* $(x_i^k)_{k\in K}$ *converges to a point* $x^\star$ *iff so does* $(x_i^{k+1})_{k\in K}$.

(vi) *If* $C = \mathbb{R}^n$*, then* $\Phi^{\hat{H}}$ *is constantly equal to* $\varphi_\star$ *as above on the limit set of* $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$.

Proof. It follows from Lemma 3.1(ii) that $u^k \in C$ holds for every $k \in \mathbb{N}$. Notice that for every $i \in [N]$ and $k \in \mathbb{N}$, either $x_i^k = x^{\mathrm{init}} \in C$ (by initialization), or there exists $k_i \leq k$ such that $x_i^k = z^{k_i} \in C$. It readily follows that $\boldsymbol{x}^k \in C \times \cdots \times C \subseteq \operatorname{int}\operatorname{dom} H = \operatorname{int}\operatorname{dom} \hat{H}$, hence, that $\operatorname{prox}_\Phi^{\hat{H}}(\boldsymbol{x}^k) \neq \emptyset$ for all $k \in \mathbb{N}$ by Lemma 3.1(ii), whence comes the well definedness of the algorithm. We now show the numbered claims.

4.2(i) It follows from Facts 2.7(i) and 2.7(ii) that $\Phi^{\hat{H}}(\boldsymbol{x}^{k+1}) \leq \Phi(\boldsymbol{u}^k) + \mathrm{D}_{\hat{H}}(\boldsymbol{u}^k, \boldsymbol{x}^{k+1}) - c_k$, where $c_k \geq 0$ can be taken as $c_k = \mathrm{D}_{\hat{H}}(\boldsymbol{u}^k, \boldsymbol{u}^{k+1})$ when Φ is convex. Therefore,

$$\begin{aligned}\Phi^{\hat{H}}(\boldsymbol{x}^{k+1}) &\leq \Phi(\boldsymbol{u}^k) + \mathrm{D}_{\hat{H}}(\boldsymbol{u}^k, \boldsymbol{x}^{k+1}) - c_k = \Phi^{\hat{H}}(\boldsymbol{x}^k) - \mathrm{D}_{\hat{H}}(\boldsymbol{u}^k, \boldsymbol{x}^k) + \mathrm{D}_{\hat{H}}(\boldsymbol{u}^k, \boldsymbol{x}^{k+1}) - c_k \\ &\overset{2.3(\mathrm{ii})}{=} \Phi^{\hat{H}}(\boldsymbol{x}^k) - \mathrm{D}_{\hat{H}}(\boldsymbol{x}^{k+1}, \boldsymbol{x}^k) - \langle \boldsymbol{u}^k - \boldsymbol{x}^{k+1}, \nabla\hat{H}(\boldsymbol{x}^{k+1}) - \nabla\hat{H}(\boldsymbol{x}^k)\rangle - c_k.\end{aligned}$$

The claim follows by noting that the inner product is zero:

$$\langle \boldsymbol{u}^k - \boldsymbol{x}^{k+1}, \nabla\hat{H}(\boldsymbol{x}^{k+1}) - \nabla\hat{H}(\boldsymbol{x}^k)\rangle = \sum_{j\in[N]} \langle u^k - \underbrace{x_j^{k+1}}_{=u^k \text{ for } j\in\mathcal{J}^{k+1}}, \nabla\hat{h}_j(\overbrace{x_j^{k+1}}^{=x_j^k \text{ for } j\notin\mathcal{J}^{k+1}}) - \nabla\hat{h}_j(x_j^k)\rangle = 0.$$

4.2(ii) Monotonic decrease of $(\Phi^{\hat{H}}(\boldsymbol{x}^k))_{k\in\mathbb{N}}$ follows from assertion 4.2(i). This ensures that the sequence converges to some value $\varphi_\star$, bounded below by $\min_{\overline{C}} \varphi$ in light of Lemma 4.1(ii).

4.2(iii) It follows from assertion 4.2(i) that

$$\textstyle\sum_{k\in\mathbb{N}} \mathrm{D}_{\hat{H}}(\boldsymbol{x}^{k+1}, \boldsymbol{x}^k) \leq \Phi^{\hat{H}}(\boldsymbol{x}^0) - \inf \Phi^{\hat{H}} \leq \Phi^{\hat{H}}(\boldsymbol{x}^0) - \inf_{\overline{C}} \varphi < \infty$$

owing to Lemma 4.1(ii) and Assumption I.A3. When φ is convex, the tighter bound in assertion 4.2 yields the similar claim for $(\mathrm{D}_{\hat{H}}(\boldsymbol{u}^k, \boldsymbol{u}^{k+1}))_{k\in\mathbb{N}}$.

4.2(iv) It follows from assertion 4.2(ii) that the entire sequence $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ is contained in the sublevel set $\{\boldsymbol{w} \mid \Phi^{\hat{H}}(\boldsymbol{w}) \leq \Phi^{\hat{H}}(\boldsymbol{x}^0)\}$, which is bounded provided that $\varphi + \delta_{\overline{C}}$ is level bounded as shown in Lemma 4.1(iii). In turn, boundedness of $(\boldsymbol{u}^k)_{k\in\mathbb{N}}$ then follows from local boundedness of $\mathrm{T}_\Gamma^{F,G} = \operatorname{prox}_\Phi^{\hat{H}}$; cf. (3.6) and Fact 2.6(i).

4.2(v) Follows from Fact 2.3(v), since $x_i^k \in \operatorname{int}\operatorname{dom} h_i = \operatorname{int}\operatorname{dom} \hat{h}_i$ for every k (with equality owing to Lemma 3.1), and $\mathrm{D}_{\hat{h}_i}(x_i^{k+1}, x_i^k) \to 0$ by assertion 4.2(iii).

4.2(vi) Follows from assertion 4.2(ii) and the continuity of $\Phi^{\hat{H}}$; see Fact 2.6(ii). ∎

In conclusion of this subsection we provide an overview of the ingredients that are needed to show that the limit points of the sequence $(z^k)_{k\in\mathbb{N}}$ generated by Algorithm 1 are stationary for problem (P). As will be shown in Lemma 4.4, these amount

to the vanishing of the *residual* $\mathrm{D}_{\hat{H}}(\boldsymbol{u}^k, \boldsymbol{x}^k)$ together with some assumptions on the distance-generating functions h_i. For the iterates of Algorithm 1, this translates to $\mathrm{D}_{\hat{h}_i^*}(s_i^k, \nabla\hat{h}_i(z^k)) \to 0$ for all indices $i \in [N]$, indicating that all vectors s_i^{k+1} in the table should be good estimates of $\nabla\hat{h}_i(z^{k+1}) = \frac{1}{\gamma_i}\nabla h_i(z^{k+1}) - \frac{1}{N}\nabla f_i(z^{k+1})$, as opposed to $\frac{1}{\gamma_i}\nabla h_i(z^k) - \frac{1}{N}\nabla f_i(z^k)$ and for the indices in $\mathcal{I}^{k+1}$ only (cf. step 3). As a result, we may view this property as jointly having $z^k - z^{k+1}$ vanish, desirable if any convergence of $(z^k)_{k\in\mathbb{N}}$ is expected, and the fact that a consensus is eventually reached among the sampled blocks.

In line with any result in the literature we are aware of, a complete convergence analysis for nonconvex problems will ultimately require $C = \mathbb{R}^n$. For convex problems, that is, when the cost function φ is convex without any among f_i and g being necessarily so, the following requirement will instead suffice to our purposes in the randomized sampling setting of $(\mathcal{S}_1)$.

Assumption II (requirements on the distance-generating functions). For $i \in [N]$, $\operatorname{dom} h_i$ is closed, and whenever $\operatorname{int}\operatorname{dom} h_i \ni z^k \to z \in \operatorname{bdry}\operatorname{dom} h_i$ it holds that $\mathrm{D}_{h_i}(z, z^k) \to 0$.

Remark 4.3.

(i) Assumption II is vacuously satisfied when $\operatorname{dom} h_i = \mathbb{R}^n$, having $\operatorname{bdry}\mathbb{R}^n = \emptyset$.

(ii) While Assumption II always holds on $\mathbb{R}$, it may fail in higher dimensions [8, Ex. 7.32].

(iii) For any $i \in [N]$, function h_i complies with Assumption II iff so does $\hat{h}_i$, owing to the inequalities $\frac{N-\gamma_i L_{f_i}}{N\gamma_i}\mathrm{D}_{h_i} \le \mathrm{D}_{\hat{h}_i} \le \frac{N+\gamma_i L_{f_i}}{N\gamma_i}\mathrm{D}_{h_i}$ (cf. (3.7)). ∎

LEMMA 4.4 (subsequential convergence recipe). *Suppose that Assumption* I *holds, and consider the iterates generated by Algorithm* 1. *Let* $x_i^k = \nabla\hat{h}_i^*(s_i^k)$ *and* $z^k = u^k$ *be the corresponding iterates generated by Algorithm* 3 *as in Lemma* 3.2, *and suppose that*

A1. $\mathrm{D}_{\hat{H}}(\boldsymbol{u}^k, \boldsymbol{x}^k) \to 0$ *(or equivalently,* $\mathrm{D}_{\hat{h}_i^*}(s_i^k, \nabla\hat{h}_i(z^k)) \to 0$, $i \in [N]$*).*

Then, letting $\varphi_\star$ *be as in Lemma* 4.2*(ii), the following hold:*

(i) $\varphi(z^k) = \Phi(\boldsymbol{u}^k) \to \varphi_\star$ *as* $k \to \infty$.

(ii) *If* $\operatorname{dom} h_i$ *is closed,* $i \in [N]$, *then having* (a) $(z^k)_{k\in K} \to z$, (b) $(x_i^k)_{k\in K} \to z$ $\exists i \in [N]$, *and* (c) $(z^{k+1}, x_i^{k+1})_{k\in K} \to (z, z)$ $\forall i \in [N]$, *are all equivalent conditions. In particular, if* $(z^k)_{k\in\mathbb{N}}$ *is bounded (e.g., when* $\varphi + \delta_{\overline{C}}$ *is level bounded), then* $\|z^{k+1} - z^k\| \to 0$ *holds, and the set of its limit points, be it* ω, *is thus nonempty, compact, and connected.*

(iii) *Under Assumption* II, $\varphi \equiv \varphi_\star$ *on* ω *(the set of limit points of* $(z^k)_{k\in\mathbb{N}}$*).*

(iv) *If* $C = \mathbb{R}^n$, *then every* $z^\star \in \omega$ *is stationary for* (P).

Proof. Lemma 4.4.A1 can be written as $\mathrm{D}_{\hat{h}_i}(z^k, \nabla\hat{h}_i^*(s_i^k)) \to 0$, $i \in [N]$. In turn, by the conjugate identity in Fact 2.3(iii), the equivalent expression in terms of s_i^k and z^k is obtained.

4.4(i) As shown in Lemma 4.2(ii), $(\Phi^{\hat{H}}(\boldsymbol{x}^k))_{k\in\mathbb{N}}$ monotonically decreases to $\varphi_\star$. In turn, Lemma 3.2(v) and Assumption 4.4.A1 then imply that $\varphi(z^k) = \Phi(\boldsymbol{u}^k)$ converges to $\varphi_\star$.

4.4(ii) The equivalences owe to Fact 2.3(v) and Lemma 4.2(v) (as $\operatorname{dom}\hat{h}_i = \operatorname{dom} h_i$), and imply $\|z^{k+1} - z^k\| \to 0$ if $(z^k)_{k\in\mathbb{N}}$ is bounded. The claim on ω then follows from [19, Rem. 5].

4.4(iii) Let $z^\star \in \omega$ be fixed, and let $(z^k)_{k\in K}$ be a subsequence converging to $z^\star$. Assertion 4.4(ii) ensures that $(\boldsymbol{x}^k)_{k\in K} \to \boldsymbol{z}^\star := (z^\star, \dots, z^\star)$, hence

$$\varphi_\star \xleftarrow[k\in K]{4.2(\mathrm{ii})} \Phi^{\hat H}(\boldsymbol{x}^k) - \mathrm{D}_{\hat H}(\boldsymbol{z}^\star, \boldsymbol{x}^k) \overset{2.7(\mathrm{i})}{\le} \Phi(\boldsymbol{z}^\star) \overset{\mathrm{lsc}}{\le} \liminf_{k\in K} \Phi(\boldsymbol{u}^k) \overset{4.4(\mathrm{i})}{=} \varphi_\star,$$

where Assumption II is used in the first limit.

4.4(iv) Suppose that $C = \mathbb{R}^n$ and $(z^k)_{k\in K} \to z^\star$ for some infinite $K \subseteq \mathbb{N}$ and $z^\star \in \mathbb{R}^n$, so that, by virtue of assertion 4.4(ii), $(\boldsymbol{x}^k, \boldsymbol{x}^{k+1})_{k\in K} \to (\boldsymbol{z}^\star, \boldsymbol{z}^\star)$. Since $(z^k, \dots, z^k) = \boldsymbol{u}^k \in \operatorname{prox}_\Phi^{\hat H}(\boldsymbol{x}^k)$, the osc of $\operatorname{prox}_\Phi^{\hat H}$ (Fact 2.6(i)) ensures that $\boldsymbol{z}^\star \in \operatorname{prox}_\Phi^{\hat H}(\boldsymbol{z}^\star)$, hence $0 \in \hat\partial\Phi(\boldsymbol{z}^\star)$ owing to Lemma 3.1(iii). By invoking Lemma A.1(iv) we conclude that $z^\star$ is stationary for (P). ∎

4.2. Randomized sampling rule ($\mathcal{S}_1$). The analysis for the randomized case dealt in this section will make use of the following result, known as the Robbins–Siegmund supermartingale theorem, and stated here in simplified form following [13, Prop. 2].

Fact 4.5 (supermartingale convergence theorem [51]). For $k \in \mathbb{N}$, let ξ_k and η_k be random variables, and $\mathcal{F}_k \subseteq \mathcal{F}_{k+1}$ be sets of random variables such that

A1. $\mathcal{F}_k \subseteq \mathcal{F}_{k+1}$;

A2. $0 \le \xi_k, \eta_k$ are functions of the random variables in $\mathcal{F}_k$;

A3. $\mathbb{E}[\, \xi_{k+1} \mid \mathcal{F}_k \,] \le \xi_k - \eta_k$.

Then, a.s., $\sum_{k\in\mathbb{N}} \eta_k < \infty$ and ξ_k converges to a (nonnegative) random variable.

The sets $\mathcal{F}_k$ in the above formulation will represent the information available at iteration k, and the notation $\mathbb{E}_k[\,\cdot\,]$ will be used as a shorthand for $\mathbb{E}[\,\cdot \mid \mathcal{F}_k\,]$.

THEOREM 4.6 (subsequential convergence of Algorithm 1 with randomized rule ($\mathcal{S}_1$)). *Suppose that Assumption* I *holds. Then, denoting* $p_{\min} = \min_i p_i$, *the iterates generated by Algorithm* 1 *with indices selected according to the randomized rule* ($\mathcal{S}_1$) *satisfy*

$$\mathbb{E}_k\big[\mathcal{L}(z^{k+1}, \boldsymbol{s}^{k+1})\big] \le \mathcal{L}(z^k, \boldsymbol{s}^k) - p_{\min} \sum_{i=1}^N \mathrm{D}_{\hat h_i^*}(s_i^k, \nabla \hat h_i(z^k)) \quad \forall k \in \mathbb{N}, \tag{4.1}$$

where $\mathcal{L}$ *is as in* (1.1) *(and satisfies* $\mathcal{L}(z^k, \boldsymbol{s}^k) = \Phi^{\hat H}(\boldsymbol{x}^k)$*; cf. Lemma* 3.2(vi)*). Moreover, letting* ω *denote the set of limit points of* $(z^k)_{k\in\mathbb{N}}$, *the following assertions hold a.s.*

(i) *The sequence* $(\mathrm{D}_{\hat h_i^*}(s_i^k, \nabla \hat h_i(z^k)))_{k\in\mathbb{N}}$ *has finite sum (and in particular vanishes),* $i \in [N]$.

(ii) *The sequence* $(\varphi(z^k))_{k\in\mathbb{N}}$ *converges to the finite value* $\varphi_\star \le \varphi(x^{\mathrm{init}})$ *of Lemma* 4.2(ii).

(iii) *If Assumption* II *is satisfied, then* $\varphi \equiv \varphi_\star$ *on* ω.

(iv) *If* $C = \mathbb{R}^n$, *then* $0 \in \hat\partial\varphi(z^\star)$ *for every* $z^\star \in \omega$.

When φ *is convex (without* g *or any* f_i *necessarily being so) and* $\operatorname{dom} h_i$ *is closed,* $i \in [N]$,

(v) $(\varphi(z^k))_{k\in\mathbb{N}}$ *converges to* $\min_{\overline{C}} \varphi$ *with* $\min_{\ell=0,\dots,k}\{\varphi(z^\ell)\} - \min_{\overline{C}} \varphi \le o(1/k)$.

(vi) *the limit points of* $(z^k)_{k\in\mathbb{N}}$ *all belong to* $\arg\min_{\overline{C}} \varphi$.

(vii) *if either Assumption* II *holds and* $\varphi + \delta_{\overline{C}}$ *is level bounded, or* $C = \mathbb{R}^n$, *then* $(z^k)_{k\in\mathbb{N}}$ *and* $(\nabla \hat h_i^*(s_i^k))_{k\in\mathbb{N}}$, $i \in [N]$, *converge to the same point in* $\arg\min_{\overline{C}} \varphi$.

Proof. By Lemma 3.2, we will consider the simpler setting of Algorithm 3. We have

$$\begin{aligned}\mathbb{E}_k\Big[\mathscr{L}(z^{k+1},\boldsymbol{s}^{k+1})\Big] &\overset{3.2(\mathrm{vi})}{=} \mathbb{E}_k\Big[\Phi^{\hat{H}}(\boldsymbol{x}^{k+1})\Big] \overset{4.2(\mathrm{i})}{\le} \mathbb{E}_k\Big[\Phi^{\hat{H}}(\boldsymbol{x}^{k}) - \sum_{i\in\mathcal{I}^{k+1}} \mathrm{D}_{\hat{h}_i}(u^k,x_i^k)\Big]\\ &= \Phi^{\hat{H}}(\boldsymbol{x}^k) - \sum_{i=1}^N p_i\,\mathrm{D}_{\hat{h}_i}(u^k,x_i^k)\\ &\overset{3.2(\mathrm{vi}),\ 2.3(\mathrm{iii})}{\le} \Phi^{\hat{H}}(\boldsymbol{x}^k) - p_{\min}\,\mathrm{D}_{\hat{H}}(\boldsymbol{u}^k,\boldsymbol{x}^k) = \mathscr{L}(z^k,\boldsymbol{s}^k) - p_{\min}\sum_{i=1}^N \mathrm{D}_{\hat{h}_i^*}(s_i^k,\nabla\hat{h}_i(z^k)),\end{aligned}$$

which is (4.1). We thus infer from Fact 4.5 that $(\mathrm{D}_{\hat{H}}(\boldsymbol{u}^k,\boldsymbol{x}^k))_{k\in\mathbb{N}}$ has a.s. finite sum, and the proof of assertions 4.6(i)–4.6(iv) then follows from Lemma 4.4.

In what follows, suppose that φ is convex and $\operatorname{dom} h_i$ is closed, $i\in[N]$, so that $\overline{C} = \bigcap_{i=1}^N \overline{\operatorname{int}\operatorname{dom} h_i} = \bigcap_{i=1}^N \operatorname{dom} h_i$ [14, Prop. 1.3.8], and in particular $\mathrm{D}_{\hat{h}_i}(y,x) < \infty$ holds for any $(y,x)\in \operatorname{dom} h_i \times \operatorname{int}\operatorname{dom} h_i \supseteq \overline{C}\times C$.

4.6(v) For any $x^\star \in \arg\min_{\overline{C}}\varphi$, so that $\boldsymbol{x}^\star := (x^\star,\dots,x^\star)\in\arg\min_{\overline{C}\times\cdots\times\overline{C}}\Phi$ (cf. Lemma A.1(v)), the three-point identity (Fact 2.3(ii)), convexity of Φ (Lemma A.1(vii)), and the inclusion $\nabla\hat{H}(\boldsymbol{x}^k)-\nabla\hat{H}(\boldsymbol{u}^k)\in\hat{\partial}\Phi(\boldsymbol{u}^k)$ (Lemma 3.1(iii)) give the bound

$$\begin{aligned}\mathrm{D}_{\hat{H}}(\boldsymbol{x}^\star,\boldsymbol{u}^k) &= \mathrm{D}_{\hat{H}}(\boldsymbol{x}^\star,\boldsymbol{x}^k) - \mathrm{D}_{\hat{H}}(\boldsymbol{u}^k,\boldsymbol{x}^k) + \langle\nabla\hat{H}(\boldsymbol{x}^k)-\nabla\hat{H}(\boldsymbol{u}^k),\boldsymbol{x}^\star-\boldsymbol{u}^k\rangle\\ &\le \mathrm{D}_{\hat{H}}(\boldsymbol{x}^\star,\boldsymbol{x}^k) - \mathrm{D}_{\hat{H}}(\boldsymbol{u}^k,\boldsymbol{x}^k) + \Phi(\boldsymbol{x}^\star)-\Phi(\boldsymbol{u}^k).\end{aligned}\tag{4.2}$$

Next,

$$\begin{aligned}&\mathbb{E}_k\left[\sum_{i=1}^N p_i^{-1}\,\mathrm{D}_{\hat{h}_i}(x^\star,x_i^{k+1})\right]\\ &= \sum_{i=1}^N p_i^{-1}\Big(\overbrace{\mathscr{P}_k\big[\mathcal{I}^{k+1}\ni i\big]}^{p_i}\mathrm{D}_{\hat{h}_i}(x^\star,u^k) + \overbrace{\mathscr{P}_k\big[\mathcal{I}^{k+1}\not\ni i\big]}^{1-p_i}\mathrm{D}_{\hat{h}_i}(x^\star,x^k)\Big)\\ &= \mathrm{D}_{\hat{H}}(\boldsymbol{x}^\star,\boldsymbol{u}^k) + \sum_{i=1}^N (1-p_i)p_i^{-1}\,\mathrm{D}_{\hat{h}_i}(x^\star,x_i^k)\\ &\overset{(4.2)}{\le} \mathrm{D}_{\hat{H}}(\boldsymbol{x}^\star,\boldsymbol{x}^k) - \mathrm{D}_{\hat{H}}(\boldsymbol{u}^k,\boldsymbol{x}^k) + \Phi(\boldsymbol{x}^\star)-\Phi(\boldsymbol{u}^k) + \sum_{i=1}^N (p_i^{-1}-1)\,\mathrm{D}_{\hat{h}_i}(x^\star,x_i^k)\\ &\overset{\text{Lemmas A.1(v) and 3.2}}{=} \sum_{i=1}^N p_i^{-1}\,\mathrm{D}_{\hat{h}_i}(x^\star,x_i^k) - \mathrm{D}_{\hat{H}}(\boldsymbol{u}^k,\boldsymbol{x}^k) - \big(\varphi(\boldsymbol{z}^k) - \min_{\overline{C}}\varphi\big),\end{aligned}\tag{4.3}$$

where $\boldsymbol{u}^k = (u^k,\dots,u^k)$. From Fact 4.5 we conclude that

$$\sum_{k\in\mathbb{N}} \mathrm{D}_{\hat{H}}(\boldsymbol{u}^k,\boldsymbol{x}^k) < \infty \quad\text{and}\quad \sum_{k\in\mathbb{N}}\big(\varphi(z^k) - \min_{\overline{C}}\varphi\big) =: c < \infty\tag{4.4}$$

a.s., and

$$\sum_{i=1}^N p_i^{-1}\,\mathrm{D}_{\hat{h}_i}(x^\star,x_i^k) \quad\text{converges a.s. for any } x^\star\in\arg\min_{\overline{C}}\varphi.\tag{4.5}$$

It now follows from (4.4) that $\varphi(z^k)$ converges a.s. to $\min_{\overline{C}}\varphi$. Moreover, since the sequence $(\min_{\ell=0,\dots,k}\varphi(z^\ell))_{k\in\mathbb{N}}$ is nonincreasing, (4.4) also yields the claimed rate.

4.6(vi) Suppose that $(z^k)_{k\in K}\to z^\star$. Then, $(\boldsymbol{u}^k)_{k\in K}\to\boldsymbol{u}^\star$ for $\boldsymbol{u}^k=(z^k,\dots,z^k)$ and $\boldsymbol{u}^\star=(z^\star,\dots,z^\star)$. Notice that $z^\star\in\overline{C}$, since $z^k\in C$ for all k (cf. Lemma 4.2). We have

$$\min_{\overline{C}}\varphi \overset{\mathrm{A.1(v)}}{=} \min_{\overline{C}\times\cdots\times\overline{C}}\Phi \le \Phi(\boldsymbol{u}^\star) \overset{\mathrm{lsc}}{\le} \liminf_{k\in K}\Phi(\boldsymbol{u}^k) \overset{3.2(\mathrm{v})}{=} \liminf_{k\in K}\varphi(z^k) \overset{4.6(\mathrm{v})}{=} \min_{\overline{C}}\varphi.\tag{4.6}$$

Therefore, $\boldsymbol{u}^\star$ is a minimizer of Φ on $\overline{C} \times \cdots \times \overline{C}$, and thus $z^\star$ is a minimizer of φ on $\overline{C}$ by virtue of Lemma A.1(v).

4.6(vii) If $\varphi + \delta_{\overline{C}}$ is level bounded, then by Lemma 4.2(iv) $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ and $(\boldsymbol{u}^k)_{k\in\mathbb{N}}$ are bounded. Alternatively, if $C = \mathbb{R}^n$, then boundedness of the former sequence follows from Fact 2.3(iv), (4.5), and Assumption I.A3, and in turn that of the latter from (4.4). In either case Assumption II holds, as discussed in Remark 4.3(i). Boundedness of the sequences ensures the existence of $K \subseteq \mathbb{N}$, $z^\star$ and $\boldsymbol{u}^\star$ as in the proof of assertion 4.6(vi). The vanishing of $\mathrm{D}_{\hat{H}}(\boldsymbol{u}^k, \boldsymbol{x}^k)$ shown in (4.4) implies through Lemma 4.4(ii) that $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ and $(\boldsymbol{u}^k)_{k\in\mathbb{N}}$ have the same limit points, and that $(\boldsymbol{x}^k)_{k\in K} \to \boldsymbol{u}^\star$. In turn, $(\sum_{i=1}^N p_i^{-1}\,\mathrm{D}_{\hat{h}_i}(u^\star, x_i^k))_{k\in K} \to 0$ holds by Assumption II. Hence, since the entire sequence is convergent (by (4.5)) we have $(\sum_{i=1}^N p_i^{-1}\,\mathrm{D}_{\hat{h}_i}(u^\star, x_i^k))_{k\in\mathbb{N}} \to 0$, which by Fact 2.3(v) implies $(x_i^k)_{k\in\mathbb{N}} \to u^\star$, $i \in [N]$. As discussed above, this implies that $(\boldsymbol{u}^k)_{k\in\mathbb{N}} \to \boldsymbol{u}^\star$, and the identity $\boldsymbol{u}^k = (z^k, \ldots, z^k)$ of Lemma 3.2(ii) yields the claimed convergence. ∎

In Theorem 4.6(vii) the assumption that $\varphi + \delta_{\overline{C}}$ is level bounded can be relaxed by instead requiring that for every $v \in \operatorname{dom} h_i$ and $\alpha \in \mathbb{R}$, the level set

$$\{w \in \operatorname{int}\operatorname{dom} h_i \mid \mathrm{D}_{h_i}(v, w) \le \alpha\}$$

is bounded, as this would suffice to ensure boundedness of the sequences. In fact, together with the closed-domain requirement this is a standing assumption in many works dealing with Bregman distances, specifically those involving *Bregman functions with zone S* (S being the interior of the domain); see, e.g., [56].

We conclude this subsection with an analysis of the strongly convex case, in which linear convergence (in expectation) will be shown. Remarkably, strong convexity of the cost function φ alone will suffice, without imposing any such requirement on the individual terms f_i or g which, in fact, are even allowed to be nonconvex.

THEOREM 4.7 (linear convergence with randomized rule $(\mathcal{S}_1)$ for strongly convex problems). *Consider the iterates of Algorithm* 1. *Additionally to Assumption* I, *suppose that*

A1. *φ is μ_φ-strongly convex;*

A2. *h_i has a locally Lipschitz gradient on the whole space $\mathbb{R}^n$, $i \in [N]$ (hence $C = \mathbb{R}^n$).*

Let $\mathcal{U}$ be a convex compact set containing x^{init} and the sequence $(z^k)_{k\in\mathbb{N}}$, and let $\ell_{h_i,\mathcal{U}}$ be a Lipschitz modulus for ∇h_i on $\mathcal{U}$, $i \in [N]$.[1] *Let $x^\star = \arg\min\varphi$, $\varphi_\star = \min\varphi$, and*

$$c_{\mathcal{U}} = \frac{\min_i p_i}{1 + \frac{1}{\mu_\varphi}\sum_i \left(\frac{\ell_{h_i,\mathcal{U}}}{\gamma_i} - \frac{\sigma_{f_i,\mathcal{U}}}{N}\right)}, \tag{4.7}$$

where $\sigma_{f_i,\mathcal{U}} \ge -L_{f_i}\ell_{h_i,\mathcal{U}}$ is a (weak) convexity modulus of f_i on $\mathcal{U}$.[2] *Then, for all $k \in \mathbb{N}$*

$$\mathbb{E}_k\big[\mathscr{L}(z^{k+1}, \boldsymbol{s}^{k+1}) - \varphi_\star\big] \le (1 - c_{\mathcal{U}})\big(\mathscr{L}(z^k, \boldsymbol{s}^k) - \varphi_\star\big), \quad \textit{and} \tag{4.8}$$

[1] $\mathcal{U}$ exists by Lemma 4.2(iv), owing to strong convexity and consequent level boundedness of φ. For $i \in [N]$, a finite $\ell_{h_i,\mathcal{U}}$ then exists because of Theorem 4.7.A2 and since $\mathcal{U} \subset C$ (as opposed to $\mathcal{U} \subseteq \overline{C}$), having $C = \mathbb{R}^n$.

[2] f_i is $\sigma_{f_i,\mathcal{U}}$-weakly convex on $\mathcal{U}$ if $f_i - \frac{\sigma_{f_i,\mathcal{U}}}{2}\|\cdot\|^2$ is convex on $\mathcal{U}$, thus coinciding with convexity (resp., $\sigma_{f_i,\mathcal{U}}$-strong convexity) on $\mathcal{U}$ when $\sigma_{f_i,\mathcal{U}} \ge 0$ (resp., $\sigma_{f_i,\mathcal{U}} > 0$). The lower bound on $\sigma_{f_i,\mathcal{U}}$ owes to Fact 2.5(iii).

$$\mathbb{E}\big[\tfrac{\mu_\varphi}{2}\|z^k - x^\star\|^2\big] \le \mathbb{E}\big[\varphi(z^k) - \varphi_\star\big] \le (1 - c_{\mathcal{U}})^k\big(\varphi(x^{\text{init}}) - \varphi_\star\big). \tag{4.9}$$

Proof. See Appendix B. ∎

In the Euclidean case, $h_i = \frac{1}{2}\|\cdot\|^2$ has a 1-Lipschitz gradient on $\mathbb{R}^n$, and the results of Theorem 4.7 hold with $\mathcal{U} = \mathbb{R}^n$ and $\ell_{h_i} = 1$, improving those of [37, Cor. 3.3] (limited to the Euclidean case) both by providing tighter rates and by relaxing (strong) convexity assumptions on individual f_i. For the uniform sampling strategy $p_i = 1/N$ the rate $1 - O(1/N)$ is obtained. The same arguments still hold for the Bregman extension of Algorithm 1 dealt with in this paper, as long as each h_i is Lipschitz differentiable. This fact is stated in the following corollary, where, by using the fact that $\mu_\varphi \ge \frac{1}{N}\sum_{i=1}^N \sigma_{f_i}$ under a convexity assumption on g, a simplified expression for the constant c in (4.7) is obtained. We remark that in the Euclidean case a variant of SVRG [34] has also been studied in [2] under similar assumptions.

COROLLARY 4.8 (global linear rate). *Additionally to Assumption* I, *suppose that*

- A1. *g is convex, and $f := \frac{1}{N}\sum_i f_i$ is μ_f-strongly convex (yet each f_i can be nonconvex);*
- A2. *∇h_i is Lipschitz on $\mathbb{R}^n$ (hence so is f_i with modulus ℓ_{f_i}; cf. Fact* 2.5*(iii)), $i \in [N]$.*

Set $\gamma_i = \alpha N/L_{f_i}$ with $\alpha \in (0,1)$ and $\kappa_f := \frac{\frac{1}{N}\sum_i \ell_{f_i}}{\mu_f}$. Then, (4.8) *and* (4.9) *hold with $c_{\mathbb{R}^n} \ge \frac{\alpha \min_i p_i}{\kappa_f}$.*

4.3. Essentially cyclic sampling rule ($\mathcal{S}_2$). The convergence results in this subsection require convexity of the nonsmooth term g and local strong convexity and smoothness of h_i (as is the case when $h_i \in \mathcal{C}^2(\mathbb{R}^n)$ with $\nabla^2 h_i \succ 0$). The proof of subsequential convergence is an adaptation of that of [37, Thm. 2.8].

THEOREM 4.9 (subsequential convergence with essentially cyclic rule ($\mathcal{S}_2$)). *Additionally to Assumption* I, *assume that g is convex, $C = \mathbb{R}^n$, and that either one of the following assumptions holds:*

- (A) *(either) each h_i is strongly convex and Lipschitz differentiable,*
- (B) *(or) φ is level bounded and each h_i is locally strongly convex and locally Lipschitz differentiable.*

Then, all the claims in Theorems 4.6*(i) to* 4.6*(iv) hold surely.*

Proof. See Appendix B. ∎

We remark that linear convergence in the strongly convex case may be obtained in a similar fashion to [37, Thm. 2.9] and [12, Thm. 3.9]. This however results in a rate more conservative than the one obtained for the randomized case (cf. [37, eq. (2.21)]), which is not consistent with what is observed in practice. Although a refined analysis not relying on conservative triangle inequalities may be possible, this direction is not investigated here.

Our next goal is to establish global (and linear) convergence results without convexity assumptions on f_i or their sum. To this end, we leverage the KL property [38, 36], which has become the standard tool in the analysis of nonconvex proximal methods, and most notably holds for the class of semialgebraic functions [18, 17, 3, 4, 5, 19].

DEFINITION 4.10 (KL property with exponent θ). *A proper lsc function $q : \mathbb{R}^n \to \overline{\mathbb{R}}$ has the KL property with exponent $\theta \in (0,1)$ if for every $\bar{w} \in \operatorname{dom} \partial q$ there exist*

$\varepsilon, \eta, \varrho > 0$ such that $\psi'(q(w) - q(\bar{w}))\operatorname{dist}(0, \partial q(w)) \geq 1$ holds for every w satisfying $\|w - \bar{w}\| < \varepsilon$ and $q(\bar{w}) < q(w) < q(\bar{w}) + \eta$, where $\psi(s) := \varrho s^{1-\theta}$.

As will be detailed in Theorem 4.11, global convergence is established when the model $\mathscr{M} : \mathbb{R}^{nN} \times \mathbb{R}^{nN} \to \overline{\mathbb{R}}$ defined as $\mathscr{M}(\boldsymbol{w}, \boldsymbol{x}) := \Phi(\boldsymbol{w}) + \mathrm{D}_{\hat{H}}(\boldsymbol{w}, \boldsymbol{x})$ has the KL property. The next assumption provides easily verifiable requirements in terms of f_i, h_i, and φ.

Assumption III (global convergence requirements). In problem (P),

A1. for $i \in [N]$, $f_i, h_i \in \mathscr{C}^2(\mathbb{R}^n)$ (hence $C = \mathbb{R}^n$) with $\nabla^2 h_i \succ 0$;

A2. φ has the KL property with exponent $\theta \in (0, 1)$ (e.g., when f_i and g are semialgebraic) and is level bounded.

THEOREM 4.11 (global and linear convergence with essentially cyclic rule ($\mathcal{S}_2$)). *Suppose that Assumptions* I *and* III *are satisfied and that g is convex. Then, the following holds for the iterates generated by Algorithm* 1 *with an essentially cyclic rule ($\mathcal{S}_2$):*

(i) $(z^k)_{k\in\mathbb{N}}$ *converges to a stationary point $z^\star$ for φ.*

(ii) *If $\theta > 1/2$, then there exists $c > 0$ such that $\varphi(z^k) - \varphi(z^\star) \leq ck^{-\frac{1}{2\theta-1}}$ holds for all $k \in \mathbb{N}$.*

(iii) *If $\theta \in (0, 1/2]$, then $(z^k)_{k\in\mathbb{N}}$ and $(\varphi(z^k))_{k\in\mathbb{N}}$ converge at R-linear rate.*

Proof. Notice that Assumption III.A1 along with level boundedness of φ in Assumption III.A2 ensures that the requirement in Theorem 4.9(B) is satisfied. By the first claim of Theorem 4.9, the sequence $(\mathrm{D}_{\hat{h}_i^\star}(s_i^k, \nabla\hat{h}_i(z^k)))_{k\in\mathbb{N}}$ converges to zero, and thus we may invoke Lemma 4.4 to conclude that the set ω of limit points of $(z^k)_{k\in\mathbb{N}}$ is nonempty, compact, connected, and made of stationary points for φ, with $\varphi \equiv \varphi_\star := \lim_{k\to\infty} \Phi^{\hat{H}}(\boldsymbol{x}^k)$ on ω. If $\Phi^{\hat{H}}(\boldsymbol{x}^k) = \varphi_\star$ holds for some $k \in \mathbb{N}$, then it follows from Lemma 4.2(i) that $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ is asymptotically constant, and the assertion holds trivially. In what follows we thus assume that $\Phi^{\hat{H}}(\boldsymbol{x}^k) > \varphi_\star$ holds for all k. The assumptions together with Lemma A.1(iii) ensure that Φ enjoys the KL property with exponent θ. Since $\hat{H}$ is locally strongly convex, we may invoke [68, Lem. 5.1] to infer that the function $\mathscr{M} : \mathbb{R}^{nN} \times \mathbb{R}^{nN} \to \overline{\mathbb{R}}$ defined as $\mathscr{M}(\boldsymbol{w}, \boldsymbol{x}) = \Phi(\boldsymbol{w}) + \mathrm{D}_{\hat{H}}(\boldsymbol{w}, \boldsymbol{x})$ has the KL property with exponent $\vartheta := \max\{\theta, 1/2\}$ at every point of the compact set $\boldsymbol{\Omega} := \{(\boldsymbol{z}^\star, \boldsymbol{z}^\star) \mid \boldsymbol{z}^\star \in \boldsymbol{\omega}\}$, where $\boldsymbol{\omega} := \{\boldsymbol{z} = (z, \dots, z) \mid z \in \omega\}$.[3] Notice that $\Phi^{\hat{H}}(\boldsymbol{x}^k) = \mathscr{M}(\boldsymbol{u}^k, \boldsymbol{x}^k)$ and $\mathscr{M}(\boldsymbol{z}^\star, \boldsymbol{z}^\star) = \Phi^{\hat{H}}(\boldsymbol{z}^\star) = \varphi_\star$ hold for every $k \in \mathbb{N}$ and $\boldsymbol{z}^\star \in \boldsymbol{\omega}$ (cf. Theorem 4.9), and that $\partial\mathscr{M}(\boldsymbol{w}, \boldsymbol{x}) = (\partial\Phi(\boldsymbol{w}) + \nabla\hat{H}(\boldsymbol{w}) - \nabla\hat{H}(\boldsymbol{x}), \nabla^2\hat{H}(\boldsymbol{x})(\boldsymbol{x} - \boldsymbol{w}))$. By Lemma 3.1(iii) we have $\nabla\hat{H}(\boldsymbol{x}^k) - \nabla\hat{H}(\boldsymbol{u}^k) \in \partial\Phi(\boldsymbol{u}^k)$, which in turn implies

$$\operatorname{dist}\big(0, \partial\mathscr{M}(\boldsymbol{u}^k, \boldsymbol{x}^k)\big) \leq \|\nabla^2\hat{H}(\boldsymbol{x}^k)\|\|\boldsymbol{x}^k - \boldsymbol{u}^k\| \leq C\|\boldsymbol{x}^k - \boldsymbol{u}^k\|, \tag{4.10}$$

where $C = \sup_k \|\nabla^2\hat{H}(\boldsymbol{x}^k)\|$ is finite due to boundedness of $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ and continuity of $\nabla^2\hat{H}$. Let $\psi(t) := \rho t^{1-\vartheta}$ be a desingularizing function for $\mathscr{M}$ on $\boldsymbol{\Omega}$ [3, Lem. 1(ii)], namely, such that

$$\psi'\big(\mathscr{M}(\boldsymbol{w}, \boldsymbol{x}) - \varphi_\star\big)\operatorname{dist}\big(0, \partial\mathscr{M}(\boldsymbol{w}, \boldsymbol{x})\big) \geq 1$$

[3]Consistently with the locality of the KL property and the compactness of $\boldsymbol{\omega}$, the global strong convexity requirement in [68, Lem. 5.1] can clearly be replaced by local strong convexity. Similarly, if Φ is a KL function with exponent θ, then it is trivially a KL function with exponent ϑ, thus complying with the requirement in the reference.

holds for some $\varepsilon > 0$ and all $(\boldsymbol{w}, \boldsymbol{x})$ ε-close to $\boldsymbol{\Omega}$ such that $0 < \mathscr{M}(\boldsymbol{w}, \boldsymbol{x}) - \varphi_\star < \varepsilon$. Since $\mathscr{M}(\boldsymbol{u}^k, \boldsymbol{x}^k) = \Phi^{\hat{H}}(\boldsymbol{x}^k) \searrow \varphi_\star$ (cf. Lemma 4.2(ii)) and $(\boldsymbol{u}^k, \boldsymbol{x}^k)_{k\in\mathbb{N}}$ is bounded and accumulates on $\boldsymbol{\Omega}$, by discarding early iterates we may assume that the inequality above holds for $(\boldsymbol{w}, \boldsymbol{x}) = (\boldsymbol{u}^k, \boldsymbol{x}^k)$, $k \in \mathbb{N}$, which combined with (4.10) results in

$$\rho^{-1}(1-\theta)^{-1}\big(\Phi^{\hat{H}}(\boldsymbol{x}^k) - \varphi_\star\big)^\theta = \psi'\big(\Phi^{\hat{H}}(\boldsymbol{x}^k) - \varphi_\star\big)^{-1} \le C\|\boldsymbol{x}^k - \boldsymbol{u}^k\|. \tag{4.11}$$

Let $\Delta_k := \psi(\Phi^{\hat{H}}(\boldsymbol{x}^k) - \varphi_\star)$, so that $\Phi^{\hat{H}}(\boldsymbol{x}^k) - \varphi_\star = (\Delta_k/\rho)^{1/1-\theta}$. By concavity of ψ we have

$$\begin{aligned}\Delta_{T(\nu+1)} - \Delta_{T\nu} &\le \psi'(\Phi^{\hat{H}}(\boldsymbol{x}^{T\nu}) - \varphi_\star)\big(\Phi^{\hat{H}}(\boldsymbol{x}^{T(\nu+1)}) - \Phi^{\hat{H}}(\boldsymbol{x}^{T\nu})\big)\\ &\overset{(4.11)}{\le} \frac{\Phi^{\hat{H}}(\boldsymbol{x}^{T(\nu+1)}) - \Phi^{\hat{H}}(\boldsymbol{x}^{T\nu})}{C\|\boldsymbol{u}^{T\nu} - \boldsymbol{x}^{T\nu}\|}\\ &\overset{(B.6)}{\le} -c\|\boldsymbol{u}^{T\nu} - \boldsymbol{x}^{T\nu}\|,\end{aligned} \tag{4.12}$$

for some constant $c > 0$. As argued in the proof of Theorem 4.9, by suitably shifting, we conclude that for all $k \in \mathbb{N}$

$$\Delta_{k+T} - \Delta_k \le -c\|\boldsymbol{u}^k - \boldsymbol{x}^k\|. \tag{4.13}$$

The rest of the proof is standard (see [3, Thm. 2]), and is provided for completeness. Since $\Delta_k \ge 0$, by telescoping we conclude that $(\|\boldsymbol{u}^k - \boldsymbol{x}^k\|)_{k\in\mathbb{N}}$ has finite sum, and since $\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\| \le \|\boldsymbol{u}^k - \boldsymbol{x}^k\|$ for all $k \in \mathbb{N}$, $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ has finite length. Therefore, $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ and $(\boldsymbol{u}^k)_{k\in\mathbb{N}}$ converge to the same stationary point, be it $\boldsymbol{z}^\star$, owing to Theorem 4.9.

We now show the convergence rates. It follows from (4.11) that

$$C\rho^{1/1-\vartheta}(1-\vartheta)\|\boldsymbol{x}^k - \boldsymbol{u}^k\| \ge \rho^{\vartheta/1-\vartheta}\big(\Phi^{\hat{H}}(\boldsymbol{x}^k) - \varphi_\star\big)^\vartheta = \Delta_k^{\vartheta/1-\vartheta}.$$

Combined with (4.13), it results in $c_1\Delta_k^{\vartheta/1-\vartheta} \le \Delta_k - \Delta_{k+T}$ for some $c_1 > 0$. We may now invoke Lemma A.3 to infer that, for every $t \in [T]$, $(\Delta_{t+\nu T})_{\nu\in\mathbb{N}}$ converges Q-linearly (to 0) if $\theta \le 1/2$ (which corresponds to $\vartheta = 1/2$), and $\Delta_{t+\nu T} \le c_3\nu^{-\frac{1-\theta}{2\theta-1}}$ for some $c_3 > 0$ otherwise (that is, if $\vartheta = \theta > 1/2$). Note that the former case implies that $(\Delta_k)_{k\in\mathbb{N}}$ converges R-linearly, whereas the latter implies that $\Delta_k \le c_4 k^{-\frac{1-\theta}{2\theta-1}}$ holds for every k and some $c_4 \ge c_3$. Recalling that $\Phi^{\hat{H}}(\boldsymbol{x}^k) - \varphi_\star = (\Delta_k/\rho)^{1/1-\theta}$, the claimed rates of for the cost function follow from Fact 2.7(i). Similarly, when $(\Delta_k)_{k\in\mathbb{N}}$ converges Q-linearly then so does $(\|\boldsymbol{x}^k - \boldsymbol{u}^k\|)_{k\in\mathbb{N}}$, as it follows from (4.13), and in turn so does $(\|\boldsymbol{x}^k - \boldsymbol{x}^{k+1}\|)_{k\in\mathbb{N}}$ owing to the inequality $\|\boldsymbol{x}^k - \boldsymbol{x}^{k+1}\| \le \|\boldsymbol{x}^k - \boldsymbol{u}^k\|$. These two facts imply that $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ and $(\boldsymbol{u}^k = (z^k, \dots, z^k))_{k\in\mathbb{N}}$ are R-linearly convergent. ∎

4.4. Low-memory variant. We now analyze Algorithm 2, which, as shown next, is simply a particular implementation of Algorithm 1.

LEMMA 4.12 (Algorithm 2 as an instance of Algorithm 1). *As long as the same parameters are chosen in Algorithms* 1 *and* 2*, to any sequence* $(\tilde{s}^k_{\mathrm{LM}}, z^k_{\mathrm{LM}}, \mathcal{J}^{k+1}_{\mathrm{LM}})_{k\in\mathbb{N}}$ *generated by Algorithm* 2 *there corresponds an identical sequence* $(\tilde{s}^k, z^k, \mathcal{J}^{k+1})_{k\in\mathbb{N}}$ *generated by Algorithm* 1*. Moreover, the indices* $(\mathcal{J}^{k+1}_{\mathrm{LM}})_{k\in\mathbb{N}}$ *comply with the essentially cyclic rule* $(\mathcal{S}_2)$ *with* $T = N$*.*

Proof. See Appendix B. ∎

As a consequence of Lemma 4.12, Algorithm 2 inherits all the convergence results of subsection 4.3. In addition, here, the convexity requirement of g in Theorems 4.9 and 4.11 can be lifted thanks to the periodic full sampling of the indices.

THEOREM 4.13 (subsequential convergence of Algorithm 2). *Suppose that Assumption* I *holds and let* ω *be the limit set of the sequence* $(\tilde{z}^k)_{k\in\mathbb{N}}$ *generated by Algorithm* 2. *Then,*

(i) *the sequence* $(\varphi(\tilde{z}^k))_{k\in\mathbb{N}}$ *converges to the finite value* $\varphi_\star \le \varphi(x^{\text{init}})$ *as in Lemma* 4.2(ii).

(ii) *if Assumption* II *is satisfied, then* $\varphi \equiv \varphi_\star$ *on* ω.

(iii) *if* $C = \mathbb{R}^n$, *then* $0 \in \hat{\partial}\varphi(z^\star)$ *for every* $z^\star \in \omega$.

Proof. As shown in Lemma 4.12, Algorithm 2 coincides with Algorithm 1 with an essentially cyclic sampling rule $(\mathcal{S}_2)$, and there exists an indexing subsequence $(k_r)_{r\in\mathbb{N}}$ with $0 < k_{r+1} - k_r \le N+1$ such that $\mathcal{K}^{k_r} = \emptyset$. Then, the $\tilde{z}$-update rule (cf. steps 4 and 9) yields

$$z^{k_r} = \tilde{z}^{k_r} = \tilde{z}^{k_r+1} = \cdots = \tilde{z}^{k_{r+1}-1} \quad \forall r \in \mathbb{N}. \tag{4.14}$$

We have

$$\begin{aligned}
\Phi^{\hat{H}}(\boldsymbol{x}^{k_r+1}) &\le \Phi^{\hat{H}}(\boldsymbol{x}^{k_r}) - \sum_{i\in\mathcal{I}^{k_r+1}} \mathrm{D}_{\hat{h}_i}(u^{k_r}, x_i^{k_r}) && \text{(Lemma 4.2(i))}\\
&= \Phi^{\hat{H}}(\boldsymbol{x}^{k_r}) - \mathrm{D}_{\hat{H}}(\boldsymbol{u}^{k_r}, \boldsymbol{x}^{k_r}) && (\mathcal{I}^{k_r+1} = [N])\\
&\le \Phi^{\hat{H}}(\boldsymbol{x}^{k_{r-1}+1}) - \mathrm{D}_{\hat{H}}(\boldsymbol{u}^{k_r}, \boldsymbol{x}^{k_r}) && \text{(Lemma 4.2(i), } k_r \ge k_{r-1}+1)
\end{aligned} \tag{4.15}$$

holding for every $r \in \mathbb{N}$. By telescoping and by using the fact that $\Phi^{\hat{H}} \ge \min\varphi > -\infty$, it follows that $(\mathrm{D}_{\hat{H}}(\boldsymbol{u}^{k_r}, \boldsymbol{x}^{k_r}))_{r\in\mathbb{N}}$ has finite sum and in particular vanishes. Since $z^{k_r} = \tilde{z}^{k_r}$,

$$\varphi_\star \xleftarrow[r\to\infty]{4.2(\text{ii})} \Phi^{\hat{H}}(\boldsymbol{x}^{k_r}) = \varphi(z^{k_r}) + \mathrm{D}_{\hat{H}}(\boldsymbol{u}^{k_r}, \boldsymbol{x}^{k_r}) = \varphi(\tilde{z}^{k_r}) + \mathrm{D}_{\hat{H}}(\boldsymbol{u}^{k_r}, \boldsymbol{x}^{k_r}),$$

whence assertion 4.13(i) follows. Assertions 4.13(ii) and 4.13(iv) follow by patterning the arguments of Lemma 4.4(iii) and 4.4(iv). ∎

In the next theorem global convergence results are provided under Assumption III in a fully nonconvex setting. Moreover, in the spirit of Theorem 4.11, linear and sublinear convergence rates are obtained according to the KL exponent.

THEOREM 4.14 (global and linear convergence of Algorithm 2). *Suppose that Assumptions* I *and* III *hold. Then, the following hold for the iterates generated by Algorithm* 2*:*

(i) $(\tilde{z}^k)_{k\in\mathbb{N}}$ *converges to a stationary point* $x^\star$ *for* φ.

(ii) *If* $\theta > 1/2$, *then there exists* $c > 0$ *such that* $\varphi(\tilde{z}^k) - \varphi(x^\star) \le ck^{-\frac{1}{2\theta-1}}$ *holds for all* $r \in \mathbb{N}$.

(iii) *If* $\theta \in (0, 1/2]$, *then* $(\tilde{z}^k)_{k\in\mathbb{N}}$ *and* $(\varphi(\tilde{z}^k))_{k\in\mathbb{N}}$ *converge at an R-linear rate.*

Proof. Notice that Assumption III entails local strong convexity and Lipschitz differentiability of each h_i. Thus, as discussed in the proof of Theorem 4.9, $\hat{H}$ is $\mu_{\hat{H},\boldsymbol{\mathcal{U}}}$-strongly convex on a convex compact set $\boldsymbol{\mathcal{U}}$ containing the iterates. Let the indexing subsequence $(k_r)_{r\in\mathbb{N}}$ be as in the proof of Theorem 4.13, and observe that $k_r \le (N+1)r$ holds for every $r \in \mathbb{N}$. The assertions are established with the same

arguments as in Theorem 4.11 with the difference that here we examine the generated iterates at subindices k_r. That is, (4.12) is replaced by

$$\Delta_{k_{r+1}} - \Delta_{k_r} \leq \frac{\Phi^{\hat{H}}(\boldsymbol{x}^{k_{r+1}}) - \Phi^{\hat{H}}(\boldsymbol{x}^{k_r})}{C\|\boldsymbol{u}^{k_r} - \boldsymbol{x}^{k_r}\|} \underset{4.2(\mathrm{i})}{\leq} \frac{\Phi^{\hat{H}}(\boldsymbol{x}^{k_r+1}) - \Phi^{\hat{H}}(\boldsymbol{x}^{k_r})}{C\|\boldsymbol{u}^{k_r} - \boldsymbol{x}^{k_r}\|} \leq -C\mu_{\hat{H},\boldsymbol{u}}\|\boldsymbol{u}^{k_r} - \boldsymbol{x}^{k_r}\|, \tag{4.16}$$

where the last inequality follows from (4.15) and Fact 2.3(vi). Subsequently, by patterning the arguments of Theorem 4.11, we obtain that $(\Delta_{k_r})_{r\in\mathbb{N}}$ converges Q-linearly if $\theta \leq 1/2$, and $\Delta_{k_r} \leq cr^{-\frac{1-\theta}{2\theta-1}}$ for some $c > 0$, if $\theta > 1/2$. The claims follow by noting that $\tilde{z}^k = z^{k_r} = u^{k_r}$ for all k satisfying $k_r \leq k < k_{r+1}$ (cf. (4.14)), and arguing as in the last part of Theorem 4.11. ∎

5. Application to phase retrieval and numerical simulations. In this section we study two examples related to the phase retrieval problem, which consists of recovering a signal based on *intensity measurements*, and arises in many important applications including X-ray crystallography, speech processing, electron microscopy, astronomy, and optical imaging; see, e.g., [21, 41, 55, 58]. Here, we consider phase retrieval problems with real-valued data, that is, given $a_i \in \mathbb{R}^n \setminus \{0\}$ and scalars $b_i \in \mathbb{R}_+$, $i \in [N]$, the goal is to find $x \in \mathbb{R}^n$ such that

$$b_i \approx \langle a_i, x\rangle^2, \quad i \in [N], \tag{5.1}$$

accounting for the fact that in real-world applications the recorded intensities are likely corrupted by noise, and may involve outliers due to measurement errors. To tackle such problems, we consider the following *sparse phase retrieval* formulation,

$$\text{minimize}_{x\in\mathbb{R}^n} \ \tfrac{1}{N}\textstyle\sum_{i=1}^N \mathcal{L}(b_i, \langle a_i, x\rangle^2) + g(x), \tag{5.2}$$

where $\mathcal{L}$ is a loss function, and g is a sparsity inducing function (e.g., l_1- or l_0-norm). In particular, we study the case of squared loss $\mathcal{L}(y, z) = \frac{1}{4}(y - z)^2$ [21, 58], and Poisson loss $\mathcal{L}(y, z) = z - y\log(z)$ [23, 69], suitable when measurements follow the Poisson model ($b_i \approx \text{Poisson}(\langle a_i, x\rangle^2)$). Other formulations with l_1-loss have been studied in the literature [29, 26].

5.1. Sparse phase retrieval with squared loss. Consider the nonconvex minimization (5.2) with squared loss $\mathcal{L}(y, z) = \frac{1}{4}(y - z)^2$, and either $g = \lambda\|\cdot\|_1$, $\lambda \geq 0$, or $g = \delta_{\mathbb{B}_\kappa}$, where $\mathbb{B}_\kappa$ is the l_0-norm ball of radius κ. This problem is written in the form of (P) with

$$f_i(x) = \tfrac{1}{4}(\langle a_i, x\rangle^2 - b_i)^2 \quad \text{and} \quad h_i(x) = \tfrac{1}{4}\|x\|^4 + \tfrac{1}{2}\|x\|^2. \tag{5.3}$$

The next lemma is a simple adaptation from [20] for finding the smoothness moduli of f_i relative to the Legendre kernel h_i, and for computing the solutions to the subproblem (2.1). For l_1-regularization, the inner subproblems amount to computations involving the soft-thresholding operator, whereas in the case of l_0-norm ball they amount to computing projections onto $\mathbb{B}_\kappa$, that is, setting to zero $n-\kappa$ elements among the smallest in absolute value.

LEMMA 5.1 ([20, Lem. 5.1, Props. 5.1 and 5.2]). *Let f_i and h_i be as in* (5.3). *Then, f_i is L_{f_i}-smooth relative to h_i with $L_{f_i} = (3\|a_i\|^4 + \|a_i\|^2|b_i|)$. Moreover, denoting $\bar{\gamma} = (\sum_{i=1}^N 1/\gamma_i)^{-1}$, for any $y(s) \in \operatorname{prox}_{\bar{\gamma}g}(\bar{\gamma}s)$ the operator T as defined in* (2.1) *may be computed as follows:*

(i) *If* $g = \lambda\|\cdot\|_1$, $\lambda \geq 0$, *then* $T(s) = t^\star y(s)$, *where* $t^\star$ *is the real positive root of the equation* $\|y(s)\|^2 t^3 + t - 1 = 0$.[4]

(ii) *If* $g = \delta_{\mathbb{B}_\kappa}(x)$, *then* $T(s) \ni -t^\star \|y(s)\|^{-1} y(s)$, *where* $t^\star$ *is the real nonnegative root of* $t^3 + t - \|y(s)\| = 0$ *(see footnote 4).*

5.2. Sparse phase retrieval with Poisson loss. We now assume that the recorded intensities follow the Poisson model ($b_i \sim \text{Poisson}(\langle a_i, x\rangle^2)$). In this setting we adapt the Poisson loss $\mathcal{L}(y, z) = z - y\log(z)$ and consider the regularized problem (5.2) with $g = \lambda\|\cdot\|_1$, $\lambda \geq 0$. This problem may be written in the form of (P) by setting

$$f_i(x) = -b_i \log(\langle a_i, x\rangle^2) + \langle a_i, x\rangle^2 \quad \text{and} \quad h_i(x) = \|a_i\|^2\|x\|^2 - 2b_i \sum_{j=1}^n \log(x_j). \tag{5.4}$$

As shown next, the nonconvex function f_i is smooth relative to h_i, and the operator T as in (2.1) is easily computable.

LEMMA 5.2. *Let* f_i *and* h_i *be as in* (5.4), *with* $a_i \in \mathbb{R}^n_+ \setminus \{0\}$ *and* $b = (b_1, \dots, b_N) \in \mathbb{R}^n_+ \setminus \{0\}$. *Then,* h_i *is a* $2\|a_i\|^2$*-strongly convex Legendre kernel (with* $\operatorname{dom} h_i = \mathbb{R}^n_{++}$*), and* f_i *is* L_{f_i}*-smooth relative to* h_i *with* $L_{f_i} = 1$. *Moreover, denoting* $c_a = \sum_{i=1}^N \frac{4}{\gamma_i}\|a_i\|^2$ *and* $c_b = \sum_{i=1}^N \frac{4}{\gamma_i} b_i$, *the operator* T *as defined in* (2.1) *with* $g = \lambda\|\cdot\|_1$, $\lambda \geq 0$, *is given by*

$$T(s) = (w_1, \dots, w_M) \quad \textit{with} \quad w_j = \tfrac{1}{c_a}\Big(s_j - \lambda + \big((s_j - \lambda)^2 + c_a c_b\big)^{1/2}\Big). \tag{5.5}$$

Proof. To avoid clutter, we drop the subscripts i. The assertion on h is of immediate verification. Since both f and h are $\mathscr{C}^2$ on $\operatorname{int}\operatorname{dom} h$, once we show that $\nabla^2 h - \nabla^2 f \succeq 0$ on $\operatorname{int}\operatorname{dom} h$ the claim will follow. From direct computations, $\nabla^2 h(x) = 2\|a\|^2 + 2b\operatorname{diag}(x_1^{-2}, \dots, x_n^{-2})$ and $\nabla^2 f(x) = 2\big(1 + \frac{b}{\langle a, x\rangle^2}\big)aa^\top$. In particular, $M(x) := \nabla^2 h(x) - \nabla^2 f(x) = 2b\big(\operatorname{diag}(x_1^{-2}, \dots, x_n^{-2}) - \frac{1}{\langle a, x\rangle^2}aa^\top\big) + 2\|a\|^2 - 2aa^\top$. For every $y \in \mathbb{R}^n$ it holds that (here a_k is the kth coordinate of a)

$$\langle y, M(x)y\rangle \geq 2b\sum_{j=1}^n \frac{y_j^2}{x_j^2} - 2b\frac{\langle a, y\rangle^2}{\langle a, x\rangle^2} \geq 2b\sum_{j=1}^n \frac{y_j^2}{x_j^2} - 2b\sum_{j=1}^n \frac{a_j x_j}{\langle a, x\rangle}\frac{y_j^2}{x_j^2} \geq 0,$$

where the first inequality follows by the Cauchy–Schwarz inequality, the second one from Jensen's inequality $\langle a, y\rangle^2 = \sum_{j=1}^n \big(a_j x_j (\frac{y_j}{x_j})\big)^2 \leq \sum_{i=1}^n a_i x_i \sum_{j=1}^n a_j x_j (\frac{y_j}{x_j})^2$, and the third one from the fact that $\sum_i \alpha_i \sum_j \beta_j \geq \sum_i \alpha_i \beta_i$ for every $\alpha_i, \beta_i \geq 0$, $i \in [n]$. The closed-form solution for the proximal mapping $T(s)$ follows directly from its first-order optimality conditions. ∎

5.3. Experimental setup. We test Algorithm 1 with sampling rules ($\mathcal{S}_1$) (using single-index selection with uniform sampling), ($\mathcal{S}_2^{\text{SHUF}}$), ($\mathcal{S}_2^{\text{CYCL}}$), and the low-memory Algorithm 2 with a cyclic inner loop (corresponding to $\mathcal{I}^{k+1} = [N]$ if $\text{mod}(k, N+1) = 0$, and $\mathcal{I}^{k+1} = \{\text{mod}(k, N+1)\}$ otherwise). We also consider the full mirror descent (MD) algorithm for nonconvex problems under the relative smoothness assumption [20], and SMD, its stochastic extension [25, 32]. The incremental method PLIAG [70] was not tested, as the problem setting does not comply with the requirements therein;

[4]Nonnegative real roots of the cubic equation $t^3 + pt + q = 0$ for some $p > 0$ and $q \leq 0$ are given by Cardano's formula $t^\star = (c - q/2)^{1/3} - (c + q/2)^{1/3}$, where $c = (q^2/4 + p^3/27)^{1/2}$; see, e.g., [57].

cf. [70, Assump. 8]. For the problem of subsection 5.1 with $g = \delta_{\mathbb{B}_\kappa}$, the (shuffled) cyclic rules in Algorithm 1 do not comply with Theorem 4.9 (since g is nonconvex), but are, however, provided as empirical evidence.

Parameters selection. For Algorithms 1 and 2, we always use $\gamma_i = {}^{0.99N}/_{L_{f_i}}$. For SMD we used the popular square-summable stepsize $\gamma_k = {}^{\alpha}/_{(L_f k)}$, where k is the iteration counter, L_f is the smoothness modulus of $f = {}^{1}/_{N} \sum_{i=1}^N f_i$ relative to a suitable Bregman kernel h, and $\alpha > 0$ is tuned for performance. In particular, for SMD in the problems described above, $L_f = \sum_{i=1}^N \frac{1}{N} L_{f_i}$ and $h(x) = \frac{1}{4}\|x\|^4 + \frac{1}{2}\|x\|^2$ for simulations related to subsection 5.1, and $h(x) = \frac{1}{N}\sum_{i=1}^N \|a_i\|^2\|x\|^2 - \frac{2}{N}\sum_{i=1}^N b_i \sum_{j=1}^n \log(x_j)$ for those related to subsection 5.2.

Optimality criteria. As a measure of suboptimality, we consider

$$\mathcal{D}(z^k) := \|z^k - v^k\| \quad \text{for some} \quad v^k \in T(\textstyle\sum_{i=1}^N \nabla \hat{h}_i(z^k)), \tag{5.6}$$

since it satisfies

$$\begin{aligned} \tfrac{1}{N}\operatorname{dist}(0, \hat{\partial}\varphi(v^k)) &\overset{\text{A.1(ii)}}{=} \inf_{\boldsymbol{w}\in\hat{\partial}\Phi(\boldsymbol{v}^k)} \tfrac{1}{N}\|\textstyle\sum_{i=1}^N w_i\| \overset{\text{3.1(iii)}}{\le} \tfrac{1}{N}\big\|\textstyle\sum_{i=1}^N \big(\nabla\hat{h}_i(z^k) - \nabla\hat{h}_i(v^k)\big)\big\| \\ &\le \tfrac{1}{N}\textstyle\sum_{i=1}^N \|\nabla\hat{h}_i(z^k) - \nabla\hat{h}_i(v^k)\| \le \eta\|z^k - v^k\| = \eta\mathcal{D}(z^k), \end{aligned} \tag{5.7}$$

where $\boldsymbol{v}^k = (v_1, \dots, v_N)$, and η is some positive constant that exists by virtue of local Lipschitz continuity of h_i, Lemma 3.1(iv), and boundedness of the sequences $(z^k)_{k\in\mathbb{N}}$ and $(v^k)_{k\in\mathbb{N}}$. Although, in a similar fashion to (5.7), $\operatorname{dist}(0, \hat{\partial}\varphi(z^k))$ may be upper bounded by $\|\tilde{s}^k - \sum_i \nabla\hat{h}_i(z^k)\|$ which would be an equally good estimate of $\operatorname{dist}(0, \hat{\partial}\varphi(z^k))$, this quantity is not readily available in other methods such as SMD. The introduction of v^k, instead, offers a viable algorithm-independent alternative that only requires access to output variables.[5]

Simulations. In the first set of simulations we consider 16×16 gray-scale images from a digits dataset [30][6] and a QR code dataset [43]. The images are vectorized resulting in the signal $x \in \mathbb{R}^n$ with $n = 256$. The data matrix $A \in \mathbb{R}^{N\times n}$ (a_i being the ith row) with $N = nd$, $d = 5$, is generated following the procedure described in [29, sect. 6.3]. Let $M \in \mathbb{R}^{n\times n}$ be a normalized Hadamard matrix. We generate d many independent and identically distributed diagonal sign matrices S_i with diagonal elements in $\{-1, 1\}$ selected uniformly at random, and set $A = [MS_1, \dots, MS_d]$. Typically $d \ge 3$ is sufficient for near complete recovery on noiseless data. In our simulations, we corrupted a fraction of the measurements $b_i = \langle a_i, x\rangle^2$ independently by setting $b_i = 0$ with probability $p_c = {}^{1}/_{50}$. All of the plotted algorithms are initialized using the initialization scheme described in [29, sect. 3].

For the l_1-regularized problem we performed tests with different values of the regularization parameter λ and found $\lambda = {}^{0.1}/_{N}$ to lead to a visually favorable recovery. When $g = \delta_{\mathbb{B}_\kappa}$, we set $\kappa = 160$ and $\kappa = 125$ for the digit and QR data, respectively. The convergence behavior in terms of $\mathcal{D}(z^k)$ (see (5.6)) is plotted in Figure 1 for a representative digit 8 image. With the above described initialization, the algorithms converge to the same cost. In our simulations SMD had the slowest performance,

[5]Using *almost sure nondegeneracy* of the fixed points of $\operatorname{prox}_\Phi^{\hat{H}}$ (cf. [1, Def. 3.5 and Lem. 3.6] and discussion therein), hence that of $T \circ \sum_i \nabla h_i$ by Lemma 3.1(ii), it can be deduced that $\mathcal{D}(z^k)$ vanishing is necessary for optimality of the limit point(s) of z^k. We, however, omit the technical details, as this behavior is anyway confirmed in the plots.

[6]https://web.stanford.edu/~hastie/ElemStatLearn/data.html.

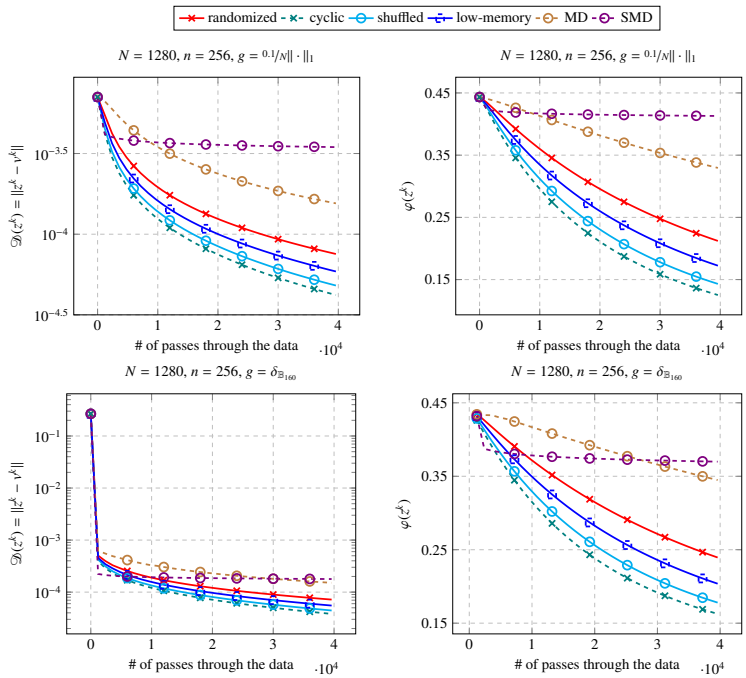
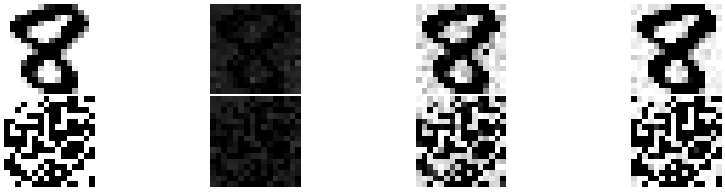


FIG. 1. *Representative convergence plots for problem* (5.2) *with squared loss on a digits image: (first row)* ℓ_1*-regularization; (second row)* ℓ_0*-norm ball constraint. The related plots for the QR code images follow a very similar trend and are therefore omitted.*



(a) original image. (b) initialization. (c) tolerance 10^{-5}. (d) tolerance 10^{-7}.

FIG. 2. *Image recovery with corrupted measurements for tolerances* $\{10^{-5}, 10^{-7}\}$*. The sparsity parameters* $\kappa = 160$ *and* $\kappa = 125$ *are used for the digit and the QR code, respectively.*

and the cyclic rule ($\mathcal{S}_2^{\mathrm{CYCL}}$) in Algorithm 1 was observed to consistently outperform all others. The low-memory Algorithm 2 has a comparable performance, almost always superior to the randomized variant. As is evident from Figure 2, despite corrupted measurements a reasonably good recovery is achieved with the l_0-norm ball. A similar recovery is observed with l_1 regularization.

In the last set of simulations we consider synthetic data. We generate a standard random Gaussian matrix $A \in \mathbb{R}^{N\times n}$ with $n = 200$, $N \in \{400, 1000\}$. The data vector a_i, $i \in [N]$, is set equal to the absolute value of the ith row of A. We also drew a random vector from $\mathcal{N}(0, I_n)$ and set the signal x equal to its absolute value. We generated the measurements according to the Poisson model $b_i \sim \mathrm{Poisson}(\langle a_i, x\rangle^2)$, $i \in [N]$, and further corrupted the measurements b_i by setting them equal to the nearest integer to the absolute value of $\|x\|^2\mathcal{N}(0,1)$ with probability $p_c = 1/10$. All methods were initialized at the same random point. We ran simulations with regularization

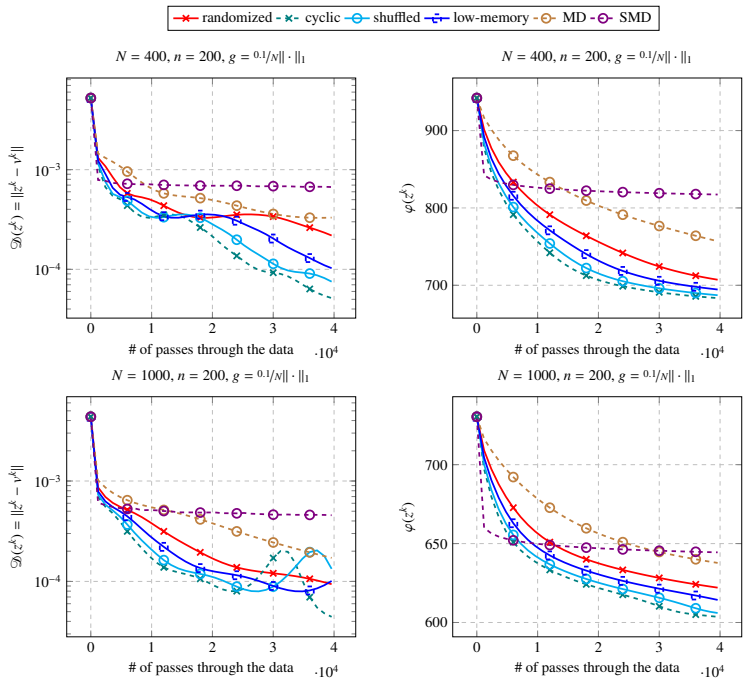


FIG. 3. *Representative convergence plots for the l_1-regularized problem with Poisson loss.*

parameter $\lambda \in \{0.01/N, 0.1/N, 1/N\}$. We only report the results for $\lambda = 0.1/N$ due to space limitations, nevertheless remarking that for other values similar plots were observed. The results are illustrated in Figure 3. Similarly to the previous experiments, SMD performed the worst, while the best results are observed for Algorithm 1 with cyclic rule (δ_2^{CYCL}). The low-memory Algorithm 2 usually outperforms the randomized variant of Algorithm 1.

6. Conclusions. A Bregman incremental aggregated method was developed that extends Finito/MISO [28, 42] to non-Lipschitz and nonconvex settings. The basic algorithm was studied under randomized and essentially cyclic sampling strategies. Furthermore, a variant with $O(n)$ memory requirements is developed that is novel even in the Euclidean case. A sure descent property established on a Bregman–Moreau envelope leads to a surprisingly simple convergence analysis. As one particularly interesting result, in the randomized setting linear convergence is established under strong convexity of the cost function without requiring convexity of the individual functions f_i or g. Future research directions include extending the analysis to the framework of the Douglas–Rachford splitting, momentum-type schemes, as well as applications to nonconvex distributed asynchronous optimization.

Appendix A. Auxiliary results.

LEMMA A.1 (equivalence between (3.1) and (P)). *Let Φ and Δ be as in* (3.1). *Then the following hold:*

(i) Cost Function: $\Phi(\boldsymbol{x}) = \varphi(x)$ *if* $\boldsymbol{x} = (x, \dots, x)$, *and* $\Phi(\boldsymbol{x}) = \infty$ *otherwise.*

(ii) Subdifferential:

$$\hat{\partial}(\Phi + \delta_{\overline{C}\times\cdots\times\overline{C}})(\boldsymbol{x}) = \left\{ \boldsymbol{v} = (v_1, \ldots, v_N) \in \mathbb{R}^{nN} \mid \sum_{i=1}^{n} v_i \in \hat{\partial}(\varphi + \delta_{\overline{C}})(x) \right\}$$

if $\boldsymbol{x} = (x, \cdots, x)$ *for some* $x \in \mathbb{R}^n$*, and is empty otherwise; the same relation still holds if the regular subdifferential* $\hat{\partial}$ *is replaced by the limiting subdifferential* ∂*.*

(iii) KL property: φ *has the KL property at* x *iff so does* Φ *at* $\boldsymbol{x} = (x, \ldots, x)$*, in which case the desingularizing functions are the same up to a positive scaling.*

(iv) Stationary points: *a point* $\boldsymbol{x}^\star$ *is stationary for problem* (3.1) *iff* $\boldsymbol{x}^\star = (x^\star, \ldots, x^\star)$ *for some* $x^\star \in \mathbb{R}^n$ *which is stationary for problem* (P)*.*

(v) Minimizers: $\boldsymbol{x}^\star$ *is a (local) minimizer of problem* (3.1) *iff* $\boldsymbol{x}^\star = (x^\star, \ldots, x^\star)$ *for some* $x^\star \in \mathbb{R}^n$ *which is a (local) minimizer for problem* (P)*; in fact,* $\inf_{\overline{C}\times\cdots\times\overline{C}} \Phi = \inf_{\overline{C}} \varphi$*.*

(vi) Level boundedness: $\varphi + \delta_{\overline{C}}$ *is level bounded iff so is* $\Phi + \delta_{\overline{C}\times\cdots\times\overline{C}}$*.*

(vii) Convexity: $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ *is convex iff so is* $\Phi : \mathbb{R}^{Nn} \to \overline{\mathbb{R}}$*.*

Proof. For notational convenience, up to possibly replacing g with $g + \delta_{\overline{C}}$ we may assume without loss of generality that $C = \mathbb{R}^n$.

A.1(i) Trivial consequence of the fact that $\operatorname{dom}\Phi \subseteq \Delta$ (the consensus set; cf. (3.1)).

A.1(ii) In light of the previous point, having $\boldsymbol{x} = (x, \ldots, x)$ for some $x \in \mathbb{R}^n$ is necessary for the nonemptiness of $\hat{\partial}\Phi(\boldsymbol{x})$. Let $\boldsymbol{x} = (x, \ldots, x)$ and $\boldsymbol{v} \in \hat{\partial}\Phi(\boldsymbol{x})$ be fixed. Then,

(A.1)
$$0 \leq \liminf_{\boldsymbol{x}\neq\boldsymbol{y}\to\boldsymbol{x}} \frac{\Phi(\boldsymbol{y}) - \Phi(\boldsymbol{x}) - \langle \boldsymbol{v}, \boldsymbol{y} - \boldsymbol{x}\rangle}{\|\boldsymbol{y} - \boldsymbol{x}\|} = \liminf_{x\neq y\to x} \frac{\varphi(y) - \varphi(x) - \langle \sum_i v_i, y - x\rangle}{\sqrt{N}\|y - x\|},$$

where the equality comes from the fact that $\operatorname{dom}\Phi \subseteq \Delta$ together with assertion A.1(i). This shows that $\sum_i v_i \in \hat{\partial}\varphi(x)$. Conversely, let $u \in \hat{\partial}\varphi(x)$ and $\boldsymbol{v} \in \mathbb{R}^{nN}$ be such that $\sum_i v_i = u$. By reading (A.1) from right to left we obtain that $\boldsymbol{v} \in \hat{\partial}\Phi(\boldsymbol{x})$. Having shown the identity of the regular subdifferential, the same claim with the limiting subdifferential follows by definition.

A.1(iii) It follows from assertion A.1(ii) that

$$\tfrac{1}{N}\operatorname{dist}(0, \partial\varphi(x))^2 = \inf_{\boldsymbol{v}\in\partial\Phi(\boldsymbol{x})} \tfrac{1}{N}\|\textstyle\sum_{i=1}^N v_i\|^2 \leq \inf_{\boldsymbol{v}\in\partial\Phi(\boldsymbol{x})} \textstyle\sum_{i=1}^N \|v_i\|^2 = \operatorname{dist}(0, \partial\Phi(\boldsymbol{x}))^2.$$

On the other hand, for any $v \in \partial\varphi(x)$ one has $(v, 0, \ldots, 0) \in \partial\Phi(\boldsymbol{x})$. Thus implying that $\operatorname{dist}(0, \partial\Phi(\boldsymbol{x}))^2 \leq \inf_{v\in\partial\varphi} \|v\|^2 = \operatorname{dist}(0, \partial\varphi(x))^2$.

A.1(iv)–A.1(vii) Directly follow from assertions A.1(i) and A.1(ii). ∎

LEMMA A.2. *Let* $\boldsymbol{\mathcal{U}} := \mathcal{U}_1 \times \cdots \times \mathcal{U}_N$ *with* $\mathcal{U}_i \subseteq \operatorname{int}\operatorname{dom} h_i$ *nonempty and convex,* $i \in [N]$*. Additionally to Assumption* I*, suppose that* g *is convex, and* h_i*,* $i \in [N]$*, is* $\ell_{h_i,\mathcal{U}_i}$*-Lipschitz differentiable and* $\mu_{h_i,\mathcal{U}_i}$*-strongly convex on* $\mathcal{U}_i$*. Then, the following hold for function* $\hat{H}$ *as in* (3.4) *with* $\gamma_i \in (0, {}^N/_{L_{f_i}})$*,* $i \in [N]$*:*

(i) $\operatorname{prox}_\Phi^{\hat{H}}$ *is Lipschitz continuous on* $\boldsymbol{\mathcal{U}}$*.*

If in addition f_i *and* h_i *are twice continuously differentiable on* $\mathcal{U}_i$*,* $i \in [N]$*, then*

(ii) $\Phi^{\hat{H}}$ *is continuously differentiable on* $\boldsymbol{\mathcal{U}}$ *with* $\nabla\Phi^{\hat{H}} = \nabla^2\hat{H} \circ (\mathrm{id} - \mathrm{prox}_{\Phi}^{\hat{H}})$.

(iii) $\mathrm{dist}(0, \partial\Phi^{\hat{H}}(\boldsymbol{x})) = \|\nabla\Phi^{\hat{H}}(\boldsymbol{x})\| \le C_{\boldsymbol{\mathcal{U}}}\|\boldsymbol{x}-\boldsymbol{z}\|$ *for any* $\boldsymbol{x} \in \boldsymbol{\mathcal{U}}$, *where* $\boldsymbol{z} = \mathrm{prox}_{\Phi}^{\hat{H}}(\boldsymbol{x})$ *and* $C_{\boldsymbol{\mathcal{U}}} = \max_i \big\{ \big(1 + \frac{\gamma_i L_{f_i}}{N}\big) \frac{\ell_{h_i,\mathcal{U}_i}}{\gamma_i} \big\}$.

Proof. It follows from Lemma 3.1(iv) and 3.1(v) that $\hat{h}_i$ is $\ell_{\hat{h}_i,\mathcal{U}_i}$-Lipschitz differentiable and $\mu_{\hat{h}_i,\mathcal{U}_i}$-strongly convex on $\mathcal{U}_i$ with $\ell_{\hat{h}_i,\mathcal{U}_i} = \big(1 + \frac{\gamma_i L_{f_i}}{N}\big) \frac{\ell_{h_i,\mathcal{U}_i}}{\gamma_i}$ and $\mu_{\hat{h}_i,\mathcal{U}_i} = \big(1 - \frac{\gamma_i L_{f_i}}{N}\big) \frac{\mu_{h_i,\mathcal{U}_i}}{\gamma_i}$. Then, $\hat{H}$ is Lipschitz differentiable and strongly convex on $\boldsymbol{\mathcal{U}}$ (with respective moduli $\ell_{\hat{H},\boldsymbol{\mathcal{U}}} = \max_i \ell_{\hat{h}_i,\mathcal{U}_i}$ and $\mu_{\hat{H},\boldsymbol{\mathcal{U}}} = \min_i \mu_{\hat{h}_i,\mathcal{U}_i}$), and therefore so is its conjugate $\hat{H}^*$ on $\hat{H}(\boldsymbol{\mathcal{U}})$ (with respective moduli $\ell_{\hat{H}^*,\boldsymbol{\mathcal{U}}} = \mu_{\hat{H},\boldsymbol{\mathcal{U}}}^{-1}$ and $\mu_{\hat{H}^*,\boldsymbol{\mathcal{U}}} = \ell_{\hat{H},\boldsymbol{\mathcal{U}}}^{-1}$). Notice that convexity of g, this being equivalent to that of G, implies that $\hat{H}(\boldsymbol{x}) + \Phi(\boldsymbol{x}) = G(\boldsymbol{x}) + \sum_{i=1}^N \frac{1}{\gamma_i} h_i(x_i)$ is strongly convex on $\boldsymbol{\mathcal{U}}$. We may thus invoke [35, Thm. 4.2] and Fact 2.3(i) to conclude that $\mathrm{prox}_{\Phi}^{\hat{H}} = \partial(\hat{H} + \Phi)^* \circ \nabla\hat{H} = \nabla(\hat{H} + \Phi)^* \circ \nabla\hat{H}$ is the composition of Lipschitz-continuous mappings on $\boldsymbol{\mathcal{U}}$, which shows assertion A.2(i). In turn, assertion A.2(ii) follows from [35, Cor. 3.1]. Finally, $\ell_{\hat{H},\boldsymbol{\mathcal{U}}}$-Lipschitz continuity of $\nabla\hat{H}$ on $\boldsymbol{\mathcal{U}}$ entails the bound $\|\nabla^2\hat{H}\| \le \ell_{\hat{H},\boldsymbol{\mathcal{U}}}$ on $\boldsymbol{\mathcal{U}}$, leading to assertion A.2(iii). ∎

LEMMA A.3. *Let* $(\alpha_k)_{k\in\mathbb{N}} \subset \mathbb{R}_+$ *be a sequence, and suppose that there exist* $c > 0$ *and* $\delta \in [1, \infty)$ *such that* $\alpha_{k+1}^{\delta} \le c(\alpha_k - \alpha_{k+1})$ *holds for every* $k \in \mathbb{N}$.

(i) *If* $\delta = 1$, *then* $(\alpha_k)_{k\in\mathbb{N}}$ *is Q-linearly convergent (to* 0*).*

(ii) *If* $\delta \in (1, \infty)$, *then there exists* $c' > 0$ *such that* $\alpha_k \le c' k^{-\frac{1}{\delta-1}}$ *holds for all* $k \in \mathbb{N}$.

B. Omitted proofs of section 4.

Proof of Theorem 4.7 *(linear convergence with randomized rule* $(\mathcal{S}_1)$*).* We will use the equivalent BC reformulation of Algorithm 3, through the identities shown in Lemma 3.2. We start by observing that $x_i^k \in \{x^{\mathrm{init}}, z^k \mid k \in \mathbb{N}\} \subseteq \mathcal{U}$ holds for any $k \in \mathbb{N}$ and $i \in [N]$, as it follows from the $\boldsymbol{x}$-update at step 3 and the fact that $u_i^k = z^k$; cf. Lemma 3.2(ii). Let $\boldsymbol{x}^\star = (x^\star, \dots, x^\star)$ be the unique minimizer of Φ (cf. Lemma A.1(v)). As shown in (4.2), denoting $\boldsymbol{v}^k := \nabla\hat{H}(\boldsymbol{x}^k) - \nabla\hat{H}(\boldsymbol{u}^k) \in \hat{\partial}\Phi(\boldsymbol{u}^k)$ we have

$$
\begin{aligned}
\Phi^{\hat{H}}(\boldsymbol{x}^k) - \min\Phi &= \Phi(\boldsymbol{u}^k) + \mathrm{D}_{\hat{H}}(\boldsymbol{u}^k, \boldsymbol{x}^k) - \Phi(\boldsymbol{x}^\star) \\
\text{Fact 2.3(ii)} &= \Phi(\boldsymbol{u}^k) + \mathrm{D}_{\hat{H}}(\boldsymbol{x}^\star, \boldsymbol{x}^k) - \mathrm{D}_{\hat{H}}(\boldsymbol{x}^\star, \boldsymbol{u}^k) + \langle \boldsymbol{v}^k, \boldsymbol{x}^\star - \boldsymbol{u}^k\rangle - \Phi(\boldsymbol{x}^\star) \\
\text{Lemma A.1(i)} &= \varphi(u^k) + \mathrm{D}_{\hat{H}}(\boldsymbol{x}^\star, \boldsymbol{x}^k) - \mathrm{D}_{\hat{H}}(\boldsymbol{x}^\star, \boldsymbol{u}^k) + \langle \textstyle\sum_{i=1}^N v_i^k, x^\star - u^k\rangle - \varphi(x^\star) \\
\text{Lemma A.1(ii)} &\le \mathrm{D}_{\hat{H}}(\boldsymbol{x}^\star, \boldsymbol{x}^k) - \mathrm{D}_{\hat{H}}(\boldsymbol{x}^\star, \boldsymbol{u}^k) - \tfrac{\mu_\varphi}{2}\|u^k - x^\star\|^2 \\
\text{Fact 2.3(ii)} &= \mathrm{D}_{\hat{H}}(\boldsymbol{u}^k, \boldsymbol{x}^k) + \langle \nabla\hat{H}(\boldsymbol{u}^k) - \nabla\hat{H}(\boldsymbol{x}^k), \boldsymbol{x}^\star - \boldsymbol{u}^k\rangle - \tfrac{\mu_\varphi}{2}\|u^k - x^\star\|^2 \\
&= \mathrm{D}_{\hat{H}}(\boldsymbol{u}^k, \boldsymbol{x}^k) + \textstyle\sum_{i=1}^N \langle \nabla\hat{h}_i(u^k) - \nabla\hat{h}_i(x_i^k), x^\star - u^k\rangle - \tfrac{\mu_\varphi}{2}\|u^k - x^\star\|^2, \qquad \text{(B.1)}
\end{aligned}
$$

where the inequality follows from strong convexity of φ. For any $\varepsilon_i > 0$, $i \in [N]$, one has

$$
\langle \nabla\hat{h}_i(u^k) - \nabla\hat{h}_i(x_i^k), x^\star - u^k\rangle \le \tfrac{\varepsilon_i}{2}\|x^\star - u^k\|^2 + \tfrac{1}{2\varepsilon_i}\|\nabla\hat{h}_i(u^k) - \nabla\hat{h}_i(x_i^k)\|^2.
$$

Plugged into (B.1) with $\varepsilon_i > 0$ such that $\sum_{i=1}^N \varepsilon_i = \mu_\varphi$, so as to cancel the square norm therein,

$$\begin{aligned}\Phi^{\hat{H}}(\boldsymbol{x}^k) - \min\Phi &\le \mathrm{D}_{\hat{H}}(\boldsymbol{u}^k, \boldsymbol{x}^k) + \textstyle\sum_{i=1}^N \frac{1}{2\varepsilon_i}\|\nabla\hat{h}_i(u^k) - \nabla\hat{h}_i(x_i^k)\|^2 \\ &\overset{\text{Fact 2.3(vii)}}{\le} \mathrm{D}_{\hat{H}}(\boldsymbol{u}^k, \boldsymbol{x}^k) + \textstyle\sum_{i=1}^N \frac{\ell_{\hat{h}_i,\mathcal{U}}}{\varepsilon_i}\mathrm{D}_{\hat{h}_i}(u^k, x_i^k) \\ &= \textstyle\sum_{i=1}^N \left(1 + \frac{\ell_{\hat{h}_i,\mathcal{U}}}{\varepsilon_i}\right)\mathrm{D}_{\hat{h}_i}(u^k, x_i^k),\end{aligned}$$

where $\ell_{\hat{h}_i,\mathcal{U}}$ is a Lipschitz constant for $\nabla\hat{h}_i$ on $\mathcal{U}$ as in Lemma 3.1(iv). By choosing $\varepsilon_i = \ell_{\hat{h}_i,\mathcal{U}}/\kappa$ with $\kappa := \frac{\sum_j \ell_{\hat{h}_j,\mathcal{U}}}{\mu_\varphi}$ (which satisfies $\sum_{i=1}^N \varepsilon_i = \mu_\varphi$), we obtain

$$\Phi^{\hat{H}}(\boldsymbol{x}^k) - \min\Phi \le (1+\kappa)\textstyle\sum_{i=1}^N \mathrm{D}_{\hat{h}_i}(u^k, x_i^k) = (1+\kappa)\,\mathrm{D}_{\hat{H}}(\boldsymbol{u}^k, \boldsymbol{x}^k).$$

Combining this with (4.1) (recall the equivalences in Lemma 3.2) yields

$$\mathbb{E}_k\left[\Phi^{\hat{H}}(\boldsymbol{x}^{k+1}) - \min\Phi\right] \le \left(1 - \tfrac{p_{\min}}{1+\kappa}\right)\left(\Phi^{\hat{H}}(\boldsymbol{x}^k) - \min\Phi\right) = (1 - c_{\mathcal{U}})\left(\Phi^{\hat{H}}(\boldsymbol{x}^k) - \min\Phi\right),$$

where $c_{\mathcal{U}}$ as in the statement is obtained by using the estimates of Lemma 3.1(iv) for the moduli $\ell_{\hat{h}_i,\mathcal{U}}$ appearing in the constant κ (since $\hat{h}_i = h_i/\gamma_i - f_i/N$ and $\ell_{f_i,\mathcal{U}} := \ell_{h_i,\mathcal{U}} L_{f_i}$ is a Lipschitz modulus for ∇f_i on $\mathcal{U}$ by Fact 2.5(iii), one has $\sigma_{f_i,\mathcal{U}} \ge -\ell_{f_i,\mathcal{U}}$ and $\ell_{\hat{h}_i,\mathcal{U}} \le \ell_{h_i,\mathcal{U}}/\gamma_i - \sigma_{f_i,\mathcal{U}}/N$). This concludes the proof of (4.8). In turn, (4.9) follows by taking unconditional expectation and using the fact that $\varphi(z^k) = \Phi(\boldsymbol{u}^k) \le \Phi^{\hat{H}}(\boldsymbol{x}^k)$, owing to Lemma 3.2(v) and Fact 2.7(i). ∎

***Proof of Theorem* 4.9**. (*subsequential convergence with essentially cyclic rule* ($\mathcal{S}_2$)). We use the simpler setting of Algorithm 3 owing to the equivalence between the algorithms shown in Lemma 3.2. Note that if φ is level bounded, then by Lemma 4.2(iv) a bounded convex set $\boldsymbol{\mathcal{U}}$ exists that contains $(\boldsymbol{x}^k)_{k\in\mathbb{N}}$ and $(\boldsymbol{u}^k)_{k\in\mathbb{N}}$. Local Lipschitz differentiability and local strong convexity thus imply through Lemmas 3.1(iv) and 3.1(v) that $\hat{H}$ is $\mu_{\hat{H},\boldsymbol{\mathcal{U}}}$-strongly convex and $\ell_{\hat{H},\boldsymbol{\mathcal{U}}}$-Lipschitz differentiable on $\boldsymbol{\mathcal{U}}$ for some constants $\ell_{\hat{H},\boldsymbol{\mathcal{U}}} \ge \mu_{\hat{H},\boldsymbol{\mathcal{U}}} > 0$. If, instead, those properties hold globally, then the same claims can hold with $\boldsymbol{\mathcal{U}} = \mathbb{R}^{nN}$. Therefore, since g is assumed to be convex, either one among Theorems 4.9(A) and (B) is enough to invoke Lemma A.2(i), implying that $\mathrm{prox}_\Phi^{\hat{H}}$ is λ-Lipschitz continuous on $\boldsymbol{\mathcal{U}}$ for some $\lambda > 0$.

Since all indices are updated at least once every T iterations, one has that

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

is well defined for each index $i \in [N]$ and $\nu \in \mathbb{N}$. In other words, since i is sampled at iteration $\nu T + t_\nu(i) - 1$ and not in any one between νT and $\nu T + t_\nu(i) - 2$, it holds that

(B.2)
$$x_i^{\nu T} = x_i^{\nu T+1} = \cdots = x_i^{\nu T + t_\nu(i) - 1} \quad \text{and} \quad x_i^{\nu T + t_\nu(i)} = u^{\nu T + t_\nu(i) - 1} \quad \forall i \in [N], \nu \in \mathbb{N},$$

recalling $\boldsymbol{u}^k = (u^k, \dots, u^k)$. We now proceed to establish a descent inequality for $\Phi^{\hat{H}}$ holding every interval of T iterations. First,

$$\begin{aligned}\Phi^{\hat{H}}(\boldsymbol{x}^{T(\nu+1)}) - \Phi^{\hat{H}}(\boldsymbol{x}^{\nu T}) &= \textstyle\sum_{\tau=1}^{T}\Big(\Phi^{\hat{H}}(\boldsymbol{x}^{\nu T+\tau}) - \Phi^{\hat{H}}(\boldsymbol{x}^{\nu T+\tau-1})\Big)\\ &\overset{4.2(\mathrm{i})}{\le} -\textstyle\sum_{\tau=1}^{T} \mathrm{D}_{\hat{H}}(\boldsymbol{x}^{\nu T+\tau}, \boldsymbol{x}^{\nu T+\tau-1})\\ &\le -\mathrm{D}_{\hat{H}}(\boldsymbol{x}^{\nu T+t}, \boldsymbol{x}^{\nu T+t-1}) \overset{2.3}{\le} -\tfrac{\mu_{\hat{H},\boldsymbol{u}}}{2}\|\boldsymbol{x}^{\nu T+t} - \boldsymbol{x}^{\nu T+t-1}\|^2 \end{aligned} \tag{B.3}$$

holds for all $t \in [T]$. Next, for every $i \in [N]$ it holds that

$$\begin{aligned}\|u^{\nu T+t_\nu(i)-1} - u^{\nu T}\| &= \tfrac{1}{\sqrt{N}}\|\boldsymbol{u}^{\nu T+t_\nu(i)-1} - \boldsymbol{u}^{\nu T}\| \le \tfrac{\lambda}{\sqrt{N}}\|\boldsymbol{x}^{\nu T+t_\nu(i)-1} - \boldsymbol{x}^{\nu T}\|\\ &\le \tfrac{\lambda}{\sqrt{N}}\textstyle\sum_{\tau=1}^{t_\nu(i)-1}\|\boldsymbol{x}^{\nu T+\tau} - \boldsymbol{x}^{\nu T+\tau-1}\|\\ &\overset{(\mathrm{B.3})}{\le} \tfrac{\lambda T}{\sqrt{N}}\sqrt{\tfrac{2}{\mu_{\hat{H},\boldsymbol{u}}}\big(\Phi^{\hat{H}}(\boldsymbol{x}^{\nu T}) - \Phi^{\hat{H}}(\boldsymbol{x}^{T(\nu+1)})\big)}, \end{aligned} \tag{B.4}$$

where the first inequality uses the λ-Lipschitz continuity of $\mathrm{prox}_{\Phi}^{\hat{H}}$, the second one the triangular inequality, and the last one the fact that $t_\nu(i) \le T$. For all $i \in [N]$, it follows from Fact 2.3(vii) and the triangular inequality that

$$\begin{aligned}\|x_i^{\nu T} - u^{\nu T}\| &\le \|x_i^{\nu T} - u^{\nu T+t_\nu(i)-1}\| + \|u^{\nu T+t_\nu(i)-1} - u^{\nu T}\|\\ &\overset{(\mathrm{B.2})}{=} \|x_i^{\nu T+t_\nu(i)-1} - x_i^{\nu T+t_\nu(i)}\| + \|u^{\nu T+t_\nu(i)-1} - u^{\nu T}\|\\ &\overset{(\mathrm{B.3})\ (\mathrm{B.4})}{\le} \sqrt{\tfrac{2}{\mu_{\hat{H},\boldsymbol{u}}}}\Big(1 + \tfrac{\lambda T}{\sqrt{N}}\Big)\sqrt{\Phi^{\hat{H}}(\boldsymbol{x}^{\nu T}) - \Phi^{\hat{H}}(\boldsymbol{x}^{T(\nu+1)})}. \end{aligned} \tag{B.5}$$

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

$$\Phi^{\hat{H}}(\boldsymbol{x}^{T(\nu+1)}) - \Phi^{\hat{H}}(\boldsymbol{x}^{\nu T}) \le -\tfrac{\mu_{\hat{H},\boldsymbol{u}}}{2(1+\lambda T/\sqrt{N})^2}\|\boldsymbol{x}^{\nu T} - \boldsymbol{u}^{\nu T}\|^2. \tag{B.6}$$

Since by $(\mathcal{S}_2)$ in any interval of length T every index is updated at least once, by suitably shifting, for every $t \in [T]$ the same holds for the sequences $(\boldsymbol{x}^{\nu T+t})_{\nu\in\mathbb{N}}$ and $(\boldsymbol{u}^{\nu T+t})_{\nu\in\mathbb{N}}$. Thus,

$$\begin{aligned}\Phi^{\hat{H}}(\boldsymbol{x}^{k+T}) - \Phi^{\hat{H}}(\boldsymbol{x}^k) &\le -\tfrac{\mu_{\hat{H},\boldsymbol{u}}}{2(1+\lambda T/\sqrt{N})^2}\|\boldsymbol{x}^k - \boldsymbol{u}^k\|^2\\ &\overset{\text{Fact 2.3(vii)}}{\le} -\tfrac{\mu_{\hat{H},\boldsymbol{u}}}{L_{\hat{H},\boldsymbol{u}}(1+\lambda T/\sqrt{N})^2}\mathrm{D}_{\hat{H}}(\boldsymbol{u}^k, \boldsymbol{x}^k) \quad \forall k \in \mathbb{N}. \end{aligned} \tag{B.7}$$

By telescoping the inequality and using the fact that the envelope is lower bounded (Lemma 4.1(ii) and Assumption I.A3), all the assertions follow from Lemma 4.4. ∎

***Proof of Lemma* 4.12** *(Algorithm* 2 *as an instance of Algorithm* 1*).* Note that, in Algorithm 1, $\sum_{i=1}^N s_i^0 = \tilde{s}^0$. By induction, suppose that $\sum_{i=1}^N s_i^k = \tilde{s}^k$ for some $k \ge 0$. Then,

$$\tilde{s}^{k+1} = \tilde{s}^k + \textstyle\sum_{i\in\mathcal{J}^{k+1}}(s_i^{k+1} - s_i^k) = \sum_{i=1}^N s_i^k + \sum_{i\in\mathcal{J}^{k+1}}(s_i^{k+1} - s_i^k) = \sum_{i=1}^N s_i^{k+1}, \tag{B.8}$$

where the first equality follows by step 3, the second one from the induction hypothesis, and the last one from the fact that $s_i^{k+1} = s_i^k$ for $i \notin \mathcal{J}^{k+1}$ as in step 3.

In what follows, let $\mathcal{N}_{\text{full}} := \{k \mid \mathcal{K}^k = \emptyset\}$, and observe that $k \in \mathcal{N}_{\text{full}}$ iff the if statement at step 2 is true. In particular, it follows from step 3 that

$$\mathcal{N}_{\text{full}} \subseteq \{k \mid \mathcal{J}_{\text{LM}}^{k+1} = [N]\}. \tag{B.9}$$

We now proceed by induction on k to establish that $(z_{\text{LM}}^k)_{k\in\mathbb{N}}$ and $(\tilde{s}_{\text{LM}}^k)_{k\in\mathbb{N}}$ are sequences generated by Algorithm 1 with index sets being chosen as

$$\mathcal{J}^{k+1} := \mathcal{J}_{\text{LM}}^{k+1} \quad \forall k \in \mathbb{N}. \tag{B.10}$$

The claim is true for $k = 0$; suppose it holds up to iteration $k \geq 0$. We consider two cases:

Case 1*:* $k \in \mathcal{N}_{\text{full}}$. We have

$$\tilde{s}_{\text{LM}}^{k+1} = \textstyle\sum_{i=1}^N \nabla\hat{h}_i(z_{\text{LM}}^k) = \sum_{i=1}^N \nabla\hat{h}_i(z^k) = \sum_{i=1}^N s_i^{k+1} = \tilde{s}^{k+1}, \tag{B.11}$$

where the first equality follows by step 5, the second one by induction, the third one by the fact that $\mathcal{J}^{k+1} = \mathcal{J}_{\text{LM}}^{k+1} = [N]$ (cf. (B.9) and (B.10)), and the last one from (B.8). It follows that the minimization problems defining z^{k+1} and z_{LM}^{k+1} (at step 1 of Algorithm 1 and step 1 of Algorithm 2, respectively) coincide, thus ensuring that $z^{k+1} = z_{\text{LM}}^{k+1}$ is a feasible update for Algorithm 1.

Case 2*:* $k \notin \mathcal{N}_{\text{full}}$. Let $t(k) := \max\{t \leq k \mid t \in \mathcal{N}_{\text{full}}\}$ be the last iteration before k at which the condition at step 2 holds, so that, according to steps 4 and 9, $\tilde{z}_{\text{LM}}^k = z_{\text{LM}}^{t(k)}$. We have

$$\begin{aligned}
\tilde{s}_{\text{LM}}^{k+1} &= \tilde{s}_{\text{LM}}^k + \sum_{i\in\mathcal{J}_{\text{LM}}^{k+1}} \Big[\nabla\hat{h}_i(z_{\text{LM}}^k) - \nabla\hat{h}_i(\tilde{z}_{\text{LM}}^k)\Big] && \text{(step 10)}\\
&= \tilde{s}^k + \sum_{i\in\mathcal{J}^{k+1}} \Big[\nabla\hat{h}_i(z^k) - \nabla\hat{h}_i(z^{t(k)})\Big] && \text{(induction and (B.10))}\\
&= \tilde{s}^k + \sum_{i\in\mathcal{J}^{k+1}} \Big[s_i^{k+1} - s_i^{t(k)+1}\Big] && \text{(step 3, (B.9), and (B.10))},
\end{aligned} \tag{B.12}$$

where (B.9) was used to infer that $s_i^{t(k)+1} = \nabla\hat{h}_i(z^{t(k)})$ for all $i \in [N]$ (hence for all $i \in \mathcal{J}^{k+1}$). To conclude, note that the selection rule for $\mathcal{J}_{\text{LM}}^{k+1}$ (cf. steps 7 and 8) ensures through (B.10) that the index sets $\mathcal{J}^{t(k)+1}, \mathcal{J}^{t(k)+2}, \ldots, \mathcal{J}^{k+1}$ are all pairwise disjoint, hence that $s_i^{t(k)} = s_i^{t(k)+1} = \cdots = s_i^k$ for all $i \in \mathcal{J}^{k+1}$, as is apparent from step 3. We may thus replace $s_i^{t(k)+1}$ with s_i^k in (B.12) to obtain the $\tilde{s}$-update of step 3, and conclude that $\tilde{s}_{\text{LM}}^{k+1} = \tilde{s}^{k+1}$. As discussed in the last part of Case 1, this in turn shows that $z^{k+1} = z_{\text{LM}}^{k+1}$ is a feasible update for Algorithm 1. ∎

REFERENCES

[1] M. Ahookhosh, A. Themelis, and P. Patrinos, *A Bregman forward-backward linesearch algorithm for nonconvex composite optimization: Superlinear convergence to nonisolated local minima*, SIAM J. Optim., 31 (2021), pp. 653–685.

[2] Z. Allen-Zhu and Y. Yuan, *Improved SVRG for non-strongly-convex or sum-of-non-convex objectives*, in Int. Conf. Mach. Learn., 2016, pp. 1080–1089.

[3] H. Attouch and J. Bolte, *On the convergence of the proximal algorithm for nonsmooth functions involving analytic features*, Math. Program., 116 (2009), pp. 5–16.

[4] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran, *Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality*, Math. Oper. Res., 35 (2010), pp. 438–457.

[5] H. Attouch, J. Bolte, and B. F. Svaiter, *Convergence of descent methods for semi-algebraic and tame problems: Proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods*, Math. Program., 137 (2013), pp. 91–129.

[6] H. H. Bauschke, J. Bolte, J. Chen, M. Teboulle, and X. Wang, *On linear convergence of non-Euclidean gradient methods without strong convexity and Lipschitz gradient continuity*, J. Optim. Theory Appl., 182 (2019), pp. 1068–1087.

[7] H. H. Bauschke, J. Bolte, and M. Teboulle, *A descent lemma beyond Lipschitz gradient continuity: First-order methods revisited and applications*, Math. Oper. Res., 42 (2017), pp. 330–348.

[8] H. H. Bauschke and J. M. Borwein, *Legendre functions and the method of random Bregman projections*, J. Convex Anal., 4 (1997), pp. 27–67.

[9] H. H. Bauschke, J. M. Borwein, and P. L. Combettes, *Essential smoothness, essential strict convexity, and Legendre functions in Banach spaces*, Commun. Contemp. Math., 03 (2001), pp. 615–647.

[10] A. Beck, *First-Order Methods in Optimization*, SIAM, Philadelphia, 2017.

[11] A. Beck and M. Teboulle, *Mirror descent and nonlinear projected subgradient methods for convex optimization*, Oper. Res. Lett., 31 (2003), pp. 167–175.

[12] A. Beck and L. Tetruashvili, *On the convergence of block coordinate descent type methods*, SIAM J. Optim., 23 (2013), pp. 2037–2060.

[13] D. P. Bertsekas, *Incremental proximal methods for large scale convex optimization*, Math. Program., 129 (2011), pp. 163–195.

[14] D. P. Bertsekas, *Convex Optimization Theory*, Athena Scientific, Belmont, MA, 2009.

[15] D. P. Bertsekas and J. N. Tsitsiklis, *Gradient convergence in gradient methods with errors*, SIAM J. Optim., 10 (2000), pp. 627–642.

[16] D. Blatt, A. O. Hero, and H. Gauchman, *A convergent incremental gradient method with a constant step size*, SIAM J. Optim., 18 (2007), pp. 29–51.

[17] J. Bolte, A. Daniilidis, and A. Lewis, *The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems*, SIAM J. Optim., 17 (2007), pp. 1205–1223.

[18] J. Bolte, A. Daniilidis, A. Lewis, and M. Shiota, *Clarke subgradients of stratifiable functions*, SIAM J. Optim., 18 (2007), pp. 556–572.

[19] J. Bolte, S. Sabach, and M. Teboulle, *Proximal alternating linearized minimization for nonconvex and nonsmooth problems*, Math. Program., 146 (2014), pp. 459–494.

[20] J. Bolte, S. Sabach, M. Teboulle, and Y. Vaisbourd, *First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems*, SIAM J. Optim., 28 (2018), pp. 2131–2151.

[21] E. J. Candes, X. Li, and M. Soltanolkotabi, *Phase retrieval via Wirtinger flow: Theory and algorithms*, IEEE Trans. Inform. Theory, 61 (2015), pp. 1985–2007.

[22] G. Chen and M. Teboulle, *Convergence analysis of a proximal-like minimization algorithm using Bregman functions*, SIAM J. Optim., 3 (1993), pp. 538–543.

[23] Y. Chen and E. Candes, *Solving random quadratic systems of equations is nearly as easy as solving linear systems*, in Advances in Neural Information Processing Systems 28, Curran Associates, Red Hook, NY, 2015, pp. 739–747.

[24] Y. T. Chow, T. Wu, and W. Yin, *Cyclic coordinate-update algorithms for fixed-point problems: Analysis and applications*, SIAM J. Sci. Comput., 39 (2017), pp. A1280–A1300.

[25] D. Davis, D. Drusvyatskiy, and K. J. MacPhee, *Stochastic Model-based Minimization under High-order Growth*, preprint, arXiv:1807.00255, 2018.

[26] D. Davis, D. Drusvyatskiy, and C. Paquette, *The nonsmooth landscape of phase retrieval*, IMA J. Numer. Anal., 40 (2020), pp. 2652–2695.

[27] A. Defazio, F. Bach, and S. Lacoste-Julien, *SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives*, in Advances in Neural Information Processing Systems 27, Curran Associates, Red Hook, NY, 2014, pp. 1646–1654.

[28] A. Defazio and J. Domke, *Finito: A faster, permutable incremental gradient method for big data problems*, Proc. Mach. Learn. Res. (PMLR), 32 (2014), pp. 1125–1133.

[29] J. C. Duchi and F. Ruan, *Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval*, Inf. Inference, 8 (2019), pp. 471–529.

[30] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*, Vol. 1, Springer Ser. Statist. New York, 2001.

[31] T. Gao, S. Lu, J. Liu, and C. Chu, *Randomized Bregman Coordinate Descent Methods for Non-Lipschitz Optimization*, preprint, arXiv:2001.05202, 2020.

[32] F. Hanzely and P. Richtárik, *Fastest rates for stochastic mirror descent methods*, Comput. Optim. Appl., 79 (2021), pp. 717–766.

[33] M. Hong, X. Wang, M. Razaviyayn, and Z.-Q. Luo, *Iteration complexity analysis of block coordinate descent methods*, Math. Program., 163 (2017), pp. 85–114.

[34] R. Johnson and T. Zhang, *Accelerating stochastic gradient descent using predictive variance reduction*, in Advances in Neural Information Processing Systems 26, Curran Associates, Red Hook, NY, 2013, pp. 315–323.

[35] C. Kan and W. Song, *The Moreau envelope function and proximal mapping in the sense of the Bregman distance*, Nonlinear Anal., 75 (2012), pp. 1385–1399.

[36] K. Kurdyka, *On gradients of functions definable in o-minimal structures*, Ann. Institut Fourier, 48 (1998), pp. 769–783.

[37] P. Latafat, A. Themelis, and P. Patrinos, *Block-coordinate and incremental aggregated proximal gradient methods for nonsmooth nonconvex problems*, Math. Program., 193 (2021), pp. 195–224.

[38] S. Łojasiewicz, *Sur la géométrie semi- et sous- analytique*, Ann. Institut Fourier, 43 (1993), pp. 1575–1595.

[39] H. Lu, *"Relative continuity" for non-Lipschitz nonsmooth convex optimization using stochastic (or deterministic) mirror descent*, INFORMS J. Optim., 1 (2019), pp. 288–303.

[40] H. Lu, R. M. Freund, and Y. Nesterov, *Relatively smooth convex optimization by first-order methods, and applications*, SIAM J. Optim., 28 (2018), pp. 333–354.

[41] D. R. Luke, J. V. Burke, and R. G. Lyon, *Optical wavefront reconstruction: Theory and numerical methods*, SIAM Rev., 44 (2002), pp. 169–224.

[42] J. Mairal, *Incremental majorization-minimization optimization with application to large-scale machine learning*, SIAM J. Optim., 25 (2015), pp. 829–855.

[43] C. A. Metzler, M. K. Sharma, S. Nagesh, R. G. Baraniuk, O. Cossairt, and A. Veeraraghavan, *Coherent inverse scattering via transmission matrices: Efficient phase retrieval algorithms and a public dataset*, in Proceedings of the 2017 IEEE International Conference Computational Photography (ICCP), IEEE, Piscatway, NJ, 2017, pp. 1–16.

[44] A. Mokhtari, M. Gúrbüzbalaban, and A. Ribeiro, *Surpassing gradient descent provably: A cyclic incremental method with linear convergence rate*, SIAM J. Optim., 28 (2018), pp. 1420–1447.

[45] A. Nedić and S. Lee, *On stochastic subgradient mirror-descent algorithm with weighted averaging*, SIAM J. Optim., 24 (2014), pp. 84–107.

[46] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, *Robust stochastic approximation approach to stochastic programming*, SIAM J. Optim., 19 (2009), pp. 1574–1609.

[47] Yu. Nesterov, *Efficiency of coordinate descent methods on huge-scale optimization problems*, SIAM J. Optim., 22 (2012), pp. 341–362.

[48] L. M. Nguyen, J. Liu, K. Scheinberg, and M. Takáč, *SARAH: A novel method for machine learning problems using stochastic recursive gradient*, in Proc. Mach. Learn. Res. (PMLR), 70 (2017), pp. 2613–2621.

[49] P. Ochs, J. Fadili, and T. Brox, *Non-smooth non-convex Bregman minimization: Unification and new algorithms*, J. Optim. Theory Appl., 181 (2019), pp. 244–278.

[50] X. Qian, A. Sailanbayev, K. Mishchenko, and P. Richtárik, *MISO is Making a Comeback with Better Proofs and Rates*, preprint, arXiv:1906.01474, https://arxiv.org/abs/1906.01474 (2019).

[51] H. Robbins and D. Siegmund, *A convergence theorem for non negative almost supermartingales and some applications*, in Herbert Robbins Selected Papers, Springer, New York, 1985, pp. 111–135.

[52] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[53] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Grundlehren Math. Wiss. 317, Springer, Berlin, 2009.

[54] M. Schmidt, N. Le Roux, and F. Bach, *Minimizing finite sums with the stochastic average gradient*, Math. Program., 162 (2017), pp. 83–112.

[55] Y. Shechtman, Y. C. Eldar, O. Cohen, H. N. Chapman, J. Miao, and M. Segev, *Phase retrieval with application to optical imaging: A contemporary overview*, IEEE Signal Process. Mag., 32 (2015), pp. 87–109.

[56] M. V. Solodov and B. F. Svaiter, *An inexact hybrid generalized proximal point algorithm and some new results on the theory of Bregman functions*, Math. Oper. Res., 25 (2000), pp. 214–230.

[57] M. R. Spiegel, *Mathematical Handbook of Formulas and Tables*, McGraw-Hill, New York, 1999.

[58] J. Sun, Q. Qu, and J. Wright, *A geometric analysis of phase retrieval*, Found. Comput. Math., 18 (2018), pp. 1131–1198.

[59] M. Teboulle, *A simplified view of first order methods for optimization*, Math. Program., 170 (2018), pp. 67–96.

[60] P. Tseng, *Convergence of a block coordinate descent method for nondifferentiable minimization*, J. Optim. Theory Appl., 109 (2001), pp. 475–494.

[61] P. Tseng, *On Accelerated Proximal Gradient Methods for Convex-Concave Optimization*, manuscript.

[62] P. Tseng and D. P. Bertsekas, *Relaxation methods for problems with strictly convex separable costs and linear constraints*, Math. Program., 38 (1987), pp. 303–321.

[63] P. Tseng and S. Yun, *A coordinate gradient descent method for nonsmooth separable minimization*, Math. Program., 117 (2009), pp. 387–423.

[64] P. Tseng and S. Yun, *Incrementally updated gradient methods for constrained and regularized optimization*, J. Optim. Theory Appl., 160 (2014), pp. 832–853.

[65] N. D. Vanli, M. Gürbüzbalaban, and A. Ozdaglar, *Global convergence rate of proximal incremental aggregated gradient methods*, SIAM J. Optim., 28 (2018), pp. 1282–1300.

[66] L. Xiao and T. Zhang, *A proximal stochastic gradient method with progressive variance reduction*, SIAM J. Optim., 24 (2014), pp. 2057–2075.

[67] Y. Xu and W. Yin, *A globally convergent algorithm for nonconvex optimization based on block coordinate update*, J. Sci. Comput., 72 (2017), pp. 700–734.

[68] P. Yu, G. Li, and T. K. Pong, *Deducing Kurdyka-Łojasiewicz Exponent via Inf-projection*, preprint, arXiv:1902.03635, https://link.springer.com/article/10.1007/s10208-021-09528-6 (2019).

[69] H. Zhang, Y. Chi, and Y. Liang, *Provable non-convex phase retrieval with outliers: Median truncated Wirtinger flow*, Proc. Mach. Learn. Res., 48 (2016), pp. 1022–1031.

[70] H. Zhang, Y.-H. Dai, L. Guo, and W. Peng, *Proximal-like incremental aggregated gradient method with linear convergence under Bregman distance growth conditions*, Math. Oper. Res., 46 (2021), pp. 61–81.

[71] S. Zhang and N. He, *On the Convergence Rate of Stochastic Mirror Descent for Nonsmooth Nonconvex Optimization*, preprint, arXiv:1806.04781, 2018.