SMART: The Stochastic Monotone Aggregated Root-Finding Algorithm

Damek Davis

January 6, 2016

Abstract We introduce the Stochastic Monotone Aggregated Root-Finding (SMART) algorithm, a new randomized operator-splitting scheme for finding roots of finite sums of operators. These algorithms are similar to the growing class of incremental aggregated gradient algorithms, which minimize finite sums of functions; the difference is that we replace gradients of functions with black-boxes called operators, which represent subproblems to be solved during the algorithm. By replacing gradients with operators, we increase our modeling power, and we simplify the application and analysis of the resulting algorithms. The operator point of view also makes it easy to extend our algorithms to allow arbitrary sampling and updating of blocks of coordinates throughout the algorithm. Implementing and running an algorithm like this on a computing cluster can be slow if we force all computing nodes to be synced up at all times. To take better advantage of parallelism, we allow computing nodes to delay updates and break synchronization.

This paper has several technical and practical contributions. We prove the weak, almost sure convergence of a new class of randomized operator-splitting schemes in separable Hilbert spaces; we prove that this class of algorithms convergences linearly in expectation when a weak regularity property holds; we highlight connections to other algorithms; and we introduce a few new algorithms for large-scale optimization.

1 Introduction

The guiding problems in optimization are evolving. While all optimization problems can be reduced to minimizing functions over sets, prototypical optimization problems, like $\min_{x \in C} f(x)$, hide structure that is found throughout modern applications, and this structure is useful for designing algorithms for large-scale problems (e.g., problems with gigabytes to terabytes of data). Among these structures, the finite sum is the most pervasive:

$$\text{minimize}_{x \in C} \frac{1}{n} \sum_{i=1}^{n} f_i(x). \tag{1.1}$$

D. Davis
Department of Mathematics, University of California, Los Angeles
Los Angeles, CA 90025, USA
E-mail: damek@math.ucla.edu

Large sums ($n \gg 0$) are common in statistical and machine learning, where the functions f_i often, but not always, correspond one-to-one with points in a dataset. When the f_i are summed together, the minimization problem (1.1) grows in complexity, but the prevailing and realistic assumption in applications is that it is drastically simpler (in terms memory or computational complexity) to perform operations, like differentiation, on the f_i than it is to perform the same operations on the entire sum.

Large sums, like (1.1), come from large datasets, even low dimensional ones. But modern applications often involve high dimensional datasets. When the decision variable $x \in \mathbb{R}^m$ is high-dimensional ($m \gg 0$), the prevailing and realistic assumption in applications is that it is drastically simpler to compute partial derivatives of the f_i, or other componentwise quantities, than it is to compute full derivatives of the f_i.

These two structural assumptions, and the host of algorithms that adopt them, have led to big improvements in large-scale algorithm design (for examples, see [28,17,33,35,34,16,20,15,25,29,9,5,8,36,4, 27,6,31]). And we do not have to look far to find problems for which these assumptions make sense. A simple problem with $n \gg 0$ and $m \gg 0$ comes from regularized least squares problems with matrix $A = \left[A_1^T, \ldots, A_n^T\right]^T \in \mathbb{R}^{n \times m}$, vector $b \in \mathbb{R}^n$, and regularizer induced by $K \in \mathbb{R}^{m \times m}$:

$$\underset{x \in \mathbb{R}^m}{\text{minimize}} \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{2} \left\| A_i \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix} - b_i \right\|^2 + \frac{1}{2} \langle Kx, x \rangle \right). \tag{1.2}$$

The two prevailing assumptions are clearly satisfied for this example when each row of A_i is sparse: each term in the sum is simple and partial derivatives of the terms are easier to compute than full derivatives. Absent other special structure in the matrices A and K (like bandlimitedness), a conceptually simple operation like differentiating the objective function requires $O(\max\{mn, m^2\})$ operations. With problems of this scale, full gradient computations cannot lie at the heart of a practical iterative algorithm for solving (1.2).

The algorithms that have sprung from these structural assumptions all look alike. They all form a sequence of points $\{x^k\}_{k \in \mathbb{N}}$ that converge (in some appropriate sense of the word) to a minimizer of (1.1). They all make a series of choices at each time step: given x^k

1. choose (randomly or otherwise) indices $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, m\}$;
2. get x^{k+1} (in some undetermined way) from x^k and $\nabla_j f_i(x^k)$ by only changing a single component of x^k (possibly using some combination of the previous terms $\{x^l\}_{l \in \{0, \ldots, k\}}$).

At this moment, the differences between these algorithms is immaterial. What matters is the contrast between these algorithms and the overwhelmingly costlier algorithms that form sequences $\{x^k\}_{k \in \mathbb{N}}$ by getting x^{k+1} from x^k and $\nabla \left[\sum_{i=1}^{n} f_i\right](x^k)$; we simply cannot afford to compute these full gradients for every $k \in \mathbb{N}$.

When looking at this 2-step prototypical algorithm, there is a looming temptation to generalize and to replace the partial derivatives of the f_i with something else. To find a good generalization, we only need to look at the optimality conditions of (1.1):

$$\text{Find } x^* \text{ such that:} \quad \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(x^*) = 0. \tag{1.3}$$

At once we reduce our cognitive load by changing notation and forgetting that there ever was a function called f_i. We simply replace every occurrence of ∇f_i with a mapping called S_i:

$$\text{Find } x^* \text{ such that:} \quad S(x^*) := \frac{1}{n} \sum_{i=1}^{n} S_i(x^*) = 0. \tag{1.4}$$

The S_i take, as input, points in a space $\mathcal{H}$, like $\mathbb{R}^n$, and output points in $\mathcal{H}$. We follow mathematical tradition and call the S_i *operators*. We do not, however, forget about our structural assumptions: for any given $x \in \mathcal{H}$, we still expect $S_i(x)$ to be drastically simpler to compute than $n^{-1} \sum_{i=1}^n S_i(x)$, and we still expect the jth component, denoted by $(S_i(x))_j$, to be drastically simpler to compute than the full operator $S_i(x)$. We gain a lot of flexibility from this generalization.

Each $S_i(x)$ can, in principle, be any mapping. But we would never be so optimistic. Instead, a coherence condition suggests itself: we assume that for each root x^* of (1.4), we have

$$(\exists \beta_{ij} > 0) : (\forall x \in \mathcal{H}) \qquad \sum_{j=1}^m \sum_{i=1}^n \beta_{ij} \|(S_i(x))_j - (S_i(x^*))_j\|_j^2 \leq \langle S(x), x - x^* \rangle. \tag{1.5}$$

This condition[1] is weaker than what can be guaranteed when each $S_i(\cdot)$ is the gradient of a convex function. In that case, the x^* in (1.5) can even vary beyond the roots of $n^{-1} \sum_{i=1}^n S_i(\cdot)$ to any point in $\mathcal{H}$; the β_{ij} are akin to inverse Lipschitz constants of gradients. But (1.4) is not limited to smooth optimization problems because (1.5) will be satisfied for S_i that are, for example, compositions of proximal and gradient operators. Beyond the benefits of increased modeling power, though, we make this generalization because it is easy to design algorithms for (1.4) that hide all of the complicated details that appear in algorithms which solve specific models, for example, it can be quite technical to design algorithms for (1.1) when two of the terms, say f_1 and f_2, are nonsmooth. Instead, we treat S_i just like we treat ∇f_i.

A need for increased modeling power and algorithms with smaller per-iteration complexity has led us to Problem (1.4) and a class of algorithms that solve it. We want to parallelize these algorithms. But they are inherently sequential because each step updates just a single coordinate using just a single function or operator, and the other coordinates cannot be updated until the active one finishes its work. A big slowdown occurs if some partial derivatives or operator coordinates are much more complicated to evaluate than all the others because we will spend most of our time updating just a few coordinates and leaving the others fixed. A solution is to eliminate the stalling between coordinate updates and allow multiple processors to work at their own pace, updating whenever they complete their work. We take this approach in this paper and allow *asynchronous* updates in our algorithm.

In total, we have taken the basic template for incremental gradient and block-coordinate optimization algorithms and increased its modeling power by introducing the root finding Problem (1.4) and an algorithm to solve it. This stochastic monotone aggregated root-finding (SMART)[2] algorithm is inspired by another algorithm called SAGA [15]. In this work we have taken SAGA and generalized it: the SMART algorithm applies to operators, allows block-coordinate updates, allows asynchronous updates, requires less memory than SAGA, and requires less computation per iteration than SAGA. The theoretical guarantees for SMART are also stronger: the sequence of points formed by the algorithm, and not just the function values, will converge with probability 1 when a solution exists (even in infinite dimensional separable Hilbert spaces, but in this case SMART converges weakly). Like SAGA, SMART converges linearly when $n^{-1} \sum_{i=1}^n f_i$ is strongly convex, and beyond that, it converges linearly when $n^{-1} \sum_{i=1}^n S_i$ is essentially quasi-strongly monotone. It even converges linearly in the asynchronous, block-coordinate case. The rest of this paper describes SMART and proves that it converges.

---

[1] The technical name might be quasi-cocoercivity; we could not find it discussed elsewhere.

[2] We use the term *monotone* because S is a quasi-monotone operator.

2 Assumptions and Notation

The SMART algorithm solves (1.4) in a separable Hilbert space, like $\mathbb{R}^n$, which we call $\mathcal{H}$. We assume the Hilbert space $\mathcal{H} = \mathcal{H}_1 \oplus \cdots \oplus \mathcal{H}_m$ is a direct sum of $m \in \mathbb{N}$ other Hilbert spaces $\mathcal{H}_1, \ldots, \mathcal{H}_m$. Given a vector $x \in \mathcal{H}$, we denote its jth component by $x_j \in \mathcal{H}_j$. Given a sequence $\{x^k\}_{k \in \mathbb{N}}$ and a vector $h \in \mathbb{N}^m$, we define

$$(\forall k \in \mathbb{N}) \qquad x^{k-h} = (x_1^{k-h_1}, \ldots, x_m^{k-h_m})$$

and use the convention that $x_j^k = x_j^0$ if $k \leq 0$. For $j \in \{1, \ldots, m\}$, we let $\langle \cdot, \cdot \rangle_j : \mathcal{H}_j \times \mathcal{H}_j \to \mathbb{R}$ denote the inner product on $\mathcal{H}_j$, and we let $\|\cdot\|_j$ be the corresponding norm. For all $x, y \in \mathcal{H}$, we let $\langle x, y \rangle_{\text{prod}} = \sum_{j=1}^m \langle x_j, y_j \rangle_j$ and $\|x\|_{\text{prod}} := \sqrt{\langle x, x \rangle_{\text{prod}}}$ be the standard inner product and norm on $\mathcal{H}$. We also fix an inner product $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ and denote the corresponding norm by $\|\cdot\|$. We make one assumption about this norm:

$$\left(\exists \underline{M}_j, \overline{M}_j > 0\right) : (\forall x \in \mathcal{H}) \qquad \sum_{i=1}^m \underline{M}_j \|x_j\|_j^2 \leq \|x\|^2 \leq \sum_{i=1}^m \overline{M}_j \|x_j\|_j^2.$$

We often choose inner products associated to self-adjoint linear maps, P, which are defined for all $x, y \in \mathcal{H}$, by $\langle x, y \rangle = \langle x, y \rangle_P := \langle Px, y \rangle_{\text{prod}} = \langle x, Py \rangle_{\text{prod}}$. We work with an underlying probability space denoted by $(\Omega, \mathcal{F}, P)$, and we assume that the space $\mathcal{H}$ is equipped with the Borel σ-algebra. We always let $\sigma(X) \subseteq \mathcal{F}$ denote the sub σ-algebra generated by a random variable X. We use the shorthand a.s. to denote almost sure convergence of a sequence of random variables. We also assume that the operators S_i are measurable. We say that a map $T : \mathcal{H} \to \mathcal{H}$ is *nonexpansive* if it is 1-Lipschitz continuous, i.e., $\|Tx - Ty\| \leq \|x - y\|$ for all $x, y \in \mathcal{H}$. A map $T : \mathcal{H} \to \mathcal{H}$ is called *demiclosed at* 0 if whenever a sequence $\{x^k\}_{k \in \mathbb{N}}$ converges weakly to a point $x \in \mathcal{H}$ and $x^k - Tx^k$ converges strongly to $0 \in \mathcal{H}$, then $Tx = x$; all nonexpansive operators are demiclosed at 0 [2, Corollary 4.8]. Given any nonempty, closed, convex set C, let $P_C(x) := \arg\min_{x^* \in C} \|x - x^*\|$ denote its projection operator, and let $d_C(x) := \min_{x^* \in C} \|x - x^*\|$ denote its distance function. For any closed convex set, $C \subseteq \mathcal{H}$, we let N_C denote the normal cone of C [2, Definition 6.37]. We define

$$S := \frac{1}{n} \sum_{i=1}^n S_i \qquad \text{and} \qquad \mathcal{S} := \text{zer}(S) = \{x \in \mathcal{H} \mid S(x) = 0\}.$$

Beyond (1.5), we use a single regularity property

$$(\exists \mu > 0) : (\forall x \in \mathcal{H}) \qquad\qquad \langle S(x), x - P_{\mathcal{S}}(x) \rangle \geq \mu \|x - P_{\mathcal{S}}(x)\|^2. \tag{2.1}$$

Operators that satisfy (2.1) are called *essentially strongly quasi-monotone*. In order for our most general results to hold, S need not satisfy (2.1), but when it is satisfied, our algorithm converges linearly. Although it is hardly comprehensive, one example is noteworthy: property (2.1) holds if $S = A^* \circ \nabla f \circ A$ for a strongly convex function f and a matrix A [25, p. 287]. See [7,37,24] for information on convex error bounds.

Most of the concepts that we use in this paper can be found in [2]. See Table C.1 for a list of symbols.

3 The SMART Algorithm

We develop an iterative algorithm that solves (1.4). The algorithm forms a sequence of *primal variables* $\{x^k\}_{k \in \mathbb{N}} \subseteq \mathcal{H}$ that converges to a root of (1.4). The algorithm also maintains a sequence of *dual variables*, one per operator, which is denoted by $\{(y_1^k, \ldots, y_n^k)\}_{k \in \mathbb{N}} \subseteq \mathcal{H}^n$.

We assume that at least one of m, the number of components, and n, the number of operators, is large, and so it costs a lot to obtain x^{k+1} from x^k by updating all m of its components, using all n of the operators.

To lower the cost, we introduce an IID sequence of $2^{\{1,\dots,m\}}$-valued (subsets of $\{1,\dots,m\}$) random variables $\{\mathsf{S}_k\}_{k\in\mathbb{N}}$ that determines which components of x^k we update at the kth iteration. The component-choosing variable S_k is coupled with an IID sequence of $\{1,\dots,n\}$-valued random variables $\{i_k\}_{k\in\mathbb{N}}$ that determine which one of the n operators $S_1,\dots,S_n$ are evaluated at the kth iteration. If $i_k = i$ and $j \in \mathsf{S}_k$, then the jth component of x^k is updated at iteration k using an evaluation of $(S_i(\cdot))_j$; the other operators and components are left alone. The user can choose $\{i_k\}_{k\in\mathbb{N}}$ and $\{\mathsf{S}_k\}_{k\in\mathbb{N}}$ however they like as long as

$$q_j := P(j \in \mathsf{S}_0) > 0 \text{ for all } j; \text{ and } p_{ij} := P(i_0 = i | j \in \mathsf{S}_0) > 0 \text{ exactly when } (S_i(\cdot))_j \not\equiv 0.$$

Unlike the point sequence $\{x^k\}_{k\in\mathbb{N}}$, the dual variables need not be updated at every iteration. Instead, we introduce an IID sequence of $\{0,1\}$-valued random variables $\{\epsilon_k\}_{k\in\mathbb{N}}$ that determines if and when we update the dual variables. If $\epsilon_k = 1$, then the dual variables are updated at iteration k; otherwise, the dual variables are left alone. The user can choose $\{\epsilon_k\}_{k\in\mathbb{N}}$ however they like as long as

$$\rho := P(\epsilon_0 = 1) > 0.$$

If $\epsilon_k = 1$, and thus the dual variables must update at iteration k, we only require, at the absolute minimum, that $y_{i_k}^k$ be updated; the rest of the variables may stay fixed. However, we allow the update of $y_{i_k}^k$ to *trigger* the update of any subset of the other dual variables. The *trigger graph* $G = (V, E)$ with vertices $V = \{1,\dots,n\}$ uses the *edge set* $E \subseteq \{1,\dots,n\}^2$ to encode, for each $i \in V$, the set of dual variables that must be updated when $i_k = i$:

$$(i, i') \in E \text{ if, and only if, } i_k = i \text{ triggers the update of dual variable } y_{i'}^k.$$

When $(i, i') \in E$, we simply say that i *triggers* i'. We require that for all i, $(i, i) \in E$, but otherwise there are no constraints on G; it can be absolutely any graph, from a completely disconnected graph, to a complete graph on n vertices. And one quantity, figuring only in our linear rate of convergence, is important: the probability that i is triggered *and* coordinate j is sampled simultaneously

$$p_{ij}^T := P((i_0, i) \in E, j \in \mathsf{S}_0) = P((i_0, i) \in E \mid j \in \mathsf{S}_0)q_j = \sum_{(i',i)\in E} p_{i'j}q_j,$$

which is easily computable, but it need not be known to guarantee convergence.

Often, the matrix of optimal operator values

$$\mathbf{S}^* := ((S_i(x^*))_j)_{ij}, \qquad x^* \in \mathcal{S}$$

has some entries which are zero (by (1.5), $\mathbf{S}^*$ is independent of $x^* \in \mathcal{S}$). For these zero entries, we simply set the corresponding dual variable to zero: $y_{i,j}^k \equiv 0$ if $\mathbf{S}_{ij}^* = 0$. In the extreme case that $\mathbf{S}^* = 0$, all operators are zero at the solution, (1.4) reduces to the common zero problem, and the dual variables $\{(y_1^k,\dots,y_n^k)\}_{k\in\mathbb{N}} \subseteq \mathcal{H}^n$ are all zero. Setting these particular dual variables to zero is not necessary, but by doing so, we use less memory than we otherwise would.

An algorithm that solves (1.4) might use the following 3-step process: given x^k

1. sample $\mathsf{S}_k, i_k,$ and ϵ_k;
2. get x^{k+1} from x^k using $\{(S_{i_k}(x^k))_j \mid j \in \mathsf{S}_k\}$ and $(y_1^k,\dots,y_n^k)$;
3. if $\epsilon_k = 1$, get $(y_1^{k+1},\dots,y_n^{k+1})$ using x^k and $(y_1^k,\dots,y_n^k)$; otherwise set $(y_1^{k+1},\dots,y_n^{k+1}) = (y_1^k,\dots,y_n^k)$.

After i_k and S_k are sampled, the inactive operators and the inactive components stall until the active ones finish Steps 2 and 3. If we have access to a parallel computing device, stalling is wasteful, so in our algorithm we let all operators work in parallel and update x^k whenever they finish their work. Mathematically, we form sequences of *delays* $\{d_k\}_{k\in\mathbb{N}} \subseteq \{0,1,\dots,\tau_p\}^m$ and $\{e_k^i\}_{k\in\mathbb{N}} \subseteq \{0,1,\dots,\tau_d\}^m$, and we replace several instances of x^k and y_i^k with iterates from the past whose coordinates were formed in the last τ_p or τ_d iterations, respectively. The final algorithm is below:

Algorithm 1 (SMART) *Let $\{\lambda_k\}_{k\in\mathbb{N}}$ be a sequence of stepsizes. Choose $x^0 \in \mathcal{H}$ and $y_1^0, \ldots, y_n^0 \in \mathcal{H}$ arbitrarily except that $y_{i,j}^0 = 0$ if $\mathbf{S}_{ij}^* = 0$. Then for $k \in \mathbb{N}$, perform the following three steps:*

1. **Sampling.** *choose a set of coordinates S_k, an operator index i_k, and dual update decision ϵ_k.*
2. **Primal update:** *set*

$$(\forall j \in \mathsf{S}_k) \qquad x_j^{k+1} = x_j^k - \frac{\lambda_k}{q_j m}\left(\frac{1}{np_{ij}}(S_{i_k}(x^{k-d_k}))_j - \frac{1}{np_{ij}}y_{i_k,j}^{k-e_k^{i_k}} + \frac{1}{n}\sum_{i=1}^n y_{i,j}^{k-e_k^i}\right);$$

$$(\forall j \notin \mathsf{S}_k) \qquad x_j^{k+1} = x_j^k.$$

3. **Dual update:** *If i_k triggers i, set*

$$\left(\forall j \in \mathsf{S}_k \text{ with } \mathbf{S}_{ij}^* \neq 0\right) \qquad y_{i,j}^{k+1} = y_{i,j}^k + \epsilon_k\left((S_i(x^{k-d_k}))_j - y_{i,j}^k\right);$$

$$(\forall j \notin \mathsf{S}_k) \qquad y_{i,j}^{k+1} = y_{i,j}^k.$$

Otherwise, set $y_{i,j}^{k+1} = y_{i,j}^k$. □

The x^{k-d_k} iterate in SMART can be *totally synchronous*, in which case $d_{k,j} = 0$ for all j; it can be *consistent-read asynchronous*, in which case $d_{k,j} = d_{k,j'}$ for all j and j'; or it can be *inconsistent-read asynchronous*, in which case $d_{k,j} \neq d_{k,j'}$ for some j and j'. Totally synchronous iterates are not delayed at all, so $x^{k-d_k} = x^k$ for all k; consistent-read asynchronous iterates are delayed, but all of their coordinates are delayed by the same amount, so $x^{k-d_k} \in \{x^k, x^{k-1}, \ldots, x^{k-\tau_p}\}$ for all k; inconsistent-read asynchronous iterates are delayed, and their coordinates can be delayed by different amounts, so $x_j^{k-d_{k,j}} \in \{x_j^k, x_j^{k-1}, \ldots, x_j^{k-\tau_p}\}$ for all j, but x^{k-d_k} is not necessarily an element of $\{x^k, x^{k-1}, \ldots, x^{k-\tau_p}\}$ for all k.

Likewise, the $y_i^{k-e_k^i}$ iterate in SMART can be totally synchronous, consistent-read asynchronous, or inconsistent-read asynchronous, corresponding to the cases $e_{k,j}^i = 0$ for all j, $e_{k,j}^i = e_{k,j'}^i$ for all j and j', or $e_{k,j}^i \neq e_{k,j'}^i$ for some j and j', respectively.

The delays $\{d_k\}_{k\in\mathbb{N}}$ and $\{e_k^i\}_{k\in\mathbb{N}}$ come with no particular order. They can induce maximal delays at all times, e.g., $d_{k,j} = \tau_p$ for all j and k, in which case the oldest information possible is used in every iteration; they can be cyclic, e.g., $d_{k,j} = k \mod (\tau_p + 1)$ for all j and k, in which case the same information is used for τ_p consecutive iterations in a row, and all the intermediate information is thrown away; but in general, the delays can be arbitrary. These delays are artificially imposed, but we can also incur delays that are beyond our control.

Uncontrolled delays can occur in the x^k iterate when a processor, called Proc_1, attempts to read coordinates $x_1^k, \ldots, x_m^k$ while another processor, called Proc_2, attempts to replace x^k with x^{k+1}. It can happen that Proc_1 successfully reads coordinate x_1^k before Proc_2 replaces it with x_1^{k+1}, but that Proc_2 replaces $x_2^k, \ldots, x_m^k$ with $x_2^{k+1}, \ldots, x_m^{k+1}$ before Proc_1 attempts to read this group of coordinates. When Proc_1 finishes reading, it will have the iterate $(x_1^k, x_2^{k+1}, \ldots, x_m^{k+1})$, which is not necessarily equal to any previous iterate x^t with $t \leq k$. This effect is exacerbated if multiple processors attempt to write and read simultaneously, but in SMART, x^{k-d_k} is inconsistent-read asynchronous, so these uncontrolled delays cause no trouble.

In general, including inconsistent-read asynchronous iterates leads to tedious convergence proofs with many-term recursive identities and complicated stepsizes. The recursive identities are necessary to control algorithm progress in a chaotic environment. The stepsizes, on the other hand, can be be optimized and have a clear dependence on the delays, sampling probabilities, and problem data. In both the inconsistent-read

and consistent-read cases, the algorithm will converge with the same range of parameters. However, the rates of convergence depend on a measure of inconsistency, which we call δ

$$\delta := \sup_{\substack{k \in \mathbb{N} \\ j,j' \in \{1,\ldots,m\}}} |d_{k,j} - d_{k,j'}|. \tag{3.1}$$

When $\delta = 0$, the x^{k-d_k} are consistent-read asynchronous, and the convergence rates improve. Otherwise $\delta \in \{0, \ldots, \tau_p\}$ and the convergence rates degrade with increasing δ. Of course, when δ is not known explicitly, it can be replaced by its upper bound τ_p.

If all of the sampling variables are statistically independent and the stepsizes are chosen small enough, SMART converges:

Theorem 3.1 (Convergence of SMART) *For all $k \geq 0$, let $\mathcal{I}_k = \sigma((i_k, \mathsf{S}_k))$, let $\mathcal{E}_k = \sigma(\varepsilon_k)$, let*

$$\mathcal{F}_k := \sigma(\{x^l\}_{l=0}^k \cup \{y_1^l, \ldots, y_n^l\}_{l=0}^k),$$

and suppose that $\{\mathcal{I}_k, \mathcal{E}_k, \mathcal{F}_k\}$ are independent. Suppose that (1.5) holds. Finally, suppose that there are constants $\underline{\lambda}, \overline{\lambda} > 0$ such that $\{\lambda_k\}_{k \in \mathbb{N}} \subseteq [\underline{\lambda}, \overline{\lambda}]$, and

$$\overline{\lambda}^2 < \begin{cases} \min_{i,j} \left\{ \dfrac{2\underline{\lambda}n^2 p_{ij}\beta_{ij}}{\frac{3\overline{M}_j \tau_p}{m\sqrt{\underline{q}}} + \frac{2\overline{M}_j}{\underline{q}m}} \right\} & \text{if } \mathsf{S}^* = 0; \\[2em] \min_{i,j} \left\{ \dfrac{\underline{\lambda}n^2 p_{ij}\beta_{ij}}{\frac{\overline{M}_j \tau_p}{m\sqrt{\underline{q}}}\sqrt{2(\tau_d+2)} + \frac{\overline{M}_j(\tau_d+2)}{m\underline{q}}} \right\} & \text{otherwise.} \end{cases} \tag{3.2}$$

Then

1. ***Convergence of operator values.*** *For $i \in \{1, \ldots, n\}$, the sequence of $\mathcal{H}$-valued random variables $\{S_i(x^{k-d_k})\}_{k \in \mathbb{N}}$ a.s. converges strongly to $S_i(x^*)$.*
2. ***Weak convergence.*** *Suppose that $I - S$ is demiclosed at 0. Then the sequence of $\mathcal{H}$-valued random variables $\{x^k\}_{k \in \mathbb{N}}$ a.s. weakly converges to an $\mathcal{S}$-valued random variable.*
3. ***Linear convergence.*** *Let $\eta < \min_{i,j}\{\rho p_{ij}^T\}$, let $\alpha \in [0,1)$, let $\underline{q} = \min_j\{q_j\}$, let $\overline{M} = \min_j\{\overline{M}_j\}$, and let*

$$\lambda \leq \min_{i,j} \left\{ \frac{2\eta(1-\alpha)n^2 p_{ij}\beta_{ij}}{\frac{2\overline{M}_j \eta(\tau_d+2)}{q_j m}\left(1 + \frac{\delta\eta}{\tau_d+1} + \frac{5\sqrt{2(\tau_d+2)}\alpha^2\delta}{m\overline{M}\left(\sqrt{2(\tau_d+2)}+\tau_p\sqrt{\underline{q}}\right)}\right) + \frac{\overline{M}_j \eta\tau_p\sqrt{2(\tau_d+2)}}{m\sqrt{\underline{q}}}\left(2 + \frac{\eta}{1-\eta}\right) + 4\mu(\tau_d+1)\alpha(1-\alpha)n^2 p_{ij}\beta_{ij}} \right\}. \tag{3.3}$$

Then if (2.1) holds, there exists a constant $C(z^0, \phi^0) \in \mathbf{R}_{\geq 0}$ depending on x^0 and ϕ^0 such that for all $k \in \mathbf{N}$,

$$\mathbb{E}\left[d_{\mathcal{S}}^2(x^k)\right] \leq \left(1 - \frac{2\alpha\mu\lambda}{\tau_p + 1}\right)^{k/(\tau_p+1)} \left(d_{\mathcal{S}}^2(x^0) + C(x^0, \phi^0)\right).$$

The proof of Theorem 3.1 without delay (i.e., $d_k \equiv 0$ and $e_k^{ij} \equiv 0$) is presented in Section 7. The proof of the full theorem is in Appendix A

Algorithm	Stepsize λ		Rate for best λ
	that is largest possible	that gives best rate	
SAGA (4.2)	$\frac{1}{2L}$	$\frac{1}{4L+\mu N}$	$1 - \frac{\mu}{4L+\mu N}$
SVRG (4.3) average update frequency τ	$\frac{1}{2L}$	$\frac{1}{4L+\mu\tau}$	$1 - \frac{\mu}{4L+\mu\tau}$
SVRG (4.4) scheduled update frequency τ	$\frac{1}{(\tau+2)L}$	$\frac{1}{2L(\tau+2)+\mu(\tau+1)}$	$1 - \frac{\mu}{2L(\tau+2)+\mu(\tau+1)}$
Finito (4.5)	$\frac{1}{2}; \gamma = \frac{2}{L}$	$\frac{1}{4}; \gamma = \frac{1}{L}$	$1 - \frac{1}{4N}\left(1 - \sqrt{1 - \frac{\hat{\mu}}{L}}\right)$
SDCA (4.8)	$\frac{3}{4}$	$\frac{3}{8}$	$1 - \frac{3\mu_0}{8(L+\mu_0 N)}$
Alternating Projections (4.10)	1	$\frac{1}{2}$	$1 - \frac{\min\{1,\varepsilon^2 L^{-2}\}}{2N\hat{\mu}}$
Kaczmarz (4.11)	1	$\frac{1}{2}$	$1 - \frac{1}{2N\|A^{-1}\|_2^2}$

Table 4.1: The stepsizes and convergence rates for the special cases of SMART introduced in Section 4

4 Connections with Other Algorithms

4.1 SAGA, SVRG, and S2GD

In the simplest case of (1.4)

$$\underset{x \in \mathcal{H}_0}{\text{minimize}} \ \frac{1}{N} \sum_{i=1}^{N} f_i(x) \tag{4.1}$$

where each f_i is convex and differentiable, and ∇f_i is L-Lipschitz, we set $S_i := \nabla f_i$. (We also set $\mathcal{H} = \mathcal{H}_0$, use the canonical norm $\|\cdot\| = \|\cdot\|_0$, and ignore coordinates and partial derivatives for the moment.) By Proposition C.1, we can set $\beta_{ij} = (LN)^{-1}$ for all i and j. With this choice of operators, the condition (2.1) is, of course, implied by the μ-strong convexity of $f := N^{-1} \sum_{i=1}^{N} f_i$. But whether or not f is strongly convex, if λ is set according to Theorem 3.1 or Table 4.1, SMART will converge.

The SAGA Algorithm [15] applied to (4.1) selects a function uniformly at random, performs a primal update, and updates a single dual variable (i.e., the trigger graph is completely disconnected). When, for all i, we set $y_i^0 = \nabla f_i(\phi_i^0)$ with $\phi_i^0 \in \mathcal{H}$, SAGA takes the form:[3]

$$x^{k+1} = x^k - \lambda \left(\nabla f_{i_k}(x^k) - y_{i_k}^k + \frac{1}{N} \sum_{i=1}^{N} y_i^k \right);$$
$$y_i^{k+1} = \begin{cases} \nabla f_i(x^k) & \text{if } i_k = i; \\ y_i^k & \text{otherwise.} \end{cases} \tag{4.2}$$

---

[3] Use $n = N, m = 1, d_k \equiv 0, e_k^i \equiv 0, q_1 \equiv 1, \tau_p = \tau_d = 0, p_{ij} \equiv N^{-1}, \rho = 1, E = \{(i,i) \mid i \in V\}, p_{ij}^T = N^{-1}$.

In the SAGA algorithm, each dual variable is just a gradient, $y_i^k := \nabla f_i(\phi_i^k)$, for a past iterate ϕ_i^k. The SAGA algorithm stores the stale gradients $\{\nabla f_i(\phi_i^k) \mid i = 1, \ldots, N\}$, which, if $x \in \mathbb{R}^d$, is the size of a $d \times N$ matrix. However, in logistic and least squares regression problems, the functions f_i have a simple form $f_i(x) = \psi_i(\langle a_i, x\rangle)$ where the $\psi_i : \mathbb{R} \to \mathbb{R}$ are differentiable functions and the $a_i \in \mathbb{R}^d$ are datapoints; in this case, $\nabla f_i(x) = \psi_i'(\langle a_i, x\rangle)a_i$, so the cost of storing $\{\nabla f_i(\phi_i^k) \mid i = 1, \ldots, N\}$ can be reduced to that of $(\psi_1'(\langle a_1, \phi_1^k\rangle), \ldots, \psi_N'(\langle a_N, \phi_N^k\rangle))^T \in \mathbb{R}^d$—a d-dimensional vector. But not all problems have this parametric form, so the SAGA algorithm is somewhat limited in scope.

The Stochastic Variance Reduced Gradient (SVRG) algorithm [20], and the similar S2GD algorithm [23], solve (4.1), but in contrast to SAGA, these algorithms store just a single vector, namely $\nabla f(\widetilde{x}^k)$. As a consequence, SVRG and S2GD must make repeated, though infrequent, evaluations of the full gradient ∇f—in addition to one extra evaluation of ∇f_{i_k} per-iteration:

$$x^{k+1} = x^k - \lambda \left(\nabla f_{i_k}(x^k) - \nabla f_{i_k}(\widetilde{x}^k) + \frac{1}{N}\sum_{i=1}^N \nabla f_i(\widetilde{x}^k) \right);$$

$$\widetilde{x}^{k+1} = \begin{cases} x^{k-t_k} & \text{if } k \equiv 0 \mod \tau - 1; \\ \widetilde{x}^{k+1} & \text{otherwise;} \end{cases}$$

where $\{i_k\}_{k\in\mathbb{N}}$ is an IID sequence of uniformly distributed $\{1, \ldots, N\}$-valued random variables and $\{t_k\}_{k\in\mathbb{N}}$ is an IID sequence of uniformly distributed $\{0, \ldots, \tau-1\}$-valued random variables. The full gradient $\nabla f(\widetilde{x}^k)$ and the point $\widetilde{x}^k$ are only updated once every $\tau \in \mathbb{N}$ iterations.

The SVRG algorithm[4] solves the SAGA storage problem, but it requires f to be strongly convex, so it is also somewhat limited in scope. SMART can mimic SVRG—even without strong convexity—by selecting a function uniformly at random, performing a primal update, and updating *all* dual variables with probability τ^{-1} (i.e., the trigger graph is the complete graph). When, for all i, we set $y_i^0 = \nabla f_i(\phi^0)$ with $\phi^0 \in \mathcal{H}$, our SVRG clone takes the form:[5]

$$x^{k+1} = x^k - \lambda \left(\nabla f_{i_k}(x^k) - y_{i_k}^k + \frac{1}{N}\sum_{i=1}^N y_i^k \right);$$

$$(\forall i) \qquad y_i^{k+1} = y_i^k + \epsilon_k(\nabla f_i(x^k) - y_i^k). \tag{4.3}$$

As in SAGA, each dual variable is just a gradient $y_i^k = \nabla f_i(\phi^k)$, for a past iterate ϕ^k, but unlike SAGA, the past iterate is the same for all i. On average, all dual variables y_i^k, and hence the full gradient $\nabla f(\phi^k)$, are only updated once every τ iterations.

Another clone of SVRG, this time with dual variables that update once every τ iterations, comes from SMART as applied in (4.3), but with a cyclic uniform dual variable delay $e_k^i = e_k := k \mod (\tau + 1)$:[6]

$$x^{k+1} = x^k - \lambda \left(\nabla f_{i_k}(x^k) - y_{i_k}^{k-e_k} + \frac{1}{N}\sum_{i=1}^N y_i^{k-e_k} \right);$$

$$(\forall i) \qquad y_i^{k+1} = \nabla f_i(x^k). \tag{4.4}$$

The dual variable $y_i^k = \nabla f_i(x^k)$, and hence the full gradient $\nabla f(x^k)$, is only updated once every τ iterations.

---

[4] From here on, we will ignore the distinction between SVRG and S2GD.

[5] Use $n = N, m = 1, \overline{M}_1 = 1, d_k \equiv 0, e_k^i \equiv 0, q_j \equiv 1, \tau_p = \tau_d = 0, p_{ij} \equiv N^{-1}, \rho = \tau^{-1}, E = V \times V$, and $p_{ij}^T = 1$.

[6] Use $n = N, m = 1, \overline{M}_1 = 1, d_k \equiv 0, e_k^i = k \mod \tau, q_j \equiv 1, \tau_p = 0, \tau_d = \tau, p_{ij} \equiv N^{-1}, \rho = 1, E = V \times V$, and $p_{ij}^T = 1$.

4.2 Finito

The Finito algorithm [16] solves (4.1), but unlike SAGA and SVRG, Finito stores one point and one gradient per function, or a superposition of the two:

$$x_i^{k+1} = \begin{cases} \frac{1}{N} \sum_{l=1}^N (x_l^k - \frac{1}{2\hat{\mu}} \nabla f_l(x_l^k)) & \text{if } i = i_k; \\ x_i^k & \text{otherwise}, \end{cases}$$

where each f_i is $\hat{\mu}$-strongly convex. For each function f_i, Finito stores $x_i^k - (2\hat{\mu})^{-1} \nabla f_i(x_i^k)$, which can be substantially costlier than storing a matrix of gradients. In addition, only when each function $f_i(x)$ is $\hat{\mu}$-strongly convex and the bound $N \geq 2L\hat{\mu}^{-1}$ holds, is Finito known to converge.

SMART can mimic SAGA and SVRG with multiple operators, but with one operator $S = S_1$, SMART recovers the Finito algorithm. To get Finito, recast (4.1) into an equivalent form with duplicated variables

$$\underset{(x_1,\ldots,x_N) \in \mathcal{H}_0^N}{\text{minimize}} \quad \frac{1}{N} \sum_{i=1}^N f_i(x_i) \qquad \text{subject to: } x_1 = x_2 = \cdots = x_N,$$

let $D := \{(x,\ldots,x) \in \mathcal{H}_0^N \mid x \in \mathcal{H}\}$ denote the *diagonal* set, and define the operator (for a fixed $\gamma > 0$)

$$\left(\forall x \in \mathcal{H}_0^N\right) \qquad S(x) := S_1(x) = x - P_D\left(x_1 - \gamma\nabla f_1(x_1),\ldots,x_N - \gamma\nabla f_N(x_N)\right).$$

Then the Finito algorithm, selects a coordinate (and hence a function) uniformly at random and performs a primal update; the sole dual variable is set to zero at all iterations:[7]

$$x_j^{k+1} = \begin{cases} (1-\lambda)x_j^k + \frac{\lambda}{N} \sum_{l=1}^N \left(x_l^k - \gamma\nabla f_l(x_l^k)\right) & \forall j \in \mathsf{S}_k; \\ x_j^k & \text{otherwise}. \end{cases} \tag{4.5}$$

Unlike the standard Finito algorithm [16], Algorithm (4.5) converges with or without strong convexity.

The constants β_{1j} depend on the constant γ and the inner products that we place on $\mathcal{H}_j := \mathcal{H}_0$ and $\mathcal{H} := \mathcal{H}_0^N$. The projection operator P_D also depends on these inner products. As long as[8] $\gamma \leq 2L^{-1}$, the operator S satisfies (1.5) with $\beta_{1j} \equiv 4^{-1}\gamma L$, and if $\mu := 1 - \sqrt{1 - 2\gamma\hat{\mu} + \gamma^2\hat{\mu}L}$, it is μ-essentially strongly quasi-monotone—provided that each f_i is $\hat{\mu}$-strongly convex and we make the choices $\langle x_j, z_j \rangle_j := \langle x_j, z_j \rangle_0$ and $\|\cdot\| = \|\cdot\|_{\text{prod}}$; the operator S need not be strongly monotone unless each f_i is strongly convex. See Proposition C.2.

The space $\mathcal{H} = \mathcal{H}_0^N$ is high-dimensional, so even storing a single vector x^k is expensive. And in practice, the gradients should also be stored—unless they are simple to recompute. Thus, it is clear that Finito, like SAGA, but unlike SVRG, is impracticable if m is too large and memory is limited. Nevertheless, Finito performs well in practice, often better than other incremental gradient methods.

---

[7] Use $n = 1, m = N, \overline{M}_j \equiv 1, d_k \equiv 0, e_k^i \equiv 0, \mathbf{S}^* = 0, q_j \equiv N^{-1}, \tau_p = 0, \tau_d = 0, p_{ij} \equiv 1, \rho = 1, E = \{(1,1)\}$, and $p_{ij}^T = N^{-1}$.

[8] See Proposition C.2.

4.3 SDCA

The Stochastic Dual Coordinate Ascent (SDCA) algorithm [34] solves a problem different from (4.1):[9]

$$\underset{z \in \mathcal{H}_0}{\text{minimize}} \frac{1}{N} \sum_{j=1}^{N} f_j(z) + \frac{\mu_0}{2} \|z\|^2 \tag{4.6}$$

SDCA does not solve this primal problem directly; instead, SDCA solves the dual problem.

$$\underset{(x_1,\ldots,x_N) \in \mathcal{H}_0^N}{\text{minimize}} \frac{1}{N} \sum_{j=1}^{N} f_j^*(-x_j) + \frac{\mu_0}{2} \left\| \frac{1}{\mu_0 N} \sum_{j=1}^{N} x_j \right\|^2.$$

If we define $\{\mathsf{S}_k\}_{k \in \mathbb{N}}$ as in SMART and restrict all S_k to have at most one element, SDCA repeatedly does the following:

$$x_j^{k+1} = \begin{cases} x_j^k + \arg\min_{x_j \in \mathcal{H}_0} \left\{ f_j^*(-x_j^k - x_j) + \frac{\mu_0 N}{2} \left\| \frac{1}{\mu_0 N} \sum_{l=1}^{N} x_l^k + \frac{1}{\mu_0 N} x_j \right\|^2 \right\} & \text{if } j \in \mathsf{S}_k; \\ x_j^k & \text{otherwise.} \end{cases} \tag{4.7}$$

For each function f_j, SDCA stores the vector x_j^k. This can be cheap because, as the optimality conditions of (4.7) show, there is always a point z_j^k such that $x_j^k \in \partial f(x_j^k)$. The drawback to (4.7) is that only when each f_i is differentiable, is SDCA known to converge. In addition, the form (4.7) is opaque, but with a bit of polishing, it can be made transparent.

Evidently, the first line of (4.7) is a *forward-backward* step:

$$x_j^{k+1} = \mathbf{prox}_{\mu_0 N f_j^*(-\cdot)} \left(x_j^k - \mu_0 N \left[\nabla g(x^k) \right]_j \right)$$

where $g(x) := 2^{-1} \mu_0 N \left\| (\mu_0 N)^{-1} \sum_{l=1}^{N} x_l \right\|^2$. Thus, SDCA is an instance of SMART with $f^* := \sum_{j=1}^{N} f_j^*(x_j)$,

$$(\forall x \in \mathcal{H}_0^N) \qquad S(z) = S_1(x) := x - \mathbf{prox}_{\mu_0 N f^*(-\cdot)}(x - \mu_0 N \nabla g(x)),$$

that selects, at each iteration, a single coordinate (hence, a function) uniformly at random and performs a primal update; the sole dual variable is set to zero at all iterations:[10]

$$x_j^{k+1} = \begin{cases} x_j^k - \lambda \left(x_j^k - \mathbf{prox}_{\mu_0 N f_j^*(-\cdot)} \left(x_j^k - \mu_0 N \left[\nabla g(x^k) \right]_j \right) \right) & \forall j \text{ with } j \in \mathsf{S}_k \\ x_j^k & \text{otherwise.} \end{cases} \tag{4.8}$$

Unlike (4.7), Algorithm (4.8) converges whether or not any f_j is differentiable.

The constants β_{1j} depend on the inner products that we place on $\mathcal{H}_j := \mathcal{H}_0$ and $\mathcal{H} := \mathcal{H}_0^N$; as we did for Finito, we set $\langle x_j, z_j \rangle_j := \langle x_j, z_j \rangle_0$, and $\|\cdot\| = \|\cdot\|_{\text{prod}}$. With this choice, the operator S satisfies (1.5) with $\beta_{1j} \equiv 3/4$, and it is $\mu_0 N (\mu_0 N + L)^{-1}$-strongly quasi-monotone. See Proposition C.3.

Although the space $\mathcal{H}$ is high-dimensional, the condition $x_j^0 \in \text{range}(\partial f_j)$ ensures that each component x_j^k will lie in the linear span of $\text{range}(\partial f_j)$, which is often one-dimensional.[11]

---

[9] In the standard SDCA problem, $f_i(z) := \psi_i(A_i^T z)$ for a convex, differentiable function ψ_i and a matrix A_i. Furthermore, the squared 2-norm is replaced with a general strongly convex function g.

[10] Use $n = 1, m = N, \overline{M}_j \equiv 1, d_k \equiv 0, e_k^i \equiv 0, \mathbf{S}^* = 0, q_j \equiv N^{-1}, \tau_p = \tau_d = 0, p_{ij} \equiv 1, \rho = 1, E = \{(1,1)\}$, and $p_{ij}^T = N^{-1}$.

[11] The Moureau identity implies that for all x_j, we have $\mathbf{prox}_{N \mu_0 f_j^*(-\cdot)}(x_j) \in \text{range}(\partial f_j)$.

4.4 Randomized Projection Algorithms

Besides minimization problems, SMART also solves feasibility problems:

$$\text{Find } x \in \bigcap_{i=1}^{s_1} C_i \qquad \text{subject to:} \qquad f_i(x) \leq 0 \text{ for } i = 1, \ldots, s_2, \qquad (N = s_1 + s_2) \tag{4.9}$$

where $C_1, \ldots, C_{s_1} \subseteq \mathcal{H}$ are closed, convex sets, and $f_i : \mathcal{H} \to (-\infty, \infty)$ are continuous, convex functions. To align (4.9) with (1.4), each set and each function are assigned an operator. The sets C_i are assigned the familiar projection operator P_{C_i}. The functions are assigned a *subgradient projector*

$$(\forall x \in \mathcal{H}) \qquad G_{f_i}(x) := x - \frac{f_i(x)}{\|g_i(x)\|^2} g_i(x)$$

where $g_i : \mathcal{H} \to \mathcal{H}$ is a *measurable subgradient selector*; i.e., for all $x \in \mathcal{H}$, $g_i(x) \in \partial f_i(x)$.

Then Problem (4.9) is an instance of (1.4) with $n = s_1 + s_2$ and

$$(\forall x \in \mathcal{H}) \qquad S_i(x) := \begin{cases} x - P_{C_i} x & \text{if } i = 1, \ldots, s_1; \\ x - G_{f_{i-s_1}}(x) & \text{if } i = s_1 + 1, \ldots, s_1 + s_2. \end{cases}$$

With SMART, we can select a function or a set uniformly at random and perform a primal update; all operators are zero at points in $\text{zer}(S)$, so all dual variables are zero:[12]

$$x^{k+1} := x^k - \lambda \begin{cases} x^k - P_{C_{i_k}} x^k & \text{if } i_k \in \{1, \ldots, s_1\}; \\ x^k - G_{f_{i_k - s_1}}(x^k) & \text{if } i_k \in \{s_1 + 1, \ldots, s_1 + s_2\}. \end{cases} \tag{4.10}$$

In general, the operator S in (4.9) has several nice properties: $\text{zer}(S)$ is the set of solutions to (4.9); S satisfies (1.5) with $\beta_{i1} \equiv N^{-1}$; S is demiclosed; and if

1. $\{C_i \mid i = 1, \ldots, s_1\} \cup \{\{x \mid f_i(x) \leq 0\} \mid i = 1, \ldots, s_2\}$ are $\hat{\mu}$-linearly regular,[13]
2. there is an $\varepsilon > 0$ such that $f_i(x) \geq \varepsilon d_{\{f_i(x) \leq 0\}}(x)$ for all $x \in \mathcal{H}$,
3. and there is an $L > 0$ such that $\|g_i(x)\| \leq L$ for all $x \in \mathcal{H}$,

then, with $\mu := (N\hat{\mu})^{-1} \max\{1, \varepsilon^2 L^{-2}\}$, the operator S is μ-essentially strongly quasi-monotone. Thus, x^k converges linearly, which is a new result for (4.10).[14]

The randomized Kaczmarz algorithm [35, 26], which solves overdetermined linear systems $Ax = b$, is a special case of (4.10): if $a_1, \ldots, a_N$ are the rows of A, which we assume, without loss of generality, are normalized and $C_i := \{x \mid \langle a_i, x \rangle = b_i\}$, then the projection is $P_{C_i}(x) = x + (b_i - \langle a_i, x \rangle)a_i$, and

$$x^{k+1} = x^k + \lambda(b_{i_k} - \langle a_{i_k}, x \rangle)a_i \tag{4.11}$$

The Kaczmarz operator is $(\|A^{-1}\|_2^{-2} N^{-1})$-essentially strongly quasi-monotone, and so x^k linearly converges to a solution of $Ax = b$.[15]

---

[12] Use $n = N, m = 1, \overline{M}_1 \equiv 1, d_k \equiv 0, e_k^i \equiv 0, \mathbf{S}^* = 0, q_j \equiv 1, \tau_p = \tau_d = 0, p_{ij} \equiv N^{-1}, \rho = 1, E = \{(i,i) \mid i \in V\}$, and $p_{ij}^T = N^{-1}$.

[13] A set family $\{D_1, \ldots, D_N\}$ is $\hat{\mu}$-linearly regular if $\forall x \in \mathcal{H}, d_{D_1 \cap \cdots \cap D_N}(x) \leq \hat{\mu} \max\{d_{D_1}(x), \ldots, d_{D_N}(x)\}$.

[14] See Proposition C.4.

[15] We define $\|A^{-1}\|_2 := \inf\{M \mid (\forall x \in \mathcal{H}) \ M\|Ax\|_2 \geq \|x\|_2\}$; see Corollary C.1.

4.5 ARock

When $n = 1$, and hence, $S = S_1$, SMART recovers the ARock Algorithm [29], which at iteration k, samples a coordinate j and updates

$$x_j^{k+1} = x_j^k - \frac{\lambda}{q_j m}(S(x^{k-d_k}))_j;$$

$$(\forall j' \neq j) \qquad x_j^{k+1} = x_j^k.$$

When we specialize SMART to this simple single operator case, which is the only case that ARock applies to, there is no difference between ARock and SMART; but Theorem 3.1 guarantees that ARock will converge under conditions weaker than those presented in [29].

In [29], the underlying Hilbert space $\mathcal{H} = \mathcal{H}_1 \times \cdots \times \mathcal{H}_m$ is a product of several other Hilbert spaces, and for ARock to converge, the norm $\| \cdot \|$ on $\mathcal{H}$ must be equal to the standard product norm: $\|x\|_{\text{prod}}^2 = \sum_{j=1}^m \|x_j\|_j^2$. We removed this assumption, and this significantly extends the problems that ARock can solve; in Sections 6.3, 6.4, and 6.5 we present new algorithms that use nonstandard norms.

In [29], the operator S must satisfy

$$(\forall x \in \mathcal{H}), (\forall y \in \mathcal{H}) \qquad \langle S(x) - S(y), x - y \rangle \geq \frac{1}{2}\|S(x) - S(y)\|^2,$$

where the inner product on left and the norm on the right are, again, both the standard ones on $\mathcal{H}$. This cocoercivity condition is plainly stronger than (1.5), and for example, fails for the subgradient projector of Section 4.4.

ARock converges linearly only when

$$(\exists \mu > 0) : (\forall x \in \mathcal{H}), (\forall x^* \in \mathcal{S}) \qquad \langle S(x), x - x^* \rangle \geq \mu\|x - x^*\|^2,$$

which is, of course, plainly stronger than (2.1), and for example, requires that $\mathcal{S}$ is a singleton.

But the biggest limitation of ARock, a limitation that we remove in SMART, is that to solve simple problems, such as (4.1), an extra primal variable must be introduced for each smooth term, and these primal variables are not low dimensional, unlike the dual variables of SMART, which tend to be gradients of functions of the form $f(\langle a_i, x \rangle)$; this difference is comparable to the difference between the low memory methods SAGA/SVRG and and high memory method Finito.

5 What's New: Improving Existing Algorithms

In Section 4 we introduced a few algorithms and described how to recover them with SMART, but we did not discuss new features which are obtainable from SMART. We do that now.

5.1 Weaker Conditions for Convergence

Only when the objectives are differentiable and strongly convex are SVRG, Finito, and SDCA known to converge. But with SMART, strong convexity can be dropped in all cases. Furthermore, in SDCA, the objectives, f_j, need only be convex, proper, and closed.

SAGA and SVRG are known to converge linearly when f is strongly convex. By Theorem 1, they converge linearly when $N^{-1} \sum_{i=1}^{N} \nabla f_i(x_i)$ is essentially strongly quasi-monotone, which occurs, for example, when there is a strongly convex function g_i and a linear map A_i such that $f_i(x) = g_i(A_i x)$.

Finito is also known to converge linearly when f_j is strongly convex. Again, by Theorem 1, Finito converges linearly when its operator S is essentially strongly quasi-monotone; the weakest conditions under which this occurs appears to be an open problem in the study of error bounds.

SDCA is known to converge linearly when each f_j is strongly convex and differentiable with Lipschitz continuous gradient. SDCA will still converge linearly if each f_j is just differentiable, but not necessarily strongly convex; see Lemma C.3 for a proof of this simple fact.

5.2 Proximable Terms

SAGA and SVRG also solve problems in which a single nonsmooth term is added to (4.1):

$$\underset{x \in \mathcal{H}_0}{\text{minimize}}\ g(x) + \frac{1}{N} \sum_{i=1}^{N} f_i(x), \tag{5.1}$$

where $g : \mathcal{H}_0 \to [-\infty, \infty)$ is closed, proper, and convex. In this composite case, the update rule is only slightly changed, for example, in SAGA (4.2), we replace the primal update with

$$x^{k+1} = \mathbf{prox}_{\lambda g}\left(x^k - \lambda \left(\nabla f_{i_k}(x^k) - y_{i_k}^k + \frac{1}{N} \sum_{i=1}^{N} y_i^k \right) \right).$$

This update rule is not a special case of SMART, but with almost no extra work, we can extend our proof of convergence in the synchronous case (Theorem 7.1) to show that the update rule works. However, when SMART is asynchronous, we hit a wall; progress seems unlikely.

Instead of pursuing asynchronous versions of this update, we introduce a new update rule: Let

$$S_i = \frac{1}{LN} \nabla f_i \circ \mathbf{prox}_{L^{-1}g} \qquad i = 1, \ldots, N \qquad \text{and} \qquad S_{N+1} = (I - \mathbf{prox}_{L^{-1}g}),$$

Then, for all $k \in \mathbb{N}$, we get the proximal SAGA update:[16] select $\lambda < (N+1)8^{-1}$ and iterate

$$x^{k+1} = x^k - \lambda \begin{cases} \left(\frac{2}{L(N+1)} \nabla f_i(\mathbf{prox}_{L^{-1}g}(x)) - \frac{2N}{N+1} y_i^k + \frac{1}{N+1} \sum_{i=1}^{N+1} y_i^k \right) & \text{if } i_k = i \text{ and } i < N+1; \\ \left(\frac{2}{N+1}(I - \mathbf{prox}_{L^{-1}g})(x^k) - \frac{2}{N+1} y_{N+1}^k + \frac{1}{N+1} \sum_{i=1}^{N+1} y_i^k \right) & \text{if } i_k = N+1; \end{cases}$$

$$y_i^{k+1} = \begin{cases} \frac{1}{LN} \nabla f_i(\mathbf{prox}_{L^{-1}g}(x^k)) & \text{if } i_k = i \text{ and } i < N+1; \\ \left(I - \mathbf{prox}_{L^{-1}g}\right)(x^k) & \text{if } i = N+1; \\ y_i^k & \text{otherwise.} \end{cases} \tag{5.2}$$

This special case of SMART has a trigger graph that is not completed disconnected, as it is in the standard SAGA algorithm; instead all vertices i in the graph contain a directed edge starting at i and ending at $N+1$.

---

[16] Use $n = N, m = 1, d_k \equiv 0, e_k^i \equiv 0, q_1 \equiv 1, \tau_p = \tau_d = 0, p_{ij} \equiv (2N)^{-1}$ for $i < N+1$ and $p_{(N+1)1} = 2^{-1}, \rho = 1,$
$E = \{(i,i) \mid i \in V\} \cup \{(i, N+1) \mid i \in V\}, p_{i1}^T = (2N)^{-1}$ for $i < N+1, p_{(N+1)1}^T = 1.$

In a similar fashion, proximal SVRG algorithms arise from our choice of S_i, but as in (4.3) and (4.4), the dual variable update probability should be $\rho = \tau^{-1}$ and the trigger graph should be the complete graph on $N + 1$ vertices.

The zero set $\mathcal{S}$ of $S = (N+1)^{-1} \sum_{i=1}^{N+1} S_i$ does not contain the set of minimizers of $f + g$, but it is related to the minimizers (if any exist) through the proximal operator of g:

$$\text{For all } x^* \in \mathcal{S}, \text{ the point } \mathbf{prox}_{\gamma g}(x^*) \text{ minimizes } f + g.$$

Unlike in SAGA and SVRG, the operators S_i do not satisfy $\langle S_i(x) - S_i(x^*), x - x^* \rangle \geq L^{-1} \| S_i(x) - S_i(x^*) \|^2$; but the sum S satisfies (1.5) with

$$\beta_{i1} = \frac{N}{2(N+1)} \qquad i = 1, \ldots, N \qquad \text{and} \qquad \beta_{(N+1)1} = \frac{1}{2(N+1)};$$

in Proposition C.5, we give different possibilities for different γ. But with these parameters, we have

$$\lambda \leq \frac{(N+1)(1 + \mu_g L^{-1})}{(16 + 2N)(1 + \mu_g L^{-1}) - 2N(\sqrt{1 - \mu_f L^{-1}})} \implies \text{linear rate:} \quad 1 - \frac{1 + \mu_g L^{-1} - \sqrt{1 - \mu_f L^{-1}}}{(8 + N)(1 + \mu_g L^{-1}) - N(\sqrt{1 - \mu_f L^{-1}})},$$

where $N^{-1} \sum_{i=1}^{N} f_i$ is μ_f-strongly convex and g is μ_g-strongly convex.[17]

5.3 Coordinate Updates

We recover Finito and SDCA with SMART by performing coordinate updates on an operator S. These operators are block-separable, but all algorithms in Section 4 will still converge if we perform finer, nonseparable updates.

For example, SAGA will converge all the same if we replace all full derivatives ∇f_i with partial derivatives $\nabla_j f_i$ (where f_i is now viewed as a function on a space $\mathcal{H}_1 \times \cdots \times \mathcal{H}_m$, and the partial derivative is taken with respect to the coordinates in $\mathcal{H}_j$, which could be an infinite dimensional space): given x^k, sample $j_k \in \{1, \ldots, m\}$ uniformly at random, and set

$$
\begin{aligned}
x_j^{k+1} &= \begin{cases} x_j^k - \lambda \left(\nabla_j f_{i_k}(x^k) - y_{i_k, j}^k + \frac{1}{N} \sum_{i=1}^{N} y_{i,j}^k \right) & \text{if } j = j_k; \\ x_j^k & \text{otherwise.} \end{cases} \\
y_{i,j}^{k+1} &= \begin{cases} \nabla_j f_i(x^k) & \text{if } i_k = i \text{ and } j_k = j; \\ y_{i,j}^k & \text{otherwise.} \end{cases}
\end{aligned}
\tag{5.3}
$$

This algorithm converges for the same range of λ as (4.2) (namely, for $\lambda < (2L)^{-1}$), but in the strongly convex case, the step size λ that gives the best rate of convergence rate changes to $\lambda = m(4Lm + \mu N)^{-1}$, and correspondingly, the convergence rate changes to $1 - \mu(4Lm + \mu N)^{-1}$.

The story is similar for all algorithms in Section 4.

---

[17] In this case, S is $((N+1)(1 + \mu_g L^{-1}))^{-1}(1 + L^{-1}\mu_g - \sqrt{1 - L^{-1}\mu_f})$ essentially strongly quasi-monotone.

5.4 Importance Sampling and Better β_{ij}

Until now, the constants β_{ij} have been constant and equal to L^{-1}. But by making the finer distinction that ∇f_i is L_i-Lipschitz continuous, can choose larger λ by sampling i_k nonuniformly.

For example, in the SAGA algorithm (4.2), we have

$$p_{i1} = P(i_k = i) = \frac{L_i}{\sum_{i=1}^{n} L_i} \implies \lambda < \frac{1}{\frac{2}{n} \sum_{i=1}^{n} L_i}. \tag{5.4}$$

Compared to $(2 \max\{L_i\})^{-1}$, which is the stepsize obtained with uniform i_k sampling, the above stepsize can be much larger. Similarly

$$i_k \text{ sampled according to (5.4)} \quad \implies \quad \text{linear rate:} \quad 1 - \frac{\mu}{\frac{4}{N} \sum_{i=1}^{N} L_i + \mu N};$$

$$i_k \text{ sampled uniformly} \quad \implies \quad \text{linear rate:} \quad 1 - \frac{\mu}{4 \max_i\{L_i\} + \mu N}.$$

Thus, importance sampling replaces maximums of Lipschitz constants by averages of Lipschitz constants.

For smooth functions, the constants β_{ij} appearing in (1.5) are related to inverse coordinatewise Lipschitz constants, which are the minimal values L_{ij} such that

$$f_i(x + \hat{y}_j) \le f_i(x) + \langle \nabla f_i(x), \hat{y}_j \rangle + \frac{L_{ij}}{2} \|\hat{y}_j\|_j^2; \qquad \hat{y}_j = (0, \ldots, 0, y_j, 0, \ldots, 0);$$

for all $x \in \mathcal{H}$ and $y_j \in \mathcal{H}_j$. From the above inequality, comes the relationship[18]

$$\frac{1}{L_{ij}} \|\nabla_j f_i(x) - \nabla_j f_i(y)\|_j^2 \le \langle \nabla f_i(x) - \nabla f_i(y), x - y \rangle,$$

which is not a favorable one because the right hand side depends on the full derivative ∇f_i.

However, the relationship improves if, say, at most $s \ll n$ partial derivatives $\nabla_j f(x_1, \ldots, x_n)$ depend on each coordinate; from that assumption, comes the relationship

$$\sum_{i=1}^{n} \frac{1}{sL_{ij}} \|\nabla_j f_i(x) - \nabla_j f_i(y)\|_j^2 \le \langle \nabla f_i(x) - \nabla f_i(y), x - y \rangle.$$

But notice that if L_i is the minimal Lipschitz constant ∇f_i, then $L_i^{-1} \ge \min_j\{(sL_{ij})^{-1}\}$.

5.5 Mini Batching

The *pre-update* mini batching method adjusts problem (4.1) by grouping functions together according to $\mathcal{B} \subseteq 2^{\{1, \ldots, N\}}$:

$$\underset{x \in \mathcal{H}}{\text{minimize}} \; \frac{1}{N} \sum_{B \in \mathcal{B}} \sum_{i \in B} \frac{1}{N(i)} f_i; \qquad (\forall i) \; N(i) := |\{B \in \mathcal{B} \mid i \in B\}|.$$

Then it runs one of the algorithms from Section 4 with the $n = |\mathcal{B}|$ functions $\sum_{i \in B} N(i)^{-1} f_i$.

---

[18] See Proposition B.4 for a proof in our general case, and [22, Lemma 4] for a proof in the case that $\mathcal{H}_j = \mathbb{R}$.

The parameter n and hence, the number of dual variables in the pre-update mini batching method can be impractically large. To save memory, only our SVRG clone (4.3) should be used with the pre-update mini batching method; with this method, a mini batch $B_{i_k} \in \mathcal{B}$ of gradients is computed at every iteration, and on average all of the dual variables are, and hence, the full gradient is, updated once per τ^{-1} iterations: label the elements of $\mathcal{B} = \{B_1, \ldots, B_n\}$, and iterate[19]

$$x^{k+1} = x^k - \lambda \left(\sum_{l \in B_{i_k}} (\nabla f_l(x^k) - \nabla f_l(\phi^k)) + \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(\phi^k) \right);$$

$$(\forall i) \qquad \phi^{k+1} = \phi^k + \epsilon_k(x^k - \phi^k).$$

In this SVRG clone, we eliminate all of the dual variables, and as a result, we save a lot of memory.

Unlike the pre-update mini batching method, the *post-update* mini batching method does not adjust problem (4.1); it only adjusts the trigger graph. Consequently, there is no grouping $\mathcal{B}$ which pairs functions together. In place of a grouping, for each function, the trigger graph G *triggers* a gradient computation for some other set of functions:

$$x^{k+1} = x^k - \lambda \left(\nabla f_{i_k}(x^k) - y_{i_k}^k + \frac{1}{N} \sum_{i=1}^{N} y_i^k \right);$$

$$y_i^{k+1} = \begin{cases} \nabla f_i(x^k) & \text{if } i_k \text{ triggers } i; \\ y_i^k & \text{otherwise.} \end{cases}$$

Here the mini batching is arbitrary, but we have already seen three specific examples of trigger graphs: the completely disconnected graph (SAGA), the completely connected graph (SVRG), and an internally directed star graph (Proximal SAGA).

Convergence rates improve with post-update mini batching because the parameter $\eta := \min_{i,j}\{\rho p_{ij}^T\}$, increases. For example, if every node $i \in \{1, \ldots, N\}$ in the trigger graph is triggered by N_T other nodes, then $p_{ij}^T := N^{-1}N_T$, and the convergence rate of SAGA improves:

$$\lambda \leq \frac{1}{4L + \frac{8\mu N}{N_T}} \implies \text{linear rate:} \quad 1 - \frac{\mu}{4L + \frac{8\mu N}{N_T}}.$$

Compared to $1 - \mu(4L + \mu N)^{-1}$, which is the convergence rate for standard SAGA, the above convergence rate can be much better. See [19] for similar convergence rate improvements from mini batching.

5.6 Asynchronous Updates

Besides ARock, all algorithms in Section 4 are synchronous because the primal and dual updates are not delayed. On paper, delaying is just a matter of changing indices, as in the *asynchronous SAGA* algorithm:

$$x^{k+1} = x^k - \lambda \left(\nabla f_{i_k}(x^{k-d_k}) - y_{i_k}^{k-e_k^i} + \frac{1}{N} \sum_{i=1}^{N} y_i^{k-e_k^i} \right);$$

$$y_i^{k+1} = \begin{cases} \nabla f_i(x^{k-d_k}) & \text{if } i_k = i; \\ y_i^k & \text{otherwise.} \end{cases}$$

---

[19] The pre-update mini batching SVRG clone is similar to the algorithm in [21].

However, on a computer, implementing an asynchronous algorithm can be difficult. For a brief discussion on implementation issues, see [29].

6 What's New: Algorithms

On the surface, the shape of an operator plainly resembles the shape of a gradient, and this is already a powerful observation, leading to the SAGA and SVRG algorithms. But operators also model nonsmooth optimization problems and even monotone inclusions; choosing these operators just requires a bit of experience, which anyone can acquire, for example, by understanding the five examples presented in this section.

The examples presented here isolate five common optimization problems, reformulate these problems with operators S_i, and then solve these problems with SMART. Once we choose the operators S_i, we can apply SMART in exponentially many ways, for example, by choosing arbitrary delays, sampling probabilities, and trigger graphs; in this section, we avoid endless customization, and instead, we apply SMART in a simple, arbitrary manner.

6.1 LinSAGA and LinSVRG

LinSAGA and LinSVRG add a linear constraint to the proximal SAGA and proximal SVRG problems:

$$\underset{x \in \mathcal{H}}{\text{minimize}} \; g(x) + \frac{1}{N} \sum_{i=1}^{N} f_i(x);$$

$$\text{subject to: } x \in V, \tag{6.1}$$

where the function $g : \mathcal{H} \to (-\infty, \infty]$ is closed, proper, and convex; the functions $f_i : \mathcal{H} \to (-\infty, \infty)$ are differentiable and the gradients $P_V \circ \nabla f_i \circ P_V$ are $\hat{L}_i$-Lipschitz continuous;[20] and the set $V \subseteq \mathcal{H}$ is a vector space.[21] We assume that P_V and $\mathbf{prox}_{\gamma g}$ are both easy to evaluate.

We model this problem with $N + 1$ operators (for some $\gamma > 0$):

$$(\forall i < N + 1) \qquad S_i := \frac{\gamma}{N} P_V \circ \nabla f_i \circ P_V \circ \mathbf{prox}_{\gamma g};$$

$$S_{N+1} := (I - 2P_V) \circ \mathbf{prox}_{\gamma g} + P_V.$$

The roots of $S := (N + 1)^{-1} \sum_{i=1}^{N+1} S_i$ are not solutions of (6.1), but in general,

$$x \in \text{zer}(S) \implies \mathbf{prox}_{\gamma g}(x) \text{ solves (6.1)},$$

and $\text{zer}(S) \neq \emptyset$ if, and only if, $\text{zer}(\partial g + N^{-1} \sum_{i=1}^{N} \nabla f_i + N_V) \neq \emptyset$.

The following iterative algorithm is a special case of SMART.

---

[20] The operator $P_V \circ \nabla f_i \circ P_V$ is the gradient of the convex function $f_i \circ P_V$, and its Lipschitz constant, which we denote by $\hat{L}_i$, is generally smaller than the Lipschitz constant of ∇f_i (see the discussion surrounding [12, Lemma 1.3].)

[21] Affine constraints, say, $Ax = b$ can replace the linear constraint $x \in V$ provided we choose any $c \in \mathcal{H}$ such that $Ac = b$, change $f(x)$ and $g(x)$ to $f(x + c)$, and $g(x + c)$, respectively, and set $V = \ker(A)$.

Algorithm 2 (LinSAGA/LinSVRG) *Choose initial points* $x^0, y_1^0, \ldots, y_{N+1}^0 \in \mathcal{H}$. *Choose stepsizes satisfying*

$$\lambda \leq \frac{(N+1)}{8} \qquad\qquad and \qquad\qquad \gamma = \frac{N}{\sum_{i=1}^{N} \hat{L}_i}.$$

Then for $k \in \mathbb{N}$, *perform the following four steps:*

1. **Sampling:** *Choose dual update decision* $\epsilon_k \in \{0, 1\}$. *Choose an index* $i_k \in \{1, \ldots, N+1\}$ *with distribution*

$$P(i_k = i) = \begin{cases} \frac{1}{2} & \text{if } i_k = N+1 \\ \frac{\hat{L}_i}{2\sum_{i=1}^{N} \hat{L}_i} & \text{otherwise.} \end{cases}$$

2. **Primal update (gradient case):** *if* $i_k < N+1$, *set*

$$x^{k+1} = x^k - \lambda \left(\frac{\gamma}{N(N+1)p_{i_k 1}} P_V \nabla f_i(P_V \mathbf{prox}_{\gamma g}(x^k)) - \frac{1}{(N+1)p_{i_k 1}} y_{i_k}^k + \frac{1}{N+1} \sum_{i=1}^{N+1} y_i^k \right).$$

3. **Primal update (proximal case):** *if* $i_k = N+1$, *set*

$$x^{k+1} = x^k - \lambda \left(\frac{1}{(N+1)p_{i_k 1}} \left((I - 2P_V)\mathbf{prox}_{\gamma g}(x^k) + P_V x^k\right) - \frac{1}{(N+1)p_{i_k 1}} y_{i_k}^k + \frac{1}{N+1} \sum_{i=1}^{N+1} y_i^k \right).$$

4. **Dual update:** *set*

$$y_i^{k+1} = \begin{cases} y_i^k + \epsilon_k \left(\frac{\gamma}{N} P_V \nabla f_i(P_V \mathbf{prox}_{\gamma g}(x^k)) - y_i^k \right) & \text{if } i_k \text{ triggers } i \text{ and } i < N+1. \\ y_{N+1}^k + \epsilon_k \left((I - 2P_V)\mathbf{prox}_{\gamma g}(x^k) - y_{N+1}^k \right) & \text{if } i_k \text{ triggers } i \text{ and } i = N+1. \end{cases} \qquad \square$$

As in Section 4.1, the difference between LinSAGA and LinSVRG lies in the trigger graph, and how often the dual variables are updated: LinSAGA uses the directed star trigger graph in which every node $i < N+1$ connects to $N+1$; and LinSVRG uses the completely connected trigger graph with dual variable update frequency $P(\epsilon_k = 1) = \tau^{-1}$ for some $\tau > 0$.

The Properties of S. The operator S satisfies the coherence condition (1.5) with

$$(\forall i \leq N) \qquad \beta_{i1} = \frac{N}{2\gamma \hat{L}_i (N+1)} \qquad\qquad and \qquad\qquad \beta_{(N+1)1} = \frac{1}{N+1} \left(1 - \frac{1}{2N} \sum_{i=1}^{N} \gamma \hat{L}_i \right),$$

and with

$$\mu = \frac{1}{N+1} \left(1 - \left(\frac{1}{(1 + (\gamma L_g)^{-1})} + \frac{\sqrt{1 - 2\gamma\mu_f + \gamma^2 L\mu_f}}{(1 + \gamma\mu_g)} \right) \right)$$

the operator S is essentially strongly quasi-monotone (if $\gamma \leq 2L^{-1}$ and $\mu > 0$), where $L = N^{-1} \sum_{i=1}^{N} \hat{L}_i$, the function $N^{-1} \sum_{i=1}^{N} f_i$ is μ_f-strongly convex, the function g is differentiable and μ_g-strongly convex, and the gradient ∇g is L_g-Lipschitz continuous.

6.2 SuperSAGA and SuperSVRG

SuperSAGA and SuperSVRG add a few nonsmooth terms to the smooth problem (4.1):

$$\underset{x \in \mathcal{H}_1}{\text{minimize}} \sum_{j=1}^{M} g_j(z) + \frac{1}{N} \sum_{i=1}^{N} f_i(z), \tag{6.2}$$

where the functions $g_j : \mathcal{H}_1 \to (-\infty, \infty]$ are closed, proper, and convex; and the functions $f_i : \mathcal{H} \to (-\infty, \infty)$ are differentiable and the gradients ∇f_i are L_i-Lipschitz continuous.

It is more convenient for us to work with the reformulated problem:

$$\underset{z \in \mathcal{H}}{\text{minimize}}\ g(x) + \frac{1}{N} \sum_{i=1}^{N} f_i(x_1)$$

$$\text{subject to: } x \in D \tag{6.3}$$

where we define the spaces $\mathcal{H} := \mathcal{H}_1^M$ and $D := \{x \in \mathcal{H} \mid x_1 = \cdots = x_M\}$; and for all $x \in \mathcal{H}$, we let $g(x) := \sum_{j=1}^{M} g(x_i)$.

Problem (6.3) is evidently a special case of Problem (6.1), so we choose the operators that worked well there:

$$(\forall i < N + 1) \qquad S_i := \frac{\gamma}{N} P_D \circ \nabla f_i \circ P_D \circ \mathbf{prox}_{\gamma g};$$

$$S_{N+1} := (I - 2P_D) \circ \mathbf{prox}_{\gamma g} + P_D.$$

Provided that M is relatively small, and for each j, the operator $\mathbf{prox}_{\gamma g_j}$ is easy to evaluate, the operators S_i are easy to evaluate. For example, for all $x \in \mathcal{H}$ and $i < N + 1$, we have

$$P_D x = \frac{1}{M} \sum_{j=1}^{M} x_j; \quad \mathbf{prox}_{\gamma g}(x) = (\mathbf{prox}_{\gamma g_j}(x_j))_{j=1}^{M}; \quad S_i(x) = \left(\frac{\gamma}{NM} \nabla f_i \left(\frac{1}{M} \sum_{l=1}^{M} \mathbf{prox}_{\gamma g_l}(x_l) \right) \right)_{j=1}^{M}.$$

And as before, the roots of $S := (N + 1)^{-1} \sum_{i=1}^{N+1} S_i$ are not solutions of (6.2), but in general,

$$x^* \in \text{zer}(S) \implies (\forall j)\ \mathbf{prox}_{\gamma g_j}(x_j^*) \text{ solves } (6.2),$$

and $\text{zer}(S) \neq \emptyset$ if, and only if, $\text{zer}(\sum_{j=1}^{M} \partial g_j + N^{-1} \sum_{i=1}^{N} \nabla f_i) \neq \emptyset$.

The following iterative algorithm is a special case of SMART.

Algorithm 3 (SuperSAGA/SuperSVRG) *Choose initial points* $x^0 \in \mathcal{H}$ *and* $\overline{y}_1^0, \ldots, \overline{y}_{N+1}^0 \in \mathcal{H}_1$. *Choose stepsizes satisfying*

$$\lambda \leq \frac{(N+1)}{8} \qquad \text{and} \qquad \gamma = \frac{MN}{\sum_{i=1}^N L_i}.$$

Then for $k \in \mathbb{N}$, *perform the following four steps:*

1. **Sampling:** *Choose dual update decision* $\epsilon_k \in \{0,1\}$. *Choose an index* $i_k \in \{1, \ldots, N+1\}$ *with distribution*

$$P(i_k = i) = \begin{cases} \frac{1}{2} & \text{if } i = N+1 \\ \frac{L_i}{2\sum_{i=1}^N L_i} & \text{otherwise.} \end{cases}$$

2. **Primal update (gradient case):** *if* $i_k < N+1$, *set*

$$(\forall j) \quad x_j^{k+1} = x_j^k - \lambda \left(\frac{\gamma}{MN(N+1)p_{i_k 1}} \nabla f_i \left(\frac{1}{M} \sum_{j=1}^M w_j^k \right) - \frac{1}{(N+1)p_{i_k 1}} \overline{y}_{i_k}^k + \frac{1}{N+1} \sum_{i=1}^k \overline{y}_i^k \right);$$

$$w_j^{k+1} = \mathbf{prox}_{\gamma g_j}(x_j^{k+1}).$$

3. **Primal update (proximal case):** *if* $i_k = N+1$, *set*

$$(\forall j) \quad x_j^{k+1} = x_j^k - \lambda \left(\frac{1}{(N+1)p_{i_k 1}} \left(w_j^k - \frac{1}{M} \sum_{j=1}^M (2w_j^k - x_j^k) \right) - \frac{1}{(N+1)p_{i_k 1}} \overline{y}_{i_k}^k + \frac{1}{N+1} \sum_{i=1}^k \overline{y}_i^k \right);$$

$$w_j^{k+1} = \mathbf{prox}_{\gamma g_j}(x_j^{k+1}).$$

4. **Dual update:** *set*

$$\overline{y}_i^{k+1} = \begin{cases} \overline{y}_i^k + \epsilon_k \left(\frac{\gamma}{MN} \nabla f_i \left(\frac{1}{M} \sum_{j=1}^M w_j^k \right) - \overline{y}_i^k \right) & \text{if } i_k \text{ triggers } i \text{ and } i < N+1; \\ \overline{y}_i^k + \epsilon_k \left(\frac{1}{M} \sum_{j=1}^M (x_j^k - w_j^k) - \overline{y}_i^k \right) & \text{if } i_k \text{ triggers } i \text{ and } i = N+1. \end{cases} \qquad \square$$

The difference between SuperSAGA and SuperSVRG, again, lies in the trigger graph and the dual variable update frequency; see the comments following Algorithm 2.

The Properties of S. The operator S is an instance of the LinSAGA/LinSVRG operator defined in Section 6.1, so the two operators satisfy the coherence condition (1.5) with the same constants[22]

$$(\forall i \leq N) \qquad \beta_{i1} = \frac{NM}{2\gamma L_i(N+1)} \qquad \text{and} \qquad \beta_{(N+1)1} = \frac{1}{N+1} \left(1 - \frac{1}{2N} \sum_{i=1}^N \frac{\gamma L_i}{M} \right),$$

and both operators are essentially strongly monotone under the same conditions.

In fact, the SuperSAGA/SuperSVRG algorithms and the LinSAGA/LinSVRG algorithms differ in just one way: for SuperSAGA/SuperSVRG, the dual variables $\overline{y}_i^k$ are vectors in $\mathcal{H}_0$ rather than vectors y_i^k in

---

[22] Clearly, $\hat{L}_i = L_i M^{-1}$.

$\mathcal{H} = \mathcal{H}_0^M$, and that saves some memory. We can use $\overline{y}_i^k$ in place of y_i^k by viewing S as an operator not on the space $\mathcal{H}^M$, but as an operator on the orthogonal decomposition $\mathcal{H} = D \oplus D^\perp$, which has just two components. Then, the identity

$$S_{N+1} = P_{D^\perp}\mathbf{prox}_{\gamma g} + P_D(I - \mathbf{prox}_{\gamma g})$$

allows us to reparamterize the components of S_i into a D component and $D^\perp$ component: for all $i \leq N$,

$$(S_i)_D = \frac{\gamma}{N}P_D \circ \nabla f_i \circ P_D \circ \mathbf{prox}_{\gamma g}; \qquad\qquad (S_i(x))_{D^\perp} \equiv 0;$$

$$(S_{N+1})_D = P_D(I - \mathbf{prox}_{\gamma g}); \qquad\qquad (S_{N+1})_{D^\perp} = P_{D^\perp}\mathbf{prox}_{\gamma g}.$$

Then, with this new decomposition, we have $(S_i(x^*))_{D^\perp} = 0$ for all i, which makes saving the full vector y_i^k superfluous; we need only store the component in D, which is precisely the vector $\overline{y}_i^k$.

   The above approach saves memory when all M components of x are updated at every iteration *and* the coordinates are consistent-read asynchronous or totally synchronous. However, if only some of the coordinates x_j are updated at each iteration *or* inconsistent-read updates are performed, Theorem 3.1 ceases to apply; in either of those cases, we must use the full dual variables y_i^k.

6.3 TropicSMART: Randomized Smoothly Coupled Monotropic Programming

The TropicSMART problem is different from all problems we have seen so far:

$$\begin{aligned} \underset{x_j \in \mathcal{H}_j}{\text{minimize}} \ & \sum_{j=1}^{M} g_j(x_j) + f(x_1, \ldots, x_M); \\ \text{subject to: } & \sum_{j=1}^{M} A_j x_j = b, \end{aligned} \tag{6.4}$$

where the sets $\mathcal{H}_j$ $(j = 1, \ldots, M+1)$ are Hilbert spaces; the sets $\mathcal{H}' := \mathcal{H}_1 \times \cdots \times \mathcal{H}_M$ and $\mathcal{H} := \mathcal{H}_1 \times \cdots \times \mathcal{H}_{M+1}$ are product spaces; the functions $g_j : \mathcal{H}_j \to (-\infty, \infty]$ are closed, proper, and convex (we also let $g(x) := \sum_{j=1}^{M} g(x_j)$); the function $f : \mathcal{H}' \to (-\infty, \infty)$ is differentiable and ∇f is L-Lipschitz continuous; the linear maps $A_j : \mathcal{H}_j \to \mathcal{H}_{M+1}$ are continuous; and $b \in \mathcal{H}_{M+1}$.

   There is only one TropicSMART operator: for all $x \in \mathcal{H}$, define

$$(S(x))_j := \begin{cases} x_j - \mathbf{prox}_{\gamma_j g_j}\left(x_j - \gamma_j A_j^*\left(x_{M+1} + 2\gamma_{M+1}\left(\sum_{l=1}^{M} A_l x_l - b\right)\right) - \gamma_j \nabla_j f(x)\right) & \text{if } j < M+1; \\ -\gamma_{M+1}\left(\sum_{l=1}^{M} A_l x_l - b\right) & \text{if } j = M+1. \end{cases}$$

The roots of $S := (N+1)^{-1}\sum_{i=1}^{N+1} S_i$ are not solutions of (6.4), but in general,

$$x^* \in \text{zer}(S) \implies (x_1^*, \ldots, x_M^*) \text{ solves (6.2)},$$

and $\text{zer}(S) \neq \emptyset$ if, and only if, $\text{zer}\left(\partial g + \nabla f + N_{\{x \in \mathcal{H} | \sum_{j=1}^{M} A_j x_j = b\}}\right) \neq \emptyset$.

   The following iterative algorithm is a special case of SMART.

Algorithm 4 (TropicSMART) *Choose initial points $x^0 \in \mathcal{H}$. Choose $\delta \in (0,1)$, and choose stepsizes satisfying*

$$\gamma_{M+1}\left(\sum_{j=1}^{M}\gamma_j\|A_j\|^2\right) \le \delta; \qquad \max_j\{\gamma_j\} \le \frac{2(1-\sqrt{\delta})}{L}; \qquad \lambda \le \frac{L\max_j\{\gamma_j\}}{4(1+\sqrt{\delta})}.$$

Then for $k \in \mathbb{N}$, perform the following two steps:

1. **Sampling:** *Choose a coordinate $j_k \in \{1,\dots,M+1\}$ uniformly at random and set $\mathsf{S}_k = \{j_k\}$.*
2. **Primal update:** *set*

$$\overline{x}_{M+1}^{k+1} = x_{M+1}^k + \gamma_{M+1}\left(\sum_{l=1}^{M}A_l x_l^k - b\right);$$

$$(\forall j \in \mathsf{S}_k\backslash\{M+1\}) \qquad \overline{x}_j^{k+1} = \mathbf{prox}_{\gamma_j g_j}\left(x_j^k - \gamma_j A_j^*(2\overline{x}_{M+1}^{k+1} - x_{M+1}^k) - \gamma_j \nabla_j f(x^k)\right);$$

$$(\forall j \in \mathsf{S}_k) \qquad x_j^{k+1} = x_j^k - \lambda\left(x_j^k - \overline{x}_j^{k+1}\right);$$

$$(\forall j \notin \mathsf{S}_k) \qquad x_j^{k+1} = x_j^k. \qquad \square$$

The Properties of S. The operator S satisfies the coherence condition (1.5) with

$$(\forall j) \qquad \beta_{1j} = \frac{L\max_j\{\gamma_j\}}{4\gamma_j}.$$

But the TropicSMART operator S does not satisfy the coherence condition in the standard metric on $\mathcal{H}$; instead there is a strongly positive self-adjoint linear operator P (e.g., with respect to $\langle\cdot,\cdot\rangle_{\mathrm{prod}}$, a symmetric positive definite matrix) so that

$$(\forall x \in \mathcal{H}),(\forall x^* \in \mathrm{zer}(S)) \qquad \langle S(x), x - x^*\rangle_P \ge \sum_{j=1}^{M+1}\beta_{1j}\|(S(x))_j\|_j^2,$$

and this linear operator P satisfies $\sum_{j=1}^{M+1}\underline{M}_j\|x_j\|_j^2 \le \|x\|_P^2 \le \sum_{j=1}^{M+1}\overline{M}_j\|x_j\|_j^2$, where for all j, we have

$$\underline{M}_j := \frac{1-\sqrt{\delta}}{\gamma_j} \qquad\qquad \text{and} \qquad\qquad \overline{M}_j := \frac{1+\sqrt{\delta}}{\gamma_j}.$$

The weakest conditions under which S is essentially strongly quasi-monotone are not known.

6.4 ProxSMART: Randomized Proximable Optimization

The ProxSMART algorithm solves the following proximable optimization problem:

$$\underset{z \in \mathcal{H}_1}{\text{minimize }} g_1(z) + \sum_{j=2}^{M} g_j(A_j z), \tag{6.5}$$

where the sets $\mathcal{H}_j$ $(j = 1, \ldots, M+1)$ are Hilbert spaces; the set $\mathcal{H} := \mathcal{H}_1 \times \cdots \times \mathcal{H}_M$ is a product space; the functions $g_j : \mathcal{H}_j \to (-\infty, \infty]$ are closed, proper, convex, and nonsmooth; and the linear maps $A_j : \mathcal{H}_1 \to \mathcal{H}_j$ are continuous.

There is only one ProxSMART operator: for all $x \in \mathcal{H}$, define

$$(S(x))_j := (S_1(x))_j = \begin{cases} x_1 - \mathbf{prox}_{\gamma_1 g_1} \left(x_1 - \gamma_1 \sum_{j=2}^{M} A_j^* x_j \right) & \text{if } j = 1; \\ x_j - \mathbf{prox}_{\gamma_j g_j^*} \left(x_j + \gamma_j A_j \left(2\overline{x}_1 - x_1 \right) \right) & \text{otherwise;} \end{cases}$$

where $\overline{x}_1 = \mathbf{prox}_{\gamma_1 g_1} \left(x_1 - \gamma_1 \sum_{j=2}^{M} A_j^* x_j \right)$. The roots of $S := (N+1)^{-1} \sum_{i=1}^{N+1} S_i$ are not solutions of (6.5), but in general,

$$x^* \in \mathrm{zer}(S) \implies x_1 \text{ solves (6.5)},$$

and $\mathrm{zer}(S) \neq \emptyset$ if, and only if, $\mathrm{zer}(\partial g_1(x) + \sum_{j=2}^{M} A_j^* \partial g_j \circ A_j) \neq \emptyset$.

The following iterative algorithm is a special case of SMART.

Algorithm 5 (ProxSMART) *Choose initial point $x^0 \in \mathcal{H}$. Choose $\delta \in (0, 1)$, and choose stepsizes satisfying*

$$\gamma_1 \left(\sum_{j=2}^{M} \gamma_j \|A_j\|^2 \right) \leq \delta \qquad \text{and} \qquad \lambda \leq \frac{1 - \sqrt{\delta}}{1 + \sqrt{\delta}}.$$

Then for $k \in \mathbb{N}$, perform the following two steps:

1. **Sampling.** *Choose a set of functions (their indices) S_k.*
2. **Primal update:** *set*

$$\overline{x}_1^{k+1} = \mathbf{prox}_{\gamma_1 g_1} \left(x_1^k - \gamma_1 \sum_{j=2}^{M} A_j^* x_j^k \right);$$

$$(\forall j \in \mathsf{S}_k \backslash \{1\}) \qquad \overline{x}_j^{k+1} = \mathbf{prox}_{\gamma_j g_j^*} \left(x_j^k + \gamma_j A_j \left(2\overline{x}_1^{k+1} - x_1^k \right) \right);$$

$$(\forall j \in \mathsf{S}_k) \qquad x_j^{k+1} = x_j^k - \frac{\lambda}{q_j M} \left(x_j^k - \overline{x}_j^{k+1} \right);$$

$$(\forall j \notin \mathsf{S}_k) \qquad x_j^{k+1} = x_j^k. \qquad \square$$

The variables x_j for which $j > 1$ are different from the dual variables in SMART because here $n = 1$. However, the x_j dual variables play a similar role to the y dual variables, and like the dual variables in SMART, they are often low dimensional—if A_i is a row vector, then x_j is a scalar. Another similarity, arising in the case $\lambda = 1$, is that for $j > 1$, x_j^k is a subgradient, much like the dual variables in SMART are gradients in SAGA and SVRG.[23] But even if A_i is the identity map, and λ is arbitrary, x_j^k is always in the linear span of ∂f, which may be low dimensional.

---

[23] If $z^+ = \mathbf{prox}_{\gamma_j g^*}(z)$, then $z^+ \in \partial g(\gamma^{-1}(z - z^+))$

The Properties of S. The operator S satisfies the coherence condition (1.5) with

$$(\forall j) \qquad \beta_{1j} = \frac{(1 - \sqrt{\delta})}{\gamma_j}.$$

But like the TropicSMART operator from Section 6.3, the ProxSMART operator S does not satisfy the coherence condition in the standard metric on $\mathcal{H}$; instead there is a strongly positive self-adjoint linear operator P so that

$$(\forall x \in \mathcal{H}), (\forall x^* \in \operatorname{zer}(S)) \qquad \langle S(x), x - x^* \rangle_P \geq \sum_{j=1}^{M} \beta_{1j} \|(S(x))_j\|_j^2,$$

and this linear operator P satisfies $\sum_{j=1}^{M} \underline{M}_j \|x_j\|_j^2 \leq \|x\|_P^2 \leq \sum_{j=1}^{M} \overline{M}_j \|x_j\|_j^2$, where for all $j > 0$, we have

$$\underline{M}_j := \frac{1 - \sqrt{\delta}}{\gamma_j} \qquad \text{and} \qquad \overline{M}_j := \frac{1 + \sqrt{\delta}}{\gamma_j}.$$

essentially strongly quasi-monotone, provided that g_1 is μ_1-strongly convex and each function g_j, $j > 1$, is differentiable, the gradient ∇g_j is L_j-Lipschitz continuous, and the constants γ_j are chosen small enough; we omit the proof of this fact. The weakest conditions under which S is essentially strongly quasi-monotone are not known.

6.5 ProxSMART+: Randomized Composite Optimization

The ProxSMART+ algorithm solves the following composite optimization problem:

$$\underset{z \in \mathcal{H}_1}{\operatorname{minimize}} \sum_{j=2}^{M} g_j(A_j z) + \frac{1}{N} \sum_{i=1}^{N} f_i(z). \tag{6.6}$$

where the sets $\mathcal{H}_j$ $(j = 1, \ldots, M + 1)$ are Hilbert spaces; the set $\mathcal{H} := \mathcal{H}_1 \times \cdots \times \mathcal{H}_M$ is a product space; the functions $g_j : \mathcal{H}_j \to (-\infty, \infty]$ are closed, proper, and convex; the functions $f_i : \mathcal{H}_1 \to (-\infty, \infty)$ are differentiable and ∇f_i is L_i-Lipschitz continuous; the maps $A_j : \mathcal{H}_1 \to \mathcal{H}_j$ are continuous linear maps.

We model the ProxSMART+ problem with $N + 1$ operators: for all $x \in \mathcal{H}$, define

$$(\forall i < N + 1) \qquad (S_i(x))_j := \begin{cases} \frac{\gamma_1}{N} \nabla f_i \left(x_1 - 2\gamma_1 \sum_{j=2}^{M} A_j^* x_j \right) & \text{if } j = 1; \\ 0 & \text{otherwise.} \end{cases}$$

$$(S_{N+1}(x))_j := \begin{cases} \gamma_1 \sum_{j=2}^{M} A_j^* x_j & \text{if } j = 1; \\ x_j - \mathbf{prox}_{\gamma_j g_j^*} \left(x_j + \gamma_j A_j \left(x_1 - 2\gamma_1 \sum_{j=2}^{M} A_j^* x_j \right) \right) & \text{otherwise.} \end{cases}$$

The roots of $S := (N+1)^{-1} \sum_{i=1}^{N+1} S_i$ are not solutions of (6.6), but in general,

$$x^* \in \operatorname{zer}(S) \implies x_1^* - 2\gamma_1 \sum_{j=2}^{M} A_j^* x_j^* \text{ solves (6.6)}$$

and $\operatorname{zer}(S) \neq \emptyset$ if, and only if, $\operatorname{zer}(\sum_{j=2}^{M} A_j^* \partial g_j \circ A_j + N^{-1} \sum_{i=1}^{N} \nabla f_i) \neq \emptyset$.

The following iterative algorithm is a special case of SMART.

Algorithm 6 (ProxSMART+) *Choose initial points* $x^0 \in \mathcal{H}$ *and* $y_{11}, \ldots, y_{(N+1)1} \in \mathcal{H}_1$. *Choose* $\delta \in (0,1)$, *and choose stepsizes satisfying*

$$\gamma_1 \left(\sum_{j=2}^{M} \gamma_j \|A_j\|^2 + \frac{1}{2N} \sum_{i=1}^{N} L_i \right) \leq \delta \quad and \quad \lambda \leq \frac{(N+1)(1-\sqrt{\delta})M}{2 \left(1 + \sqrt{\delta} - \frac{1}{2\sqrt{\delta}N} \sum_{i=1}^{N} L_i \right) \left(\frac{\gamma_1}{N} \sum_{i=1}^{N} L_i} + 1 \right)}.$$

Then for $k \in \mathbb{N}$, *perform the following four steps:*

1. **Sampling.** *Choose dual update decision* $\epsilon_k \in \{0,1\}$. *Choose* $j_k \in \{1, \ldots, m\}$ *with distribution* $(j > 1)$

$$q_1 = P(j_k = 1) = \frac{1}{\frac{M-1}{\frac{\gamma_1}{N} \sum_{i=1}^{N} L_i} + 1}, \qquad\qquad q_j = P(j_k = j) = \frac{1 - q_1}{M-1},$$

and set $\mathsf{S}_k = \{j_k\}$. *Also, given* j_k, *choose* $i_k \in \{1, \ldots, N+1\}$ *with distribution* $(i < N+1$ *and* $j > 1)$

$$p_{i1} = \frac{\gamma_1 L_i}{N \left(\frac{\gamma_1}{N} \sum_{i=1}^{N} L_i + 1 \right)}; \qquad\qquad p_{(N+1)1} = 1 - \sum_{i=1}^{N} p_{i1};$$

$$p_{ij} = 0; \qquad\qquad\qquad\qquad\qquad p_{(N+1)j} = 1.$$

2. **Primal update (gradient case):** *if* $i_k < N+1$ *and* $\hat{x}_1^k = x_1^k - 2\gamma_1 \sum_{j=2}^{M} A_j^* x_j^k$, *set*

$$\overline{x}_1^{k+1} = x_1^k - \frac{\lambda}{q_1 M} \left(\frac{\gamma_1}{N(N+1)p_{i_k 1}} \nabla f_{i_k} \left(\hat{x}_1^k \right) - \frac{1}{(N+1)p_{i_k 1}} y_{i_k,1}^k + \frac{1}{N+1} \sum_{i=1}^{N+1} y_{i,1}^k \right);$$

$$(\forall j > 1) \qquad x_j^{k+1} = x_j^k.$$

3. **Primal update (proximal case):** *if* $i_k = N+1$ *and* $\hat{x}_1^k = x_1^k - 2\gamma_1 \sum_{j=2}^{M} A_j^* x_j^k$, *set*

$$\overline{x}_1^{k+1} = x_1^k - \left(\frac{\gamma_1}{(N+1)p_{i_k 1}} \sum_{j=2}^{M} A_j^* x_j^k - \frac{1}{(N+1)p_{i_k 1}} y_{i_k,1}^k + \frac{1}{N+1} \sum_{i=1}^{N+1} y_{i,1}^k \right);$$

$$(\forall j \in \mathsf{S}_k \backslash \{1\}) \qquad \overline{x}_j^{k+1} = \mathbf{prox}_{\gamma_j g_j^*} \left(x_j^k + \gamma_j A_j \hat{x}_1^k \right);$$

$$(\forall j \in \mathsf{S}_k) \qquad x_j^{k+1} = x_j^k - \frac{\lambda}{q_j M} \left(x_j^k - \overline{x}_j^{k+1} \right);$$

$$(\forall j \notin \mathsf{S}_k) \qquad x_j^{k+1} = x_j^k.$$

4. **Dual update:** *if* $\hat{x}_1^k = x_1^k - 2\gamma_1 \sum_{j=2}^{M} A_j^* x_j^k$, *set*

$$y_{i,1}^{k+1} = \begin{cases} y_{i,1}^k + \epsilon_k \left(\frac{\gamma_1}{N} \nabla f_i \left(\hat{x}_1^k \right) - y_{i,1}^k \right) & \text{if } i_k \text{ triggers } i \text{ and } i < N+1; \\ y_{i,1}^k + \epsilon_k \left(\gamma_1 \sum_{j=2}^{M} A_j^* x_j^k - y_{i,1}^k \right) & \text{if } i_k \text{ triggers } i \text{ and } i = N+1. \end{cases} \qquad \square$$

As in the ProxSMART+ algorithm, the dual variables x_j with $j > 1$, are subgradients of g_j, and they often live in low-dimensional spaces; see the comments immediately following Algorithm 6.

We save a bit of memory on the y_i dual variables, too, because for all $j > 1$ and for all i, $\mathbf{S}_{ij}^* = 0$; thus, we only maintain the first component of these dual variables.

The Properties of S. The operator S satisfies the coherence condition (1.5) with

$$(\forall 1 \leq i < N+1) \qquad \beta_{i1} = \frac{N(1-\sqrt{\delta})}{2(N+1)\gamma_1^2 L_i} \qquad \text{and} \qquad (\forall j) \qquad \beta_{(N+1)j} = \frac{1-\sqrt{\delta}}{\gamma_1}.$$

But like the TropicSMART and ProxSMART operators from Sections 6.3 and 6.4, the ProxSMART+ operator S does not satisfy the coherence condition in the standard metric on $\mathcal{H}$; instead there is a strongly positive self-adjoint linear operator P so that

$$(\forall x \in \mathcal{H}), (\forall x^* \in \text{zer}(S)) \qquad \langle S(x), x - x^* \rangle_P \geq \sum_{i=1}^{N} \sum_{j=1}^{M} \beta_{ij} \|(S_i(x))_j - (S_i(x^*))_j\|_j^2,$$

and this linear operator P satisfies $\sum_{j=1}^{M} \underline{M}_j \|x_j\|_j^2 \leq \|x\|_P^2 \leq \sum_{j=1}^{M} \overline{M}_j \|x_j\|_j^2$ where for all $j > 0$, we have

$$\underline{M}_j := \frac{1-\sqrt{\delta}}{\gamma_j} \qquad \text{and} \qquad \overline{M}_j := \frac{1+\sqrt{\delta}}{\gamma_j}.$$

The operator S is essentially strongly quasi-monotone, provided that $N^{-1}\sum_{i=1}^{N} f_i$ is μ_1-strongly convex, each function g_j, $j > 1$, is differentiable, each gradient ∇g_j is $\overline{L}_j$-Lipschitz continuous, and the constants γ_j are chosen small enough; we omit the proof of this fact. The weakest conditions under which S is essentially strongly quasi-monotone are not known.

7 Proof in the Synchronous Case

The proof that SMART converges is illuminating when the algorithm is totally synchronous, i.e., when $d_k \equiv 0$ and $e_k^i \equiv 0$. The proof of the following theorem should be read as a warmup before moving onto the proof of the full theorem, which is presented in Appendix A.

Theorem 7.1 (Convergence of SMART in the Synchronous Case) *For all $k \geq 0$, let $\mathcal{I}_k = \sigma((i_k, \mathsf{S}_k))$, let $\mathcal{E}_k = \sigma(\varepsilon_k)$, let*

$$\mathcal{F}_k := \sigma(\{x^l\}_{l=0}^{k} \cup \{y_1^l, \ldots, y_n^l\}_{l=0}^{k}),$$

and suppose that $\{\mathcal{I}_k, \mathcal{E}_k, \mathcal{F}_k\}$ are independent. Suppose that (1.5) holds. Finally assume that $d_k \equiv 0$, $e_k^{ij} \equiv 0$, and $\lambda_k := \lambda$ satisfies

$$\lambda < \begin{cases} \min_{i,j} \left\{ \frac{n^2 p_{ij} \beta_{ij} q_j m}{\overline{M}_j} \right\} & \text{if } \mathbf{S}^* = 0; \\ \min_{i,j} \left\{ \frac{n^2 p_{ij} \beta_{ij} q_j m}{2\overline{M}_j} \right\} & \text{otherwise.} \end{cases}$$

Then

1. **Convergence of operator values.** For $i \in \{1, \ldots, n\}$, the sequence of $\mathcal{H}$-valued random variables $\{S_i(x^k)\}_{k \in \mathbb{N}}$ a. s. converges strongly to $S_i(x^*)$.
2. **Weak convergence.** Suppose that $I - S$ is demiclosed at 0. Then the sequence of $\mathcal{H}$-valued random variables $\{x^k\}_{k \in \mathbb{N}}$ a. s. weakly converges to an $\mathcal{S}$-valued random variable.
3. **Linear convergence.** Let $\eta := \min_{i,j}\{\rho p_{ij}^T\}$, let $\alpha \in [0, 1)$, and let

$$\lambda \leq \begin{cases} \min_{i,j}\left\{\frac{(1-\alpha)\beta_{ij}n^2 p_{ij}q_j m}{\overline{M}_j}\right\} & \text{if } \mathbf{S}^* = 0; \\ \min_{i,j}\left\{\frac{\eta(1-\alpha)\beta_{ij}n^2 p_{ij}q_j m}{2\overline{M}_j\eta + 2\mu\alpha(1-\alpha)\beta_{ij}n^2 p_{ij}q_j}\right\} & \text{otherwise.} \end{cases} \tag{7.1}$$

Then if (2.1) holds, there exists a constant $C(z^0, \phi^0) \in \mathbf{R}_{\geq 0}$ depending on x^0 and ϕ^0 such that for all $k \in \mathbf{N}$,

$$\mathbb{E}\left[d_{\mathcal{S}}^2(x^k)\right] \leq \left(1 - \frac{2\alpha\mu\lambda}{m}\right)^k \left(d_{\mathcal{S}}^2(x^0) + C(x^0, \phi^0)\right).$$

Proof **Notation.** Let $x^* \in \mathcal{S}$. The following quantities are often repeated—in both the synchronous and asynchronous settings.

1. **The t_k variable.** For all $k \geq 0$, let t_k be the $\{0, (q_1 m)^{-1}\} \times \cdots \times \{0, (q_m m)^{-1}\}$-valued random variable, which for all $j \in \{1, \ldots, m\}$, satisfies

$$t_{k,j} := \begin{cases} \frac{1}{q_j m} & \text{if } j \in \mathsf{S}_k; \\ 0 & \text{otherwise.} \end{cases}$$

2. **The primal and dual Q_i operators.** For all i, define two new operators $Q_i^p, Q_i^d : \mathcal{H} \to \mathcal{H}$: for all $x, y \in \mathcal{H}$, set

$$(Q_i^p(x))_j := \begin{cases} \frac{1}{np_{ij}}(S_i(x))_j & \text{if } p_{ij} \neq 0; \\ 0 & \text{otherwise;} \end{cases} \quad \text{and} \quad (Q_i^d(y))_j := \begin{cases} \frac{1}{np_{ij}}y_{i,j} & \text{if } p_{ij} \neq 0; \\ 0 & \text{otherwise.} \end{cases}$$

   For all i, define $Q_i^* := Q_i^p(x^*)$. By (1.5), Q_i^* is independent of the choice of $x^* \in \operatorname{zer}(S)$.
3. **The primal and dual r_{ij}^p functions.** For each i and j, define two functions $r_{ij}^p, r_{ij}^d : \mathcal{H} \to \mathbb{R}_+$: for all $x, y \in \mathcal{H}$, let

$$r_{ij}^p(x) := \|(Q_i^p(x))_j - (Q_i^*)_j\|_j^2 \quad \text{and} \quad r_{ij}^d(y) := \|(Q_i^d(y))_j - (Q_i^*)_j\|_j^2.$$

4. **The $\mathfrak{Q}^k$ random vector.** If

$$\mathfrak{Q}^k = t_k \odot \left(Q_{i_k}^p(x^k) - Q_{i_k}^d(y_{i_k}^k) + \frac{1}{n}\sum_{i=1}^n y_i^k\right),$$

   then $x^{k+1} = x^k - \lambda\mathfrak{Q}^k$ and $\mathbb{E}\left[\mathfrak{Q}^k \mid \mathcal{F}_k\right] = m^{-1}S(x^k)$.
5. **The γ_{ij} constants.** For all i and j, choose any constants that satisfy[24]

$$\frac{2\overline{M}_j\lambda^2}{q_j m^2} < \gamma_{ij} < \frac{2\lambda\beta_{ij}n^2 p_{ij}}{m} - \frac{2\overline{M}_j\lambda^2}{q_j m^2}. \tag{7.2}$$

---

[24] The constraint (7.2) has a solution by the assumptions of the theorem.

6. **The $R_{l,ij}$ constants.** For all i and j, we define two *positive constants*[25]

$$R_{1,ij} := \gamma_{ij} - \frac{2\overline{M}_j\lambda^2}{q_j m^2} \qquad \text{and} \qquad R_{2,ij} := \frac{2\lambda\beta_{ij}n^2 p_{ij}}{m} - \frac{2\overline{M}_j\lambda^2}{q_j m^2} - \gamma_{ij}.$$

Assumption 1 *We only address the case in which* $\mathbf{S}^*$ *is not necessarily the zero matrix. From this case, and with the help of the identity* $r_{ij}^d(y_i^k) \equiv 0$*, it is straightforward to retrieve the case in which* $\mathbf{S}^* = 0$*.*

Parts 1 and 2: Two essential elements feature in our proof. The indispensable *supermartingale convergence theorem* [32, Theorem 1], with which we show that a pivotal sequence of random variables converges, is our hammer for nailing down the effect of randomness in SMART:

Theorem 7.2 (Supermartingale convergence theorem) *Let* $(\Omega, \mathcal{F}, P)$ *be a probability space. Let* $\mathfrak{F} := \{\mathcal{F}_k\}_{k\in\mathbb{N}}$ *be an increasing sequence of sub* σ*-algebras of* $\mathcal{F}$ *such that* $\mathcal{F}_k \subseteq \mathcal{F}_{k+1}$*. Let* $\{X_k\}_{k\in\mathbb{N}}$ *and* $\{Y_k\}_{k\in\mathbb{N}}$ *be sequences of* $[0,\infty)$*-valued random variables such that for all* $k \in \mathbb{N}$*,* X_k *and* Y_k *are* $\mathcal{F}_k$ *measurable, and*

$$(\forall k \in \mathbb{N}) \qquad \mathbb{E}[X_{k+1} \mid \mathcal{F}_k] + Y_k \leq X_k.$$

Then $\sum_{k=0}^{\infty} Y_k < \infty$ *a.s. and* X_k *a.s. converges to a* $[0,\infty)$*-valued random variable.*

The other equally indispensable element of our proof is the next inequality, which, when taken together with the supermartingale convergence theorem, will ultimately show that SMART converges: with

$$\kappa_k := \sum_{j=1}^{m}\sum_{i=1}^{n} \frac{\gamma_{ij}p_{ij}}{\rho p_{ij}^T} r_{ij}^d(y_i^k); \qquad \text{and} \qquad Y_k := \lambda^2 \sum_{j=1}^{m} \frac{\overline{M}_j}{m^2} \|(S(x^k))_j\|_j^2 + 2\lambda^2 \sum_{j=1}^{m} \frac{\overline{M}_j}{q_j m^2} \left\| \frac{1}{n}\sum_{i=1}^{n} y_{i,j}^k \right\|_j^2$$
$$+ \sum_{j=1}^{m}\sum_{i=1}^{n} p_{ij} R_{1,ij} r_{ij}^d(y_i^k) + \sum_{j=1}^{m}\sum_{i=1}^{n_1} p_{ij} R_{2,ij} r_{ij}^p(x^k),$$

the supermartingale inequality holds

$$(\forall k \in \mathbb{N})(\forall x^* \in \mathcal{S}) \qquad \mathbb{E}\left[\|x^{k+1} - x^*\|^2 + \kappa_{k+1} \mid \mathcal{F}_k\right] + Y_k \leq \|x^k - x^*\|^2 + \kappa_k. \tag{7.3}$$

So, by the supermartingale convergence theorem, the sequence Y_k is a.s. summable and $X_k := \|x^k - x^*\|^2 + \kappa_k$ a.s. converges to a $[0,\infty)$-valued random variable, and these conclusions hold for any element $x^* \in \mathcal{S}$.

At this point, Part 1 of the theorem follows because $\sum_{k=0}^{\infty}\sum_{j=1}^{m}\sum_{i=1}^{n} p_{ij} R_{2,ij} r_{ij}^p(x^k) \leq \sum_{k=0}^{\infty} Y_k < \infty$ a.s., and hence, $\|S_i(x^k) - S_i(x^*)\| = (np_{ij})^2 r_{ij}^p(x^k)$ a.s. converges to 0.

Moreover, the road to almost sure weak convergence of x^k is only three steps long:

1. Because $\sum_{k=0}^{\infty} Y_k < \infty$ a.s., we conclude that $\|S(x^k)\|$ and $r_{ij}^d(y_i^k)$ a.s. converge to 0.
2. Because each $r_{ij}^d(y_i^k)$ a.s. converges to 0, we conclude that κ_k a.s. converges to 0.
3. Because κ_k a.s. converges to 0 and $\|x^k - x^*\|^2 + \kappa_k$ a.s. converges to a $[0,\infty)$-valued random variable, we conclude that $\|x^k - x^*\|^2$ a.s. converges to a $[0,\infty)$-valued random variable.

At last, x^k weakly converges to an $\mathcal{S}$-valued random vector:

---

[25] By the choice of γ_{ij}, there is a constant $b > 0$ such that for all i, j and k, we have $R_{1,ij} > b$ and $R_{2,ij} > b$.

Proposition 7.1 (Weak convergence assuming demiclosedness) *Suppose that $I-S$ is demiclosed at 0. Let $\{z^k\}_{k\in\mathbb{N}}$ be a sequence of random vectors such that, for all $z^* \in \mathrm{zer}(S)$, the sequence $\{\|z^k - z^*\|^2\}_{k\in\mathbb{N}}$ a. s. converges to a $[0,\infty)$-valued random variable. In addition, assume that $\|S(z^k)\|$ a. s. converges to 0. Then z^k a. s. converges to an $\mathcal{S}$-valued random variable.*

Proof The set $\mathcal{S}$ is closed[26], the space $\mathcal{H}$ is separable, and for all $z^* \in \mathcal{S}$, the sequence $\{\|z^k - z^*\|\}_{k\in\mathbb{N}}$ a. s. converges, so the exact argument in [9, Prop. 2.3(iii)] shows that there exists $\hat{\Omega} \subseteq \Omega$ such that $P(\hat{\Omega}) = 1$ with the property that for all $z^* \in \mathrm{zer}(S)$ and for all $\omega \in \hat{\Omega}$, the sequence $\{\|z^k(\omega) - z^*\|^2\}_{k\in\mathbb{N}}$ converges. Furthermore, by assumption, there exists $\widetilde{\Omega} \subseteq \hat{\Omega}$ such that $P(\widetilde{\Omega}) = 1$ with the property that $\|S(z^k(\omega))\|$ a. s. almost surely converges to 0.

Finishing the proof with demiclosedness, let $\omega \in \widetilde{\Omega}$, and let z be a weak sequential cluster point of $\{z^k(\omega)\}_{k\in\mathbb{N}}$, say $z^{k_j}(\omega) \rightharpoonup z$ (cluster points exist because $\{z^k\}_{k\in\mathbb{N}}$ is bounded). Then $\|z^{k_j}(\omega) - (z^{k_j}(\omega) - S(z^{k_j}(\omega)))\| = \|S(z^{k_j}(\omega))\| \to 0$ so by the demiclosedness of $I-S$, the limit is a zero: $S(z) = 0$. In summary, for all $z^* \in \mathcal{S}$, the sequence $\{\|z^k(\omega) - z^*\|\}_{k\in\mathbb{N}}$ converges and every weak sequential cluster point of $\{z^k\}_{k\in\mathbb{N}}$ is an element of $\mathcal{S}$. Therefore, by [2, Lemma 2.39], the sequence $\{z^k(\omega)\}_{k\in\mathbb{N}}$ weakly converges to an element of $\mathcal{S}$. Because $\omega \in \widetilde{\Omega}$ is arbitrary and $P(\widetilde{\Omega}) = 1$, the sequence $\{z^k\}_{k\in\mathbb{N}}$ a. s. weakly converges, and by the classic result [30, Corollary 1.13], the weak limit of $\{z^k\}_{k\in\mathbb{N}}$ is measurable. $\square$

Our task is now clear: we must prove (7.3). But to do so, we must first bound several random variables. We present the three bounds we need now and defer their proofs until later.

Lemma 7.1 (Variance bound) *For $j \in \{1, \ldots, m\}$, let $\eta_j > 0$ be a positive real number. Then for all $s \in \mathbb{N}$, we have*

$$\sum_{j=1}^{m} \eta_j \mathbb{E}\left[\|\mathfrak{Q}_j^s\|_j^2 \mid \mathcal{F}_s\right] \le 2\sum_{j=1}^{m}\sum_{i=1}^{n} \frac{p_{ij}}{q_j m^2}\eta_j r_{ij}^p(x^s) + 2\sum_{j=1}^{m}\sum_{i=1}^{n} \frac{p_{ij}}{q_j m^2}\eta_j r_{ij}^d(y_i^s)$$
$$- 2\sum_{j=1}^{m} \frac{1}{q_j m^2}\eta_j \left\|\frac{1}{n}\sum_{i=1}^{n} y_{i,j}^s\right\|_j^2 - \sum_{j=1}^{m} \frac{\eta_j}{m^2}\|(S(x^s))_j\|_j^2.$$

Lemma 7.2 (Using (1.5)) *For all $s \in \mathbb{N}$ and $\lambda > 0$, we have*

$$2\lambda\langle S(x^s), x^s - x^*\rangle \ge \sum_{j=1}^{m}\sum_{i=1}^{n} 2\lambda\beta_{ij}n^2 p_{ij}^2 r_{ij}^p(x^s).$$

Lemma 7.3 (Recursive κ_k bound) *For all $s \in \mathbb{N}$, we have*

$$\mathbb{E}\left[\kappa_{s+1} \mid \mathcal{F}_k\right] \le \kappa_s - \sum_{j=1}^{m}\sum_{i=1}^{n} p_{ij}\gamma_{ij} r_{ij}^d(y_i^s) + \sum_{j=1}^{m}\sum_{i=1}^{n} p_{ij}\gamma_{ij} r_{ij}^p(x^s).$$

Lemmas in hand, we can now prove (7.3)

Lemma 7.4 (Proof of (7.3)) *Equation (7.3) holds for all $s \in \mathbb{N}$.*

---

[26] Let $y^k \in \mathcal{S}$ and suppose that $y^k \to y \in \mathcal{H}$. We claim that $y \in \mathcal{S}$. Indeed, from (1.5), there exists a constant $\beta > 0$ such that $\beta\|S(y)\|^2 \le \langle S(y), y - y^k\rangle \to 0$. Thus, $S(y) = 0$, and $y \in \mathcal{S}$.

Proof Fix $s \in \mathbb{N}$. Then

$$
\begin{aligned}
&\mathbb{E}\left[\|x^{s+1} - x^*\|^2 + \kappa_{s+1} \mid \mathcal{F}_s\right] \\
&= \|x^s - x^*\|^2 - 2\lambda\mathbb{E}\left[\langle\mathfrak{Q}^s, x^s - x^*\rangle \mid \mathcal{F}_s\right] + \lambda^2\mathbb{E}\left[\|\mathfrak{Q}^s\|^2 \mid \mathcal{F}_s\right] + \mathbb{E}\left[\kappa_{s+1} \mid \mathcal{F}_s\right] \\
&= \|x^s - x^*\|^2 - \frac{2\lambda}{m}\langle S(x^s), x^s - x^*\rangle + \lambda^2\mathbb{E}\left[\|\mathfrak{Q}^s\|^2 \mid \mathcal{F}_s\right] + \mathbb{E}\left[\kappa_{s+1} \mid \mathcal{F}_s\right].
\end{aligned}
$$

where the second equality follows from the linearity of expectation. Now, apply Lemmas 7.1, 7.2, and 7.3:

$$
\begin{aligned}
&\leq \|x^s - x^*\|^2 + \kappa_s + \lambda^2\mathbb{E}\left[\sum_{j=1}^{m}\overline{M}_j\|\mathfrak{Q}_j^s\|_j^2 \mid \mathcal{F}_s\right] \\
&\quad - \sum_{j=1}^{m}\sum_{i=1}^{n}p_{ij}\gamma_{ij}r_{ij}^d(y_i^s) - \sum_{j=1}^{m}\sum_{i=1}^{n}p_{ij}\left(\frac{2\lambda\beta_{ij}n^2p_{ij}}{m} - \gamma_{ij}\right)r_{ij}^d(x^s) \\
&\leq \|x^s - x^*\|^2 + \kappa_s - \lambda^2\sum_{j=1}^{m}\frac{\overline{M}_j}{m^2}\|(S(x^s))_j\|_j^2 - 2\lambda^2\sum_{j=1}^{m}\frac{\overline{M}_j}{q_jm^2}\left\|\frac{1}{n}\sum_{i=1}^{n}y_{i,j}^s\right\|_j^2 \\
&\quad - \sum_{j=1}^{m}\sum_{i=1}^{n}p_{ij}\left(\gamma_{ij} - \frac{2\overline{M}_j\lambda^2}{q_jm^2}\right)r_{ij}^d(y_i^s) - \sum_{j=1}^{m}\sum_{i=1}^{n}p_{ij}\left(\frac{2\lambda\beta_{ij}n^2p_{ij}}{m} - \frac{2\overline{M}_j\lambda^2}{q_jm^2} - \gamma_{ij}\right)r_{ij}^p(x^s). \\
&= \|x^s - x^*\|^2 + \kappa_s - Y_s. \quad \square
\end{aligned}
$$

We finish the proof of Parts 1 and 2 by proving Lemmas 7.1, 7.2, and 7.3.

Proof (of Lemma 7.1 (variance bound)) In the next sequence of inequalities, we use the following variance identity three times: For any random vector $X : \Omega \to \mathcal{H}_j$ and a sub σ-algebra $\mathcal{X}$ on Ω, we have $\mathbb{E}\left[\|X - \mathbb{E}[X \mid \mathcal{X}]\|_j^2 \mid \mathcal{X}\right] = \mathbb{E}\left[\|X\|_j^2 \mid \mathcal{X}\right] - \|\mathbb{E}[X \mid \mathcal{X}]\|_j^2$. By the law of iterated expectation:

$$
\begin{aligned}
&2\mathbb{E}\left[\|t_{s,j}((Q_{i_s}^d(y_{i_s}^s))_j - (Q_{i_s}^*)_j - \frac{1}{n}\sum_{i=1}^{n}y_{i,j}^s\|_j^2 \mid \mathcal{F}_s\right] \\
&= 2\mathbb{E}\left[\mathbb{E}\left[\|t_{s,j}((Q_{i_s}^d(y_{i_s}^s))_j - (Q_{i_s}^*)_j - \frac{1}{n}\sum_{i=1}^{n}y_{i,j}^s\|_j^2 \mid \sigma(t_{s,j}, \mathcal{F}_s)\right] \mid \mathcal{F}_s\right] \\
&= 2\mathbb{E}\left[\mathbb{E}\left[\|t_{s,j}((Q_{i_s}^d(y_{i_s}^s))_j - (Q_{i_s}^*)_j)\|_j^2 \mid \sigma(t_{s,j}, \mathcal{F}_s)\right] - \left\|t_{s,j}\frac{1}{n}\sum_{i=1}^{n}y_{i,j}^s\right\|_j^2 \mid \mathcal{F}_s\right] \\
&= 2\sum_{i=1}^{n}\frac{p_{ij}}{q_jm^2}\|((Q_{i_s}^d(y_i^s))_j - (Q_i^*)_j)\|_j^2 - \frac{2}{q_jm^2}\left\|\frac{1}{n}\sum_{i=1}^{n}y_{i,j}^s\right\|_j^2.
\end{aligned}
$$

Therefore,

$$\mathbb{E}\left[\|\mathfrak{Q}_j^s\|_j^2 \mid \mathcal{F}_s\right]$$

$$= \mathbb{E}\left[\|\mathfrak{Q}_j^s - \frac{1}{m}(S(x^s))_j\|_j^2 \mid \mathcal{F}_s\right] + \frac{1}{m^2}\|(S(x^s))_j\|_j^2$$

$$= \mathbb{E}\left[\|t_{s,j}\left[(Q_{i_s}^d(y_{i_s}^s))_j - (Q_{i_s}^*)_j - \frac{1}{n}\sum_{i=1}^n y_{i,j}^s\right]\right.$$

$$\left. - t_{s,j}\left[(Q_{i_s}^p(x^s))_j - (Q_{i_s}^*)_j\right] + \frac{1}{m}(S(x^s))_j\|_j^2 \mid \mathcal{F}_s\right] + \frac{1}{m^2}\|(S(x^s))_j\|_j^2$$

$$\leq 2\mathbb{E}\left[\|t_{s,j}((Q_{i_s}^d(y_{i_s}^s))_j - (Q_{i_s}^*)_j - \frac{1}{n}\sum_{i=1}^n y_{i,j}^s)\|_j^2 \mid \mathcal{F}_s\right]$$

$$+ 2\mathbb{E}\left[\|t_{s,j}((Q_{i_s}^p(x^s))_j - (Q_{i_s}^*)_j) - \frac{1}{m}(S(x^s))_j\|_j^2 \mid \mathcal{F}_s\right] + \frac{1}{m^2}\|(S(x^s))_j\|_j^2$$

$$\leq 2\mathbb{E}\left[\|t_{s,j}((Q_{i_s}^d(y_{i_s}^s))_j - (Q_{i_s}^*)_j)\|_j^2 \mid \mathcal{F}_s\right] + 2\mathbb{E}\left[\|t_{s,j}((Q_{i_s}^p(x^s))_j - (Q_{i_s}^*)_j)\|_j^2 \mid \mathcal{F}_s\right]$$

$$- \frac{2}{q_j m^2}\left\|\frac{1}{n}\sum_{i=1}^n y_{i,j}^s\right\|_j^2 - \frac{1}{m^2}\|(S(x^s))_j\|_j^2.$$

$$\leq 2\sum_{i=1}^n \frac{p_{ij}}{q_j m^2}\|((Q_i^d(y_i^s))_j - (Q_i^*)_j)\|_j^2 + 2\sum_{i=1}^n \frac{p_{ij}}{q_j m^2}\|((Q_i^p(x^s))_j - (Q_i^*)_j)\|_j^2$$

$$- \frac{2}{q_j m^2}\left\|\frac{1}{n}\sum_{i=1}^n y_{i,j}^s\right\|_j^2 - \frac{1}{m^2}\|(S(x^s))_j\|_j^2.$$

To get the claimed identity, multiply this inequality by η_j for each j and sum. $\qquad\square$

Proof (of Lemma 7.2 (using (1.5))) From (1.5),

$$\langle S(x^s), x^s - x^* \rangle \geq \sum_{j=1}^m \sum_{i=1}^n \beta_{ij}\|(S_i(x^s))_j - (S_i(x^*))_j\|_j^2 \geq \sum_{j=1}^m \sum_{i=1}^n \beta_{ij} n^2 p_{ij}^2 \|(Q_i^p(x^s))_j - (Q_i^*)_j\|_j^2.$$

Proof (of Lemma 7.3 (recursive κ_k bound)) For any i, j with $\mathbf{S}_{ij}^* \neq 0$ and $\gamma_{ij} > 0$,

$$\mathbb{E}\left[r_{ij}^d(y_i^{s+1}) \mid \mathcal{F}_k\right] = \left(1 - \rho p_{ij}^T\right) r_{ij}^d(y_i^s) + \rho p_{ij}^T r_{ij}^p(x^s)$$

because $r_{ij}^d(y_i^{s+1})$ depends only on $y_{i,j}^{s+1}$, not on its other components, and the probability of update for $(y_i^k)_j$ is $P((i_k, i) \in E, t_{k,j} \neq 0, \epsilon_k = 1) = \rho p_{ij}^T$. When $\mathbf{S}_{ij}^* = 0$, the left hand side of the equation is zero, so the above equation holds as an inequality. Thus, to get the claimed inequality, multiply both sides of the above equation by $p_{ij}\gamma_{ij}(\rho p_{ij}^T)^{-1}$ and sum over i and j. $\qquad\square$

Part 3: In this part, we no longer work with arbitrary zeros $x^* \in \mathcal{S}$. Instead we work with the sequence of zeros $P_{\mathcal{S}}(x^k)$, which, by definition, satisfy

$$d_{\mathcal{S}}^2(x^{k+1}) = \|x^{k+1} - P_{\mathcal{S}}(x^{k+1})\|^2 \leq \|x^{k+1} - P_{\mathcal{S}}(x^k)\|^2.$$

But notice that, because they are independent of the fixed zero x^* from Parts 1 and 2, we can freely use the inequalities in Lemmas 7.1, 7.2, and 7.3.

In this part, we further constrain the constants γ_{ij}. (We defer the proof of the lemma for a moment.)

Lemma 7.5 (Choosing γ_{ij}) *Let $\xi := 2m^{-1}\mu\alpha\lambda$. Then for all i and j, there exists $\gamma_{ij} > 0$ such that*

$$\frac{2\overline{M}_j\lambda^2}{q_j m^2 \left(1 - \frac{\xi}{\rho p_{ij}^T}\right)} \leq \gamma_{ij} \leq \frac{2(1-\alpha)\lambda\beta_{ij}n^2 p_{ij}}{m} - \frac{2\overline{M}_j\lambda^2}{q_j m^2}.$$

Then we finish off the proof by appealing to the lower bound

$$\langle S(x^k), x^k - P_{\mathcal{S}}(x^k)\rangle \geq \mu\|x^k - P_{\mathcal{S}}(x^k))\|^2$$

in the following sequence of inequalities

$$\mathbb{E}\left[d_{\mathcal{S}}^2(x^{k+1}) + \kappa_{k+1} \mid \mathcal{F}_k\right]$$
$$\leq \mathbb{E}\left[\|x^{k+1} - P_{\mathcal{S}}(x^k)\|^2 + \kappa_{k+1} \mid \mathcal{F}_k\right]$$
$$= \|x^k - P_{\mathcal{S}}(x^k)\|^2 - 2\lambda\mathbb{E}\left[\langle\mathfrak{Q}^k, x^k - P_{\mathcal{S}}(x^k)\rangle \mid \mathcal{F}_k\right] + \lambda^2\mathbb{E}\left[\|\mathfrak{Q}^k\|^2 \mid \mathcal{F}_k\right] + \mathbb{E}\left[\kappa_{k+1} \mid \mathcal{F}_k\right]$$
$$= \|x^k - P_{\mathcal{S}}(x^k)\|^2 - \frac{2\lambda}{m}\langle S(x^k), x^k - P_{\mathcal{S}}(x^k)\rangle + \lambda^2\mathbb{E}\left[\|\mathfrak{Q}^k\|^2 \mid \mathcal{F}_k\right] + \mathbb{E}\left[\kappa_{k+1} \mid \mathcal{F}_k\right]$$
$$\leq \left(1 - \frac{2\alpha\mu\lambda}{m}\right)\|x^k - P_{\mathcal{S}}(x^k)\|^2 - \frac{2\lambda(1-\alpha)}{m}\langle S(x^k), x^k - P_{\mathcal{S}}(x^k)\rangle + \lambda^2\mathbb{E}\left[\sum_{j=1}^m \overline{M}_j\|\mathfrak{Q}_j^k\|_j^2 \mid \mathcal{F}_k\right] + \mathbb{E}\left[\kappa_{k+1} \mid \mathcal{F}_k\right]$$
$$\leq \left(1 - \frac{2\alpha\mu\lambda}{m}\right)\|x^k - P_{\mathcal{S}}(x^k)\|^2 + \kappa_k \qquad \text{(Lemmas 7.1, 7.2, and 7.3 are used below)}$$
$$\quad - \sum_{j=1}^m\sum_{i=1}^n p_{ij}\left(\gamma_{ij} - \frac{2\overline{M}_j\lambda^2}{q_j m^2}\right)r_{ij}^d(y_i^k) - \sum_{j=1}^m\sum_{i=1}^n p_{ij}\left(\frac{2(1-\alpha)\lambda\beta_{ij}n^2 p_{ij}}{m} - \frac{2\overline{M}_j\lambda^2}{q_j m^2} - \gamma_{ij}\right)r_{ij}^p(x^k)$$
$$\leq (1-\xi)\left(\|x^k - P_{\mathcal{S}}(x^k)\|^2 + \kappa_k\right) - \sum_{j=1}^m\sum_{i=1}^n p_{ij}\left(\gamma_{ij}\left(1 - \frac{\xi}{\rho p_{ij}^T}\right) - \frac{2\overline{M}_j\lambda^2}{q_j m^2}\right)r_{ij}^d(y_i^k)$$
$$\leq (1-\xi)\left(d_{\mathcal{S}}^2(x^k) + \kappa_k\right) \tag{7.4}$$

(Apply the law of iterated expectations to get the linear convergence rate (with $C(z^0, \phi^0) := \kappa_0$).)

The only loose end, which we now tie up, is the proof of Lemma 7.5.

Proof (of Lemma 7.5 (choosing γ_{ij})) We have assumed that

$$\lambda \leq \min_{i,j}\left\{\frac{\eta(1-\alpha)\beta_{ij}n^2 p_{ij}q_j m}{2\overline{M}_j\eta + 2\mu\alpha(1-\alpha)\beta_{ij}n^2 p_{ij}q_j}\right\},$$

and consequently, if

$$w_{ij} := \frac{\frac{(1-\alpha)\beta_{ij}n^2 p_{ij}q_j m}{2\overline{M}_j}}{\frac{m\eta}{2\alpha\mu} + \frac{(1-\alpha)\beta_{ij}n^2 p_{ij}q_j m}{2\overline{M}_j}} < 1,$$

then

$$\lambda \leq \frac{w_{ij} m \eta}{2\alpha\mu} = \frac{(1-\alpha)\beta_{ij} n^2 p_{ij} q_j m (1-w_{ij})}{2\overline{M}_j} \leq \frac{(1-\alpha)\beta_{ij} n^2 p_{ij} q_j m (1-w_{ij})}{(2-w_{ij})\overline{M}_j}.$$

In particular, $\xi \leq w_{ij}\eta < 1$. Now rearrange the above inequality to get

$$\frac{\overline{M}_j \lambda^2}{q_j m^2 (1-w_{ij})} + \frac{\overline{M}_j \lambda^2}{q_j m^2} = \frac{\overline{M}_j \lambda^2 (2-w_{ij})}{q_j m^2 (1-w_{ij})} \leq \frac{2\overline{M}_j \lambda^2}{q_j m^2 (1-w_{ij})} \leq \frac{(1-\alpha)\lambda\beta_{ij} n^2 p_{ij}}{m}.$$

The last bound proves the claimed bound because $\left(1 - \frac{\xi}{\rho p_{ij}^T}\right)^{-1} \leq \left(1 - \frac{w_{ij}\eta}{\rho p_{ij}^T}\right)^{-1} \leq (1-w_{ij})^{-1}.$ $\square$

$\square$

8 Future Work

The SMART algorithm calls for five avenues of future work: numerical experiments, especially in the asynchronous setting; the creation of new operators S; the characterization of the error bound condition which we call essential strong quasi-monotonicity; the convergence rate analysis of the SMART algorithm without assuming essential strong quasi-monotonicity; and nonconvex extensions. All of these avenues present challenges.

The SAGA, SVRG, Finito, SDCA, and randomized projection algorithms presented in Section 4 perform well in the synchronous setting. Asynchronous algorithms often obtain a linear speedup (in the number of computing cores) over their synchronous implementations, so we expect these algorithms to perform even better in the asynchronous setting. However, implementing asynchronous algorithms still requires a bit of programming expertise, so we expect that good experimental results will take some time to acquire.

Any operator that satisfies the coherence condition (1.5) can be plugged into the SMART algorithm; as such, the power of SMART increases with each new operator discovered. We presented many examples of operators in this paper, and we expect there to be more in the future.

The essential strong quasi-monotonicity property (2.1) appears to be the weakest possible condition under which a first-order algorithm will converge linearly. This deep, difficult to characterize property is related to the Hoffman bound [18], the linear regularity assumption, and the Kurdyka-Łojasiewicz property [7]. We look forward to a calculus of operations that preserve this property.

Even if S is not essentially strongly quasi-monotone, we suspect that SMART converges sublinearly; to show these rates, the techniques in [11, 14, 12, 13] should be adapted to the stochastic and asynchronous settings.

We showed that SMART converges when the coherence condition is satisfied, and this tethers our results to convex optimization problems. But SMART can and should be applied to nonconvex problems. In the nonconvex case, convergence guarantees will certainly be weaker than those presented here, but SMART will likely perform well on these problems.

References

1. Baillon, J.B., Haddad, G.: Quelques propriétés des opérateurs angle-bornés et n-cycliquement monotones. Israel Journal of Mathematics **26**(2), 137–150 (1977)
2. Bauschke, H.H., Combettes, P.L.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces, 1st edn. Springer Publishing Company, Incorporated (2011)
3. Bauschke, H.H., Wang, C., Wang, X., Xu, J.: On Subgradient Projectors. SIAM Journal on Optimization **25**(2), 1064–1082 (2015). DOI 10.1137/14096267X. URL http://dx.doi.org/10.1137/14096267X
4. Bertsekas, D.P.: Incremental proximal methods for large scale convex optimization. Mathematical Programming **129**(2), 163–195 (2011). DOI 10.1007/s10107-011-0472-0. URL http://dx.doi.org/10.1007/s10107-011-0472-0
5. Bertsekas, D.P.: Incremental aggregated proximal and augmented lagrangian algorithms. arXiv preprint arXiv:1509.09257 (2015)
6. Bianchi, P.: Ergodic convergence of a stochastic proximal point algorithm. arXiv preprint arXiv:1504.05400 (2015)
7. Bolte, J., Nguyen, T.P., Peypouquet, J., Suter, B.: From error bounds to the complexity of first-order descent methods for convex functions. arXiv preprint arXiv:1510.08234 (2015)
8. Combettes, P.L., Eckstein, J.: Asynchronous Block-Iterative Primal-Dual Decomposition Methods for Monotone Inclusions. arXiv preprint arXiv:1507.03291 (2015)
9. Combettes, P.L., Pesquet, J.C.: Stochastic Quasi-Fejér Block-Coordinate Fixed Point Iterations with Random Sweeping. SIAM Journal on Optimization **25**(2), 1221–1248 (2015). DOI 10.1137/140971233. URL http://dx.doi.org/10.1137/140971233
10. Combettes, P.L., Yamada, I.: Compositions and convex combinations of averaged nonexpansive operators. Journal of Mathematical Analysis and Applications **425**(1), 55 – 70 (2015). DOI http://dx.doi.org/10.1016/j.jmaa.2014.11.044. URL http://www.sciencedirect.com/science/article/pii/S0022247X14010865
11. Davis, D.: Convergence Rate Analysis of Primal-Dual Splitting Schemes. SIAM Journal on Optimization **25**(3), 1912–1943 (2015). DOI 10.1137/151003076. URL http://dx.doi.org/10.1137/151003076
12. Davis, D.: Convergence Rate Analysis of the Forward-Douglas-Rachford Splitting Scheme. SIAM Journal on Optimization **25**(3), 1760–1786 (2015). DOI 10.1137/140992291. URL http://dx.doi.org/10.1137/140992291
13. Davis, D., Yin, W.: Convergence rate analysis of several splitting schemes. arXiv preprint arXiv:1406.4834 (2014)
14. Davis, D., Yin, W.: Faster convergence rates of relaxed Peaceman-Rachford and ADMM under regularity assumptions. arXiv preprint arXiv:1407.5210v3 (2014)
15. Defazio, A., Bach, F., Lacoste-Julien, S.: Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In: Advances in Neural Information Processing Systems, pp. 1646–1654 (2014)
16. Defazio, A., Domke, J., Caetano, T.: Finito: A faster, permutable incremental gradient method for big data problems. In: Proceedings of the 31st International Conference on Machine Learning (ICML-14), pp. 1125–1133 (2014)
17. Fercoq, O., Richtrik, P.: Accelerated, parallel, and proximal coordinate descent. SIAM Journal on Optimization **25**(4), 1997–2023 (2015). DOI 10.1137/130949993. URL http://dx.doi.org/10.1137/130949993
18. Hoffman, A.J.: On approximate solutions of systems of linear inequalities. Journal of Research of the National Bureau of Standards **49**(4), 263–265 (1952)
19. Hofmann, T., Lucchi, A., Lacoste-Julien, S., McWilliams, B.: Neighborhood Watch: Stochastic Gradient Descent with Neighbors. arXiv preprint arXiv:1506.03662v3 (2015)
20. Johnson, R., Zhang, T.: Accelerating stochastic gradient descent using predictive variance reduction. In: Advances in Neural Information Processing Systems, pp. 315–323 (2013)
21. Konečnỳ, J., Liu, J., Richtárik, P., Takáč, M.: mS2GD: Mini-Batch Semi-Stochastic Gradient Descent in the Proximal Setting. arXiv preprint arXiv:1410.4744 (2014)
22. Konečnỳ, J., Qu, Z., Richtárik, P.: Semi-stochastic coordinate descent. arXiv preprint arXiv:1412.6293 (2014)
23. Konečnỳ, J., Richtárik, P.: Semi-stochastic gradient descent methods. arXiv preprint arXiv:1312.1666 (2013)
24. Lai, M.J., Yin, W.: Augmented ℓ_1 and Nuclear-Norm Models with a Globally Linearly Convergent Algorithm. SIAM Journal on Imaging Sciences **6**(2), 1059–1091 (2013)
25. Liu, J., Wright, S.J., Ré, C., Bittorf, V., Sridhar, S.: An Asynchronous Parallel Stochastic Coordinate Descent Algorithm. Journal of Machine Learning Research **16**, 285–322 (2015)
26. Liu, J., Wright, S.J., Sridhar, S.: An Asynchronous Parallel Randomized Kaczmarz Algorithm. arXiv preprint arXiv:1401.4780 (2014)
27. Nedic, A., Bertsekas, D.P.: Incremental subgradient methods for nondifferentiable optimization. SIAM Journal on Optimization **12**(1), 109–138 (2001). DOI 10.1137/S1052623499362111. URL http://dx.doi.org/10.1137/S1052623499362111
28. Nesterov, Y.: Efficiency of coordinate descent methods on huge-scale optimization problems. SIAM Journal on Optimization **22**(2), 341–362 (2012). DOI 10.1137/100802001. URL http://dx.doi.org/10.1137/100802001

29. Peng, Z., Xu, Y., Yan, M., Yin, W.: ARock: an Algorithmic Framework for Asynchronous Parallel Coordinate Updates. arXiv preprint arXiv:1506.02396 (2015)
30. Pettis, B.J.: On integration in vector spaces. Transactions of the American Mathematical Society **44**(2), 277–304 (1938)
31. Recht, B., Re, C., Wright, S., Niu, F.: Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In: Advances in Neural Information Processing Systems, pp. 693–701 (2011)
32. Robbins, H., Siegmund, D.: A convergence theorem for non negative almost supermartingales and some applications. In: Herbert Robbins Selected Papers, pp. 111–135. Springer (1985)
33. Schmidt, M., Roux, N.L., Bach, F.: Minimizing finite sums with the stochastic average gradient. arXiv preprint arXiv:1309.2388 (2013)
34. Shalev-Shwartz, S., Zhang, T.: Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. Mathematical Programming pp. 1–41
35. Strohmer, T., Vershynin, R.: A randomized Kaczmarz algorithm with exponential convergence. Journal of Fourier Analysis and Applications **15**(2), 262–278 (2009)
36. Wang, M., Bertsekas, D.P.: Incremental constraint projection-proximal methods for nonsmooth convex optimization. SIAM Journal on Optimization (To appear)
37. Zhang, H.: The restricted strong convexity revisited: analysis of equivalence with error bound and quadratic growth. arXiv preprint arXiv:1511.01635 (2015)

Appendix

A Proof in the Asynchronous Case

Proof (Convergence of SMART)

Notation. Let $x^* \in \mathcal{S}$. Some of our notation, in particular, the definitions of $t_k, Q_i^p, Q_i^d, r_{ij}^p$, and r_{ij}^d, remains the same as in the proof of the nonasynchronous case (Theorem 7.1). Below, we adapt the rest of the notation to the present asynchronous case:

1. **Integer indexed sequences.** We extend every sequence indexed by natural numbers, say $\{z^k\}_{k \in \mathbb{N}}$, to all of $\mathbb{Z}$ by setting $z^k := z^0$ for all $k \leq 0$.

2. **The $\mathfrak{Q}^k$ random vector.** If

$$\mathfrak{Q}^k = t_k \odot \left(Q_{i_k}^p (x^{k-d_k}) - Q_{i_k}^d (y_i^{k-e_k^i}) + \frac{1}{n} \sum_{i=1}^{n} y_i^{k-e_k^i} \right),$$

then $x^{k+1} = x^k - \lambda \mathfrak{Q}^k$ and $\mathbb{E}\left[\mathfrak{Q}^k \mid \mathcal{F}_k \right] = m^{-1} S(x^{k-d_k})$.

3. **The γ_{ij} constants.** For all i and j, choose any constants that satisfy[27]

$$
\begin{aligned}
\frac{2\overline{M}_j(\tau_d+1)}{q_j m^2}\left(1 + \frac{\tau_p}{mC}\right)\lambda_k^2 &\leq \frac{2\overline{M}_j(\tau_d+1)}{\underline{q} m^2}\left(1 + \frac{\tau_p}{mC}\right)\overline{\lambda}^2 \\
&< \gamma_{ij} \\
&< \frac{2\underline{\lambda} n^2 p_{ij}\beta_{ij}}{m} - \overline{M}_j\left(\frac{2}{\underline{q} m^2} + \tau_p\left(\frac{2}{\underline{q} m^3 C} + \frac{C}{m}\right)\right)\overline{\lambda}^2 \\
&\leq \frac{2\lambda_k n^2 p_{ij}\beta_{ij}}{m} - \overline{M}_j\left(\frac{2}{q_j m^2} + \tau_p\left(\frac{2}{q_j m^3 C} + \frac{C}{m}\right)\right)\lambda_k^2. \quad \text{(A.1)}
\end{aligned}
$$

4. **The C constant.** Define the positive constant

$$C := \frac{\sqrt{2(\tau_d+2)}}{m\sqrt{\underline{q}}}.$$

---

[27] The constraint (A.1) has a solution by the assumptions of the theorem.

5. **The $R_{l,ij}^k$ constants.** For all i, j, and k, we define two *positive constants*[28]

$$R_{1,ij}^k := \frac{\gamma_{ij}}{\tau_d + 1} - \frac{2\overline{M}_j}{q_j m^2}\left(1 + \frac{\tau_p}{mC}\right)\lambda_k^2; \text{ and}$$

$$R_{2,ij}^k := \frac{2\lambda_k n^2 p_{ij}\beta_{ij}}{m} - \overline{M}_j\left(\frac{2}{q_j m^2} + \tau_p\left(\frac{2}{q_j m^3 C} + \frac{C}{m}\right)\right)\lambda_k^2 - \gamma_{ij}.$$

Assumption 2 *We only address the case in which $\mathbf{S}^*$ is not necessarily the zero matrix. From this case, and with the help of the identity $r_{ij}^d(y_i^{k-e_k^i}) \equiv 0$, it is straightforward to retrieve the case in which $\mathbf{S}^* = 0$.*

Parts 1 and 2: In the synchronous case, the supermartingale convergence theorem and a supermartingale inequality featured; the same is true in the present asynchronous case, but the supermartingale inequality now accounts for the use of delayed iterates and operator values: with

$$\kappa_k := \underbrace{\sum_{j=1}^m \sum_{i=1}^n \frac{\gamma_{ij}p_{ij}}{\rho p_{ij}^T}r_{ij}^d(y_i^k) + \sum_{h=0}^{\tau_d - 1}\sum_{j=1}^m\sum_{i=1}^n \frac{(h+1)p_{ij}\gamma_{ij}}{\tau_d + 1}r_{ij}^d(y_i^{k-\tau_d+h})}_{:=\kappa_{1,k}} + \underbrace{\sum_{t=k-\tau_p+1}^k \sum_{j=1}^m \frac{\overline{M}_j(t-k+\tau_p)}{mC}\|x_j^t - x_j^{t-1}\|_j^2}_{:=\kappa_{2,k}};$$

$$Y_k := \lambda_k^2 \sum_{j=1}^m \frac{\overline{M}_j}{m^2}\left(1 + \frac{\tau_p}{mC}\right)\|(S(x^{k-d_k}))_j\|_j^2 + 2\lambda_k^2 \sum_{j=1}^m \frac{\overline{M}_j}{q_j m^2}\left(1 + \frac{\tau_p}{mC}\right)\left\|\frac{1}{n}\sum_{i=1}^n y_{i,j}^{k-e_k^i}\right\|_j^2$$

$$\sum_{j=1}^m\sum_{i=1}^n p_{ij}R_{1,ij}^k r_{ij}^d(y_i^{k-e_k^i}) + \sum_{j=1}^m\sum_{i=1}^n p_{ij}R_{2,ij}^k r_{ij}^p(x^{k-d_k}) + \sum_{j=1}^m\sum_{i=1}^n \sum_{\substack{h=0 \\ h\neq\tau_d-e_{k,j}^i}}^{\tau_d} \frac{p_{ij}\gamma_{ij}}{\tau_d + 1}r_{ij}^d(y_i^{k-\tau_d+h}),$$

the supermartingale inequality holds

$$(\forall k \in \mathbb{N})\,(\forall x^* \in \mathcal{S}) \qquad \mathbb{E}\left[\|x^{k+1} - x^*\|^2 + \kappa_{k+1} \mid \mathcal{F}_k\right] + Y_k \leq \|x^k - x^*\|^2 + \kappa_k. \tag{A.2}$$

So, by the supermartingale convergence theorem, the sequence Y_k is a.s. summable and $X_k := \|x^k - x^*\|^2 + \kappa_k$ a.s. converges to a $[0,\infty)$-valued random variable, and these conclusions hold for any element $x^* \in \mathcal{S}$.

At this point, as in the synchronous case, Part 1 of the theorem follows because $\sum_{k=0}^\infty \sum_{j=1}^m \sum_{i=1}^n p_{ij}R_{2,ij}^k r_{ij}^p(x^{k-d_k}) \leq \sum_{k=0}^\infty Y_k < \infty$ a.s., and hence, $\|S_i(x^{k-d_k}) - S_i(x^*)\| = (np_{ij})^2 r_{ij}^p(x^{k-d_k})$ a.s. converges to 0.

The road to almost sure weak convergence of x^k, now 5 steps long, is slightly more complicated than in the synchronous case:

1. Because $\sum_{k=0}^\infty Y_k < \infty$ a.s., we conclude that each term[29]

$$\|S(x^{k-d_k})\|; \qquad \left\|\left(\frac{1}{n}\sum_{i=1}^n y_{i,j}^{k-e_k^i}\right)_{j=1}^m\right\|; \qquad r_{ij}^d(y_i^{k-e_k^i});$$

$$r_{ij}^p(x^{k-d_k}); \qquad \sum_{\substack{h=0 \\ h\neq\tau_d-e_{k,j}^i}}^{\tau_d} \frac{p_{ij}\gamma_{ij}}{\tau_d + 1}r_{ij}^d(y_i^{k-\tau_d+h}),$$

a.s. converges to 0.

2. Because $r_{ij}^d(y_i^{k-e_k^i})$ and $\sum_{\substack{h=0; h\neq\tau_d-e_k^{ij}}}^{\tau_d}\frac{p_{ij}\gamma_{ij}}{\tau_d+1}r_{ij}^d(y_i^{k-\tau_d+h})$ a.s. converge to 0, we conclude that $\kappa_{1,k}$ a.s. converges to 0.

---

[28] By the choice of γ_{ij}, there is a constant $b > 0$ such that for all i,j and k, we have $R_{1,ij}^k > b$ and $R_{2,ij}^k > b$.

[29] Here we invoke the assumed equivalence of the two norms $\|x\|$ and $\|x\|_{\text{prod}} = \sqrt{\sum_{j=1}^m \|x_j\|_j^2}$.

3. Because $r_{ij}^d(y_i^{k-e_k^i})$, $\|n^{-1}\sum_{i=1}^n (y_i^{k-e_k^i})_j\|$, and $r_{ij}^p(x^{k-d_k})$ a.s. converge to 0 and because for all $t \in \{0, \ldots, \tau\}$, we have

$$x_j^{k-t+1} - x_j^{k-t} = -\lambda_{k-t} t_{k-t,j}\left((Q_{i_{k-t}}^p(x^{k-t-d_{k-t}}))_j - (Q_{i_{k-t}}^d(y_{i_k}^{k-t-e_{k-t}^{i_{k-t}}}))_j + \frac{1}{n}\sum_{i=1}^n y_{i,j}^{k-t-e_{k-t}^i}\right)$$

$$= -\lambda_{k-t} t_{k-t,j}\Bigg(((Q_{i_{k-t}}^p(x^{k-t-d_{k-t}}))_j - (Q_{i_{k-t}}^*)_j)$$

$$+ ((Q_{i_{k-t}}^*)_j - (Q_{i_{k-t}}^d(y_{i_k}^{k-t-e_{k-t}^{i_{k-t}}}))_j) + \frac{1}{n}\sum_{i=1}^n y_{i,j}^{k-t-e_{k-t}^i}\Bigg) \overset{k\to\infty}{\to} 0 \text{ a.s.,}$$

we conclude $\kappa_{2,k}$ and $\|x^{k-d_k} - x^k\|^2$ a.s. converge to 0:

4. Because κ_k a.s. converges to 0 and $\|x^k - x^*\|^2 + \kappa_k$ a.s. converges to a $[0,\infty)$-valued random variable, we conclude that $\|x^k - x^*\|^2$ a.s. converges to a $[0,\infty)$-valued random variable.

5. Because $\|x^k - x^*\|^2$ a.s. converges to a $[0,\infty)$-valued random variable, because $\|x^k - x^{k-d_k}\|$ a.s. converges to 0, and because

$$\|x^{k-d_k} - x^*\|^2 - \|x^k - x^*\|^2 = \|x^k - x^{k-d_k}\|^2 + 2\langle x^{k-d_k} - x^k, x^k - x^*\rangle \overset{k\to\infty}{\to} 0 \text{ a.s.,}$$

we conclude that $\|x^{k-d_k} - x^*\|^2$ a.s. converges to a $[0,\infty)$-valued random variable.

With the facts listed here, Proposition 7.1 immediately renders the sequence x^{k-d_k} a.s. weakly convergent to an $\mathcal{S}$-valued random variable. And at last, due to the limit $\|x^k - x^{k-d_k}\| \to 0$ a.s., the sequence x^k a.s. weakly converges to an $\mathcal{S}$-valued random variable.

Our task is now clear: we must prove (A.2). But to do so, we must first bound several random variables. We present the two bounds we need now and defer their proofs until later.

Lemma A.1 (Variance bound) *For $j \in \{1, \ldots, m\}$, let $\eta_j > 0$ be a positive real number. Then for all $s \in \mathbb{N}$, we have*

$$\sum_{j=1}^m \eta_j \mathbb{E}\left[\|\mathfrak{Q}_j^s\|_j^2 \mid \mathcal{F}_s\right] \leq 2\sum_{j=1}^m \sum_{i=1}^n \frac{p_{ij}\eta_j}{q_j m^2} r_{ij}^p(x^{s-d_s}) + 2\sum_{j=1}^m \sum_{i=1}^n \frac{p_{ij}\eta_j}{q_j m^2} r_{ij}^d(y_i^{s-e_s^i})$$

$$- 2\sum_{j=1}^m \frac{\eta_j}{q_j m^2}\left\|\frac{1}{n}\sum_{i=1}^n y_{i,j}^{s-e_s^i}\right\|_j^2 - \sum_{j=1}^m \frac{\eta_j}{m^2}\|S(x^{s-d_s})\|_j^2. \tag{A.3}$$

Lemma A.2 (Recursive κ_k bound) *For all $s \in \mathbb{N}$ and $\alpha \in [0,1]$, we have*

$$\mathbb{E}\left[\kappa_{s+1} \mid \mathcal{F}_s\right]$$

$$\leq \kappa_s + \frac{2\lambda_s}{m}\langle S(x^{s-d_s}), x^s - x^*\rangle - \frac{2\alpha\lambda_s}{m}\langle S(x^{s-d_s}), x^{s-d_s} - x^*\rangle + \sum_{j=1}^m \sum_{i=1}^n \frac{2\alpha\lambda_s n^2 p_{ij}^2 \beta_{ij}}{m} r_{ij}^p(x^{s-d_s})$$

$$+ \lambda_s^2 \sum_{j=1}^m \frac{\tau_p \overline{M}_j}{mC}\mathbb{E}\left[\|\mathfrak{Q}_j^s\|_j^2 \mid \mathcal{F}_s\right] - \sum_{j=1}^m \sum_{i=1}^n p_{ij}\left(\frac{2(1-\alpha)\lambda_s \beta_{ij} n^2 p_{ij}}{m} - \frac{\tau_p \overline{M}_j C_j \lambda_s^2}{m} - \gamma_{ij}\right) r_{ij}^p(x^{s-d_s})$$

$$- \sum_{j=1}^m \sum_{i=1}^n \frac{p_{ij}\gamma_{ij}}{\tau_d + 1} r_{ij}^d(y_i^{s-e_s^i}) - \sum_{j=1}^m \sum_{i=1}^n \sum_{\substack{h=0 \\ h \neq \tau_d - e_{s,j}^i}}^{\tau_d} \frac{p_{ij}\gamma_{ij}}{\tau_d + 1} r_{ij}^d(y_i^{s-\tau_d+h}).$$

Lemmas in hand, we can now prove (A.2). Compared to the synchronous case, we prove a refined bound, depending on the parameter α, which will figure into our proof of linear convergence.

Lemma A.3 (Proof of (A.2)) *Equation (A.2) holds for all $s \in \mathbb{N}$. Moreover, for every $s \in \mathbb{N}$, $\alpha \in [0,1]$ and $x^* \in \mathcal{S}$, we have*

$$\mathbb{E}\left[\|x^{s+1} - x^*\|^2 + \kappa_{s+1} \mid \mathcal{F}_k\right] + Y_s \leq \|x^s - x^*\|^2 + \kappa_s$$

$$- \frac{2\alpha\lambda_s}{m}\langle S(x^{s-d_s}), x^{s-d_s} - x^*\rangle + \sum_{j=1}^m \sum_{i=1}^n \frac{2\alpha\lambda_s n^2 p_{ij}^2 \beta_{ij}}{m} r_{ij}^p(x^{s-d_s}). \tag{A.4}$$

Proof Fix $s \in \mathbb{N}$. Then

$$\mathbb{E}\left[\|x^{s+1} - x^*\|^2 + \kappa_{s+1} \mid \mathcal{F}_s\right] = \|x^s - x^*\|^2 - 2\lambda_s \mathbb{E}\left[\langle \mathfrak{Q}^s, x^s - x^* \rangle \mid \mathcal{F}_s\right] + \lambda_s^2 \mathbb{E}\left[\|\mathfrak{Q}^s\|^2 \mid \mathcal{F}_s\right] + \mathbb{E}\left[\kappa_{s+1} \mid \mathcal{F}_s\right]$$

$$= \|x^s - x^*\|^2 - \frac{2\lambda_s}{m}\langle S(x^{s-d_s}), x^s - x^* \rangle + \lambda_s^2 \mathbb{E}\left[\|\mathfrak{Q}^s\|^2 \mid \mathcal{F}_s\right] + \mathbb{E}\left[\kappa_{s+1} \mid \mathcal{F}_s\right],$$

where the second equality follows from the linearity of expectation. Next apply Lemma A.2 to the above equality:

$$\leq \|x^s - x^*\|^2 + \kappa_s - \frac{2\alpha\lambda_s}{m}\langle S(x^{s-d_s}), x^{s-d_s} - x^* \rangle + \lambda_s^2 \mathbb{E}\left[\sum_{j=1}^m \overline{M}_j \left(1 + \frac{\tau_p}{mC}\right)\|\mathfrak{Q}_j^s\|_j^2 \mid \mathcal{F}_s\right]$$

$$- \sum_{j=1}^m \sum_{i=1}^n \frac{p_{ij}\gamma_{ij}}{\tau_d + 1} r_{ij}^d(y_i^{s-e_s^i})$$

$$- \sum_{j=1}^m \sum_{i=1}^n p_{ij} \left(\frac{2(1-\alpha)\lambda_s\beta_{ij}n^2 p_{ij}}{m} - \frac{\tau_p \overline{M}_j C \lambda_s^2}{m} - \gamma_{ij}\right) r_{ij}^p(x^{s-d_s})$$

$$- \sum_{j=1}^m \sum_{i=1}^n \sum_{\substack{h=0 \\ h \neq \tau_d - e_{s,j}^i}}^{\tau_d} \frac{p_{ij}\gamma_{ij}}{\tau_d + 1} r_{ij}^d(y_i^{s-\tau_d+h}).$$

Finally, apply Lemma A.1 and combine terms:

$$\leq \|x^s - x^*\|^2 + \kappa_s - \frac{2\alpha\lambda_s}{m}\langle S(x^{s-d_s}), x^{s-d_s} - x^* \rangle$$

$$- \lambda_s^2 \sum_{j=1}^m \frac{\overline{M}_j}{m^2}\left(1 + \frac{\tau_p}{mC}\right)\|(S(x^{s-d_s}))_j\|_j^2 - 2\lambda_s^2 \sum_{j=1}^m \frac{\overline{M}_j}{q_j m^2}\left(1 + \frac{\tau_p}{mC}\right)\left\|\frac{1}{n}\sum_{i=1}^n y_{i,j}^{s-e_s^i}\right\|_j^2$$

$$- \sum_{j=1}^m \sum_{i=1}^n p_{ij}\left(\frac{\gamma_{ij}}{\tau_d+1} - \frac{2\overline{M}_j}{q_j m^2}\left(1 + \frac{\tau_p}{mC}\right)\lambda_s^2\right) r_{ij}^d(y_i^{s-e_s^{ij}})$$

$$- \sum_{j=1}^m \sum_{i=1}^n p_{ij}\left(\frac{2(1-\alpha)\lambda_s\beta_{ij}n^2 p_{ij}}{m} - \frac{\tau_p \overline{M}_j C \lambda_s^2}{m} - \frac{2\overline{M}_j}{q_j m^2}\left(1 + \frac{\tau_p}{mC}\right)\lambda_s^2 - \gamma_{ij}\right) r_{ij}^p(x^{s-d_s})$$

$$- \sum_{j=1}^m \sum_{i=1}^n \sum_{\substack{h=0 \\ h \neq \tau_d - e_{s,j}^i}}^{\tau_d} \frac{p_{ij}\gamma_{ij}}{\tau_d + 1} r_{ij}^d(y_i^{s-\tau_d+h})$$

$$\leq \|x^s - x^*\|^2 + \kappa_s - \frac{2\alpha\lambda_s}{m}\langle S(x^{s-d_s}), x^{s-d_s} - x^* \rangle$$

$$- \lambda_s^2 \sum_{j=1}^m \frac{\overline{M}_j}{m^2}\left(1 + \frac{\tau_p}{mC}\right)\|(S(x^{s-d_s}))_j\|_j^2 - 2\lambda_s^2 \sum_{j=1}^m \frac{\overline{M}_j}{q_j m^2}\left(1 + \frac{\tau_p}{C}\right)\left\|\frac{1}{n}\sum_{i=1}^n y_{i,j}^{s-e_s^i}\right\|_j^2$$

$$- \sum_{j=1}^m \sum_{i=1}^n p_{ij}\left(\frac{\gamma_{ij}}{\tau_d+1} - \frac{2\overline{M}_j}{q_j m^2}\left(1 + \frac{\tau_p}{mC}\right)\lambda_s^2\right) r_{ij}^d(y_i^{s-e_s^i})$$

$$- \sum_{j=1}^m \sum_{i=1}^n p_{ij}\left(\frac{2(1-\alpha)\lambda_s n^2 p_{ij}\beta_{ij}}{m} - \overline{M}_j\left(\frac{2}{q_j m^2} + \tau_p\left(\frac{2}{q_j m^3 C} + \frac{C}{m}\right)\right)\lambda_s^2 - \gamma_{ij}\right) r_{ij}^p(x^{s-d_s})$$

$$- \sum_{j=1}^m \sum_{i=1}^n \sum_{\substack{h=0 \\ h \neq \tau_d - e_{s,j}^i}}^{\tau_d} \frac{p_{ij}\gamma_{ij}}{\tau_d + 1} r_{ij}^d(y_i^{s-\tau_d+h})$$

$$= \|x^s - x^*\|^2 + \kappa_s - Y_s - \frac{2\alpha\lambda_s}{m}\langle S(x^{s-d_s}), x^{s-d_s} - x^* \rangle + \sum_{j=1}^m \sum_{i=1}^n \frac{2\alpha\lambda_s n^2 p_{ij}^2 \beta_{ij}}{m} r_{ij}^p(x^{s-d_s}). \square$$

Inequality ([A.2](#)) is simply the above inequality with $\alpha = 0$.

We finish the proof of Parts [1](#) and [2](#) by proving Lemmas [A.1](#) and [A.2](#).

Proof (of Lemma [A.1](#) (variance bound)) Up to a change of indices, the proof of Lemma [A.1](#) follows the proof Lemma [7.1](#) exactly; we omit the proof and avoid repeating ourselves. □

Proof (Lemma [A.2](#) (recursive κ_k bound))

The bound is the addition of two further bounds: for all $s \in \mathbb{N}$ and $\alpha \in [0,1]$, we have

1. **Recursive $\kappa_{1,k}$ bound.**

$$\mathbb{E}\left[\kappa_{1,s+1} \mid \mathcal{F}_s\right] \le \kappa_{1,s} - \sum_{j=1}^m \sum_{i=1}^n \frac{p_{ij}\gamma_{ij}}{\tau_d+1} r_{ij}^d(y_i^{s-e_s^i}) + \sum_{j=1}^m \sum_{i=1}^n p_{ij}\gamma_{ij} r_{ij}^p(x^{s-d_s})$$
$$- \sum_{j=1}^m \sum_{i=1}^n \sum_{\substack{h=0 \\ h \ne \tau_d - e_{s,j}^i}}^{\tau_d} \frac{p_{ij}\gamma_{ij}}{\tau_d+1} r_{ij}^d(y_i^{s-\tau_d+h}); \text{ and}$$

2. **Recursive $\kappa_{2,k}$ bound.**

$$\kappa_{2,s+1} \le \kappa_{2,s} + \frac{2\lambda}{m}\langle S(x^{s-d_s}), x^s - x^* \rangle - \frac{2\alpha\lambda_s}{m}\langle S(x^{s-d_s}), x^{s-d_s} - x^* \rangle + \sum_{j=1}^m \sum_{i=1}^n \frac{2\alpha\lambda_s n^2 p_{ij}^2 \beta_{ij}}{m} r_{ij}^p(x^{s-d_s})$$
$$- \sum_{j=1}^m \sum_{i=1}^n p_{ij}\left(\frac{2\beta_{ij}\lambda_s n^2 p_{ij}}{m} - \frac{\tau_p \overline{M}_j C}{m}\lambda_s^2\right) r_{ij}^p(x^{s-d_s}) + \sum_{j=1}^m \frac{\tau_p \overline{M}_j}{mC}\|\mathfrak{Q}_j^s\|_j^2.$$

The proofs of the above bounds are orthogonal; the $\kappa_{1,k}$ bound is a consequence of how we sample variables in SMART, while the $\kappa_{2,k}$ bound is a consequence of ([1.5](#)). Thus, we separately prove each bound.

Proof of recursive $\kappa_{1,k}$ bound. For any i and j with $\mathbf{S}_{ij}^* \ne 0$, we have

$$\mathbb{E}\left[r_{ij}^d(y_i^{s+1}) \mid \mathcal{F}_s\right] = \left(1 - \rho p_{ij}^T\right) r_{ij}^d(y_i^s) + \rho p_{ij}^T r_{ij}^p(x^{s-d_s}).$$

because $r_{ij}^d(y_i^{s+1})$ depends only on $y_{i,j}^{s+1}$, not on its other components, and the probability of update for $(y_i^k)_j$ is $P((i_k,i) \in E, t_{k,j} \ne 0, \epsilon_k = 1) = \rho p_{ij}^T$. We can also break up the sum:

$$\sum_{h=0}^{\tau_d-1} \sum_{j=1}^m \sum_{i=1}^n \frac{(h+1)p_{ij}\gamma_{ij}}{\tau_d+1} r_{ij}^d(y_i^{s+1-\tau_d+h})$$
$$= \sum_{h=0}^{\tau_d-1} \sum_{j=1}^m \sum_{i=1}^n \frac{(h+1)p_{ij}\gamma_{ij}}{\tau_d+1} r_{ij}^d(y_i^{s-\tau_d+h}) + \sum_{j=1}^m \sum_{i=1}^n \frac{\tau_d p_{ij}\gamma_{ij}}{\tau_d+1} r_{ij}^d(y_i^s)$$
$$- \sum_{j=1}^m \sum_{i=1}^n \frac{p_{ij}\gamma_{ij}}{\tau_d+1} r_{ij}^d(y_i^{s-e_s^i}) - \sum_{j=1}^m \sum_{i=1}^n \sum_{\substack{h=0 \\ h \ne \tau_d - e_{s,j}^i}}^{\tau_d-1} \frac{p_{ij}\gamma_{ij}}{\tau_d+1} r_{ij}^d(y_i^{s-\tau_d+h}).$$

Therefore,

$$
\begin{aligned}
&\mathbb{E}\left[\kappa_{1,s+1} \mid \mathcal{F}_s\right] \\
&= \sum_{j=1}^{m} \sum_{i=1}^{n} \frac{\gamma_{ij} p_{ij}}{\rho p_{ij}^T} \mathbb{E}\left[r_{ij}^d(y_i^{s+1}) \mid \mathcal{F}_k\right] + \sum_{h=0}^{\tau_d-1} \sum_{j=1}^{m} \sum_{i=1}^{n} \frac{(h+1)p_{ij}\gamma_{ij}}{\tau_d+1} r_{ij}^d(y_i^{s+1-\tau_d+h}) \\
&\leq \sum_{j=1}^{m} \sum_{i=1}^{n} \left(\frac{\gamma_{ij}p_{ij}}{\rho p_{ij}^T} - p_{ij}\gamma_{ij}\right) r_{ij}^d(y_i^s) + \sum_{j=1}^{m} \sum_{i=1}^{n} p_{ij}\gamma_{ij} r_{ij}^p(x^{s-d_s}) \\
&\quad + \sum_{h=0}^{\tau_d-1} \sum_{j=1}^{m} \sum_{i=1}^{n} \frac{(h+1)p_{ij}\gamma_{ij}}{\tau_d+1} r_{ij}^d(y_d^{s-\tau_d+h}) + \sum_{j=1}^{m} \sum_{i=1}^{n} \frac{\tau_d p_{ij}\gamma_{ij}}{\tau_d+1} r_{ij}^d(y_i^s) \\
&\quad - \sum_{j=1}^{m} \sum_{i=1}^{n} \frac{p_{ij}\gamma_{ij}}{\tau_d+1} r_{ij}^d(y_i^{s-e_s^i}) - \sum_{j=1}^{m} \sum_{i=1}^{n} \sum_{\substack{h=0 \\ h \neq \tau_d - e_{s,j}^i}}^{\tau_d-1} \frac{p_{ij}\gamma_{ij}}{\tau_d+1} r_{ij}^d(y_i^{s-\tau_d+h}),
\end{aligned}
$$

where the first inequality uses the previous two blocks of equalities, and the trivial bound: $r_{ij}^p(x^{s-d_s}) \geq 0$ for all i and j with $\mathbf{S}_{ij}^* = 0$. The claimed inequality now follows because

$$
\left(\left(\frac{\gamma_{ij}p_{ij}}{\rho p_{ij}^T} - p_{ij}\gamma_{ij}\right) + \frac{\tau_d p_{ij}\gamma_{ij}}{\tau_d+1}\right) r_{ij}^d(y_i^s) \leq \left(\frac{\gamma_{ij}p_{ij}}{\rho p_{ij}^T} - \frac{p_{ij}\gamma_{ij}}{\tau_d+1}\right) r_{ij}^d(y_i^s),
$$

and hence, we can add $-\frac{p_{ij}\gamma_{ij}}{\tau_d+1} r_{ij}^d(y_i^s)$ back to the last term in the previous sum. This completes the proof of the $\kappa_{1,k}$ bound.

Proof of the recursive $\kappa_{2,k}$ bound. By Jensen's inequality, $(\sum_{i=1}^{n} p_{ij} = 1)$

$$
\begin{aligned}
\sum_{j=1}^{m} \tau_p C \|(S(x^{s-d_s}))_j\|_j^2 &= \sum_{j=1}^{m} \tau_p C \left\| \frac{1}{n} \sum_{i=1}^{n} \left[(S_i(x^{s-d_s}))_j - (S_i(x^*))_j\right] \right\|_j^2 \\
&= \sum_{j=1}^{m} \tau_p C \left\| \sum_{i=1}^{n} p_{ij} \left[(Q_i^p(x^{s-d_s}))_j - (Q_i^*)_j\right] \right\|_j^2 \leq \sum_{j=1}^{m} \sum_{i=1}^{n} p_{ij} \tau_p C r_{ij}^p(x^{s-d_s}).
\end{aligned}
$$

From (1.5),

$$
\langle S(x^{s-d_s}), x^{s-d_s} - x^* \rangle \geq \sum_{j=1}^{m} \sum_{i=1}^{n} \beta_{ij} \|(S_i(x^{s-d_s}))_j - (S_i(x^*))_j\|_j^2 \geq \sum_{j=1}^{m} \sum_{i=1}^{n} \beta_{ij} n^2 p_{ij}^2 r_{ij}^p(x^{s-d_s}).
$$

Because $x_j^t - x_j^{t+1} = \lambda_s \mathfrak{Q}_j^t$ for all $t \in \mathbb{N}$,

$$
\begin{aligned}
\sum_{j=1}^{m} \sum_{t=s-\tau_p+1}^{s} \frac{\overline{M}_j}{C} \|x_j^t - x_j^{t-1}\|_j^2 &= \sum_{j=1}^{m} \sum_{t=s-\tau_p+1}^{s} \frac{\overline{M}_j(t-s+\tau_p)}{C} \|x_j^t - x_j^{t-1}\|_j^2 \\
&\quad - \sum_{j=1}^{m} \sum_{t=s-\tau_p+2}^{s+1} \frac{\overline{M}_j(t-(s+1)+\tau_p)}{C} \|x_j^t - x_j^{t-1}\|_j^2 + \lambda_s^2 \sum_{j=1}^{m} \frac{\overline{M}_j \tau_p}{C} \|\mathfrak{Q}_j^s\|_j^2. \\
&= m\kappa_{2,s} - m\kappa_{2,s+1} + \lambda_s^2 \sum_{j=1}^{m} \frac{\overline{M}_j \tau_p}{C} \|\mathfrak{Q}_j^s\|_j^2.
\end{aligned}
$$

To get the next bound, we apply the Cauchy-Schwarz inequality, Young's inequality, the equivalence of $\|\cdot\|$ and $\|\cdot\|_{\mathrm{prod}}$, and the convexity of $\|\cdot\|^2$—in that order:

$$
\begin{aligned}
2\lambda_s\langle S(x^{s-d_s}), \sum_{t=s-d_{k,j}+1}^{s}(x^t - x^{t-1})\rangle &\geq -\|S(x^{s-d_s})\|\left\|\sum_{t=s-d_{k,j}+1}^{s}(x^t - x^{t-1})\right\| \\
&\geq -C\tau_p\lambda^2\|S(x^{s-d_s}))\|^2 - \frac{1}{C\tau_p}\left\|\sum_{t=s-d_{k,j}+1}^{s}(x^t - x^{t-1})\right\|^2 \\
&\geq -C\tau_p\lambda^2\sum_{j=1}^{n}\overline{M}_j\|S(x^{s-d_s}))_j\|_j^2 - \frac{1}{C\tau_p}\sum_{j=1}^{m}\overline{M}_j\left\|\sum_{t=s-d_{k,j}+1}^{s}(x_j^t - x_j^{t-1})\right\|_j^2 \\
&\geq -\lambda_s^2\sum_{j=1}^{m}\tau_p C\overline{M}_j\|(S(x^{s-d_s}))_j\|_j^2 - \sum_{j=1}^{m}\sum_{h=s-\tau_p+1}^{s}\frac{\overline{M}_j}{C}\|x_j^t - x_j^{t-1}\|_j^2.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
2\lambda_s\langle S(x^{s-d_s}), x^s - x^*\rangle &= 2\lambda_s\langle S(x^{s-d_s}), x^{s-d_s} - x^*\rangle \\
&\quad + 2\lambda_s\sum_{j=1}^{m}\langle (S(x^{s-d_s}))_j, \sum_{t=s-d_{k,j}+1}^{s}(x_j^t - x_j^{t-1})\rangle_j \\
&\geq 2\alpha\lambda_s\langle S(x^{s-d_s}), x^{s-d_s} - x^*\rangle + 2(1-\alpha)\lambda_s\langle S(x^{s-d_s}), x^{s-d_s} - x^*\rangle \\
&\quad - \lambda_s^2\sum_{j=1}^{m}\tau_p C\overline{M}_j\|(S(x^{s-d_s}))_j\|_j^2 - \sum_{j=1}^{m}\sum_{h=s-\tau_p+1}^{s}\frac{\overline{M}_j}{C}\|x_j^t - x_j^{t-1}\|_j^2 \\
&\geq 2\alpha\lambda_s\langle S(x^{s-d_s}), x^{s-d_s} - x^*\rangle - \sum_{j=1}^{m}\sum_{i=1}^{n}\frac{2\alpha\lambda_s n^2 p_{ij}^2\beta_{ij}}{m}r_{ij}^p(x^{s-d_s}) \\
&\quad + \sum_{j=1}^{m}\sum_{i=1}^{n}p_{ij}\left(2\lambda_s\beta_{ij}n^2 p_{ij} - \tau_p C\lambda_s^2\right)r_{ij}^p(x^{s-d_s}) \\
&\quad - \left(m\kappa_{2,s} - m\kappa_{2,s+1} + \lambda_s^2\sum_{j=1}^{m}\frac{\tau_p}{C}\|\mathfrak{Q}_j^s\|_j^2\right).
\end{aligned}
$$

After a simple rearrangement of terms, this last bound completes the proof of the $\kappa_{2,k}$ bound, and consequently, the proof of the lemma is complete. $\qquad\square$

Part 3: In this part, we introduce two constants, called $T_{1,ij}$ and $T_{2,ij}$, which play the same role as $R_{1,ij}^k$ and $R_{2,ij}^k$ play in Parts 1 and 2 (in fact, $T_{1,ij} = R_{1,ij}^k$)

$$
\begin{aligned}
T_{1,ij} &:= \frac{\gamma_{ij}}{\tau_d + 1} - \frac{2\overline{M}_j}{q_j m^2}\left(1 + \frac{\tau_p}{mC}\right)\lambda^2; \text{ and} \\
T_{2,ij} &:= \frac{2(1-\alpha)\lambda n^2 p_{ij}\beta_{ij}}{m} - \overline{M}_j\left(\frac{2}{q_j m^2} + \tau_p\left(\frac{2}{q_j m^3 C} + \frac{C}{m}\right)\right)\lambda^2 - \gamma_{ij}.
\end{aligned}
$$

Not only do we require that γ_{ij} be chosen so that $T_{1,ij}$ and $T_{2,ij}$ are positive, but γ_{ij} must be further constrained so that several other inequalities hold. (We defer the proof of the lemma for a moment.)

Lemma A.4 (Choosing γ_{ij}) *Let λ satisfy* (3.3). *Set $\xi := 2m^{-1}\alpha\mu\lambda$ and $\zeta := (2.5\alpha)^{-1}\mu\overline{M}\left(1 + \frac{\tau\sqrt{q}}{\sqrt{2(\tau_d+2)}}\right)$. Then for all i and j, there exists $\gamma_{ij} > 0$ such that*

$$T_{1,ij} - \frac{\xi\gamma_{ij}}{\rho p_{ij}^T} - \frac{4\alpha\mu\delta\overline{M}_j}{q_j m^3}\left(\frac{1}{\zeta} + \lambda\right)\lambda^2 \geq 0; \qquad\qquad T_{2,ij} - \frac{4\alpha\mu\delta\overline{M}_j}{q_j m^3}\left(\frac{1}{\zeta} + \lambda\right)\lambda^2 \geq 0;$$

$$T_{1,ij} - \frac{2\overline{M}_j\tau_p\lambda^2\xi}{Cq_j m^3(1-\xi)} \geq 0; \qquad\qquad T_{2,ij} - \frac{2\overline{M}_j\tau_p\lambda^2\xi}{Cq_j m^3(1-\xi)} \geq 0. \qquad (A.5)$$

Moreover, $1 - \xi(\tau_d + 1) \geq 0$.

As in the synchronous case, a bound like (7.4) holds. But unlike (7.4), in the present asynchronous case, the expectation $\mathbb{E}\left[\|x^k - x^*\| + \kappa_k\right]$ contracts, not every iteration, but over each series of τ_p iterations, and for this reason we will apply (A.4) multiple times, saving nonnegative summands from later iterations to absorb positive summands from earlier iterations which would otherwise, if not absorbed, prevent the expectation from contracting at all.

We split our descent into three stages: *initial descent*, in which we descend one iteration, from iteration $k + 1$ to iteration k; *intermediate descent*, in which we descend at most τ_p iterations, from iteration k to iteration $k - h$ for some $h \in \{0, \ldots, \tau_p\}$; and *final descent*, in which we descend one final iteration, from iteration $k - h$ to iteration $k - h - 1$. In each stage of descent, we apply (A.4), between 1 and τ_p times—with the law of iterated expectations being never mentioned, but always applied to turn all conditional expectation inequalities into unconditional expectation inequalities.

As in the synchronous case, we no longer work with arbitrary zeros $x^* \in \mathcal{S}$. Instead we work with the sequence of zeros $P_{\mathcal{S}}(x^k)$, which, by definition, satisfy

$$d_{\mathcal{S}}^2(x^{k+1}) = \|x^{k+1} - P_{\mathcal{S}}(x^{k+1})\|^2 \leq \|x^{k+1} - P_{\mathcal{S}}(x^{k-h})\|^2; \text{ and}$$
$$d_{\mathcal{S}}^2(x^{k-h}) = \|x^{k-h} - P_{\mathcal{S}}(x^{k-h})\|^2 \leq \|x^{k-h} - P_{\mathcal{S}}(x^{k-h-1})\|^2.$$

for any $h \in \mathbb{N}$. Below, our choice of h is precisely the maximal delay of any coordinate of x^{k-d_k}: let the integers $j_1, \ldots, j_m \in \{1, \ldots, m\}$ order the delays from smallest to largest: $d_{k,j_1} \leq \ldots \leq d_{k,j_m}$. With this notation, $\delta \geq d_{k,j_l} - d_{k,j_m}$ for $l = 1, \ldots, m$, and we set $h = d_{k,j_m}$.

Initial descent (from iteration $k + 1$ to iteration k). Apply (A.4) to descend to the kth iteration:

$$\mathbb{E}\left[d_{\mathcal{S}}^2(x^{k+1}) + \kappa_{k+1}\right]$$
$$\leq \mathbb{E}\left[\|x^{k+1} - P_{\mathcal{S}}(x^{k-d_{k,j_m}})\|^2 + \kappa_{k+1}\right] \qquad \text{(by definition of } P_{\mathcal{S}})$$
$$\leq \mathbb{E}\left[\|x^k - P_{\mathcal{S}}(x^{k-d_{k,j_m}})\|^2 + \kappa_s - \frac{2\alpha\lambda}{m}\langle S(x^{k-d_k}), x^{k-d_k} - P_{\mathcal{S}}(x^{k-d_{k,j_m}})\rangle\right]$$
$$- \lambda^2\sum_{j=1}^m \frac{\overline{M}_j}{m^2}\left(1 + \frac{\tau_p}{mC}\right)\mathbb{E}\left[\|(S(x^{k-d_k}))_j\|_j^2\right] - \sum_{j=1}^m\sum_{i=1}^n p_{ij}T_{1,ij}\mathbb{E}\left[r_{ij}^d(y_i^{k-e_k^i})\right]$$
$$- \sum_{j=1}^m\sum_{i=1}^n p_{ij}T_{2,ij}\mathbb{E}\left[r_{ij}^p(x^{k-d_k})\right] - \sum_{j=1}^m\sum_{i=1}^n \sum_{\substack{h=0 \\ h\neq\tau_d - e_{k,j}^i}}^{\tau_d} \frac{p_{ij}\gamma_{ij}}{\tau_d + 1}\mathbb{E}\left[r_{ij}^d(y_i^{k-\tau_d+h})\right]. \qquad (A.6)$$

Intermediate descent (from iteration k to iteration $k - d_{k,j_m}$). To bound our initial descent (A.6), apply (A.4) a total of d_{k,j_m} times:

$$(A.6) \leq \mathbb{E}\left[\|x^{k-d_{k,j_m}} - P_{\mathcal{S}}(x^{k-d_{k,j_m}})\|^2 + \kappa_{k-d_{k,j_m}} - \frac{2\alpha\lambda}{m}\langle S(x^{k-d_k}), x^{k-d_k} - P_{\mathcal{S}}(x^{k-d_{k,j_m}})\rangle\right]$$
$$- \lambda^2\sum_{j=1}^m \frac{\overline{M}_j}{m^2}\left(1 + \frac{\tau_p}{mC}\right)\mathbb{E}\left[\|(S(x^{k-d_k}))_j\|_j^2\right] - \sum_{t=k-d_{k,j_m}}^k \sum_{j=1}^m\sum_{i=1}^n p_{ij}T_{1,ij}\mathbb{E}\left[r_{ij}^d(y_i^{t-e_t^i})\right]$$
$$- \sum_{t=k-d_{k,j_m}}^k \sum_{j=1}^m\sum_{i=1}^n p_{ij}T_{2,ij}\mathbb{E}\left[r_{ij}^p(x^{t-d_t})\right] - \sum_{j=1}^m\sum_{i=1}^n \sum_{\substack{h=0 \\ h\neq\tau_d - e_{k-d_{k,j_m},j}^j}}^{\tau_d} \frac{p_{ij}\gamma_{ij}}{\tau_d + 1}\mathbb{E}\left[r_{ij}^d(y_i^{k-d_{k,j_m}-\tau_d+h})\right]. \qquad (A.7)$$

As we descended to the $(k-d_{k,j_m})$th iteration, we dropped a lot of negative terms including $2\alpha\lambda\langle S(x^{s-d_s}), x^{s-d_s}-P_{\mathcal{S}}(x^{k-d_{k,j_m}})\rangle$ for all $s\in\{k-1,\ldots,k-d_{k,j_m}\}$. To show that the expectation contracts, we do not need these terms, so we omit them.

We need to rearrange the terms in (A.7) in order to derive a contraction. First we squeeze a norm term out of the inner product term—this being the only time we use the essential strong quasi-monotonicity of S:

$$\langle S(x^{k-d_k}), x^{k-d_k} - P_{\mathcal{S}}(x^{k-d_{k,j_m}})\rangle = \langle S(x^{k-d_k}), x^{k-d_k} - P_{\mathcal{S}}(x^{k-d_k})\rangle + \langle S(x^{k-d_k}), P_{\mathcal{S}}(x^{k-d_k}) - P_{\mathcal{S}}(x^{k-d_{k,j_m}})\rangle$$
$$\geq \mu\|x^{k-d_k} - P_{\mathcal{S}}(x^{k-d_k})\|^2 + \langle S(x^{k-d_k}), P_{\mathcal{S}}(x^{k-d_k}) - P_{\mathcal{S}}(x^{k-d_{k,j_m}})\rangle.$$

Then we simply split up the term

$$\|x^{k-d_{k,j_m}} - P_{\mathcal{S}}(x^{k-d_{k,j_m}})\|^2 = \left(1 - \frac{2\alpha\mu\lambda}{m}\right)\|x^{k-d_{k,j_m}} - P_{\mathcal{S}}(x^{k-d_{k,j_m}})\|^2 + \frac{2\alpha\mu\lambda}{m}\|x^{k-d_{k,j_m}} - P_{\mathcal{S}}(x^{k-d_{k,j_m}})\|^2$$

and couple it with $\mu\|x^{k-d_k} - P_{\mathcal{S}}(x^{k-d_k})\|^2$:

$$\begin{aligned}
(\text{A.7}) \leq\ & \mathbb{E}\left[\left(1 - \frac{2\mu\alpha\lambda}{m}\right)\|x^{k-d_{k,j_m}} - P_{\mathcal{S}}(x^{k-d_{k,j_m}})\|^2 + \kappa_{k-d_{k,j_m}}\right]\\
& - \lambda^2 \sum_{j=1}^{m}\frac{\overline{M}_j}{m^2}\left(1+\frac{\tau_p}{mC}\right)\mathbb{E}\left[\|(S(x^{k-d_k}))_j\|_j^2\right] - \sum_{t=k-d_{k,j_m}}^{k}\sum_{j=1}^{m}\sum_{i=1}^{n}p_{ij}T_{1,ij}\mathbb{E}\left[r_{ij}^d(y_i^{t-e_t^i})\right]\\
& - \sum_{t=k-d_{k,j_m}}^{k}\sum_{j=1}^{m}\sum_{i=1}^{n}p_{ij}T_{2,ij}\mathbb{E}\left[r_{ij}^p(x^{t-d_t})\right] - \sum_{j=1}^{m}\sum_{i=1}^{n}\sum_{\substack{h=0\\h\neq\tau_d-e_{k-d_{k,j_m},j}^i}}^{\tau_d}\frac{p_{ij}\gamma_{ij}}{\tau_d+1}\mathbb{E}\left[r_{ij}^d(y_i^{k-d_{k,j_m}-\tau_d+h})\right]\\
& + \frac{2\alpha\mu\lambda}{m}\mathbb{E}\left[\|x^{k-d_{k,j_m}} - P_{\mathcal{S}}(x^{k-d_{k,j_m}})\|^2 - \|x^{k-d_k} - P_{\mathcal{S}}(x^{k-d_k})\|^2\right]\\
& - \frac{2\alpha\lambda}{m}\mathbb{E}\left[\langle S(x^{k-d_k}), P_{\mathcal{S}}(x^{k-d_k}) - P_{\mathcal{S}}(x^{k-d_{k,j_m}})\rangle\right]. & (\text{A.8})
\end{aligned}$$

The final two lines of the above equations are preventing the expectation from contracting, so we absorb them into the other terms; if δ is 0, these terms are 0, as $x^{k-d_{k,j_m}} = x^{k-d_k}$, and there is no need for the next lemma. (We defer the proof of the lemma for a moment.)

Lemma A.5 (Swapping zeros in the inconsistent case) *The following bound holds:*

$$\begin{aligned}
& \frac{2\alpha\mu\lambda}{m}\mathbb{E}\left[\|x^{k-d_{k,j_m}} - P_{\mathcal{S}}(x^{k-d_{k,j_m}})\|^2 - \|x^{k-d_k} - P_{\mathcal{S}}(x^{k-d_k})\|^2\right] - \frac{2\alpha\lambda}{m}\mathbb{E}\left[\langle S(x^{k-d_k}), P_{\mathcal{S}}(x^{k-d_k}) - P_{\mathcal{S}}(x^{k-d_{k,j_m}})\rangle\right]\\
& \leq 4\alpha\mu\lambda^3 \sum_{t=k-d_{k,j_m}}^{k}\sum_{j=1}^{m}\sum_{i=1}^{n}\frac{\delta p_{ij}\overline{M}_j}{q_j m^3}\left(\frac{1}{\zeta\lambda}+1\right)\mathbb{E}\left[r_{ij}^p(x^{t-d_t})\right] + 4\alpha\mu\lambda^3 \sum_{t=k-d_{k,j_m}}^{k}\sum_{j=1}^{m}\sum_{i=1}^{n}\frac{\delta p_{ij}\overline{M}_j}{q_j m^3}\left(\frac{1}{\zeta\lambda}+1\right)\mathbb{E}\left[r_{ij}^d(y_i^{s-e_s^i})\right]\\
& + \lambda^2 \sum_{j=1}^{m}\frac{\overline{M}_j}{m^2}\left(1+\frac{\tau_p}{mC}\right)\mathbb{E}\left[\|(S(x^{k-d_k}))_j\|_j^2\right].
\end{aligned}$$

Lemma A.5 takes care of the offending terms in (A.8):

$$\begin{aligned}
(\text{A.8}) \leq\ & \mathbb{E}\left[\left(1 - \frac{2\mu\alpha\lambda}{m}\right)\|x^{k-d_{k,j_m}} - P_{\mathcal{S}}(x^{k-d_{k,j_m}})\|^2 + \kappa_{k-d_{k,j_m}}\right]\\
& - \sum_{t=k-d_{k,j_m}}^{k}\sum_{j=1}^{m}\sum_{i=1}^{n}p_{ij}\left(T_{1,ij} - \frac{4\alpha\mu\delta\overline{M}_j}{q_j m^3}\left(\frac{1}{\zeta}+\lambda\right)\lambda^2\right)\mathbb{E}\left[r_{ij}^d(y_i^{t-e_t^i})\right]\\
& - \sum_{t=k-d_{k,j_m}}^{k}\sum_{j=1}^{m}\sum_{i=1}^{n}p_{ij}\left(T_{2,ij} - \frac{4\alpha\mu\delta\overline{M}_j}{q_j m^3}\left(\frac{1}{\zeta}+\lambda\right)\lambda^2\right)\mathbb{E}\left[r_{ij}^p(x^{t-d_t})\right]\\
& - \sum_{j=1}^{m}\sum_{i=1}^{n}\sum_{\substack{h=0\\h\neq\tau_d-e_{k-d_{k,j_m},j}^i}}^{\tau_d}\frac{p_{ij}\gamma_{ij}}{\tau_d+1}\mathbb{E}\left[r_{ij}^d(y_i^{k-d_{k,j_m}-\tau_d+h})\right]. & (\text{A.9})
\end{aligned}$$

Our intermediate descent is completed simply by writing $\kappa_{k-d_k} = (1-\xi)\kappa_{k-d_k} + \xi\kappa_{k-d_k}$ and absorbing the $\xi\kappa_{1,k-d_k}$ terms in the second and fourth lines of the above inequality (we add $\xi\gamma_{ij}(\rho p_{ij}^T)^{-1}$ to all terms on the second line even though it only appears in the $t = k - d_{k,j_m}$ term in $\xi\kappa_{1,k-d_{k,j_m}}$):

$$
\begin{aligned}
\text{(A.9)} \leq{}& \mathbb{E}\left[(1-\xi)\left(\|x^{k-d_{k,j_m}} - P_{\mathcal{S}}(x^{k-d_{k,j_m}})\|^2 + \kappa_{k-d_{k,j_m}}\right) + \xi\kappa_{2,k-d_{k,j_m}}\right] \\
&- \sum_{t=k-d_{k,j_m}}^{k} \sum_{j=1}^{m} \sum_{i=1}^{n} p_{ij}\left(T_{1,ij} - \frac{\xi\gamma_{ij}}{\rho p_{ij}^T} - \frac{4\alpha\mu\delta\overline{M}_j}{q_j m^3}\left(\frac{1}{\zeta}+\lambda\right)\lambda^2\right)\mathbb{E}\left[r_{ij}^d(y_i^{t-e_t^i})\right] \\
&- \sum_{t=k-d_{k,j_m}}^{k} \sum_{j=1}^{m} \sum_{i=1}^{n} p_{ij}\left(T_{2,ij} - \frac{4\alpha\mu\delta\overline{M}_j}{q_j m^3}\left(\frac{1}{\zeta}+\lambda\right)\lambda^2\right)\mathbb{E}\left[r_{ij}^p(x^{t-d_t})\right] \\
&- \sum_{j=1}^{m} \sum_{i=1}^{n} \sum_{\substack{h=0 \\ h\neq\tau_d-e_{k-d_{k,j_m},j}^i}}^{\tau_d} \frac{p_{ij}\gamma_{ij}}{\tau_d+1}(1-\xi(h+1))\mathbb{E}\left[r_{ij}^d(y_i^{k-d_{k,j_m}-\tau_d+h})\right]. \tag{A.10}
\end{aligned}
$$

Final descent (from iteration $k - d_{k,j_m}$ to iteration $k - d_{k,j_m} - 1$). The expectation is nearly contracting by a factor of $(1-\xi)$. The term $\xi\kappa_{2,k-d_{k,j_m}}$ is all that stands in our way, but it can, by descending one more iteration, be absorbed. (We defer the proof of the lemma for a moment.)

Lemma A.6 (Recursive $\kappa_{2,k}$ bound) *For all $s \in \mathbb{N}$,*

$$
\mathbb{E}\left[\kappa_{2,s}\right] \leq \lambda^2 \sum_{j=1}^{m} \sum_{i=1}^{n} \frac{2\overline{M}_j\tau_p p_{ij}}{Cq_j m^3}\mathbb{E}\left[r_{ij}^p(x^{s-1-d_{s-1}})\right] + \lambda^2 \sum_{j=1}^{m} \sum_{i=1}^{n} \frac{2\overline{M}_j\tau_p p_{ij}}{Cq_j m^3}\mathbb{E}\left[r_{ij}^d(y_i^{s-1-e_{s-1}^i})\right] + \frac{\tau_p}{\tau_p+1}\mathbb{E}\left[\kappa_{2,s-1}\right]
$$

By Equation (A.5), the last three lines in (A.10) are nonnegative; drop these nonnegative terms, use the bound $\|x^{k-d_{k,j_m}} - P_{\mathcal{S}}(x^{k-d_{k,j_m}})\|^2 \leq \|x^{k-d_{k,j_m}} - P_{\mathcal{S}}(x^{k-d_{k,j_m}-1})\|^2$, descend one more step with the aid of (A.4), and use Lemma A.6 to bound $\mathbb{E}\left[\xi\kappa_{2,k-d_k-1}\right]$—in that order: if $h = k - d_{k,j_m} - 1$, then

$$
\begin{aligned}
\text{(A.10)} \leq{}& \mathbb{E}\left[(1-\xi)\left(\|x^h - P_{\mathcal{S}}(x^h)\|^2 + \kappa_h\right) + \frac{\tau_p}{\tau_p+1}\xi\kappa_{2,h}\right] \\
&- (1-\xi)\sum_{j=1}^{m}\sum_{i=1}^{n} p_{ij}T_{1,ij}\mathbb{E}\left[r_{ij}^d(y_i^{h-e_h^i})\right] + \xi\sum_{j=1}^{m}\sum_{i=1}^{n}\frac{2\overline{M}_j\tau_p p_{ij}\lambda^2}{Cq_j m^3}\mathbb{E}\left[r_{ij}^d(y_i^{h-e_h^i})\right] \\
&- (1-\xi)\sum_{j=1}^{m}\sum_{i=1}^{n} p_{ij}T_{2,ij}\mathbb{E}\left[r_{ij}^p(x^{h-d_h})\right] + \xi\sum_{j=1}^{m}\sum_{i=1}^{n}\frac{2\overline{M}_j\tau_p p_{ij}\lambda^2}{Cq_j m^3}\mathbb{E}\left[r_{ij}^p(x^{h-d_h})\right] \\
\leq{}& \mathbb{E}\left[\left(1-\frac{\xi}{\tau_p+1}\right)\left(\|x^h - P_{\mathcal{S}}(x^h)\|^2 + \kappa_h\right)\right] \\
&- (1-\xi)\sum_{j=1}^{m}\sum_{i=1}^{n} p_{ij}\left(T_{1,ij} - \frac{2\overline{M}_j\tau_p\lambda^2\xi}{Cq_j m^3(1-\xi)}\right)\mathbb{E}\left[r_{ij}^d(y_i^{h-e_h^i})\right] \\
&- (1-\xi)\sum_{j=1}^{m}\sum_{i=1}^{n} p_{ij}\left(T_{2,ij} - \frac{2\overline{M}_j\tau_p\lambda^2\xi}{Cq_j m^3(1-\xi)}\right)\mathbb{E}\left[r_{ij}^p(x^{h-d_h})\right].
\end{aligned}
$$

Equation (A.5) implies that the last two lines of the equation are negative, so

$$
\begin{aligned}
\mathbb{E}\left[d_{\mathcal{S}}(x^{k+1})^2 + \kappa_{k+1}\right] &\leq \left(1 - \frac{\xi}{\tau_p+1}\right)\mathbb{E}\left[d_{\mathcal{S}}^2(x^{k-d_{k,j_m}-1}) + \kappa_{k+1}\right] \\
&\leq \left(1 - \frac{\xi}{\tau_p+1}\right)\max_{h\in\{k-\tau_p-1,\ldots,k-1\}}\left\{\mathbb{E}\left[d_{\mathcal{S}}^2(x^h) + \kappa_h\right]\right\}.
\end{aligned}
$$

holds for all k. Set $C(z^0,\phi^0) := \kappa_0$ and unfold these recursive bounds with the next lemma to get the rate.

Lemma A.7 (Linear convergence rate of sequences that contract within a fixed number of steps) *[29, Lemma 6]]*
Let $\{a_k\}_{k\in\mathbb{N}}$ be a nonnegative sequence of real numbers. Suppose that there exists $\rho \in [0,1)$ and $\kappa \in \mathbb{N}$ such that for all $k \in \mathbb{N}$, we have

$$a_{k+1} \leq \rho \max_{k-\kappa \leq h \leq k} a_k.$$

Then for all $s \in \mathbb{N}$, we have

$$a_k \leq \rho^{k/(\kappa+1)} a_0.$$

We finish the proof of Part 3 by proving Lemmas A.4, A.5, and A.6.

Proof (of Lemma A.4 (choosing γ_{ij})) We have assumed that

$$\lambda \leq \min_{i,j} \left\{ \frac{2\eta(1-\alpha)n^2 p_{ij}\beta_{ij}}{\frac{2\overline{M}_j\eta(\tau_d+2)}{q_j m}\left(1+\frac{\delta\eta}{\tau_d+1}+\frac{5\sqrt{2(\tau_d+2)}\alpha^2\delta}{m\overline{M}\left(\sqrt{2(\tau_d+2)}+\tau_p\sqrt{\underline{q}}\right)}\right)+\frac{\overline{M}_j\eta\tau_p\sqrt{2(\tau_d+2)}}{m\sqrt{\underline{q}}}\left(2+\frac{\eta}{1-\eta}\right)+4\mu(\tau_d+1)\alpha(1-\alpha)n^2 p_{ij}\beta_{ij}} \right\},$$

and consequently[30], if

$$w_{ij} := \frac{\frac{2(1-\alpha)n^2 p_{ij}\beta_{ij}}{\frac{2\overline{M}_j(\tau_d+2)}{q_j m}\left(1+\frac{\delta\eta}{\tau_d+1}+\frac{2\mu\alpha\delta}{m\zeta}\right)+\overline{M}_j\tau_p C\left(2+\frac{\eta}{1-\eta}\right)}}{\frac{m\eta}{2\alpha\mu(\tau_d+1)}+\frac{2(1-\alpha)n^2 p_{ij}\beta_{ij}}{\frac{2\overline{M}_j(\tau_d+2)}{q_j m}\left(1+\frac{\delta\eta}{\tau_d+1}+\frac{2\mu\alpha\delta}{m\zeta}\right)+\overline{M}_j\tau_p C\left(2+\frac{\eta}{1-\eta}\right)}} \leq 1,$$

then

$$\lambda \leq \frac{w_{ij}m\eta}{2\alpha\mu(\tau_d+1)} = \frac{2(1-\alpha)n^2 p_{ij}\beta_{ij}(1-w_{ij})}{\frac{2\overline{M}_j(\tau_d+2)}{q_j m}\left(1+\frac{\delta\eta}{\tau_d+1}+\frac{2\mu\alpha\delta}{m\zeta}\right)+\overline{M}_j\tau_p C\left(2+\frac{\eta}{1-\eta}\right)},$$

which immediately implies the last claimed bound in the statement of this lemma: $\xi(\tau_d+1) \leq w_{ij}\eta \leq 1$.

The four inequalities that remain hold when the bound

$$\frac{(\tau_d+1)}{\left(1-\frac{(\tau_d+1)\xi}{\rho p_{ij}^T}\right)}\left(-T_{1,ij}+\frac{\gamma_{ij}}{\tau_d+1}+\frac{2\overline{M}_j\tau_p\xi\lambda^2}{Cq_j m^3(1-\xi)}+\frac{4\alpha\mu\delta\overline{M}_j\lambda^2}{q_j m^3}\left(\frac{1}{\zeta}+\lambda\right)\right)$$

$$= \frac{(\tau_d+1)\lambda^2}{\left(1-\frac{(\tau_d+1)\xi}{\rho p_{ij}^T}\right)}\left[\left(\frac{2\overline{M}_j}{q_j m^2}\left(1+\frac{\tau_p}{mC}\right)+\frac{2\overline{M}_j\tau_p\xi}{Cq_j m^3(1-\xi)}+\frac{4\alpha\mu\delta\overline{M}_j}{q_j m^3}\left(\frac{1}{\zeta}+\lambda\right)\right)\right]$$

$$\leq \frac{2(1-\alpha)n^2\lambda p_{ij}\beta_{ij}}{m}-\overline{M}_j\left(\frac{2}{q_j m^2}+\tau_p\left(\frac{2}{q_j m^3 C}+C\right)\right)\lambda^2-\frac{2\overline{M}_j\tau_p\lambda^2\xi}{Cq_j m^3(1-\xi)}-\frac{4\alpha\mu\delta\overline{M}_j}{q_j m^3}\left(\frac{1}{\zeta}+\lambda\right)$$

$$= T_{2,ij}-\frac{2\overline{M}_j\tau_p\lambda^2\xi}{Cq_j m^3(1-\xi)}-\frac{4\alpha\mu\delta\overline{M}_j}{q_j m^3}\left(\frac{1}{\zeta}+\lambda\right)+\gamma_{ij} \tag{A.11}$$

is satisfied because in this case γ_{ij}, which will necessarily be positive, can be taken as the upper or lower bound in (A.11), yielding

$$\frac{(\tau_d+1)}{\left(1-\frac{(\tau_d+1)\xi}{\rho p_{ij}^T}\right)}\left(-T_{1,ij}+\frac{\gamma_{ij}}{\tau_d+1}+\frac{2\overline{M}_j\tau_p\xi\lambda^2}{Cq_j m^3(1-\xi)}+\frac{4\alpha\mu\delta\overline{M}_j\lambda^2}{q_j m^3}\left(\frac{1}{\zeta}+\lambda\right)\right) \leq \gamma_{ij}; \text{ and}$$

$$T_{2,ij}-\frac{2\overline{M}_j\tau_p\lambda^2\xi}{Cq_j m^3(1-\xi)}-\frac{4\alpha\mu\delta\overline{M}_j}{q_j m^3}\left(\frac{1}{\zeta}+\lambda\right)+\gamma_{ij} \geq \gamma_{ij}.$$

---

[30] Use the identity $2\mu\alpha\delta(\zeta m)^{-1} = 5\alpha^2\delta\left(m\overline{M}\left(1+\frac{\tau\sqrt{\underline{q}}}{\sqrt{2(\tau_d+2)}}\right)\right)^{-1} = \frac{5\sqrt{2(\tau_d+2)}\alpha^2\delta}{m\overline{M}\left(\sqrt{2(\tau_d+2)}+\tau\sqrt{\underline{q}}\right)}$.

After rearranging, these conservative inequalities imply all four bounds in (A.5).

Thus, we finish the proof with (A.11): Because $\xi \le w_{ij}\eta < 1$, $\xi/(1-\xi) \le w_{ij}\eta/(1 - w_{ij}\eta) \le \eta/(1-\eta)$, and $2(\tau_d + 2)(m^2 \underline{q} C)^{-1} = C$, we have

$$
\begin{aligned}
&\lambda^2 \left[(\tau_d + 1) \left(\frac{2\overline{M}_j}{q_j m^2} \left(1 + \frac{\tau_p}{mC} \right) + \frac{2\overline{M}_j \tau_p \xi}{Cq_j m^3 (1-\xi)} + \frac{4\alpha\mu\delta\overline{M}_j}{q_j m^3}\left(\frac{1}{\zeta} + \lambda\right) \right) \right. \\
&\qquad \left. + (1 - w_{ij}) \left(\overline{M}_j \left(\frac{2}{q_j m^2} + \tau_p \left(\frac{2}{q_j m^3 C} + \frac{C}{m} \right) \right) + \frac{2\overline{M}_j \tau_p \lambda^2 \xi}{Cq_j m^3(1-\xi)} + \frac{4\alpha\mu\delta\overline{M}_j}{q_j m^3}\left(\frac{1}{\zeta} + \lambda\right) \right) \right] \\
&\le \lambda^2 \left[(\tau_d + 1) \left(\frac{2\overline{M}_j}{\underline{q} m^2} \left(1 + \frac{\tau_p}{mC} \right) + \frac{2\overline{M}_j \tau_p \xi}{C\underline{q} m^3 (1-\xi)} + \frac{4\alpha\mu\delta\overline{M}_j}{\underline{q} m^3}\left(\frac{1}{\zeta} + \lambda\right) \right) \right. \\
&\qquad \left. + (1 - w_{ij}) \left(\overline{M}_j \left(\frac{2}{\underline{q} m^2} + \tau_p \left(\frac{2}{\underline{q} m^3 C} + \frac{C}{m} \right) \right) + \frac{2\overline{M}_j \tau_p \lambda^2 \xi}{C\underline{q} m^3(1-\xi)} + \frac{4\alpha\mu\delta\overline{M}_j}{\underline{q} m^3}\left(\frac{1}{\zeta} + \lambda\right) \right) \right] \\
&\le \overline{M}_j \lambda^2 \left[(\tau_d + 2) \left(\frac{2}{\underline{q} m^2} \left(1 + \frac{\tau_p}{mC} \right) + \frac{2\tau_p \xi}{C\underline{q} m^3(1-\xi)} + \frac{4\alpha\mu\delta}{\underline{q} m^3}\left(\frac{1}{\zeta} + \lambda\right) \right) \right] \\
&\le \overline{M}_j \lambda^2 \left[(\tau_d + 2) \left(\frac{2}{\underline{q} m^2} + \frac{4\alpha\mu\delta}{\underline{q} m^3}\left(\frac{1}{\zeta} + \lambda\right) \right) + \frac{\tau_p}{m} C \left(2 + \frac{\eta}{1-\eta} \right) \right] \\
&\overset{(A.3)}{\le} \overline{M}_j \lambda^2 \left[\frac{2(\tau_d + 2)}{\underline{q} m^2} \left(1 + \frac{\delta\eta}{\tau_d + 1} + \frac{2\mu\alpha\delta}{m\zeta} \right) + \frac{\tau_p}{m} C \left(2 + \frac{\eta}{1-\eta} \right) \right] \\
&\le \frac{2(1-\alpha)n^2 \lambda p_{ij} \beta_{ij} (1 - w_{ij})}{m}.
\end{aligned}
$$

(We bound $(q_j)^{-1} \le \overline{q}^{-1}$ in the first inequality, and we bound $1 - w_{ij} \le 1$ in the second.) Divide both sides of this inequality by $1 - w_{ij}$, rearrange, and use the bound $\left(1 - \frac{(\tau_d + 1)\xi}{\rho p_{ij}^T} \right)^{-1} \le \left(1 - \frac{w_{ij}\eta}{\rho p_{ij}^T} \right)^{-1} \le (1 - w_{ij})^{-1}$ to get (A.11). □

Proof (of Lemma A.5 (swapping zeros in the inconsistent case)) We split the proof according to the bounds we apply:

1. **Firm nonexpansiveness.** The map $P_\mathcal{S}$ is firmly nonexpansive in $\|\cdot\|$, so for all $x, y \in \mathcal{H}$, $\|P_\mathcal{S}(x) - P_\mathcal{S}(y)\|^2 + \|x - P_\mathcal{S}(x) - (y - P_\mathcal{S}(y))\|^2 \le \|x - y\|^2$. Therefore,

$$
\left\| \left(x^{k-d_{k,jm}} - P_\mathcal{S}(x^{k-d_{k,jm}}) \right) - \left(x^{k-d_k} - P_\mathcal{S}(x^{k-d_k}) \right) \right\|^2 \le \|x^{k-d_{k,jm}} - x^{k-d_k}\|^2 - \|P_\mathcal{S}(x^{k-d_{k,jm}}) - P_\mathcal{S}(x^{k-d_k})\|^2.
\tag{A.12}
$$

2. **Young's inequality.** Apply the bound $\|a + b\|^2 \le (1+\varepsilon)\|a\|^2 + (1 + \varepsilon^{-1})\|b\|^2$ to get

$$
\begin{aligned}
&\|x^{k-d_{k,jm}} - P_\mathcal{S}(x^{k-d_{k,jm}})\|^2 - \|x^{k-d_k} - P_\mathcal{S}(x^{k-d_k})\|^2 \\
&\le \left(\frac{1}{\zeta\lambda} + 1 \right) \left\| \left(x^{k-d_{k,jm}} - P_\mathcal{S}(x^{k-d_{k,jm}}) \right) - \left(x^{k-d_k} - P_\mathcal{S}(x^{k-d_k}) \right) \right\|^2 + \zeta\lambda \|x^{k-d_k} - P_\mathcal{S}(x^{k-d_k})\|^2 \\
&\overset{(A.12)}{\le} \left(\frac{1}{\zeta\lambda} + 1 \right) \left[\|x^{k-d_{k,jm}} - x^{k-d_k}\|^2 - \|P_\mathcal{S}(x^{k-d_{k,jm}}) - P_\mathcal{S}(x^{k-d_k})\|^2 \right] + \zeta\lambda \|x^{k-d_k} - P_\mathcal{S}(x^{k-d_k})\|^2.
\end{aligned}
$$

3. **Equivalence of norms, convexity of $\|\cdot\|^2$, and the variance bound.**

$$\mathbb{E}\left[\|x^{k-d_{k},jm} - x^{k-d_{k}}\|^2\right]$$

$$\leq \sum_{j=1}^{m} \mathbb{E}\left[\overline{M}_j \|x_j^{k-d_{k},jm} - x_j^{k-d_{k}}\|_j^2\right]$$

$$\leq \sum_{j=1}^{m} \mathbb{E}\left[\overline{M}_j \delta \sum_{t=k-d_{k,jm}+1}^{k-d_{k,j}} \|x_j^t - x_j^{t-1}\|_j^2\right]$$

$$\leq \sum_{j=1}^{m} \mathbb{E}\left[\overline{M}_j \delta \sum_{t=k-d_{k,jm}+1}^{k+1} \|x_j^t - x_j^{t-1}\|_j^2\right]$$

$$\overset{\text{(A.3)}}{\leq} 2\lambda^2 \sum_{t=k-d_{k,jm}}^{k} \sum_{j=1}^{m}\sum_{i=1}^{n} \frac{\delta p_{ij}\overline{M}_j}{q_j m^2}\mathbb{E}\left[r_{ij}^p(x^{t-d_t})\right] + 2\lambda^2 \sum_{t=k-d_{k,jm}}^{k}\sum_{j=1}^{m}\sum_{i=1}^{n}\frac{\delta p_{ij}\overline{M}_j}{q_j m^2}\left[r_{ij}^d(y_i^{s-e_s^i})\right].$$

4. **Essential strong quasi-monotonicity.** The next two bounds hold

$$\langle S(x^{k-d_k}), x^{k-d_k} - P_{\mathcal{S}}(x^{k-d_k})\rangle \geq \mu\|x^{k-d_k} - P_{\mathcal{S}}(x^{k-d_k})\|^2;$$

$$\frac{1}{\mu^2}\|S(x^{k-d_k})\|^2 \geq \|x^{k-d_k} - P_{\mathcal{S}}(x^{k-d_k})\|^2.$$

(After an application of the Cauchy-Schwarz inequality, the second bound follows from the first.) Thus,

$$\frac{2\alpha\mu\zeta\lambda^2}{m}\|x^{k-d_k} - P_{\mathcal{S}}(x^{k-d_k})\|^2 \leq \frac{2\alpha\zeta\lambda^2}{m\mu}\|S(x^{k-d_k})\|^2.$$

5. **Cauchy-Schwarz and Young's Inequality.**

$$\frac{2\alpha\lambda}{m}|\langle S(x^{k-d_k}), P_{\mathcal{S}}(x^{k-d_k}) - P_{\mathcal{S}}(x^{k-d_k,jm})\rangle| \leq \frac{\alpha\zeta\lambda^2}{2m\mu}\|S(x^{k-d_k})\|^2 + \frac{2\alpha\mu}{m\zeta}\|P_{\mathcal{S}}(x^{k-d_k}) - P_{\mathcal{S}}(x^{k-d_k,jm})\|^2.$$

6. **The ζ bound.**

$$\frac{2\alpha\mu\zeta\lambda^2}{m}\mathbb{E}\left[\|x^{k-d_k} - P_{\mathcal{S}}(x^{k-d_k})\|^2\right] - \frac{2\alpha\mu\lambda}{m}\left(\frac{1}{\zeta\lambda}+1\right)\mathbb{E}\left[\|P_{\mathcal{S}}(x^{k-d_k,jm}) - P_{\mathcal{S}}(x^{k-d_k})\|^2\right]$$

$$\quad - \frac{2\alpha\lambda}{m}\mathbb{E}\left[\langle S(x^{k-d_k}), P_{\mathcal{S}}(x^{k-d_k}) - P_{\mathcal{S}}(x^{k-d_k,jm})\rangle\right]$$

$$\leq \frac{2.5\alpha\zeta\lambda^2}{m\mu}\mathbb{E}\left[\|S(x^{k-d_k})\|^2\right] - \left(\frac{2\alpha\mu}{m\zeta} + \frac{2\alpha\mu\lambda}{m} - \frac{2\alpha\mu}{m\zeta}\right)\mathbb{E}\left[\|P_{\mathcal{S}}(x^{k-d_k,jm}) - P_{\mathcal{S}}(x^{k-d_k})\|^2\right]$$

$$\leq \lambda^2 \sum_{j=1}^{m}\frac{\overline{M}_j}{m^2}\left(1+\frac{\tau_p}{mC}\right)\mathbb{E}\left[\|(S(x^{k-d_k}))_j\|_j^2\right].$$

(The last line follows from the third because (i) $2.5\alpha\zeta(m\mu)^{-1} \leq \overline{M}_j\left(1+\tau(mC)^{-1}\right)$ and (ii) the second term on line three is negative.)

We conclude by putting all the bounds together:

$$
\frac{2\alpha\mu\lambda}{m}\mathbb{E}\left[\|x^{k-d_{k,jm}} - P_{\mathcal{S}}(x^{k-d_{k,jm}})\|^2 - \|x^{k-d_k} - P_{\mathcal{S}}(x^{k-d_k})\|^2\right] - \frac{2\alpha\lambda}{m}\mathbb{E}\left[\langle S(x^{k-d_k}), P_{\mathcal{S}}(x^{k-d_k}) - P_{\mathcal{S}}(x^{k-d_{k,jm}})\rangle\right]
$$

$$
\leq \frac{2\alpha\mu\lambda}{m}\left(\frac{1}{\zeta\lambda}+1\right)\mathbb{E}\left[\|x^{k-d_{k,jm}} - x^{k-d_k}\|^2\right] - \frac{2\alpha\lambda}{m}\mathbb{E}\left[\langle S(x^{k-d_k}), P_{\mathcal{S}}(x^{k-d_k}) - P_{\mathcal{S}}(x^{k-d_{k,jm}})\rangle\right]
$$

$$
+ \frac{2\alpha\mu\zeta\lambda^2}{m}\mathbb{E}\left[\|x^{k-d_k} - P_{\mathcal{S}}(x^{k-d_k})\|^2\right] - \frac{2\alpha\mu\lambda}{m}\left(\frac{1}{\zeta\lambda}+1\right)\mathbb{E}\left[\|P_{\mathcal{S}}(x^{k-d_{k,jm}}) - P_{\mathcal{S}}(x^{k-d_k})\|^2\right]
$$

$$
\leq 4\alpha\mu\lambda^3 \sum_{t=k-d_{k,jm}}^{k}\sum_{j=1}^{m}\sum_{i=1}^{n}\frac{\delta p_{ij}\overline{M}_j}{q_j m^3}\left(\frac{1}{\zeta\lambda}+1\right)\mathbb{E}\left[r_{ij}^p(x^{t-d_t})\right] + 4\alpha\mu\lambda^3 \sum_{t=k-d_{k,jm}}^{k}\sum_{j=1}^{m}\sum_{i=1}^{n}\frac{\delta p_{ij}\overline{M}_j}{q_j m^3}\left(\frac{1}{\zeta\lambda}+1\right)\mathbb{E}\left[r_{ij}^d(y_i^{s-e_s^i})\right]
$$

$$
+ \lambda^2 \sum_{j=1}^{m}\frac{\overline{M}_j}{m^2}\left(1+\frac{\tau_p}{mC}\right)\mathbb{E}\left[\|(S(x^{k-d_k}))_j\|_j^2\right]. \qquad \square
$$

Proof (of Lemma A.6 (recursive $\kappa_{2,k}$ bound)) For any $s \in \mathbb{N}$, we have

$$
\mathbb{E}\left[\kappa_{2,s}\right] = \sum_{t=s-\tau_p+1}^{s}\sum_{j=1}^{m}\frac{\overline{M}_j(t-s+\tau_p)}{mC}\mathbb{E}\left[\|x_j^t - x_j^{t-1}\|_j^2\right]
$$

$$
= \sum_{j=1}^{m}\frac{\overline{M}_j\tau_p}{mC}\mathbb{E}\left[\|x_j^s - x_j^{s-1}\|_j^2\right] + \sum_{t=s-\tau_p+1}^{s-1}\sum_{j=1}^{m}\frac{\overline{M}_j(t-s+\tau_p)}{mC}\mathbb{E}\left[\|x_j^t - x_j^{t-1}\|_j^2\right]
$$

$$
\leq \lambda^2 \sum_{j=1}^{m}\frac{\overline{M}_j\tau_p}{mC}\mathbb{E}\left[\|\mathfrak{Q}_j^{s-1}\|_j^2\right] + \frac{\tau_p}{\tau_p+1}\mathbb{E}\left[\kappa_{2,s-1}\right]
$$

$$
\overset{(A.3)}{\leq} \lambda^2 \sum_{j=1}^{m}\sum_{i=1}^{n}\frac{2\overline{M}_j\tau_p p_{ij}}{Cq_j m^3}\mathbb{E}\left[r_{ij}^p(x^{s-1-d_{s-1}})\right] + \lambda^2 \sum_{j=1}^{m}\sum_{i=1}^{n}\frac{2\overline{M}_j\tau_p p_{ij}}{Cq_j m^3}\mathbb{E}\left[r_{ij}^d(y_i^{s-1-e_{s-1}^i})\right] + \frac{\tau_p}{\tau_p+1}\mathbb{E}\left[\kappa_{2,s-1}\right]. \square
$$

<div align="right">$\square$</div>

B Auxiliary Results

B.1 Properties of Operators

Definition B.1 (Cocoercive Operators) An operator $S : \mathcal{H} \to \mathcal{H}$ is called β-*cocoercive* if

$$(\forall x, y \in \mathcal{H}) \qquad \langle Sx - Sy, x - y \rangle \geq \beta \|Sx - Sy\|^2.$$

Definition B.2 (Averaged Operators) Let $\alpha \in [0, 1]$. An operator $T : \mathcal{H} \to \mathcal{H}$ is called α-*averaged* if there is a nonexpansive map $N : \mathcal{H} \to \mathcal{H}$ such that

$$T = (1 - \alpha)I_{\mathcal{H}} + \alpha N.$$

The map T is is called *firmly nonexpansive* if it is $(1/2)$-averaged.

Proposition B.1 (Cocoercivness from Averagedness) *Let* $\alpha \in [0, 1]$, *and let* $T : \mathcal{H} \to \mathcal{H}$ *be an* α-*averaged operator. Then for all* $\beta > 0$, $\beta(I - T)$ *is* $(1/(2\alpha\beta))$-*cocoercive.*

Proof There is a nonexpansive map $N : \mathcal{H} \to \mathcal{H}$ such that $T = (1 - \alpha)I + \alpha N$, and by definition, $T' := (1/2)I + (1/2)N$ is firmly nonexpansive. Thus, $I - T' = (1/2)(I - N)$ is firmly nonexpansive, and hence, 1-cocoercive [2, Remark 4.24(iii)]. Therefore, for all $x, y \in \mathcal{H}$, the next bound holds and proves that $\beta(I - T)$ is $(1/\alpha\beta)$-cocoercive:

$$\langle \alpha\beta(I - N)x - \alpha\beta(I - N)y, x - y \rangle = 2\alpha\beta\langle \frac{1}{2}(I - N)x - \frac{1}{2}(I - N)y, x - y \rangle \geq 2\alpha\beta \left\| \frac{1}{2}(I - N)x - \frac{1}{2}(I - N)y \right\|^2$$

$$= \frac{1}{2\alpha\beta} \|\alpha\beta(I - N)x - \alpha\beta(I - N)y\|^2. \qquad \square$$

Proposition B.2 (The Composition of Averaged Operators is Averaged [10]) *Let* $\alpha_1, \alpha_2 \in [0, 1]$, *and let* $T_1, T_2 : \mathcal{H} \to \mathcal{H}$ *be* α_1- *and* α_2-*averaged maps, respectively. Then* $T_1 \circ T_2$ *is* $\alpha_{1,2}$ *averaged where*

$$\alpha_{1,2} := \frac{\alpha_1 + \alpha_2 - 2\alpha_1\alpha_2}{1 - \alpha_1\alpha_2}.$$

Proposition B.3 (Strongly monotone operators from Lipschitz operators) *Let* $L \in (0, 1)$ *and suppose that* $T : \mathcal{H} \to \mathcal{H}$ *be* L-*Lipschitz continuous. Then* $I - T$ *is* $(1 - L)$-*strongly monotone.*

Proposition B.4 (Coordinate Lipschitz Constants and Coordinate Cocoercivity.) *Let* $f : \mathcal{H} \to (-\infty, \infty]$ *be a Fréchet differentiable convex function. If* $j \in \{1, \ldots, m\}$ *and*

$$(\forall x \in \mathcal{H}), (\forall y_j \in \mathcal{H}_j) \qquad f(x + \hat{y_j}) \leq f(x) + \langle \nabla f(x), \hat{y_j} \rangle + \frac{L_j}{2}\|\hat{y_j}\|^2; \qquad \hat{y_j} = (0, \ldots, 0, y_j, 0, \ldots, 0), \tag{B.1}$$

then

$$(\forall x \in \mathcal{H}), (\forall y \in \mathcal{H}) \qquad \frac{1}{L_j}\|\nabla f(x) - \nabla f(y)\|_j^2 \leq \langle \nabla f(x) - \nabla f(y), x - y \rangle.$$

Proof Fix $y \in \mathcal{H}$, and let define a function g, which continues to satisfy (B.1),

$$g(x) = f(x) - f(y) - \langle \nabla f(y), x - y \rangle.$$

Then $g \geq 0$, and $g(y) = 0$, so g is minimized at y. In addition,

$$(\forall x \in \mathcal{H}), (\forall y_j \in \mathcal{H}_j) \qquad g(y) \leq g(x + \hat{y_j}) \leq g(x) + \langle \nabla f(x), \hat{y_j} \rangle + \frac{L_j}{2}\|\hat{y_j}\|^2; \qquad \hat{y_j} = (0, \ldots, 0, y_j, 0, \ldots, 0).$$

Thus, we minimize the right hand side of this inequality, over all $\hat{y_j} \in \mathcal{H}$, and get the next bound:

$$f(x) - f(y) - \langle \nabla f(y), x - y \rangle = g(x) - g(y) \geq \frac{1}{2L_j}\|\nabla_j f(x) - \nabla_j f(y)\|_j^2.$$

We likewise get the opposite inequality in which the points x and y are exchanged; add both inequalities, and get the result.

Proposition B.5 (Contractive Forward-Gradient Operator) *Suppose that the function $f : \mathcal{H} \to (-\infty, \infty)$ is μ_f-strongly convex, differentiable, and that the gradient ∇f is L-Lipschitz continuous. Then $I - \gamma \nabla f$ is*

$$\sqrt{1 - 2\gamma\mu_f + \gamma^2 L\mu_f}$$

Lipschitz continuous whenever $\gamma \leq 2L^{-1}$.

Proof In the next sequence of inequalities, we use the bound $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq L^{-1}\|\nabla f(x) - \nabla f(y)\|^2$ once:

$$(\forall x, y \in \mathcal{H}) \qquad \|(I - \gamma\nabla f)x - (I - \gamma\nabla f)y\|^2 \leq \|x - y\|^2 - 2\gamma\langle \nabla f(x) - \nabla f(y), x - y \rangle + \gamma^2\|\nabla f(x) - \nabla f(y)\|^2$$
$$\leq \|x - y\|^2 + \gamma^2\|\nabla f(x) - \nabla f(y)\|^2$$
$$- 2\gamma\left(1 - \frac{\gamma L}{2}\right)\langle \nabla f(x) - \nabla f(y), x - y \rangle - \gamma^2 L\langle \nabla f(x) - \nabla f(y), x - y \rangle$$
$$\leq \left(1 - 2\gamma\mu_f\left(1 - \frac{\gamma L}{2}\right)\right)\|x - y\|^2. \qquad \square$$

C Examples of S

C.1 Old Operators

Proposition C.1 (SAGA/SVRG/S2GD Operator Properties) *Assume the setting of Section 4.1, and in particular, that*

$$S_i = \nabla f_i$$

Then

1. **Coherence:** *S satisfies the coherence condition*

$$(\forall x \in \mathcal{H}), (\forall x^* \in \mathrm{zer}(S)) \qquad \langle S(x), x - x^* \rangle \geq \sum_{i=1}^{N} \beta_{i1} \|S_i(x) - S_i(x^*)\|^2.$$

   with $\beta_{i1} \equiv nL^{-1}$.
2. **Essential strong quasi-monotonicity:** *S is μ-essentially strongly monotone whenever $N^{-1} \sum_{i=1}^{N} f_i$ is μ-strongly convex.*
3. **Roots:** *$\mathrm{zer}(S)$ is precisely the set of minimizers of the dual problem to (4.6).*
4. **Demiclosedness:** *$I - S$ is demiclosed at 0.*

Proof Part 1 (coherence): The Baillon-Haddad Theorem [1] guarantees that for all $i \in \{1, \dots, N\}$ and $x \in \mathcal{H}$,

$$\frac{1}{L} \|\nabla f_i(x) - \nabla f_i(x^*)\|^2 \leq \langle \nabla f_i(x) - \nabla f_i(x^*), x - x^* \rangle,$$

which implies that

$$\frac{1}{N} \sum_{i=1}^{N} \frac{1}{L} \|S_i(x) - S_i(x^*)\|^2 = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{L} \|\nabla f_i(x) - \nabla f_i(x^*)\|^2 \leq \frac{1}{N} \sum_{i=1}^{N} \langle \nabla f_i(x) - \nabla f_i(x^*), x_i - x^* \rangle = \langle S(x), x - x^* \rangle.$$

Parts 2 (essential strong quasi-monotonicity) and 3 (roots) are simple, so we omit the proofs.

Part 4 (demiclosedness): The operator $I - S$ is nonexpansive by [2, Proposition 4.33]; thus it is demiclosed at 0. $\qquad\square$

Proposition C.2 (Finito Operator Properties) *Assume the setting of Section 4.2, and in particular, that*

$$\left(\forall x \in \mathcal{H}_0^N \right) \qquad S(x) = x - P_D \left(x_1 - \gamma \nabla f_1(x_1), \dots, x_N - \gamma \nabla f_N(x_N) \right). \tag{C.1}$$

In addition, let $\gamma \leq 2L^{-1}$. Then

1. **Coherence:** *S satisfies the coherence condition*

$$(\forall x \in \mathcal{H}), (\forall x^* \in \mathrm{zer}(S)) \qquad \langle S(x), x - x^* \rangle \geq \sum_{j=1}^{N} \beta_{1j} \|(S(x))_j\|_j^2.$$

   with $\beta_{1j} \equiv 4^{-1} \gamma L$.
2. **Essential strong quasi-monotonicity:** *S is*

$$\mu := 1 - \sqrt{1 - 2\gamma\hat{\mu} + \gamma^2 \hat{\mu} L}$$

   essentially strong quasi-monotone whenever each f_i is $\hat{\mu}$-strongly convex.
3. **Roots:** *$\mathrm{zer}(S) = \{(x_0^*, \dots, x_0^*) \in \mathcal{H}_0^N \mid \text{minimizes } (4.1)\}$*
4. **Demiclosedness:** *$I - S$ is demiclosed at 0.*

Proof Part 1 (coherence): Let $\mathbf{f} : \mathcal{H} \to \mathcal{H}$ be the function $\mathbf{f}(x_1, \ldots, x_N) = \sum_{j=1}^{N} f_j(x_j)$. The operator $\nabla \mathbf{f} : x \mapsto (\nabla f_1(x_1), \ldots, \nabla f_N(x_N))$ is L^{-1}-cocoercive [1] and L-Lipschitz, and thus, $I_{\mathcal{H}} - \gamma \nabla \mathbf{f}$ is $(2^{-1}\gamma L)$-averaged. In addition, P_D is 2^{-1}-averaged [2, Proposition 4.8], so by Proposition B.2, the composition $P_D \circ (I_{\mathcal{H}} - \gamma \nabla \mathbf{f})$ is $2/(4 - \gamma L)$-averaged. Therefore, $S = (I_{\mathcal{H}} - P_D \circ (I_{\mathcal{H}} - \gamma \nabla \mathbf{f}))$ is $(4^{-1}\gamma L)$-cocoercive, and so $\beta_{1j} = 4^{-1}\gamma L$.

Part 2 (essential strong quasi-monotonicity): by B.5, if $\gamma \leq 2L^{-1}$, then the map $I_{\mathcal{H}} - \gamma \nabla \mathbf{f}$ is $(1 - \mu)$-Lipschitz, so the composition $P_D(I_{\mathcal{H}} - \gamma \nabla \mathbf{f})$ is $(1 - \mu)$-Lipschitz. Therefore, by Proposition B.3, the operator S is μ-strongly quasi-monotone.

Part 3 (roots) is simple so we omit the proof.

Part 4 (demiclosedness): By [2, Proposition 4.33 and Proposition 4.8], the operator $I - S$ is nonexpansive because it is the composition of two nonexpansive maps. Thus, $I - S$ is demiclosed at 0. □

Proposition C.3 (SDCA Operator Properties) *Assume the setting of section 4.3, and in particular, that*

$$S = I - \mathbf{prox}_{\mu_0 N f^*(-\cdot)} \circ (I - \mu_0 N \nabla g).$$

1. **Coherence:** S *satisfies the coherence condition*

$$(\forall x \in \mathcal{H}), (\forall x^* \in \mathrm{zer}(S)) \qquad \langle S(x), x - x^* \rangle \geq \sum_{j=1}^{N} \beta_{1j} \| (S(x))_j \|_j^2.$$

with $\beta_{1j} \equiv 3/4$

2. **Essential strong quasi-monotonicity:** S *is*

$$\mu = \frac{\mu_0 N}{(\mu_0 N + L)}$$

essentially strongly quasi-monotone—whether or not any f_j is strongly convex.

3. **Roots:** $\mathrm{zer}(S)$ *is precisely the set of dual solutions to* (4.7).

4. **Demiclosedness:** $I - S$ *is demiclosed at* 0.

Proof Part 1 (coherence): The gradient ∇g is $(\mu_0 N)^{-1}$-Lipschitz, so $I_{\mathcal{H}} - \mu_0 N \nabla g$ is $(1/2)$-averaged [2, Proposition 4.33]. Thus, by Proposition B.2, the composition of the operators $\mathbf{prox}_{\mu_0 N f^*(-\cdot)} \circ (I_{\mathcal{H}} - \mu_0 N \nabla g)$, both of which are 2^{-1}-averaged, is $(2/3)$-averaged. Thus, by Lemma B.1, $S = I_{\mathcal{H}} - \mathbf{prox}_{\mu_0 N f^*(-\cdot)} \circ (I_{\mathcal{H}} - \mu_0 N \nabla g)$ is $(3/4)$-cocoercive.

Part 2 (essential strong quasi-monotonicity): All the strong monotonicity comes from the $\mathbf{prox}_{\mu_0 N f^*(-\cdot)}$ operator: The gradient of $f(x_1, \ldots, x_N) = \sum_{i=1}^{N} f_i(x_i)$ is L-Lipschitz continuous on $\mathcal{H}$, so the conjugate function f^* is L^{-1}-strongly convex [2, Theorem 18.15], and this, in turn, implies that $\mathbf{prox}_{\mu_0 N f^*(-\cdot)}$ is $L(\mu_0 N + L)^{-1}$-Lipschitz continuous. Thus, $\mathbf{prox}_{\mu_0 N f^*(-\cdot)} \circ (I_{\mathcal{H}} - \mu_0 N \nabla g)$ is $L(\mu_0 N + L)^{-1}$-Lipschitz continuous [2, Proposition 23.11], so by Proposition B.3, S is $\mu_0 N(\mu_0 N + L)^{-1}$-essentially strongly quasi-monotone (indeed, strongly monotone).

Part 3 (roots) follows from [2, Proposition 25.1(iv)].

Part 4 (demiclosedness): In part 1, we showed that $I - S$ is averaged, and hence nonexpansive. Thus, $I - S$ is demiclosed at 0. □

Proposition C.4 (Randomized Projection Operator Properties) *Assume the setting of Section 4.4, and in particular, that*

$$S_i = \begin{cases} I - P_{C_i} & if \ i = 1, \ldots, s_1; \\ I - G_{f_{i-s_1}} & if \ i = s_1 + 1, \ldots, s_1 + s_2. \end{cases}$$

Then

1. **Coherence:** S satisfies the coherence condition

$$(\forall x \in \mathcal{H}), (\forall x^* \in \mathrm{zer}(S)) \qquad \langle S(x), x - x^* \rangle \geq \sum_{i=1}^{N} \beta_{i1} \|S_i(x) - S_i(x^*)\|^2.$$

with $\beta_{i1} \equiv 1$.
2. **Essential strong quasi-monotonicity:** S is

$$\mu = \frac{\max\{1, \varepsilon^2 L^{-2}\}}{N\hat{\mu}^2}$$

essentially strongly quasi-monotone whenever
(a) $\{C_i \mid i = 1, \ldots, s_1\} \cup \{\{x \mid f_i(x) \leq 0\} \mid i = 1, \ldots, s_2\}$ are $\hat{\mu}$-linearly regular,[31]
(b) there is an $\varepsilon > 0$ such that $f_i(x) \geq \varepsilon d_{\{f_i(x) \leq 0\}}(x)$ for all $x \in \mathcal{H}$,
(c) and there is an $L > 0$ such that $\|g_i(x)\| \leq L$ for all $x \in \mathcal{H}$.
3. **Roots:** $\mathrm{zer}(S)$ is precisely the set of solutions to the feasibility problem (4.9).
4. **Demiclosedness:** $I - S$ is demiclosed at 0.

Proof Parts 1 (coherence) and 3 (roots): Each projection operator P_{C_i} is 2^{-1}-averaged by [2, Proposition 4.8], and so $I - P_{C_i}$ is 1-cocoercive by Proposition B.1. Similarly, by [3, Fact 2.1(v)], each subgradient projector 2^{-1} quasi-averaged, i.e.,

$$(\forall x \in \mathcal{H}), (\forall x^* \in \{z \mid f(z) \leq 0\}) \qquad \|G_{f_i}(x) - x^*\|^2 \leq \|x - x^*\|^2 - \|x - G_{f_i}(x)\|^2.$$

Thus, by rearranging this inequality, we have

$$(\forall x \in \mathcal{H}), (\forall x^* \in \{z \mid f(z) \leq 0\}) \qquad \langle x - G_{f_i}(x), x - x^* \rangle \geq \|x - G_{f_i}(x)\|^2.$$

Altogether, with C equal to the set of points in the intersection (4.9), we have

$$(\forall x \in \mathcal{H}), (\forall x^* \in C) \qquad \langle S(x), x - x^* \rangle \geq \frac{1}{N}\sum_{i=1}^{N}\|S_i(x)\|^2.$$

To complete Parts 1 and 3, we need only show that $C = \mathrm{zer}(S)$. But the operator $(I-S) = \frac{1}{N}\sum_{i=1}^{N}(I - S_i)$ is the average of quasi-nonexpansive operators, and all of the fixed-point sets of these operators overlap (because $\mathrm{Fix}(G_{f_i}) = \{x \mid f_i(x) \leq 0\}$ [3, Fact 2.1(ii)] and there is a point in the intersection (4.9)). Thus, by [2, Proposition 4.34], the set of fixed points of $I - S$ is equal to the set of all points in the intersection; in other words, $C = \mathrm{zer}(S)$.

Part 2 (essential strong quasi-monotonicity): Let $x^* \in \mathrm{zer}(S)$. Then for all $x \in \mathcal{H}$, with $f(x) > 0$, we have

$$\langle x - G_{f_i}(x), x - x^* \rangle = \frac{f_i(x)}{\|g_i(x)\|^2}\langle g_i(x), x - x^* \rangle \geq \frac{f_i(x)}{\|g_i(x)\|^2}(f_i(x) - f_i(x^*)) \geq \frac{f_i(x)^2}{\|g_i(x)\|^2} \geq \frac{\epsilon^2}{L^2}d^2_{\{f_i(z) \leq 0\}}(x).$$

Similarly, because $I - P_{C_i}$ is 1-cocoercive, for all $x \in \mathcal{H}$, we have

$$\langle x - P_{C_i}(x), x - x^* \rangle \geq \|x - P_{C_i}(x)\|^2 = d^2_{C_i}(x).$$

By adding these bounds all together, we have

$$\begin{aligned}(\forall x \in \mathcal{H}) \qquad \langle S(x), x - x^* \rangle &\geq \frac{1}{N}\min\left\{1, \frac{\epsilon^2}{L^2}\right\}\left(\sum_{i=1}^{s_1} d^2_{C_i}(x) + \sum_{i=s_1+1}^{s_1+s_2} d^2_{\{f_{i-s_1}(z) < 0\}}(x)\right)\\ &\geq \frac{1}{N}\min\left\{1, \frac{\epsilon^2}{L^2}\right\}\max\left(\left\{d^2_{C_i}(x) \mid i = 1, \ldots, s_1\right\} \cup \left\{d^2_{\{f_i(z) \leq 0\}}(x) \mid i = 1, \ldots, s_2\right\}\right)\\ &\geq \frac{d^2_{\mathrm{zer}(S)}(x)}{N\hat{\mu}^2}\min\left\{1, \frac{\epsilon^2}{L^2}\right\},\end{aligned}$$

which proves that S is μ essentially strongly quasi-monotone.

---

[31] A set family $\{D_1, \ldots, D_N\}$ is $\hat{\mu}$-linearly regular if $\forall x \in \mathcal{H}$, $d_{D_1 \cap \cdots \cap D_N}(x) \leq \hat{\mu}\max\{d_{D_1}(x), \ldots, d_{D_N}(x)\}$.

Part 4 (demiclosedness): Let $\{x^k\}_{k\in\mathbb{N}}\subseteq\mathcal{H}$ be a sequence of points, let $x\in\mathcal{H}$, and suppose that two limits hold: $x^k\rightharpoonup x$ and $S(x^k)\to 0$. To show that $I-S$ is demiclosed at 0, we need to show that $S(x)=0$. We prove this in two parts: first we show that for all i, $S_i(x^k)\to 0$; and second we show that these limits imply that $x\in\operatorname{zer}(S)$.

"$S_i(x^k)\to 0$:" For all i, let $T_i=I-S_i$. These operators T_i are quasinonexpansive (as noted in Part 1). Thus, for any $y\in\operatorname{zer}(S)$, we have

$$2\langle T_ix^k-x^k,x^k-y\rangle=\|T_ix^k-y\|^2-\|T_ix^k-x^k\|^2-\|x^k-y\|^2\le-\|T_ix^k-x^k\|^2.$$

Thus,

$$\frac{1}{N}\sum_{i=1}^{N}-\|T_ix^k-x^k\|^2\ge\frac{1}{N}\sum_{i=1}^{N}2\langle T_ix^k-x^k,x^k-y\rangle=\langle S(x^k),x^k-y\rangle\to 0,$$

and so $S_i(x^k)=T_ix^k-x^k\to 0$,

"$x\in\operatorname{zer}(S)$:" All the projection operators P_{C_i} are nonexpansive, and moreover, $P_{C_i}x^k-x^k\to 0$. Therefore, because all nonexpansive operators are demiclosed at 0, we have $P_{C_i}x=x$ and $x\in C_i$.

The subgradient projectors are not nonexpansive, but we can still show that $G_{f_i}x=x$, and hence, $f_i(x)\le 0$. Indeed, $\|G_{f_i}(x^k)-x^k\|=f_i(x^k)\|g_i(x^k)\|^{-1}\to 0$. And because subdifferential operators of continuous convex functions are locally bounded, because the sequence $\{x^k\}_{k\in\mathbb{N}}$ is bounded, and because the subgradients $\{g_i(x^k)\}_{k\in\mathbb{N}}$ are bounded, it follows that $f_i(x^k)\to 0$. Then because convex functions are weakly lower semicontinuous, we have $f_i(x)\le\liminf_k f_i(x^k)\le 0$. Altogether, $x\in\operatorname{zer}(S)$, and consequently, S is demiclosed at 0. $\qquad\square$

Corollary C.1 (Kaczmarz Operator Properties) *Assume the setting of Section 4.4, and in particular, that*

$$(\forall i)\qquad S_i=(I-P_{C_i})=(\langle a_i,\cdot\rangle-b_i)a_i.$$

Then

1. **Coherence:** *S satisfies the coherence condition*

$$(\forall x\in\mathcal{H}),(\forall x^*\in\operatorname{zer}(S))\qquad\langle S(x),x-x^*\rangle\ge\sum_{i=1}^{N}\beta_{i1}\|S_i(x)-S_i(x^*)\|^2.$$

   with $\beta_{i1}\equiv 1$.

2. **Essential strong quasi-monotonicity:** *S is*

$$\mu=\frac{1}{N\|A^{-1}\|_2^2}.$$

   essentially strong quasi-monotone, where

$$\|A^{-1}\|_2:=\inf\{M\mid(\forall x\in\mathcal{H})\ M\|Ax\|_2\ge\|x\|_2\}.$$

3. **Roots:** *$\operatorname{zer}(S)$ is precisely the set of solutions to the linear equation $Ax=b$.*
4. **Demiclosedness:** *$I-S$ is demiclosed at 0.*

Proof Parts 1 (coherence), 3 (roots), and 4 (demiclosedness) follow from Proposition C.4.
Part 2: Observe that

$$\langle S(x),x-x^*\rangle=\frac{1}{N}\sum_{i=1}^{N}\langle(\langle a_i,x\rangle-b_i)a_i,x-x^*\rangle=\frac{1}{N}\sum_{i=1}^{N}\langle(\langle a_i,x\rangle-\langle a_i,x^*\rangle)a_i,x-x^*\rangle=\frac{1}{N}\sum_{i=1}^{N}\langle a_i,x-x^*\rangle^2$$

$$=\frac{1}{N}\|A(x-x^*)\|_2^2$$

$$\ge\frac{1}{N\|A^{-1}\|_2^2}\|x-x^*\|^2.\qquad\square$$

C.2 New Operators

Proposition C.5 (Proximal SAGA/SVRG Operator Properties) *Assume the setting of Section 5.2, and in particular, that*

$$(\forall i < N + 1) \qquad S_i = \frac{\gamma}{N} \nabla f_i \circ \mathbf{prox}_{\gamma g};$$
$$S_{N+1} = (I - \mathbf{prox}_{\gamma g}),$$

Then

1. **Coherence:** *S satisfies the coherence condition*

$$(\forall x \in \mathcal{H}) , (\forall x^* \in \mathrm{zer}(S)) \qquad \langle S(x), x - x^* \rangle \geq \sum_{i=1}^{N} \beta_{i1} \|S_i(x) - S_i(x^*)\|^2.$$

with

$$\beta_{i1} := \frac{N}{2\gamma L(N+1)} \qquad i = 1, \ldots, N \qquad \qquad and \qquad \qquad \beta_{(N+1)1} := \frac{1}{(N+1)} \left(1 - \frac{\gamma L}{2}\right)$$

2. **Essential strong quasi-monotonicity:** *S is*

$$\mu = \frac{1 + \gamma \mu_g - \sqrt{1 - 2\gamma \mu_f + \gamma^2 L \mu_f}}{(N+1)(1 + \gamma \mu_g)}$$

essentially strongly quasi-monotone, where $N^{-1} \sum_{i=1}^{N} f_i$ is μ_f-strongly convex and g is μ_g-strongly convex.

3. **Roots:** *with any choice of γ, we have*

$$x \in \mathrm{zer}(S) \implies \mathbf{prox}_{\gamma g}(x) \text{ solves } (5.1),$$

and $\mathrm{zer}(S) \neq \emptyset$ if, and only if, $g + N^{-1} \sum_{j=1}^{N} f_j$ has a minimizer.

4. **Demiclosedness:** *$I - S$ is demiclosed at 0.*

Proof Part 1 (coherence): The operator $S_{N+1} = (I - \mathbf{prox}_{\gamma g})$ is 1-cocoercive and ∇f_i is $N(\gamma L)^{-1}$-cocoercive. Thus, for all $x \in \mathcal{S}$, we have

$$
\begin{aligned}
\langle S(x), x - x^* \rangle &= \frac{1}{N+1} \sum_{i=1}^{N+1} \langle S_i(x) - S_i(x^*), x - x^* \rangle \\
&= \frac{1}{N+1} \sum_{i=1}^{N} \langle S_i(x) - S_i(x^*), \mathbf{prox}_{\gamma g}(x) - \mathbf{prox}_{\gamma g}(x^*) \rangle + \frac{1}{N+1} \langle S_{N+1}(x) - S_{N+1}(x^*), x - x^* \rangle \\
&\quad + \frac{1}{(N+1)} \sum_{i=1}^{N} \langle S_i(x) - S_i(x^*), S_{N+1}(x) - S_{N+1}(x^*) \rangle \\
&\geq \frac{N}{\gamma L(N+1)} \sum_{i=1}^{N} \|S_i(x) - S_i(x^*)\|^2 + \frac{1}{(N+1)} \|S_{N+1}(x) - S_{N+1}(x^*)\|^2 \\
&\quad - \frac{N}{2\gamma L(N+1)} \sum_{i=1}^{N} \|S_i(x) - S_i(x^*)\|^2 - \frac{\gamma L}{2(N+1)} \|S_{N+1}(x) - S_{N+1}(x^*)\|^2 \\
&\geq \frac{N}{2\gamma L(N+1)} \sum_{i=1}^{N} \|S_i(x) - S_i(x^*)\|^2 + \frac{1}{(N+1)} \left(1 - \frac{\gamma L}{2}\right) \|S_{N+1}(x) - S_{N+1}(x^*)\|^2.
\end{aligned}
$$

Part 2 (essential strong quasi-monotonicity): Let

$$T = \left(I - \frac{\gamma}{N} \sum_{i=1}^{N} \nabla f_i\right) \circ \mathbf{prox}_{\gamma g}.$$

The operator $\mathbf{prox}_{\gamma g}$ is $(1+\gamma\mu_g)^{-1}$-Lipschitz continuous [2, Proposition 23.11]. In addition, by Proposition B.5, $\left(I - \gamma N^{-1} \sum_{i=1}^{N} \nabla f_i\right)$ is $\sqrt{1 - 2\gamma\mu_f + \gamma^2 L\mu_f}$-Lipschitz continuous, whenever $\gamma \leq 2L^{-1}$. Thus, T, the composition of the two Lipschitz operators, is $(1 + \gamma\mu_g)^{-1}\sqrt{1 - 2\gamma\mu_f + \gamma^2 L\mu_f}$ Lipschitz continuous. Therefore, $S = (N+1)^{-1}(I - T)$ is

$$\mu = \frac{1}{N+1}\left(1 - \frac{\sqrt{1 - 2\gamma\mu_f + \gamma^2 L\mu_f}}{1 + \gamma\mu_g}\right) = \frac{1 + \gamma\mu_g - \sqrt{1 - 2\gamma\mu_f + \gamma^2 L\mu}}{(N+1)(1 + \gamma\mu_g)}$$

essentially strongly quasi-monotone.

Part 3 (roots):

$$x \in \mathrm{zer}(S) \Leftrightarrow 0 = (I - \mathbf{prox}_{\gamma g}) + \frac{\gamma}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{prox}_{\gamma g}(x)) \in \gamma\partial g(\mathbf{prox}_{\gamma g}(x)) + \frac{\gamma}{N}\sum_{i=1}^{N}\nabla f_i(\mathbf{prox}_{\gamma g}(x))$$

$$\Leftrightarrow \mathbf{prox}_{\gamma g}(x) \text{ minimizes } g + \frac{1}{N}\sum_{j=1}^{N}f_j. \qquad \square$$

Part 4 (demiclosedness): The operator $T = I - S$ in (displayed in Part 2) is the composition of two nonexpansive maps, and thus, it is nonexpansive. Therefore, T is demiclosed at 0.

Proposition C.6 (LinSAGA/LinSVRG/SuperSAGA/SuperSVRG Operator Properties) *Assume the setting of Section 6.1, and in particular, that*

$$(\forall i < N + 1) \qquad S_i = \frac{\gamma}{N}P_V \circ \nabla f_i \circ P_V \circ \mathbf{prox}_{\gamma g};$$

$$S_{N+1} = (I - 2P_V) \circ \mathbf{prox}_{\gamma g} + P_V.$$

Then

1. ***Coherence:*** *S satisfies the coherence condition*

$$(\forall x \in \mathcal{H}), (\forall x^* \in \mathrm{zer}(S)) \qquad \langle S(x), x - x^* \rangle \geq \sum_{i=1}^{N}\beta_{i1}\|S_i(x) - S_i(x^*)\|^2.$$

with

$$\beta_{i1} := \frac{N}{2\gamma\hat{L}_i(N+1)} \qquad i = 1, \dots, N \qquad and \qquad \beta_{(N+1)1} := \frac{1}{N+1}\left(1 - \frac{1}{2N}\sum_{i=1}^{N}\gamma\hat{L}_i\right).$$

2. ***Essential strong quasi-monotonicity:*** *S is*

$$\mu = \frac{1}{N+1}\left(1 - \left(\frac{1}{(1 + (\gamma L_g)^{-1})} + \frac{\sqrt{1 - 2\gamma\mu_f + \gamma^2 L\mu_f}}{(1 + \gamma\mu_g)}\right)\right)$$

essentially strongly quasi-monotone (when $\gamma \leq 2L^{-1}$ and $\mu > 0$), where $L = N^{-1}\sum_{i=1}^{N}\hat{L}_i$, the function $N^{-1}\sum_{i=1}^{N}f_i$ is μ_f-strongly convex, the function g is differentiable and μ_g-strongly convex, and the gradient ∇g is L_g-Lipschitz continuous.

3. ***Roots:*** *with any choice of γ, we have*

$$x^* \in \mathrm{zer}(S) \implies \mathbf{prox}_{\gamma g}(x^*) \text{ solves } (6.2),$$

and $\mathrm{zer}(S) \neq \emptyset$ if, and only if, $\mathrm{zer}\left(\partial g + N^{-1}\sum_{i=1}^{N}\nabla f_i + N_V\right) \neq \emptyset$.

4. ***Demiclosedness:*** *$I - S$ is demiclosed at 0.*

Proof Part 1 (coherence): The operator $S_{N+1} = (I - 2P_V) \circ \mathbf{prox}_{\gamma g} + P_V$ is 1-cocoercive because

$$S_{N+1} = I - \left[P_V (2\mathbf{prox}_{\gamma g} - I) + I - \mathbf{prox}_{\gamma g} \right]$$

and by [2, Proposition 4.21], $P_V(2\mathbf{prox}_{\gamma g} - I) + I - \mathbf{prox}_{\gamma g}$ is 2^{-1} averaged. Not only is the operator 1-cocoercive, it also nicely partitions into V and $V^\perp$ components:

$$S_{N+1} = P_{V^\perp} \mathbf{prox}_{\gamma g} + P_V (I - \mathbf{prox}_{\gamma g})$$

Therefore, $\langle S_i(x), S_{N+1}(y) \rangle = \langle S_i(x), P_V(y - \mathbf{prox}_{\gamma g}(y)) \rangle$ for all $i < N+1$ and any $x, y \in \mathcal{H}$. In addition, $\frac{\gamma}{N} P_V \circ \nabla f_i \circ P_V$ is $N(\gamma L_i)^{-1}$-cocoercive. Thus, for all $x \in \mathcal{S}$, we have

$$
\begin{aligned}
\langle S(x), x - x^* \rangle &= \frac{1}{N+1} \sum_{i=1}^{N+1} \langle S_i(x) - S_i(x^*), x - x^* \rangle \\
&= \frac{1}{N+1} \sum_{i=1}^{N} \langle S_i(x) - S_i(x^*), P_V \mathbf{prox}_{\gamma g}(x) - P_V \mathbf{prox}_{\gamma g}(x^*) \rangle + \frac{1}{N+1} \langle S_{N+1}(x) - S_{N+1}(x^*), x - x^* \rangle \\
&\quad + \frac{1}{N+1} \sum_{i=1}^{N} \langle S_i(x) - S_i(x^*), S_{N+1}(x) - S_{N+1}(x^*) \rangle \\
&\geq \frac{N}{N+1} \sum_{i=1}^{N} \frac{1}{\gamma \hat{L}_i} \| S_i(x) - S_i(x^*) \|^2 + \frac{1}{N+1} \| S_{N+1}(x) - S_{N+1}(x^*) \|^2 \\
&\quad - \frac{N}{N+1} \sum_{i=1}^{N} \frac{1}{2\gamma \hat{L}_i} \| S_i(x) - S_i(x^*) \|^2 - \frac{\frac{1}{N} \sum_{i=1}^{N} \gamma \hat{L}_i}{2(N+1)} \| S_{N+1}(x) - S_{N+1}(x^*) \|^2 \\
&\geq \frac{N}{2(N+1)} \sum_{i=1}^{N} \frac{1}{\gamma \hat{L}_i} \| S_i(x) - S_i(x^*) \|^2 + \frac{1}{N+1} \left(1 - \frac{1}{2N} \sum_{i=1}^{N} \gamma \hat{L}_i \right) \| S_{N+1}(x) - S_{N+1}(x^*) \|^2.
\end{aligned}
$$

Part 2 (essential strong quasi-monotonicity): Let

$$T := P_{V^\perp} \circ (I - \mathbf{prox}_{\gamma g}) + P_V \circ \left(I - \frac{\gamma}{N} \sum_{i=1}^{N} \nabla f_i \right) \circ P_V \circ \mathbf{prox}_{\gamma g}$$

The operator T is Lipschitz continuous because it is built from Lipschitz continuous pieces. For example, by Proposition B.5, the operator $P_V \circ \left(I - \gamma N^{-1} \sum_{i=1}^{N} \nabla f_i \right) \circ P_V$ is $\sqrt{1 - 2\gamma \mu_f + \gamma^2 L \mu_f}$-Lipschitz continuous; by B.5 the operator $\mathbf{prox}_{\gamma g}$ is $(1 + \mu_g \gamma)^{-1}$-Lipschitz continuous; by [2, Remark 23.19], we have $I - \mathbf{prox}_{\gamma g} = \mathbf{prox}_{(\gamma g)^*}$, and by [2, Theorem 18.15], the function $(\gamma g)^*$ is $(\gamma L_g)^{-1}$-strongly convex; thus, by [2, Proposition 23.11], $I - \mathbf{prox}_{\gamma g}$ is $(1 + (\gamma L_g)^{-1})^{-1}$-Lipschitz continuous. When taken together, these properties yield

$$
\begin{aligned}
(\forall x, y \in \mathcal{H}) \, \| T(x) - T(y) \| &\leq \| P_{V^\perp}(I - \mathbf{prox}_{\gamma g})(x) - P_{V^\perp}(I - \mathbf{prox}_{\gamma g})(y) \| \\
&\quad + \left\| P_V \circ \left(I - \frac{\gamma}{N} \sum_{i=1}^{N} \nabla f_i \right) \circ P_V \circ \mathbf{prox}_{\gamma g}(x) - P_V \circ \left(I - \frac{\gamma}{N} \sum_{i=1}^{N} \nabla f_i \right) \circ P_V \circ \mathbf{prox}_{\gamma g}(y) \right\| \\
&\leq \left(\frac{1}{(1 + (\gamma L_g)^{-1})} + \frac{\sqrt{1 - 2\gamma \mu_f + \gamma^2 L \mu_f}}{(1 + \gamma \mu_g)} \right) \| x - y \|.
\end{aligned}
$$

Thus, by Proposition B.3, $S = (N+1)^{-1}(I - T)$ is

$$\frac{1}{N+1} \left(1 - \left(\frac{1}{(1 + (\gamma L_g)^{-1})} + \frac{\sqrt{1 - 2\gamma \mu_f + \gamma^2 L \mu_f}}{(1 + \gamma \mu_g)} \right) \right)$$

strongly monotone.

Part 3 (roots):

$$x \in \mathrm{zer}(S) \Leftrightarrow 0 = P_{V^\perp} \mathbf{prox}_{\gamma g}(x); \text{ and}$$

$$0 = P_V(x - \mathbf{prox}_{\gamma g}(x)) + \frac{\gamma}{N} \sum_{i=1}^{N} P_V \nabla f_i(P_V \mathbf{prox}_{\gamma g}(x))$$

$$\in P_V \partial g(\mathbf{prox}_{\gamma g}(x)) + \frac{\gamma}{N} \sum_{i=1}^{N} P_V \nabla f_i(P_V \mathbf{prox}_{\gamma g}(x))$$

$$\Leftrightarrow 0 \in \partial g(\mathbf{prox}_{\gamma}(x)) + \frac{\gamma}{N} \sum_{i=1}^{N} \nabla f_i(P_V \mathbf{prox}_{\gamma g}(x)) + N_V(\mathbf{prox}_{\gamma g}(x)),$$

where the last line follows because $\mathbf{prox}_{\gamma g}(x) \in V$, and so $N_V(\mathbf{prox}_{\gamma g}(x)) = V^\perp$ absorbs the nonzero $V^\perp$ component of $(x - \mathbf{prox}_{\gamma g}(x)) + \gamma N^{-1} \sum_{i=1}^{N} \nabla f_i(\mathbf{prox}_{\gamma g}(x))$.

Part 4 (demiclosedness): The operator $T = I - S$ in (displayed in Part 2) is nonexpansive:

$$(\forall x, y \in \mathcal{H}) \, \|T(x) - T(y)\|^2 = \|P_{V^\perp}(I - \mathbf{prox}_{\gamma g})(x) - P_{V^\perp}(I - \mathbf{prox}_{\gamma g})(y)\|^2$$

$$+ \left\| P_V \circ \left(I - \frac{\gamma}{N} \sum_{i=1}^{N} \nabla f_i\right) \circ P_V \circ \mathbf{prox}_{\gamma g}(x) - P_V \circ \left(I - \frac{\gamma}{N} \sum_{i=1}^{N} \nabla f_i\right) \circ P_V \circ \mathbf{prox}_{\gamma g}(y) \right\|^2$$

$$\leq \|(I - \mathbf{prox}_{\gamma g})(x) - (I - \mathbf{prox}_{\gamma g})(y)\|^2 + \|\mathbf{prox}_{\gamma g}(x) - \mathbf{prox}_{\gamma g}(y)\|^2$$

$$\leq \|x - y\|^2,$$

where the first "=" follows because $V \perp V^\perp$ and the last line follows because $\mathbf{prox}_{\gamma g}$ is 2^{-1}-averaged. Therefore T is nonexpansive and demiclosed at 0. $\square$

Proposition C.7 (TropicSMART Operator Properties) *Assume the setting of Section 6.3, and in particular, that for all $x \in \mathcal{H}$, we have*

$$(S(x))_j = \begin{cases} x_j - \mathbf{prox}_{\gamma_j g_j} \left(x_j - \gamma_j A_j^* \left(x_{M+1} + 2\gamma_{M+1} \left(\sum_{l=1}^{M} A_l x_l - b \right) \right) - \gamma_j \nabla_j f(x) \right) & \text{if } j < M+1; \\ -\gamma_{M+1} \left(\sum_{l=1}^{M} A_l x_l - b \right) & \text{if } j = M+1. \end{cases}$$

Then

1. **Coherence:** *there is a strongly positive self-adjoint linear operator $P : \mathcal{H} \to \mathcal{H}$ such that if $\delta \in (0, 1)$;*

$$\gamma_{M+1} \left(\sum_{j=1}^{M} \gamma_j \|A_j\|^2 \right) \leq \delta; \qquad \text{and} \qquad \max_j\{\gamma_j\} \leq \frac{2(1 - \sqrt{\delta})}{L},$$

then the linear map P satisfies $\sum_{j=1}^{M+1} \underline{M}_j \|x_j\|_j^2 \leq \|x\|_P^2 \leq \sum_{j=1}^{M+1} \overline{M}_j \|x_j\|_j^2$, where for all j, we have

$$\underline{M}_j := \frac{1 - \sqrt{\delta}}{\gamma_j} \qquad \text{and} \qquad \overline{M}_j := \frac{1 + \sqrt{\delta}}{\gamma_j}.$$

With these parameters, S also satisfies the coherence condition (1.5)

$$(\forall x \in \mathcal{H}), (\forall x^* \in \mathrm{zer}(S)) \qquad \langle S(x), x - x^* \rangle_P \geq \sum_{j=1}^{M+1} \beta_{1j} \|(S(x))_j\|_j^2 \qquad \text{with} \qquad \beta_{1j} := \frac{L \max_j\{\gamma_j\}}{4\gamma_j}.$$

2. **Roots:** *with any choice of γ_j, we have*

$$x^* \in \text{zer}(S) \implies (x_1^*, \ldots, x_M^*) \text{ solves } (6.2),$$

and $\text{zer}(S) \neq \emptyset$ *if, and only if,* $\text{zer}\left(\partial g + \nabla f + N_{\{x \in \mathcal{H} | \sum_{j=1}^M A_j x_j = b\}}\right) \neq \emptyset$.

3. **Demiclosedness:** $I - S$ *is demiclosed at* 0.

Proof Part 1 (coherence): Here is the linear map P in block matrix form

$$P = \begin{bmatrix} \frac{1}{\gamma_1} I_{\mathcal{H}_1} & 0 & \cdots & 0 & A_1^* \\ 0 & \frac{1}{\gamma_2} I_{\mathcal{H}_2} & \cdots & 0 & A_2^* \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & \cdots & \frac{1}{\gamma_M} I_{\mathcal{H}_M} & A_M^* \\ A_1 & A_2 & \cdots & A_M & \frac{1}{\gamma_{M+1}} I_{\mathcal{H}_{M+1}} \end{bmatrix}.$$

The lower bound is below

$$\begin{aligned}
\langle x, x \rangle_P &= \sum_{j=1}^{M+1} \frac{1}{\gamma_j} \|x_j\|_j^2 + \sum_{j=1}^M 2\langle A_j x_j, x_{M+1} \rangle_{M+1} \\
&\geq \sum_{j=1}^{M+1} \frac{1}{\gamma_j} \|x_j\|_j^2 - \sum_{j=1}^M 2\|A_j\| \|x_j\|_j \|x_{M+1}\|_{M+1} \\
&\geq \sum_{j=1}^{M+1} \frac{1}{\gamma_j} \|x_j\|_j^2 - \sum_{j=1}^M \left(\frac{\sqrt{\delta} \|x_j\|_j^2}{\gamma_j} + \frac{\|A_j\|^2 \|x_{M+1}\|_{M+1}^2}{\sqrt{\delta}} \right) \\
&\geq \sum_{j=1}^M \frac{1 - \sqrt{\delta}}{\gamma_j} \|x_j\|_j^2 + \frac{1}{\gamma_{M+1}} \left(1 - \gamma_{M+1} \sum_{j=1}^M \frac{\|A_j\|^2}{\sqrt{\delta}} \right) \|x_{M+1}\|_{M+1}^2 \\
&\geq \sum_{j=1}^{M+1} \frac{1 - \sqrt{\delta}}{\gamma_j} \|x_j\|_j^2.
\end{aligned}$$

The upper bound follows the exact same argument but all "$-$" signs are changed to "$+$" signs.

To get the β_{ij}, we find an α-averaged operator T in the norm $\| \cdot \|_P$ such that $S = (I - T)$; then we apply Proposition B.1 together with the lower on $\| \cdot \|_P^2$ that we just derived.

Let $A : \mathcal{H} \to 2^{\mathcal{H}}$ be the monotone operator[32]

$$(\forall x \in \mathcal{H}) \qquad A(x_1, \ldots, x_{M+1}) := \begin{bmatrix} \partial g_1(x_1) \\ \vdots \\ \partial g_M(x_M) \\ 0 \end{bmatrix};$$

let $B : \mathcal{H} \to \mathcal{H}$ be the skew symmetric linear map:

$$B := \begin{bmatrix} 0 & \cdots & 0 & A_1 \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & A_{M+1} \\ -A_1^* & \cdots & -A_M^* & 0 \end{bmatrix};$$

---

[32] This matrix notation is an intuitive visual form for the product of the subdifferentials.

and let $C : \mathcal{H} \to \mathcal{H}$ be the cocoercive operator

$$(\forall x \in \mathcal{H}) \qquad C(x_1, \ldots, x_{M+1}) := \begin{bmatrix} \nabla f(x_1, \ldots, x_M) \\ 0 \end{bmatrix}.$$

Then define T to be the forward-backward operator

$$T := J_{P^{-1}(A+B)} \circ (I - P^{-1}C).$$

Given x, we compute Tx:

$$x^+ = Tx \Leftrightarrow P(x - x^+) \in (A+B)x^+ + Cx$$

$$\Leftrightarrow \begin{bmatrix} \frac{1}{\gamma_1}(x_1 - x_1^+) + A_1^*(x_{M+1} - x_{M+1}^+) \\ \vdots \\ \frac{1}{\gamma_M}(x_M - x_M^+) + A_M^*(x_{M+1} - x_{M+1}^+) \\ \frac{1}{\gamma_{M+1}}(x_{M+1} - x_{M+1}^+) + \sum_{j=1}^M A_j(x_j - x_j^+) \end{bmatrix} \in \begin{bmatrix} \partial g_1(x_1^+) + A_1^* x_{M+1}^+ + \nabla_1 f(x_1, \ldots, x_M) \\ \vdots \\ \partial g_M(x_M^+) + A_M^* x_{M+1}^+ + \nabla_M f(x_1, \ldots, x_M) \\ -\sum_{j=1}^M A_j x_j^+ \end{bmatrix}$$

$$\Leftrightarrow (\forall j < M+1) \qquad x_j^+ = \mathbf{prox}_{\gamma_j g_j}\left(x_j - \gamma_j A_j^*\left(x_{M+1} + 2\gamma_{M+1}\left(\sum_{l=1}^M A_l x_l - b \right) \right) - \gamma_j \nabla_j f(x_1, \ldots, x_m) \right);$$

$$x_{M+1}^+ = x_{M+1} + \gamma_{M+1}\left(\sum_{l=1}^M A_l x_l - b \right). \tag{C.2}$$

Thus, $S = I - T$.

The operator T is a forward-backward operator, and such operators are always α-averaged. But computing α requires the averagedness coefficient of $I - P^{-1}C$; this constant can be computed from existing work:

Lemma C.1 ([11, **Proposition 1.5**]) *Let $(\mathcal{G}, \langle, \rangle)$ be a Hilbert space, and let $U : \mathcal{G} \to \mathcal{G}$ be a strongly positive linear map such that for all $x \in \mathcal{G}$, we have $\|x\|_U^2 = \langle Ux, x \rangle \geq \xi \|x\|^2$. Then if $C : \mathcal{H} \to C$ is β-cocoercive in the norm $\| \cdot \|$, it is $\beta\xi$ cocoercive in the norm $\| \cdot \|_U$.*

Thus, from the already computed lower bound $\| \cdot \|_P^2 \geq (1 - \sqrt{\delta})(\max_j\{\gamma_j\})^{-1} \| \cdot \|_{\text{prod}}^2$, and from the knowledge that C is L^{-1}-cocoercive in $\| \cdot \|_{\text{prod}}$, the composition $P^{-1}C$ is $(1-\sqrt{\delta})(L\max_j\{\gamma_j\})^{-1}$-cocoercive in $\| \cdot \|_P$. Then the standard result [2, Proposition 4.33] shows that $I - P^{-1}C$ is $\alpha_1 := 2^{-1}(1-\sqrt{\delta})^{-1}L\max_j\{\gamma_j\}$ averaged as long as

$$\max_j\{\gamma_j\} \leq \frac{2(1-\sqrt{\delta})}{L},$$

which we assume to be true. Therefore, because $J_{P^{-1}(A+B)}$ is $\alpha_2 := 2^{-1}$ averaged in $\| \cdot \|_P^2$, Proposition B.2 shows that T is

$$\alpha := \frac{\alpha_1 + \alpha_2 - 2\alpha_1\alpha_2}{1 - \alpha_1\alpha_2} = \frac{\frac{1}{2}}{1 - \frac{L\max_j\{\gamma_j\}}{4(1-\sqrt{\delta})}} = \frac{2}{4 - \frac{L\max_j\{\gamma_j\}}{1-\sqrt{\delta}}}$$

averaged in the norm $\| \cdot \|_P$. Thus, from Proposition B.1, the operator $S = I - T$ is

$$\beta := 1 - \frac{1}{2\alpha} = \frac{L\max_j\{\gamma_j\}}{4(1-\sqrt{\delta})}$$

cocoercive in the norm $\| \cdot \|_P$; with this property, we obtain β_{ij}

$$(\forall x \in \mathcal{H}) \qquad \langle S(x), x - x^* \rangle_P \geq \beta \|S(x)\|_P^2 \geq \sum_{j=1}^{M+1} \frac{\beta(1-\sqrt{\delta})}{\gamma_j} \|(S(x))_j\|_j^2 = \sum_{j=1}^{M+1} \frac{L\max_j\{\gamma_j\}}{4\gamma_j} \|(S(x))_j\|_j^2.$$

Part 2 (roots): From the first line of (C.2), the roots of S are precisely the zeros of $A + B + C$. Then it is straightforward to check that $x^* \in \mathrm{zer}(A + B + C)$, implies that $(x_1^*, \ldots, x_m^*) \in \mathrm{zer}(\partial g + \nabla f + N_{\{x \in \mathcal{H} | \sum_{j=1}^M A_j x_j = b\}})$, and if $(x_1^*, \ldots, x_m^*) \in \mathrm{zer}(\partial g + \nabla f + N_{\{x \in \mathcal{H} | \sum_{j=1}^M A_j x_j = b\}})$, then there exists $x_{M+1}^* \in \mathcal{H}_{M+1}$ such that $x^* = (x_1^*, \ldots, x_{M+1}^*) \in \mathrm{zer}(A + B + C)$. Both inclusions imply that $(x_1^*, \ldots, x_M^*)$ minimizes (6.4).

Part 3 (demiclosedness): The operator $T = I - S$ in (displayed in Part 1) is the composition of two nonexpansive maps (in the norm $\| \cdot \|_P$), and thus, it is nonexpansive. Therefore, T is demiclosed at 0. $\square$

Proposition C.8 (ProxSMART Operator Properties) *Assume the setting of Section 6.4, and in particular, that for all $x \in \mathcal{H}$, we have*

$$
(S(x))_j = \begin{cases} x_1 - \mathbf{prox}_{\gamma_1 g_1}\left(x_1 - \gamma_1 \sum_{j=2}^M A_j^* x_j\right) & \text{if } j = 1; \\ x_j - \mathbf{prox}_{\gamma_j g_j^*}\left(x_j + \gamma_j A_j\left(2\overline{x}_1 - x_1\right)\right) & \text{otherwise;} \end{cases}
$$

where $\overline{x}_1 = \mathbf{prox}_{\gamma_1 g_1}\left(x_1 - \gamma_1 \sum_{j=2}^M A_j^* x_j\right)$. Then

1. *Coherence: there is a strongly positive self-adjoint linear operator $P : \mathcal{H} \to \mathcal{H}$ such that if $\delta \in (0, 1)$;*

$$
\gamma_1 \left(\sum_{j=2}^M \gamma_j \|A_j\|^2\right) \leq \delta.
$$

Then the linear map P satisfies $\sum_{j=1}^M \underline{M}_j \|x_j\|_j^2 \leq \|x\|_P^2 \leq \sum_{j=1}^M \overline{M}_j \|x_j\|_j^2$, where for all j, we have

$$
\underline{M}_j := \frac{1 - \sqrt{\delta}}{\gamma_j} \qquad\qquad and \qquad\qquad \overline{M}_j := \frac{1 + \sqrt{\delta}}{\gamma_j}.
$$

With these parameters, S also satisfies the coherence condition (1.5):

$$
(\forall x \in \mathcal{H}), (\forall x^* \in \mathrm{zer}(S)) \qquad \langle S(x), x - x^* \rangle_P \geq \sum_{j=1}^M \beta_{1j} \|(S(x))_j\|_j^2 \qquad with \qquad \beta_{1j} := \frac{1 - \sqrt{\delta}}{\gamma_j}.
$$

2. *Roots: with any choice of γ_j, we have*

$$
x^* \in \mathrm{zer}(S) \implies x_1 \text{ solves } (6.5),
$$

and $\mathrm{zer}(S) \neq 0$ if, and only if, $\mathrm{zer}(\partial g_1(x) + \sum_{j=2}^M A_j^* \partial g_j \circ A_j) \neq \emptyset$.
3. *Demiclosedness: $I - S$ is demiclosed at 0.*

Proof Part 1 (coherence): Here is the linear map P in block matrix form

$$
P = \begin{bmatrix} \frac{1}{\gamma_1} I_{\mathcal{H}_1} & -A_2^* & \cdots \cdots & -A_M^* \\ -A_2 & \frac{1}{\gamma_2} I_{\mathcal{H}_2} & 0 & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ -A_M & 0 & \cdots & 0 & \frac{1}{\gamma_M} I_{\mathcal{H}_M} \end{bmatrix}.
\tag{C.3}
$$

The lower bound is below

$$\langle x, x \rangle_P = \sum_{j=1}^{M} \frac{1}{\gamma_j} \|x_j\|_j^2 - \sum_{j=1}^{M} 2\langle A_j x_j, x_1 \rangle_1$$

$$\geq \sum_{j=1}^{M} \frac{1}{\gamma_j} \|x_j\|_j^2 - \sum_{j=1}^{M} 2\|A_j\|\|x_j\|_j\|x_1\|_1$$

$$\geq \sum_{j=1}^{M} \frac{1}{\gamma_j} \|x_j\|_j^2 - \sum_{j=1}^{M} \left(\frac{\sqrt{\delta}\|x_j\|_j^2}{\gamma_j} + \frac{\|A_j\|^2\|x_1\|_1^2}{\sqrt{\delta}} \right)$$

$$\geq \sum_{j=2}^{M} \frac{1 - \sqrt{\delta}}{\gamma_j} \|x_j\|_j^2 + \frac{1}{\gamma_1} \left(1 - \gamma_1 \sum_{j=1}^{M} \frac{\|A_j\|^2}{\sqrt{\delta}} \right) \|x_1\|_1^2$$

$$\geq \sum_{j=1}^{M} \frac{1 - \sqrt{\delta}}{\gamma_j} \|x_j\|_j^2.$$

The upper bound follows the exact same argument but all "$-$" signs (except for the one on the first line) are changed to "$+$" signs.

To get the β_{ij}, we find an α-averaged operator T in the norm $\|\cdot\|_P$ such that $S = (I - T)$; then we apply Proposition B.1 together with the lower on $\|\cdot\|_P^2$ that we just derived.

Let $A : \mathcal{H} \to 2^{\mathcal{H}}$ be the monotone operator

$$A(x_1, \ldots, x_M) := \begin{bmatrix} \partial g_1(x_1) \\ \partial g_2^*(x_2) \\ \vdots \\ \partial g_M^*(x_M) \end{bmatrix};$$

and let $B : \mathcal{H} \to \mathcal{H}$ be the skew symmetric linear map:

$$B := \begin{bmatrix} 0 & \cdots & 0 & A_1 \\ 0 & \cdots & 0 & A_M \\ -A_1^* & \cdots & -A_M^* & 0 \end{bmatrix}.$$

Then define T to be the resolvent operator

$$T := J_{P^{-1}(A+B)}$$

Given x, we compute Tx:

$$x^+ = Tx \Leftrightarrow P(x - x^+) \in (A + B)x^+$$

$$\Leftrightarrow \begin{bmatrix} \frac{1}{\gamma_1}(x_1 - x_1^+) - \sum_{j=2}^{m} A_j^*(x_j - x_j^+) \\ \frac{1}{\gamma_2}(x_2 - x_2^+) - A_2(x_1 - x_1^+) \\ \vdots \\ \frac{1}{\gamma_M}(x_M - x_M^+) - A_2(x_1 - x_1^+) \end{bmatrix} \in \begin{bmatrix} \sum_{j=2}^{m} A_j^* x_j^+ \\ \partial g_2^*(x_2^+) - A_2^* x_1^+ \\ \vdots \\ \partial g_M^*(x_M^+) - A_2^* x_1^+ \end{bmatrix}$$

$$\Leftrightarrow \qquad x_1^+ = \mathbf{prox}_{\gamma_1 g_1} \left(x_1 - \gamma_1 \sum_{j=2}^{M} A_j^* x_j \right);$$

$$(\forall j > 1) \qquad x_j^+ = \mathbf{prox}_{\gamma_j g_j^*} \left(x_j + \gamma_j A_j \left(2x_1^+ - x_1 \right) \right). \tag{C.4}$$

Thus, $S = I - T$.

The operator T is a resolvent, and such operators are always 2^{-1}-averaged. Thus, from Proposition B.1, the operator $S = I - T$ is 1-cocoercive in the norm $\|\cdot\|_P$; with this property, we obtain β_{ij}

$$(\forall x \in \mathcal{H}) \qquad \langle S(x), x - x^* \rangle_P \geq \|S(x)\|_P^2 \geq \sum_{j=1}^{M} \frac{1 - \sqrt{\delta}}{\gamma_j} \|(S(x))_j\|_j^2. \qquad \square$$

Part 2 (roots): From the first line of (C.4), the roots of S are precisely the zeros of $A + B$. Then it is straightforward to check that $x^* \in \mathrm{zer}(A + B)$, implies that $x_1^* \in \mathrm{zer}(\partial g_1(x) + \sum_{j=2}^{M} A_j^* \partial g_j \circ A_j)$, and if $x_1^* \in \mathrm{zer}(\partial g_1(x) + \sum_{j=2}^{M} A_j^* \partial g_j \circ A_j)$, then there exists $(x_2^*, \ldots, x_M^*) \in \mathcal{H}_2 \times \cdots \times \mathcal{H}_M$ such that $x^* = (x_1^*, \ldots, x_M^*) \in \mathrm{zer}(A + B)$. Both inclusions imply that x_1^* minimizes (6.5).

Part 3 (demiclosedness): The operator $T = I - S$ in (displayed in Part 1) is a nonexpansive map (in the norm $\|\cdot\|_P$). Therefore, T is demiclosed at 0. $\qquad \square$

Proposition C.9 (ProxSMART+ Operator Properties) *Assume the setting of Section 6.5, and in particular, that for all $x \in \mathcal{H}$, we have*

$$(\forall i < N + 1) \qquad (S_i(x))_j = \begin{cases} \frac{\gamma_1}{N} \nabla f_i \left(x_1 - 2\gamma_1 \sum_{j=2}^{M} A_j^* x_j \right) & \text{if } j = 1; \\ 0 & \text{otherwise.} \end{cases}$$

$$(S_{N+1}(x))_j = \begin{cases} \gamma_1 \sum_{j=2}^{M} A_j^* x_j & \text{if } j = 1; \\ x_j - \mathbf{prox}_{\gamma_j g_j^*} \left(x_j + \gamma_j A_j \left(x_1 - 2\gamma_1 \sum_{j=2}^{M} A_j^* x_j \right) \right) & \text{otherwise,} \end{cases}$$

Then

1. **Coherence:** *there is a strongly positive self-adjoint linear operator $P : \mathcal{H} \to \mathcal{H}$ such that if $\delta \in (0, 1)$;*

$$\gamma_1 \left(\sum_{j=2}^{M} \gamma_j \|A_j\|^2 + \frac{1}{2N} \sum_{i=1}^{N} L_i \right) \leq \delta.$$

Then the linear map P satisfies $\sum_{j=1}^{M} \underline{M}_j \|x_j\|_j^2 \leq \|x\|_P^2 \leq \sum_{j=1}^{M} \overline{M}_j \|x_j\|_j^2$, where for all $j > 0$, we have

$$\underline{M}_j := \left(1 - \sqrt{\delta} + \frac{1}{2\sqrt{\delta}N} \sum_{i=1}^{N} L_i \right) \qquad \text{and} \qquad \overline{M}_j := \left(1 + \sqrt{\delta} - \frac{1}{2\sqrt{\delta}N} \sum_{i=1}^{N} L_i \right).$$

With these parameters, S also satisfies the coherence condition (1.5):

$$(\forall x \in \mathcal{H}), (\forall x^* \in \mathrm{zer}(S)) \qquad \langle S(x), x - x^* \rangle_P \geq \sum_{i=1}^{N+1} \sum_{j=1}^{M} \beta_{ij} \|(S_i(x))_j - (S_i(x^*))_j\|_j^2$$

with

$$(\forall 1 \leq i < N + 1) \qquad \beta_{i1} := \frac{N(1 - \sqrt{\delta})}{2(N+1)\gamma_1^2 L_i}; \quad (\forall i < N + 1), (\forall j > 1) \qquad \beta_{ij} \equiv 0; \quad (\forall j) \qquad \beta_{(N+1)j} := \frac{1 - \sqrt{\delta}}{\gamma_j}.$$

2. **Roots:** *with any choice of γ_j, we have*

$$x^* \in \mathrm{zer}(S) \implies x_1^* - 2\gamma_1 \sum_{j=2}^{M} A_j^* x_j^* \text{ solves (6.6)},$$

and $\mathrm{zer}(S) \neq \emptyset$ if, and only if, $\mathrm{zer}(\sum_{j=2}^{M} A_j^* \partial g_j \circ A_j + N^{-1} \sum_{i=1}^{N} \nabla f_i) \neq \emptyset$.

3. **Demiclosedness:** $I - S$ *is demiclosed at* 0.

Proof Part 1 (coherence): The linear map P is given in (C.3). The lower bound is below

$$
\begin{aligned}
\langle x, x \rangle_P &= \sum_{j=1}^{M} \frac{1}{\gamma_j} \|x_j\|_j^2 - \sum_{j=1}^{M} 2\langle A_j x_j, x_1 \rangle_1 \\
&\geq \sum_{j=1}^{M} \frac{1}{\gamma_j} \|x_j\|_j^2 - \sum_{j=1}^{M} 2\|A_j\| \|x_j\|_j \|x_1\|_1 \\
&\geq \sum_{j=1}^{M} \frac{1}{\gamma_j} \|x_j\|_j^2 - \sum_{j=1}^{M} \left(\frac{\sqrt{\delta} \|x_j\|_j^2}{\gamma_j} + \frac{\|A_j\|^2 \|x_1\|_1^2}{\sqrt{\delta}} \right) \\
&\geq \sum_{j=2}^{M} \frac{1 - \sqrt{\delta}}{\gamma_j} \|x_j\|_j^2 + \frac{1}{\gamma_1} \left(1 - \gamma_1 \sum_{j=1}^{M} \frac{\|A_j\|^2}{\sqrt{\delta}} \right) \|x_1\|_1^2 \\
&\geq \sum_{j=2}^{M} \frac{1 - \sqrt{\delta}}{\gamma_j} \|x_j\|_j^2 + \frac{1}{\gamma_1} \left(1 - \sqrt{\delta} + \frac{1}{2\sqrt{\delta}N} \sum_{i=1}^{N} L_i \right) \|x_1\|_1^2.
\end{aligned}
$$

The upper bound follows the exact same argument but all "$-$" signs are changed to "$+$" signs; the exception to this rule is the last line, in which the $-\frac{1}{2\sqrt{\delta}N}\sum_{i=1}^{N} L_i$ must be changed to $\frac{1}{2\sqrt{\delta}N}\sum_{i=1}^{N} L_i$.

Let $A : \mathcal{H} \to 2^{\mathcal{H}}$ be the monotone operator

$$
(\forall x \in \mathcal{H}) \qquad A(x_1, \ldots, x_M) := \begin{bmatrix} 0 \\ \partial g_2^*(x_2) \\ \vdots \\ \partial g_M^*(x_M) \end{bmatrix};
$$

let $B : \mathcal{H} \to \mathcal{H}$ be the skew symmetric linear map:

$$
B := \begin{bmatrix} 0 & \cdots & 0 & A_1 \\ 0 & \cdots & 0 & A_M \\ -A_1^* & \cdots & -A_M^* & 0 \end{bmatrix}.
$$

Then from the proof of Theorem C.8 (with $g_1 \equiv 0$), we find that $S_{N+1} = I - J_{P^{-1}(A+B)}$ is 1-cocoercive in $\|\cdot\|_P$. Thus,

$$
\begin{aligned}
\langle S_{N+1}(x) - S_{N+1}(x^*), x - x^* \rangle_P &\geq \|S_{N+1}(x) - S_{N+1}(x^*)\|_P^2 \\
&\geq \sum_{j=2}^{M} \frac{1 - \sqrt{\delta}}{\gamma_j} \|(S_{N+1}(x))_j - (S_{N+1}(x^*))_j\|_j^2 \\
&\quad + \left(1 - \sqrt{\delta} + \frac{1}{2\sqrt{\delta}N} \sum_{i=1}^{N} L_i \right) \|(S_{N+1}(x))_1 - (S_{N+1}(x^*))_1\|_1^2 \qquad \text{(C.5)}
\end{aligned}
$$

Now we take care of the S_i operators for $i < N + 1$: set $\hat{x}_1 = x_1 - 2\gamma_1 \sum_{j=2}^M A_j^* x_j$ and $\hat{x}_1^* = x_1^* - 2\gamma_1 \sum_{j=2}^M A_j^* x_j^*$.

$$
\begin{aligned}
\langle S_i(x) - S_i(x^*), x - x^* \rangle_P &= \langle \frac{1}{N}(\nabla f_i(\hat{x}_1) - \nabla f_i(\hat{x}_1^*)), x_1 - x_1^* \rangle_1 \\
&\quad - \gamma_1 \langle \frac{1}{N}((\nabla f_i(\hat{x}_1) - \nabla f_i(\hat{x}_1^*)), \sum_{j=2}^M A_j^*(x_j - x_j^*) \rangle_1 \\
&\geq \langle \frac{1}{N}(\nabla f_i(\hat{x}_1) - \nabla f_i(\hat{x}_1^*)), \hat{x}_1 - \hat{x}_1^* \rangle_1 \\
&\quad + \gamma_1 \langle \frac{1}{N}((\nabla f_i(\hat{x}_1) - \nabla f_i(\hat{x}_1^*)), \sum_{j=2}^M A_j^*(x_j - x_j^*) \rangle_1 \\
&\geq \frac{N}{\gamma_1^2 L_i} \left\| \frac{\gamma_1}{N}((\nabla f_i(\hat{x}_1) - \nabla f_i(\hat{x}_1^*)) \right\|_1^2 \\
&\quad - \frac{N\sqrt{\delta}}{L_i \gamma_1^2} \left\| \frac{\gamma_1}{N}((\nabla f_i(\hat{x}_1) - \nabla f_i(\hat{x}_1^*)) \right\|_1^2 - \frac{L_i}{4N\sqrt{\delta}} \left\| \gamma_1 \sum_{j=2}^M A_j^*(x_j - x_j^*) \right\|_1^2 \\
&= \frac{N(1 - \sqrt{\delta})}{\gamma_1^2 L_i} \| (S_i(x))_1 - (S_i(x^*))_1 \|_1^2 - \frac{L_i}{4N\sqrt{\delta}} \| (S_{N+1}(x))_1 - (S_{N+1}(x^*))_1 \|_1^2 \qquad \text{(C.6)}
\end{aligned}
$$

Therefore, to show prove the coherence condition, we average (C.5) and (C.6):

$$
\begin{aligned}
\langle S(x), x - x^* \rangle_P &= \frac{1}{N+1} \sum_{i=1}^{N+1} \langle S_i(x) - S_i(x^*), x - x^* \rangle_P \\
&\geq \sum_{i=1}^N \frac{N(1 - \sqrt{\delta})}{\gamma_1^2 L_i} \| (S_i(x))_1 - (S_i(x^*))_1 \|_1^2 + \sum_{j=1}^M \frac{1 - \sqrt{\delta}}{\gamma_1} \| (S_{N+1}(x))_j - (S_{N+1}(x^*))_j \|_j^2 .
\end{aligned}
$$

Part 2 (roots): First suppose that $x^* \in \text{zer}(S)$. Then $N^{-1} \sum_{i=1}^N \nabla f_i(x_1 - 2\gamma_1 \sum_{j=2}^M A_j^* x_j) + \sum_{j=2}^M A_j^* x_j^* = 0$, and moreover, the points x_j^* for $j > 2$ satisfy:

$$
\begin{aligned}
x_j^* = \mathbf{prox}_{\gamma_j g_j^*} \left(x_j^* + \gamma_j A_j \left(x_1^* - 2\gamma_1 \sum_{j=2}^M A_j^* x_j^* \right) \right) &\Leftrightarrow A_j \left(x_1^* - 2\gamma_1 \sum_{j=2}^M A_j^* x_j^* \right) \in \partial g_j^*(x_j^*) \\
&\Leftrightarrow x_j^* \in \partial g_j \left(A_j \left(x_1^* - 2\gamma_1 \sum_{j=2}^M A_j^* x_j^* \right) \right)
\end{aligned}
$$

Therefore,

$$
0 = \frac{1}{N} \sum_{i=1}^N \nabla f_i \left(x_1^* - 2\gamma_1 \sum_{j=2}^M A_j^* x_j^* \right) + \sum_{j=2}^M A_j^* x_j^* \in \frac{1}{N} \sum_{i=1}^N \nabla f_i \left(x_1^* - 2\gamma_1 \sum_{j=2}^M A_j^* x_j^* \right) + \sum_{j=2}^M A_j^* \partial g_j \left(A_j \left(x_1^* - 2\gamma_1 \sum_{j=2}^M A_j^* x_j^* \right) \right);
$$

in particular, $x_1^* - 2\gamma_1 \sum_{j=2}^M A_j^* x_j^*$ minimizes (6.6).

To show that $\text{zer}(\sum_{j=1}^M A_j^* \partial g_j \circ A_j + N^{-1} \sum_{i=1}^N \nabla f_i) \neq \emptyset$ implies that $\text{zer}(S) \neq \emptyset$, reverse the above argument: choose $\hat{x}_1^* \in \text{zer}(\sum_{j=1}^M A_j^* \partial g_j \circ A_j + N^{-1} \sum_{i=1}^N \nabla f_i)$ and write $0 = N^{-1} \sum_{i=1}^N \nabla f_i + \sum_{j=1}^M A_j^* x_j^*$, where $x_j^* \in \partial g_j(A_j \hat{x}_1^*)$. Then $A_j x_1^* \in \partial g_1^*(x_j^*)$, and consequently, we have $x_j^* = \mathbf{prox}_{\gamma_j g_j^*}(x_j^* + A_j \hat{x}_1^*)$. Thus, if $x_1^* = \hat{x}_1^* + 2\gamma_1 \sum_{j=2}^M A_j x_j^*$, then we have $x^* = (x_1^*, \ldots, x_M^*) \in \text{zer}(S)$.

Part 3 (demiclosedness): In Part 1, our assumption that $x^* \in \text{zer}(S)$ was not necessary, and if we instead assume that $x^* = y$, where y is an arbitrary point in $\mathcal{H}$, the exact same argument yields that

$$
\langle S(x) - S(y), x - y \rangle_P \geq \sum_{i=1}^{N+1} \sum_{j=1}^M \beta_{ij} \| (S_i(x))_j - (S_i(y))_j \|_j^2 \geq C \sum_{i=1}^{N+1} \| S_i(x) - S_i(y) \|_P^2 \geq C(N+1) \| S(x) - S(y) \|_P^2,
$$

where $C > 0$ is a constant; in other words, the operator S is $C(N+1)$-cocoercive. Thus, by [2, Proposition 4.33],

$$T = I - \frac{1}{C(N+1)}S = \left(1 - \frac{1}{C(N+1)}\right)I + \frac{1}{C(N+1)}(I - S)$$

is nonexpansive. Therefore, because T is demiclosed at 0, it is easy to see that $C(N+1)(T - (1 - (C(N+1))^{-1})I) = I - S$ is demiclosed at 0. $\qquad \square$

Table C.1: Symbols

Indices

n	$\in$	$\mathbb{N}$; number of functions/operators
m	$\in$	$\mathbb{N}$; number of coordinates
i	$\in$	$\{1,\ldots,n\}$; index value for operators
j	$\in$	$\{1,\ldots,m\}$; index value for coordinates
k	$\in$	$\mathbb{N}$; index value for sequences

Spaces

$\mathcal{H}_j$		A separable Hilbert space
$\mathcal{H}$	$:=$	$\mathcal{H}_1 \times \cdots \times \mathcal{H}_m$

Variables

x, x_j	Primal variables: $x \in \mathcal{H}$ and x_j is the jth coordinate of x
$y_i, y_{i,j}$	Dual variables: $y_i \in \mathcal{H}$ and $y_{i,j}$ is the jth coordinate of y_i

Norms and inner products

$\|\cdot\|_j,\ \langle\cdot,\cdot\rangle_j$	A norm and an inner product on $\mathcal{H}_j$
$\|\cdot\|,\ \langle\cdot,\cdot\rangle$	A norm and an inner product on $\mathcal{H}$
$\|\cdot\|_{\mathrm{prod}}$	$(\forall x \in \mathcal{H})\ \|x\|_{\mathrm{prod}}^2 := \sum_{j=1}^m \|x_j\|_j^2.$
$\underline{M}_j, \overline{M}_j$	$(\forall x \in \mathcal{H})\ \sum_{i=1}^m \underline{M}_j \|x_j\|_j^2 \le \|x\|^2 \le \sum_{i=1}^m \overline{M}_j \|x_j\|_j^2$

Operators/Functions

f_i		A convex function with a Lipschitz continuous gradient
g_j		A nonsmooth/proximable function
S_i		A map from $\mathcal{H}$ to $\mathcal{H}$
S	$:=$	$n^{-1}\sum_{i=1}^n S_i$
$\mathcal{S}$	$:=$	$\mathrm{zer}(S)$
$\mathbf{S}^*$	$:=$	$((S_i(x^*))_j)_{ij}, \qquad (x^* \in \mathcal{S})$

Operators properties

β_{ij}	Coherence constants:

$$(\forall x \in \mathcal{H}),(\forall x^* \in \mathcal{S})\ \sum_{j=1}^m \sum_{i=1}^n \beta_{ij}\|(S_i(x))_j - (S_i(x^*))_j\|_j^2 \le n^{-1}\sum_{i=1}^n \langle S_i(x), x - x^*\rangle$$

μ	Essential strong quasi-monotonicity constant:

$$(\forall x \in \mathcal{H})\ \mu\|x - P_{\mathcal{S}}(x)\|^2 \le \langle S(x), x - P_{\mathcal{S}}(x)\rangle$$

Random Variables

i_k	$\in$	$\{1,\ldots,n\}$; an IID sequence of random operator indices
S_k	$\subseteq$	$\{1,\ldots,m\}$; an IID sequence of random sets of coordinates
ϵ_k	$\in$	$\{0,1\}$; an IID sequence of random binary decisions, which indicate whether to update the current dual variable ($\epsilon_k = 1$) or to leave it fixed ($\epsilon_k = 0$)

Graphs

$G = (V, E)$	Trigger graph: vertices $V = \{1,\ldots,n\}$ and edges $E \subseteq \{1,\ldots,n\}^2$;

$$(i, i') \in E \text{ if, and only if, } i_k = i \text{ triggers update of dual variable } y_{i'}^k.$$

Probabilities

q_j	$:=$	$P(j \in \mathsf{S}_k)$
p_{ij}	$:=$	$P(i_k = i \mid j \in \mathsf{S}_k)$
ρ	$:=$	$P(\epsilon_k = 1)$
p_{ij}^T	$:=$	The probability that the ith dual variable is updated and $j \in \mathsf{S}^k$:

$$P((i_k, i) \in E, j \in \mathsf{S}_k) = P((i_k, i) \in E \mid j \in \mathsf{S}_k)q_j = \sum_{(i', i) \in E} p_{i'j}q_j$$

Delays

τ_p	$\in$	$\mathbb{N}$; maximum primal variable delay
τ_d	$\in$	$\mathbb{N}$; maximum dual variable delay
d_k	$\in$	$\{0, 1, \ldots, \tau_p\}^m$; the vector of primal variable delays at the kth iteration
δ	$:=$	$\sup\{|d_{k,j} - d_{k,j'}| \mid k \in \mathbb{N} \text{ and } j, j' \in \{1,\ldots,m\}\}$
e_k^i	$\in$	$\{0, 1, \ldots, \tau_d\}^m$; the vector of dual variable delays for the ith dual variable at the kth iteration